**CellCharter: a scalable framework to chart and compare cell niches across multiple samples and spatial -omics technologies.**

Marco Varrone[1,2,3], Daniele Tavernari[1,2,3,4], Albert Santamaria-Martínez[2,4], Giovanni Ciriello[1,2,3,*]

[1]Department of Computational Biology, University of Lausanne, Lausanne, Switzerland;

[2]Swiss Cancer Center Léman, Lausanne, Switzerland;

[3]Swiss Institute of Bioinformatics, Lausanne, Switzerland;

[4]Swiss Institute for Experimental Cancer Research, École Polytechnique Fédérale de Lausanne (EPFL), Lausanne, Switzerland;

*Correspondence should be addressed to: giovanni.ciriello@unil.ch.

## ABSTRACT

Tissues are organized in niches where cell types interact to implement specific functions. Spatial -omics technologies allow to decode the molecular features and spatial interactions that determine such niches. However, computational approaches to process and interpret spatial molecular profiles are challenged by the scale and diversity of this data. Here, we present CellCharter, an algorithmic framework for the identification, characterization, and comparison of cellular niches from heterogeneous spatial transcriptomics and proteomics datasets comprising multiple samples. CellCharter outperformed existing methods, identified biologically meaningful cellular niches in different contexts, and discovered spatial cancer cell states, characterized by cell-intrinsic features and spatial interactions between tumor and immune cells. In non-small cell lung cancer, CellCharter revealed a cellular niche composed of neutrophils and tumor cells expressing markers of hypoxia and cell migration. Expression of these markers determined a spatial gradient associated with cancer cell state transition and neutrophil infiltration. Moreover, CellCharter showed that similar compositions of immune cell types can exhibit remarkably different spatial organizations in different tumors, highlighting the need for exploring spatial cell interactions to decipher intratumor heterogeneity. Overall, CellCharter is a flexible and scalable framework to explore and compare the spatial organization of normal and malignant tissues.

## INTRODUCTION

High-throughput molecular profiling of tissue samples allows to decipher heterogeneity and biological activity among species, organs, and cells. Technological advances in the past few decades have made it possible to characterize a great variety of molecular features unbiasedly and, often, at single-cell resolution[1]. Spatial -omics approaches represent the latest revolution of this progress[2]. These technologies allow not only to quantify molecular features in single cells, but also to retain information on the physical location of each cell within the tissue, so as to map molecular data on the tissue architecture[3,4]. Spatial molecular profiles were shown to recapitulate tissue anatomy and reveal cellular niches characterized by specific admixing of different cell types[5–7]. Recently, these technologies have been applied to cancer tissues[8] to dissect, for example, the association between cellular patterns and disease aggressiveness[9–11], or to investigate whether the oncogenic potential of mutated cells depends on their location within a tissue[12,13]. However, these technologies are often still limited in terms of resolution and/or scalability. Various types of spatial -omics technologies are currently being developed, the most advanced of which can be broadly categorized into spatial proteomics and spatial transcriptomics. Beyond the type of molecule that is assayed (protein or mRNA), spatial proteomics and transcriptomics approaches mostly differ in terms of *resolution* and *coverage*. Spatial proteomics largely relies on multiple cycles of multiplexed immunofluorescence or on imaging mass cytometry/spectrometry[3,14]. As such, they provide single-cell resolution data but low coverage, with only tens or a few hundred proteins that can be quantified. Spatial transcriptomics can be divided into image-based and sequencing-based approaches[15,16]. The former typically assays up to a thousand genes but achieves single-cell or even sub-cellular resolution. The latter can cover the entire transcriptome but within fixed-size spots that usually comprise between 10 and 100 cells. High-resolution implementations of sequencing-based spatial transcriptomics have been proposed, but they haven't yet reached standardization and commercialization[17–21].

In parallel with the development of these technologies, new computational approaches have emerged to process and analyze spatial molecular profiles. One approach for characterizing the spatial cellular architecture is *spatial clustering*, which assigns cells to clusters based on both their intrinsic features, such as protein or mRNA abundance, and the features of neighboring cells in the tissue. Hence, whereas clustering approaches that are exclusively based on cell-intrinsic features determine populations of molecularly similar cells, spatial clustering determines cellular *niches* characterized by specific admixing of these populations. Methods to perform spatial clustering are mostly based on three general approaches: Hidden Markov Random Field (HMRF[22]), Graph Neural Networks (GNN[23]), and Neighborhood Composition (NC). HMRF is an extension of Hidden Markov Models that determines the state (i.e., cluster) of a cell based on its features and the state of an unbounded number of neighboring cells. A Bayesian version of HMRF is used in BayesSpace[24], whereas DR.SC[25] combines dimensionality reduction of the feature space and HMRF into a single operation. GNN is a machine learning technique that generates vector-based representation of a node (e.g., a cell) from the convolution of features of its neighbors in a network[26]. The number of layers of the GNN determines the number of feature convolutions, allowing to propagate information beyond the immediate neighborhood. Tools like SEDR[27] and STAGATE[28] use a GNN to determine network representations of spatial transcriptomics datasets, where cells are nodes, and links are drawn based on spatial proximity. Lastly, by NC, we refer to methods that cluster cells based on the proportion of cell types within their neighborhood[5]. These

approaches require manually curated cell type annotations and do not represent truly systematic clustering approaches. Nonetheless, in specific contexts, they represent a scalable solution, especially for spatial proteomics datasets that, in a single tissue slide, may reach millions of cells.

Outstanding challenges in the field concern scalability and portability. Indeed, given the currently scarce availability of spatial -omics datasets, most methods have not been designed to simultaneously cluster, characterize, and compare large numbers of samples. GNN approaches have high computing and memory requirements, even when using Graphics Processing Units (GPUs), as they require loading the entire network in memory. This becomes rapidly unfeasible for large tissue slides or when clustering multiple slides together, either preventing the use of the tool or requiring breaking the sample into subsections. Beyond scalability, clustering multiple samples requires correcting for potential batch effects. Currently, only BayesSpace is predisposed to include a batch effect correction procedure, while the other tools leave to the user to implement and integrate one. Moreover, with rapidly evolving technologies, tools need to be usable with data generated with different techniques. None of the current approaches satisfies this criterion, either because they require specific cell layouts or because they cannot scale with high-throughput single-cell assays. Finally, with expected increasing data availability, there is a need for tools capable not only to identify spatial clusters, but also to determine their biological features and understand how these features change across tissue types and conditions. Here, we introduce CellCharter (https://github.com/CSOgroup/cellcharter), a new algorithmic framework to address these challenges.

## RESULTS

### Inference, characterization, and comparison of spatial clusters

We designed CellCharter to achieve three main objectives: 1) to analyze large cohorts of spatially profiled samples, 2) to be agnostic of the underlying technology used to generate spatial molecular profiles, and 3) to implement not only a spatial clustering algorithm, but also a suite of approaches for cluster characterization and comparison. CellCharter starts by taking in input a spatial transcriptomics or proteomics dataset represented as a matrix of features, corresponding to the abundance of a given gene or protein in each cell or spot, and spatial coordinates of each cell/spot (**Fig. 1a** - left). Dimensionality reduction and batch effect correction of the feature space are then performed using variational autoencoders[29–31] (VAE) specifically designed for either transcriptomics or proteomics data. VAE-derived embeddings define the new set of features for each cell/spot. Notably, this is the only step of our method that depends on the type of spatial -omics data used in input and, given the modular architecture of CellCharter, switching from one VAE to another will not affect the rest of the analyses. Next, CellCharter builds a network of cells/spots based on their spatial proximity and, for each cell/spot $A$, we define the $l$-neighborhood of $A$ as the set of cells/spots that are at most $l$ steps away from $A$ in the network, where $l$ is a user-defined parameter (in **Fig. 1a** - center, $l$ = 3). To aggregate the features of each cell/spot with those of its $l$-neighborhood, we assign to each cell/spot $A$ the concatenation of its own vector of features and a series of vectors each containing the averages of each feature among the set of cells/spots at distance $i$ from $A$, for each $i \in [1, l]$. In principle, aggregation metrics other than the average can be used (see Methods). Lastly, cells/spots are clustered based on this vector of aggregated features using a Gaussian Mixture Model (GMM) approach. To determine the final number of

clusters, CellCharter introduces an approach that assesses the stability of a given number of clusters based on the Fowlkes-Mallows index[32] (**Fig. 1a** - right). Briefly, GMM is run multiple times (n = 10 in this manuscript) for each number of possible clusters within a user-defined range. A solution with *n* clusters is considered "stable" when cluster assignments are highly reproducible across multiple GMM runs for *n*, *n-1*, and *n+1*. By incorporating a batch effect correction step and using a highly scalable approach to encode spatial information, CellCharter is particularly suited to simultaneously determine spatial clusters among multiple samples, which is desirable to validate spatial niches across independent experiments and compare them across different conditions.

In addition, CellCharter improves existing downstream analyses and implements new ones to: (1) determine cluster proportions for each sample; (2) compute cell type enrichment for each cluster (when cell type annotations are available); (3) estimate significant spatial proximity among clusters (cluster neighborhood enrichment or cluster NE) and how it varies among conditions (*differential* cluster NE); (4) characterize and compare cluster *shapes* (**Fig. 1b**). CellCharter introduces an analytical approach to compute *asymmetric* cluster NE, which is more efficient than currently available permutation-based methods and allows discriminating when the neighborhood of one cluster is enriched in another cluster but not vice versa. For example, this condition is well illustrated by spatial clusters derived by CellCharter on a tissue section of a mouse normal spleen analyzed using the CODEX spatial proteomics platform[5] (**Fig. 1c**). Here, the germinal center-enriched (GC) cluster (orange) is exclusively in contact with the marginal zone-enriched cluster (light blue) and a cluster found at the boundary between GC and periarterial lymphatic sheaths (PALS, purple). However, both the marginal zone and the GC-PALS boundary clusters make several interactions with other clusters, hence the enrichments of interactions between different clusters are asymmetric and this relationship is well captured by the cluster NE approach implemented in CellCharter (**Fig. 1c**). In addition, CellCharter is, to the best of our knowledge, the only methodology that allows to characterize cluster shapes. Specifically, it analyzes the shape of each *cluster component*, i.e., the set of cells belonging to the same cluster and within a connected component of the cell network that CellCharter uses to encode spatial proximity. For each cluster component, CellCharter computes its *boundary*, using an approach based on alpha shapes[33], from which it derives the area and perimeter of the component, and its *bounding box*, i.e., the smallest rectangle encasing the cluster component, from which it derives the minor and major axes. Based on these measures, we compute 4 scores: **curl**, which expresses how curved or twisted a shape is, **elongation**, which is the ratio of the major and minor axes, **linearity**, which assesses how well a shape can be approximated by a linear path, and **purity**, which quantifies the fraction of cells within a cluster boundary that belong to that cluster (**Fig. 1d**). Representative cluster components from the mouse spleen dataset demonstrate how combinations of these scores allow to describe shapes as linear, round, circular, and irregular (**Fig. 1e**). Overall, CellCharter is a comprehensive suite to identify, characterize, and compare spatial cell clusters.

To demonstrate the effectiveness and efficiency of CellCharter, we first assessed its spatial clustering performance against state-of-the-art approaches. To this purpose, we used an annotated spatial transcriptomics dataset (10x Genomics Visium) comprising 12 samples (4 samples from 3 donors) of human dorsolateral prefrontal cortex (DLPFC)[34]. Within each sample, spots have been manually assigned to one of six cortical layers (L1-L6) or to white matter (WM) (**Suppl. Fig. 1a**). These labels are shared among samples and were used as the true spatial clusters. For each spatial clustering algorithm, we performed hyperparameter

tuning on 3 samples (one for each donor) and we clustered the spots of the remaining 9 samples, either on each sample individually or on all of them jointly. Spatial clusters obtained with each method were compared to the true clusters using the Adjusted Rand Index (ARI) and each method was run 10 times with different seeds to assess the robustness of the solutions. Whereas STAGATE obtained the best performance when clustering individual samples (**Suppl. Fig. 1b**), CellCharter outperformed existing tools when jointly clustering all samples, both in terms of best ARI (**Fig. 1f**) and average ARI over multiple runs (**Fig. 1g, Suppl. Fig. 1c**). In addition, CellCharter was the most efficient algorithm in terms of computational and memory efficiency, both in its GPU and CPU versions (**Fig. 1h**), except for BayesSpace which was slightly faster than CellCharter in CPU, although requiring more memory. Interestingly, 99.85% of running time in CellCharter was dedicated to dimensionality reduction and batch correction (**Suppl. Fig. 1d**), highlighting the high efficiency of the clustering procedure. It is to be noted that if supplying additional samples does not provide additional information for the training of the VAE, the time required for dimensionality reduction and batch correction does not increase. Indeed, the running time of CellCharter did not increase when clustering more than 6 samples, whereas the running time of BayesSpace continued to increase (**Fig. 1h**). To further evaluate the performances of the best methods from the previous comparisons, CellCharter and STAGATE, we used a single-cell resolution spatial proteomics dataset (based on the CODEX platform) comprising 3 samples of normal mouse spleen and 6 samples of mouse spleen derived from an animal model of systemic lupus erythematosus[5]. Manual annotations from the original work classified regions in these samples as B-follicles, marginal zones, PALS, or red pulp (**Fig. 1i**). Spatial clustering was jointly performed on all samples. Even though no ground truth annotations were provided for individual cells, challenging a quantitative assessment of the clustering results (e.g., by ARI), visual comparisons of spatial clusters and manually annotated regions clearly showed that CellCharter clusters best approximated the true tissue components (**Fig. 1i**), generating clusters with higher purity (**Suppl. Fig. 1e**) and that better mimicked the spleen anatomy than STAGATE. Notably, on this large dataset (707,466 cells in total), CellCharter was four times faster than STAGATE, largely due to its more scalable approaches for spatial features encoding and clustering (**Fig. 1j**).

**CellCharter identifies changing cellular niches in the spleen of systemic lupus mice**

The mouse spleen spatial proteomics dataset gave us the opportunity of using CellCharter to characterize and compare spatial clusters between two conditions: healthy spleen (BALBc model) vs. spleen in mice affected by systemic lupus erythematosus (MRL model) (**Fig. 2a**). Using the cluster stability analysis introduced in CellCharter, we determined stable cluster solutions for n = 4 and n = 11 clusters (**Fig. 2b**) and selected the higher number of clusters to have a more fine-grained description of the tissue architecture (**Suppl. Fig. 2a**). Cells were assigned to all spatial clusters in all samples, although in different proportions (**Fig. 2c**). Indeed, while BALBc samples showed a similar distribution of cluster proportions, MRL samples appeared to be more heterogeneous, particularly among stages of the disease (early, intermediate, and late). Using available cell type annotations, CellCharter determined cell type enrichments within each cluster and, consistent with the notion of cellular niche, most clusters were enriched for specific combinations of cell types, recapitulating the spleen tissue architecture (**Fig. 2d**). For example, we identified a GC cluster (C3), composed of B cells and follicular dendritic cells (FDCs), a PALS cluster (C5) composed of $CD4^+$ and $CD8^+$ T cells, and a cluster at the GC-PALS boundary (C2), which, especially in MRL samples, was enriched in B220 positive ($B220^+$) double-negative ($CD8^-CD4^-$) T cells.

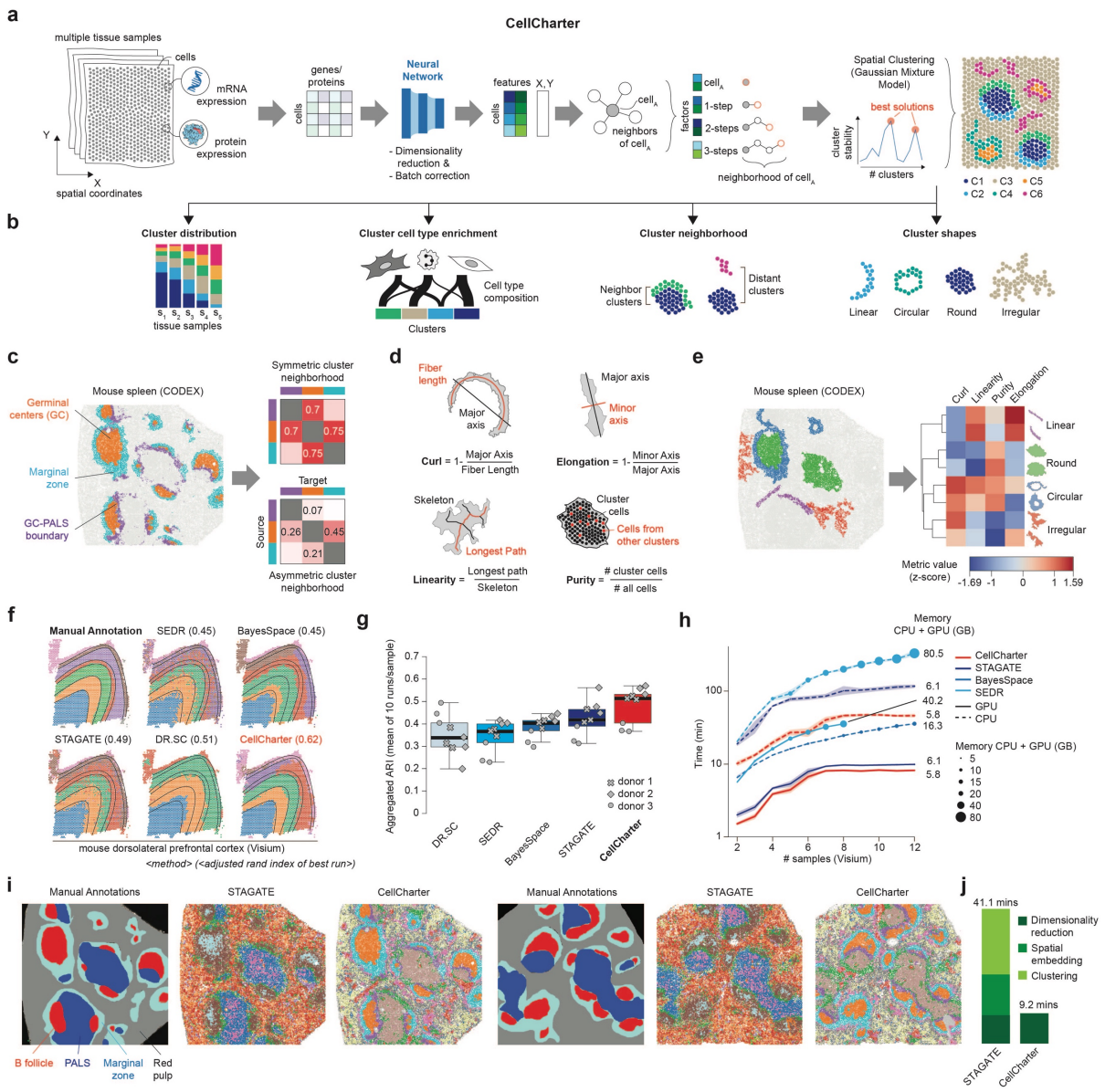


**Figure 1: CellCharter identifies, characterizes, and compares spatial clusters.**
***a)*** *Workflow of CellCharter. From left to right: Spatial molecular profiling generates a gene and/or protein expression matrix comprising the coordinates (x,y) of each cell; dimensionality reduction and batch correction are performed on this matrix, and, for each cell ($cell_A$), its features are concatenated with the features of its neighbors; lastly Gaussian Mixture Model clustering is performed and best cluster solutions are chosen based on cluster stability.* ***b)*** *Downstream analyses implemented in CellCharter for the characterization and comparison of spatial clusters.* ***c)*** *Example of 3 spatial clusters (color coded) in a tissue sample of mouse spleen analyzed by CODEX (left) and symmetric (top right) vs. asymmetric (bottom right) neighborhood enrichment analysis.* ***d)*** *Schematic representation of the four metrics (curl, elongation, linearity, and purity) implemented in CellCharter to describe shape of spatial clusters.* ***e)*** *Example of spatial cluster components (color coded) in a tissue sample of mouse spleen analyzed by CODEX (left). Heatmap representation of the shape metric values for each cluster component (right). Cluster components are grouped in representative shape classes: linear, round, circular, and irregular.* ***f)*** *Manually annotated and predicted cluster labels by all the evaluated methods for one tissue slide from the Visium DLPFC dataset. The best Adjusted Rand Index (ARI) value is reported out of 10 repetitions.* ***g)*** *Mean ARI for each DLPFC sample (over 10 repetitions, y-axis) obtained by the listed methods (x-axis) upon performing joint spatial clustering of all samples (n = 9 samples).* ***h)*** *Runtime and memory requirement for the evaluated clustering methods at increasing DLPFC dataset size.* ***i)***

*Manual annotations and spatial clusters (color coded) determined by STAGATE and CellCharter in two tissue samples of mouse spleen analyzed by CODEX.* ***j)*** *Running time (GPU) of STAGATE and CellCharter to perform spatial clustering of 9 mouse spleen samples. For both methods, the running time of the three main steps are shown separately.*

From normal spleen samples to increasing stages of the disease, the number of cells assigned to the GC cluster progressively decreased as opposed to the cells assigned to the GC-PALS cluster (**Fig. 2e-f**, insets 1-2-3). Other areas of the spleen were split into multiple clusters based on the prevalence of different cell types. For example, the large red pulp area was split into 3 clusters characterized by different dominant cell types, B cells (C7), erythroblasts (C8), and NK cells (C9); whereas 2 separate clusters were associated with trabecular structures (C10, C11), one of which was highly enriched in granulocytes, marked by Ly6G expression (C11). The granulocyte-enriched cluster exhibited a gradual expansion with the emergence and progression of the disease (**Fig. 2e-f**, insets 4-5-6). One cluster was associated with staining artifacts or missing markers (C1) (**Suppl. Fig. 2b**).

To further appreciate changes in spatial architecture from normal spleen to systemic lupus, we performed cluster neighborhood enrichment (NE) and differential cluster NE analysis, to determine preferential interactions and changes of interactions among clusters. As anticipated, we found significant NE changes between BALBc and MRL samples, concerning the GC, marginal zone, and GC-PALS clusters. Indeed, in MRL samples, the GC cluster (C3) decreased interaction enrichment with the marginal zone (C4) in favor of interactions with the GC-PALS cluster (C2) (**Fig. 2g**). Comparison of cluster shapes in these three clusters revealed that both the GC-PALS and germinal center clusters significantly increased their curl values, while GC-PALS and marginal zone clusters significantly lost linearity (**Fig. 2h**). These shape differences indicated increased irregularity of the clusters and loss of the tissue architecture that characterize normal spleen anatomy (**Fig. 2i**). Overall, changes in cluster proportions (**Fig. 2c**), differential cluster NE (**Fig. 2g**), and shape comparisons (**Fig. 2h**) reflected a gradual expansion and infiltration of T-cells from the GC-PALS cluster into the germinal center (**Suppl. Fig. 3a**), which is consistent with germinal center activation after infection[35]. Differential cluster NE analysis also showed a change in the spatial arrangement of the trabecula-associated clusters (C10 and C11) and their relationship with the PALS (C6) and B cell-enriched red pulp (C7). Indeed, in MRL samples granulocyte-enriched trabeculae (C11) were present in higher proportion than in the normal spleen, and were restricted in the red pulp, whereas the other trabecular cluster (C10) was in contact with PALS, specifically the PALS cluster that was enriched in dendritic cells (**Fig. 2g**). Interestingly, cluster C11 lacked expression of CD31 (**Suppl. Fig. 3b**), a marker shown to be absent in spleen capillaries within the red pulp but present in the central arteries which are surrounded by the PALS[36]. Hence, in MRL samples, CellCharter captured the emergence of two distinct trabecular niches establishing preferential spatial interactions that were not present in the normal spleen: a $CD31^{+}/Ly6G^{-}$ cluster in proximity of PALS and a $CD31^{-}/Ly6G^{+}$ cluster within the red pulp (**Fig. 2j**). Overall, these analyses and results well demonstrate the potential of CellCharter not only to determine biologically meaningful spatial cellular niches, but also to discover how these niches emerge and/or change in different conditions.

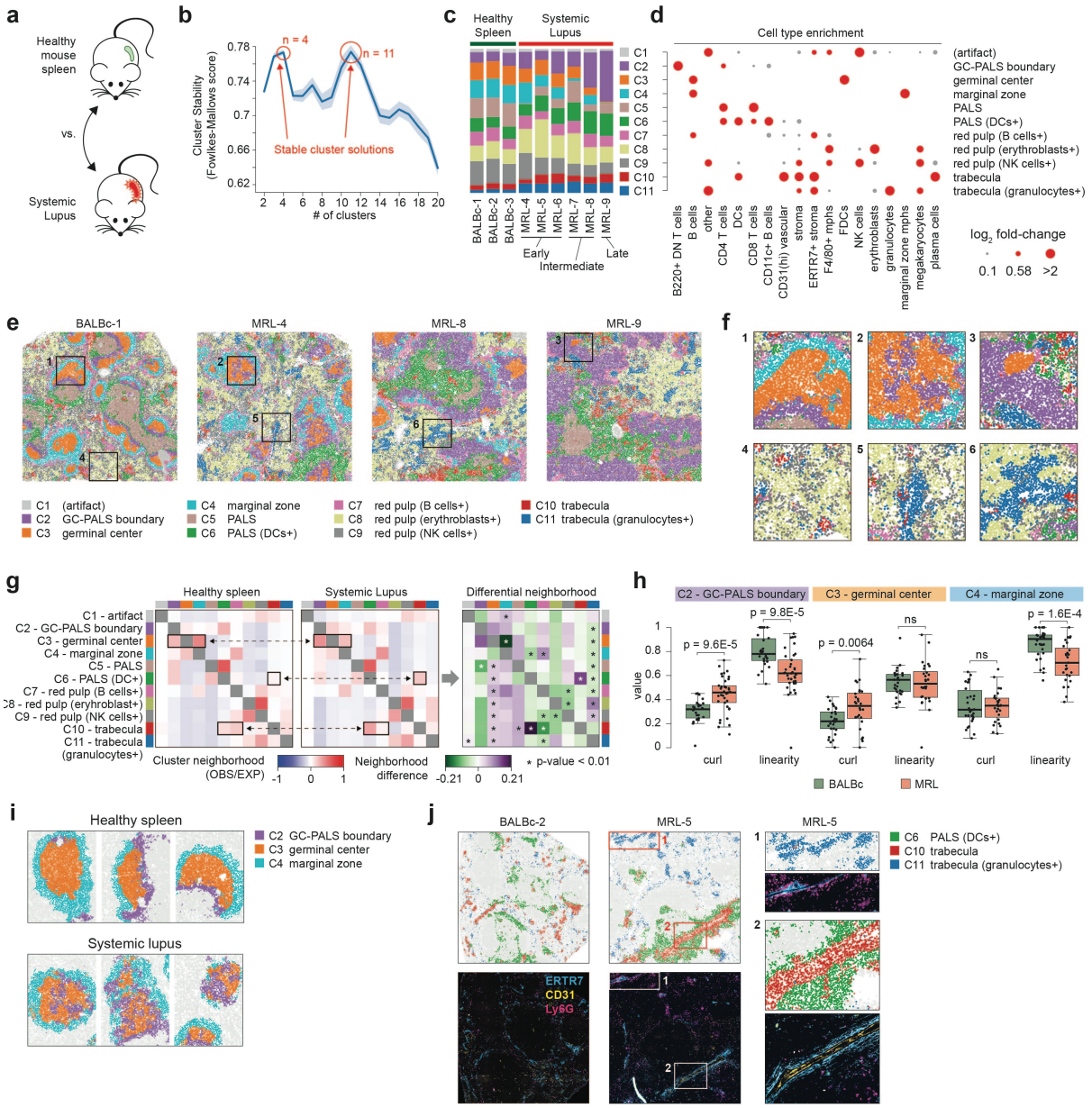


**Figure 2: Spatial cellular niches in the spleen of healthy mice and mice affected by systemic lupus.** ***a)*** *We applied CellCharter to a spatial proteomics dataset of spleens of healthy mice (BALBc) and mice with systemic lupus erythematosus (MRL).* ***b)*** *CellCharter cluster stability (y-axis) for range of numbers of cluster (x-axis). Most stable cluster solutions are highlighted.* ***c)*** *CellCharter spatial cluster proportions in each sample. MRL samples are grouped based on disease stage (early, intermediate, late).* ***d)*** *Cell type enrichment in each spatial cluster. Based on cell type enrichment and spatial location in the tissue, every cluster was associated with an anatomical area of the spleen (annotated on the right).* ***e-f)*** *Spatial clusters in one healthy sample (BALBc-1) and three systemic lupus samples at different disease stages: MRL-4 (early), MRL-8 (intermediate), MRL-9 (late). Areas 1, 2, and 3 zoom in on representative clusters associated with germinal centers. Areas 4, 5, 6 zoom in on representative clusters associated with granulocyte-enriched trabecular structures.* ***g)*** *Cluster neighborhood enrichment (left and center heatmap) and differential neighborhood enrichment (right heatmap) between spatial clusters in healthy mice and in mice affected by systemic lupus. Examples of differentially enriched neighborhoods are highlighted.* ***h)*** *Curl and linearity values (y-axis) for clusters C2 (left), C3 (center), and C4 (right) in healthy (green) and systemic lupus samples (red). p-values were computed by two-tailed Wilcoxon test.* ***i)*** *Representative examples of C2, C3, and C4 cluster component from healthy (top) and systemic lupus samples (bottom).* ***j)*** *Representative examples of C6, C10, and*

*C11 cluster components from BALBc-2 and MRL-5 samples (top) and matching immunofluorescence staining images for the indicated markers (bottom). Insets (1 and 2) highlight the differential composition of trabecular clusters C10 and C11.*

**Using CellCharter to decipher intratumor heterogeneity**

Intratumor heterogeneity is characterized by both heterogeneous cancer cell populations, or *cancer cell states*, and diverse composition of the tumor microenvironment (TME)[37–40]. In this context, spatial molecular profiles of multiple tumor samples offer the opportunity to decipher how tumor and TME cell populations organize and interact in the tissue. To exploit this opportunity, we used CellCharter to analyze data from 8 non-small cell lung cancer tissue sections derived from 5 patients: 4 lung adenocarcinoma (LUAD) and 1 lung squamous cell carcinoma (LUSC) (**Fig. 3a**). All tissue samples were previously analyzed using the Nanostring CosMx spatial transcriptomics platform[41], which assayed mRNA expression for 960 genes at single-cell resolution.

CellCharter identified three possible stable solutions with n = 3, 8, or 20 clusters (**Fig. 3b**). We examined these solutions to investigate what information was captured by clustering at different granularity. For each solution, we identified "tumor-enriched" clusters, i.e., clusters that were largely composed of tumor cells (85-90%). Notably, more than 90% of tumor cells were contained in tumor-enriched clusters (**Suppl. Table 1**). Interestingly, we found that with n = 3 clusters, tumor areas from all 8 samples were largely grouped into one main tumor-enriched cluster that was shared among all patients (**Fig. 3c** - top). With n = 8 clusters, the majority of patients exhibited one private tumor-enriched cluster (**Fig. 3c** - center). Lastly, with n = 20 clusters, each patient exhibited multiple private tumor-enriched clusters, potentially reflecting distinct cancer cell states (**Fig. 3c** – bottom and **Suppl. Fig. 4a**). Hence, to a certain extent, CellCharter stable cluster solutions reflected a hierarchy of biological entities: cancer, individual tumors, and intratumor cell states. To further explore spatial features of intratumor heterogeneity, we focused on the 20-cluster solution. Cell type enrichment analysis confirmed the presence of several tumor-enriched clusters and clusters characterized instead by distinct combinations of immune and stromal cell types, referred to as TME-enriched clusters (**Fig. 3d**). Tumor-enriched clusters were almost invariably patient-specific but shared between independent samples derived from the same patient (**Fig. 3c** – barplot). The only exception was cluster C4, which was shared among LUAD-5 and LUAD-12 and exhibited an enrichment for spatial interaction with the same TME-enriched cluster (C17) in both tumors (**Fig. 3d-e**).

Cluster NE analysis showed that tumor-enriched clusters within the same patient typically exhibited preferential interactions with distinct TME-enriched clusters (**Fig. 3e** and **Suppl. Fig. 4b**). For example, in patient LUAD-9, we found 2 major tumor-enriched clusters (C0, C12): C0 exhibited frequent interactions with a cluster composed of neutrophils and NK cells (C11), whereas C12 mostly interacted with a cluster composed of tumor cells and $CD4^+$ T cells (C18) (**Fig. 3f**). To investigate whether different spatial arrangements corresponded also to different intrinsic transcriptional features, we performed differential expression analysis between clusters C0 and C12 (**Suppl. Table 2**). Among the upregulated genes in C0 compared to C12, we found the hypoxia-inducible gene *NDRG1*, which may also promote stem-like phenotypes and epithelial-to-mesenchymal transition (EMT) in lung cancer[42,43], the angiogenic factor VEGFA, which is also induced by hypoxia[44], and several genes associated with cytokine signaling and neutrophils chemotaxis. Prominent examples of these genes were *S100A8* and *S100A9*, and chemokines *CXCL1*, *CXCL2*, and *CXCL3* (**Fig. 3g**), which beyond their role as

neutrophil attractants[45] have also been shown to promote tumor invasion and metastatic capacity[24]. Gene set enrichment analysis confirmed that proteins involved in cytokine signaling and neutrophils chemotaxis were enriched among upregulated genes in C0, which is consistent with this cluster always being in spatial proximity with the neutrophil-enriched cluster C11 (**Fig. 3e-f**). Upregulated genes in C12 compared to C0 were instead enriched for cell proliferation markers, such as *MKI67*, and comprised the fibroblast growth factor receptors *FGFR1* and *FGFR2* and the histone modifier *EZH2* (**Fig. 3g**), which is frequently over-expressed in aggressive lung adenocarcinoma[46]. Consistent with the upregulated markers in the two clusters, gene signature scores associated with cytokine signaling, response-to-hypoxia, and EMT were higher in cluster C0 than in C12 (**Suppl. Fig. 5a**) and co-localized in the tumor niche surrounding the neutrophil-enriched cluster C11 in independent tumor samples (**Fig. 3h** and **Suppl. Fig. 5b-d**). Conversely, these signatures exhibited an anti-correlated spatial gradient with a cell proliferation signature (**Fig. 3h** and **Suppl. Fig. 5e**). In particular, response-to-hypoxia signature scores decreased in tumor cells as their distance from the neutrophil-enriched cluster C11 increased and this pattern was specific for neutrophils in this spatial cluster (**Suppl. Fig. 6a**). A positive correlation between response-to-hypoxia signature scores and neutrophil infiltration was further confirmed in multiple independent lung adenocarcinoma datasets (**Suppl. Fig. 6b**). The spatial cellular niche revealed by CellCharter in these tumors is likely associated with inter-cellular interactions between neutrophil and tumor cells. Indeed, it is known that tumor growth leads to the emergence of hypoxic and necrotic regions, within which cells secrete neutrophil-recruiting chemokines such as IL-8, IL-6, CXCL1, CXCL2, CXCL5, CXCL8, and SOD2[47–49], several of which were over-expressed in cluster C0 (**Suppl. Table 2**). Neutrophil recruitment further boosts this signaling and, importantly, promotes angiogenesis, cancer cell migration, and EMT. Our results hence indicate the presence of a cancer cell “spatial state” characterized by distinct cell intrinsic features and cancer-TME interactions.

Lastly, we wondered whether cell types present in the same proportion among patients could assume different spatial arrangements. To investigate this possibility, we focused on the composition and spatial organization of the TME in the 5 lung cancer patients. Based on previously annotated cell types, LUAD-5, LUAD-9, and LUAD-12 exhibited a similar composition of immune and stromal cells, with a high fraction of neutrophils and fibroblasts (**Fig. 4a**). In these samples, epithelial cells and monocytes were both enriched in cluster C10 in LUAD-9, but clustered separately in LUAD-5 (C2 and C5) (**Fig. 4b** - left). Similarly, neutrophils and fibroblasts were both enriched in cluster C14 in LUAD-12 but were enriched in different clusters (C11 and C17) in LUAD-9 (**Fig. 4b** - right). To quantify these differences, we performed cell-type NE analyses for each of these samples among all cell types (**Suppl. Fig. 7**). Consistent with spatial cluster compositions (**Fig. 4b**), epithelial cells and monocytes were spatially closer than expected in LUAD-9, but not in LUAD-5 (**Fig. 4c** - top). Indeed, these cells were highly intermixed in LUAD-9, but spatially segregated in LUAD-5 (**Fig. 4d**). Even more striking were the different neighborhood enrichment and spatial arrangements of neutrophils and fibroblasts in LUAD-12 and in LUAD-9 (**Fig. 4c** - bottom). These cell types consistently exhibited high intermixing and co-clustering in LUAD-12, whereas they were spatially segregated in different clusters in LUAD-9 (**Fig. 4e**). These results well demonstrate the power of including spatial information in single-cell molecular profiles. Indeed, whereas classical scRNA-seq analyses would have predicted similar TME composition in these samples, spatial transcriptomics revealed drastically different admixing of immune and stromal cell types.

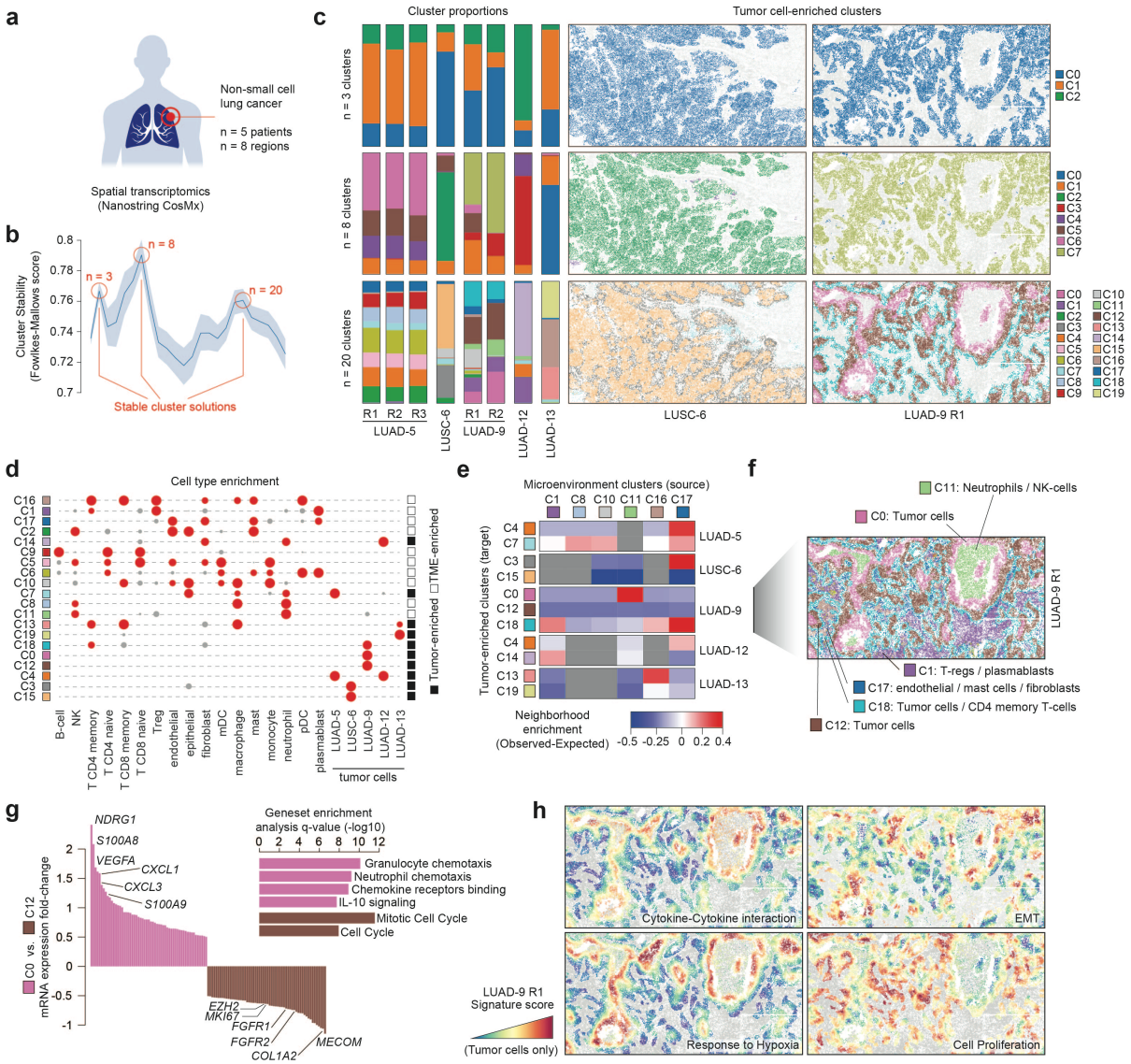


**Figure 3: Spatial cellular niches in non-small cell lung cancer.** ***a)*** *We applied CellCharter to a single-cell spatial transcriptomics dataset comprising 8 slides from 5 non-small cell lung cancer patients.* ***b)*** *CellCharter cluster stability (y-axis) for range of numbers of cluster (x-axis). Most stable cluster solutions are highlighted.* ***c)*** *CellCharter spatial cluster proportions in each sample for three numbers of clusters solutions (left) and the corresponding cell labels for tumor cell-enriched clusters in representative samples (right). (LUAD: lung adenocarcinoma, LUSC: lung squamous cell carcinoma.)* ***d)*** *Cell type enrichment in each spatial cluster (n = 20 clusters).* ***e)*** *Neighborhood enrichment analysis between the TME-enriched clusters (source) and the tumor-enriched clusters (target). Only TME clusters for which there is a positive neighborhood enrichment with at least one tumor-enriched cluster are shown.* ***f)*** *Representative example of TME-enriched and tumor-enriched clusters in LUAD-9 R1.* ***g)*** *mRNA expression fold-change (log2 – y-axis) of differentially expressed genes between the clusters C0 and C12 in patient LUAD-9 (left) and gene set enrichment analysis of genes upregulated in C0 (pink) and C12 (brown) (right).* ***h)*** *Gene expression signature scores of tumor cells in LUAD-9 R1 for four different gene signatures.*

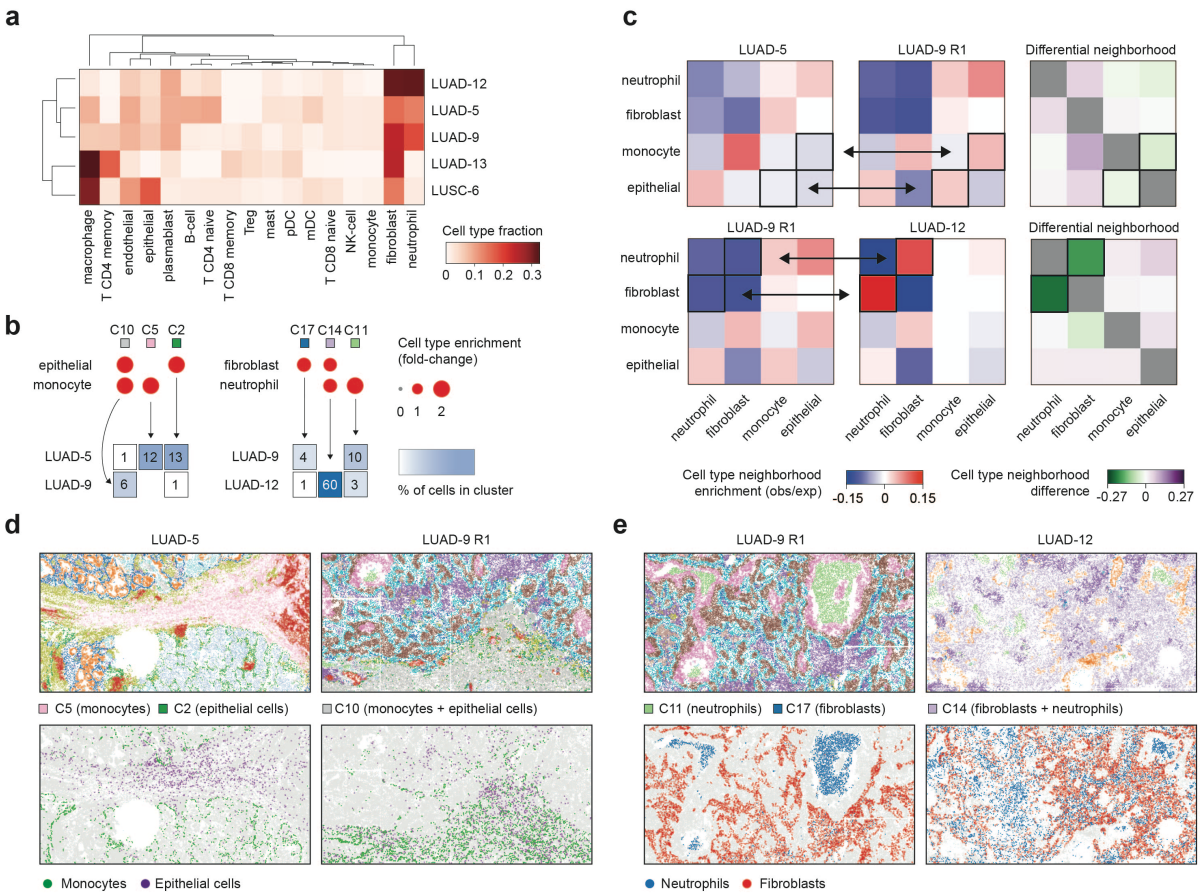


**Figure 4: Cell type composition versus cell type admixing.** ***a)*** *Immune and stromal cell type fractions in tumor samples from 5 non-small cell lung cancer patients (LUAD: lung adenocarcinoma, LUSC: lung squamous cell carcinoma). b) Cell type enrichment (top) and cell type percentages (bottom) of the indicated cell types (rows) in the indicated spatial clusters (column).* ***c)*** *Cluster neighborhood enrichment (left and center heatmaps) and differential neighborhood enrichment (right heatmap) between the indicated cell types in the indicated tumor samples. Examples of differentially enriched neighborhoods are highlighted.* ***d-e)*** *Spatial cluster assignments (top) and cell type labeling (bottom) for the indicated clusters and cell types in the indicated samples.*

## DISCUSSION

The design and development of computational tools to analyze data from emerging technologies pose two major challenges: 1) anticipating which technologies will ultimately prevail and, hence, which type of data will be most frequently generated, and 2) anticipating the needs and requirements of analyses that are not yet possible. Spatial molecular profiles are currently generated for a limited number of samples and using a wide range of technologies, most of which are upgraded or completely change from one year to the other. In the next few years, as technologies will mature, we expect that multiple layers of spatial omics data will be generated for large collections of samples. To address these challenges, we designed CellCharter to be technology-agnostic and highly scalable. We tested CellCharter on 3 different technologies comprising spatial proteomics based on cyclic immunofluorescence of fluorophore-conjugated oligonucleotides, sequencing-based spatial transcriptomics, and image-based spatial targeted transcriptomics. CellCharter generated meaningful results in all applications, recapitulating tissue morphologies and biological processes, such as germinal center activation in autoimmune disease. Importantly, we improved the benchmarking procedure by introducing a hyperparameter tuning phase to avoid biases in testing, by repeating runs multiple times to increase the reliability of the results, and by testing the quality of clustering of all samples jointly. CellCharter outperformed existing approaches and was significantly faster than existing methods, allowing to jointly cluster multiple samples simultaneously. This provides a critical advantage when it comes to matching and comparing clusters across samples or conditions and makes our tool the most suitable to analyze large datasets of spatial profiles.

To accommodate both spatial proteomics and transcriptomics datasets, CellCharter was developed with a modular architecture, where only the dimensionality reduction and batch correction step is dependent on the data type. In addition, CellCharter is fully compatible with the scverse ecosystem, a consortium of Python tools for the analysis of omics data in life sciences[50]. This design will facilitate the use and integration of additional data types. For example, emerging spatial profiling technologies have been proposed to study epigenetic cell features[51], copy number alterations[52], chromatin accessibility[53], and combined single-cell RNA-seq and protein profiles[17,54,55]. In addition, all these molecular layers could be combined with cell morphology and tissue architectural features derived from hematoxylin and eosin (H&E) staining. To analyze these data types with CellCharter, it will be sufficient to introduce the correct model to generate the feature embeddings, without modifying the other steps of the analysis. In particular, recently proposed deep neural networks trained on a large collection of H&E images[56], together with the models already included in CellCharter, could prove extremely valuable to generate embeddings that integrate transcriptional or proteomic features with tissue histology. Spatial niches determined from this combination of features could empower digital pathology by revealing histological features associated with specific molecular phenotypes, which could serve as prognostic markers in the clinic.

Lastly, our results on non-small cell lung cancer tissues effectively demonstrated the potential of spatial -omics profiles to investigate intratumor heterogeneity in terms of both cell-intrinsic features and cell-cell interactions. The recent wide application of scRNA-seq to large tumor cohorts has allowed to determine, on the one hand, stromal and immune cell subtypes with fine granularity[57] and, on the other hand, cancer cell states defined by concurrent activation of specific transcriptional programs[58]. Spatial clustering on similar cohorts will reveal where and

which of these cell subtypes and states interact, determining unique cellular niches. In this study, we determined a lung cancer cellular niche where tumor cells characterized by a response-to-hypoxia and EMT transcriptional state surrounded a neutrophil cluster, prompting mechanistic hypotheses on the emergence of this cell state and interactions. In this sense, spatial and non-spatial single-cell omics data provide complementary information that, when available across multiple samples and conditions, will provide a holistic definition of *cell state* based on both the intrinsic molecular features of a cell and its set of interactions. Moreover, in the same dataset, we showed differential admixing of immune cell populations that were present in similar proportions across different samples. With the extent of immune infiltration and immune cell type composition entering the clinic as markers of therapeutic response[59,60], it will be interesting to explore the functional and prognostic impact of cell type admixing as well. Overall, cell spatial coordinates introduce a new critical layer of information to decode phenotypic heterogeneity from molecular data. In this context, CellCharter offers a flexible and scalable solution to interpret spatial information and exploit its potential.

## METHODS

### CellCharter: Spatial clustering

Spatial clustering groups spots or cells based on the features of the spot/cell itself, and the features of the surrounding neighbors. CellCharter's approach for spatial clustering involves building a network from the coordinates of the spots/cells, performing dimensionality reduction and batch effect removal. Then, for each cell, aggregating its features with the ones of its neighbors, and finally running clustering on the aggregated matrix.

#### *Spatial network construction*

We represent spatial -omics data as networks, with spatial locations as nodes, connected by an edge if they are in close proximity to each other. Depending on the technology, we used different approaches implemented in the Squidpy library[61] (v1.2.2) to build the network. Given the regular structure of the Visium data, we assigned as neighbors, for each spot, the 6 closest surrounding spots. For the CODEX and CosMx data, we built the network using Delaunay triangulation[62]. However, Delaunay triangulation leads to long edges between cells, especially at the borders of the slide. Hence, we removed edges between nodes at a distance greater than the 99th percentile of all distances between connected nodes.

#### *Dimensionality reduction and batch effect removal*

Spatial transcriptomics allows for measuring thousands of genes simultaneously, thus generating high-dimensionality data. We used scVI[30] to perform dimensionality reduction and batch effect removal of the spatial transcriptomics (10x Visium and Nanostring CosMx) samples. Spatial proteomics data have lower dimensionality, measuring the expression of tens or hundreds of genes. In this case, dimensionality reduction may not be necessary. However, we used it to reduce noise and the running time of clustering. For this purpose, we used scArches[31].

*Neighborhood aggregation*

Incorporating into a spot or cell the features of the neighbors is done by *neighborhood aggregation*. This approach consists of concatenating the features of the spot/cell with features aggregated from neighbors at increasing layers from the considered spot/cell, up to a certain layer $L$. Aggregation functions are used to obtain a single feature vector from the vectors of multiple neighbors. Common aggregation functions are *mean, standard deviation, min*, and *max*. New aggregation functions can be defined according to the relationships with neighbors we want to capture. For example, *standard deviation* can be used to measure the variability of cell phenotypes around a certain spot/cell. Additionally, more than one aggregation feature can be simultaneously used to capture multiple types of relationship, for a total of $J$ aggregations. Given $X$ the matrix of spot/cell features (dimensionality-reduced or not), $A^{(l)}$ for $l \in \{0, \dots, L\}$ the adjacency matrix of the neighbors at layer $l$, $f_j$ for $j \in \{0, \dots, J\}$ the aggregation functions, $(X, Y)$ the matrix concatenation operation, and $X \cdot Y$ matrix inner product operation, then the output of the neighborhood aggregation step is a matrix $Z$:

$$Z = \left(X, f_1\left(A^{(1)} \cdot X\right), \dots, f_J\left(A^{(1)} \cdot X\right), \dots, f_1\left(A^{(L)} \cdot X\right), \dots, f_J\left(A^{(L)} \cdot X\right)\right)$$

*Clustering*

The matrix $Z$ contains information about the spots/cells and their neighbors. Thus, the final step is to perform clustering on $Z$ using a Gaussian Mixture Model (GMM). CellCharter uses an implementation of GMMs of the PyCave library[63] (v3.2.1) that is able to run efficiently on CPU or GPU for multiple samples simultaneously.

*Cluster stability*

Clustering with a GMM requires specifying the desired number of clusters $K$. Thus, CellCharter provides a procedure based on the stability of the clustering to identify the best candidates for $K$. The objective is to identify the values of $K$ for which the clustering result is similar among multiple runs with $K$ clusters and also similar to the runs with $K-1$ and $K+1$ clusters. The user specifies the range of values of $K$ to explore and the maximum number of repeated runs $R$ to perform for each value of $K$. For each $K$ in the specified range, CellCharter executes a single run of clustering with $K$ clusters. At the end of the set of clusterings, the average of the Fowlkes-Mallows Index (FMI)[32] is measured between the clusters at $K-1$ and $K$, and between $K$ and $K+1$. The more stable the clustering with $K$ clusters, the higher the average FMI. Then, a new set of clusterings is performed and the average FMI is computed of all combinations of results between the $r = 2$ runs at $K-1$ and $K$, and at $K$ and $K+1$. The process is repeated until the $r = R$. Moreover, to improve the computational efficiency, it is possible to set that, for each new additional run $r$, if the mean squared deviation between the average FMIs at $r-1$ and $r$ is lower than a user-defined *tolerance* value, then the process has converged and it completes without the need to reach $r = R$. Since the process generates up to $R$ models for every candidate of number of clusters, once a value for $K$ has been chosen, CellCharter selects, out of the $r$ models, the one with the highest marginal likelihood.

<u>CellCharter: Cell type enrichment</u>

Identifying the cell type composition of spatial clusters can aid in their characterization. We implemented cell type enrichment of cell type $t$ in cluster $c$ as the ratio between the proportion of cells of type $t$ in $c$ (observed value) and the proportion of cells of type $t$ in all samples

(expected value). The ratios are then $log_2$-normalized to obtain a zero value if the likelihood of $t$ to be in $c$ is equivalent to random chance, a negative value for depletion, and a positive value for enrichment. In the CODEX mouse spleen dataset, a cell type was considered enriched with a $\log_2$ fold-change greater than 0.58, equivalent to a fold-change of 1.5. In the CosMx NSCLC dataset, a cell type was considered enriched with a $\log_2$ fold-change greater than 1, equivalent to a fold-change of 2.

<u>CellCharter: Neighborhood enrichment</u>

*Analytical neighborhood enrichment*

Given cells partitioned into $K$ groups $C_k$ for $k \in \{1, \dots, K\}$ with adjacency matrix $A$ of the spatial network composed of $V$ nodes and $E$ edges. Be $k_v$ the node degree of cell $v$. The proximity between two groups can be computed from their neighborhood enrichment, which measures how likely are the cells between two groups to be connected by an edge, compared to random chance. We developed an analytical formulation of the neighborhood enrichment between groups $C_i$ and $C_j$, which is computed as the ratio between the observed and the expected number of links connecting cells belonging to $C_i$ and the cells belonging to $C_j$.

Formally, the observed value is:

$$\sum_{v \in C_i, w \in C_j} A_{vw}$$

The expected value is:

$$\sum_{v \in C_i, w \in C_j} k_w \cdot k_w \cdot \frac{1}{2|E|}$$

The final result is a matrix of size $K \times K$.

*Asymmetric neighborhood enrichment*

A symmetric formulation of the neighborhood enrichment fails to capture unbalanced connectivities between groups (**Fig. 1c**). Hence, we developed an asymmetric version of the neighborhood enrichment, that takes into account the proportion rather than the number of edges between two groups. The observed value between groups $C_i$ and $C_j$ is the number of edges between $C_i$ and $C_j$ divided by the total number of edges of the cells of the group $C_i$:

$$\frac{\sum_{v \in C_i, w \in C_j} A_{vw}}{\sum_{v \in C_i} k_v}$$

The expected value is equivalent to the expected value of the symmetric version, divided by the total number of edges of the cells of the group $C_i$:

$$\sum_{w \in C_j} k_w \cdot \frac{1}{2|E|}$$

In this context, the asymmetric neighborhood enrichment is computed as the difference between the observed and expected values, rather than the ratio.

*Differential neighborhood enrichment*

The differential neighborhood enrichment between two conditions is defined as the difference between the neighborhood enrichment matrices of the two conditions. To evaluate if a value is significant, we permute the condition label of each sample and perform bootstrapping multiple times to estimate an associated p-value. The p-value for the differential neighborhood enrichment between groups $C_i$ and $C_j$ is the proportion of cases in which the permuted value is greater than the observed value if the latter is positive and the permuted value is lower than the observed value if the latter is negative. This approach requires computing the neighborhood enrichment several times. The development of an analytical version of the neighborhood enrichment, rather than relying on permutations to calculate the expected value, makes it computationally tractable.

<u>CellCharter: Shape characterization</u>

The same cellular niche can be present in multiple tissues or even in multiple locations of the same tissue sample. This implies the existence of multiple cluster components, which are connected components of cells belonging to the same spatial cluster. To characterize the shape of a spatial cluster, it is necessary to identify its cluster components, determine their boundaries, and compute the value of the metrics that we developed to describe shape (linear, round, circular and irregular). For determining the boundaries of a cluster component, we developed a new technique based on alpha shapes[33]. For each component, we computed the alpha shape with a starting value of alpha that depends on the resolution of the data. If the alpha value was too small, the alpha shape of the component would result in multiple separated polygons. If so, we doubled the alpha value and repeated the computation of the alpha shape. The procedure is iterated until an alpha shape composed of a single polygon is obtained. Finally, the boundaries of the cluster component are the alpha shape with the minimum alpha value that results in a single polygon. To keep shapes simple, holes with an area relative to the boundary area lower than a threshold $a$ (e.g., 0.1) were removed. Thus, a boundary can contain holes with an area greater than $a$ times the area of the boundary. Once we determined the boundary of each component, we used the Shapely library[64] to compute geometric information such as its perimeter $P$, area $A$, its minor and major axes from its minimum rotated rectangle that are used to compute the following four shape metrics.

*Curl*

Curl measures how twisted or curved is a cluster. It is computed as one minus the ratio between the major axis of the minimum rotated rectangle that fits the polygon and the fiber length. Circular and irregular clusters have a high curl.

$$curl = 1 - \frac{\text{major axis length}}{\text{fiber length}} = 1 - \frac{\text{major axis length}}{\frac{4A}{P - \sqrt{P^2 - 16A}}}$$

*Elongation*

If we fit the minimum rectangle surrounding a cluster, a linear cluster will result in a rectangle with the minor edge much shorter than the major edge. Elongation is defined as one minus the ratio between the length of the minor axis and the length of the major axis.

$$elongation = 1 - \frac{\text{minor axis length}}{\text{major axis length}}$$

*Linearity*

We developed a new metric that we called linearity. It measures the ability of a cluster component to approximate a linear path. First, we computed the skeleton of the polygon using a type of skeletonization[65] implemented in the scikit-image library[66]. Then we used the sknw library[67] to obtain a weighted network in which a node is a juncture between lines of the skeleton, and two nodes are linked by an edge if they are connected by a line in the skeleton. The weight of an edge is the length of the line connecting the two junctures. Linear and circular clusters tend to have a skeleton that is composed of a single main axis, with a few short lines branching out of it, while the skeleton of round and irregular clusters has numerous bifurcations of similar length. Thus, linearity is defined as the length of the longest path in the network divided by the total length of the network. To take into account also circular clusters, we include as paths also all the possible cycles which form a basis for cycles in the network, computed using the NetworkX library[62].

$$linearity = \frac{\text{longest path length}}{\text{total skeleton length}}$$

*Purity*

Cells of a cluster may be locally intermixed with cells of other clusters. If the intermixing does not exceed the level at which the cells of the cluster form a connected component, cells of different clusters could be found within the boundaries of a cluster component.

Purity measures the degree of cluster intermixing within a component. Very compact clusters will have a higher purity than sparse and irregular clusters. Thus, given $N$ cells within the borders of a cluster component, for each cluster $k \in \{1, \dots, K\}$, $N_k$ cells are assigned to cluster $k$ such that $\sum_{k \in \{1,\dots,K\}} N_k = N$, then the purity of a cluster component of cluster $\hat{k}$ is defined as:

$$purity = \frac{N_{\hat{k}}}{N}$$

<u>Benchmarking comparison of CellCharter with existing tools</u>

We compared CellCharter against four established methods for spatial clustering: DR.SC (v.2.7), BayesSpace (v.1.4.1), SEDR (commit 18616df), and STAGATE (v1.0.0). The dataset is composed of annotated samples of six-layered human dorsolateral prefrontal cortex, processed using the 10x Genomics Visium platform[34]. The samples are divided into two pairs of adjacent slides from each of the three donors, for a total of 12 slides. Almost every Visium spot was assigned to one of the six cortical layers (L1-L6) or to white matter (WM), for a total of seven labels, with few spots for which a label couldn't be assigned. We evaluated the methods in two settings:

- individual: each sample is clustered and evaluated individually
- joint: all the samples are clustered jointly and evaluated both individually and jointly.

First, we divided the 12 samples into a validation set and a test set. The former was composed of samples 151507, 151672, and 151673, one for each donor, selected randomly and used for hyperparameter tuning. We assigned the remaining 9 samples to the test set. Once the hyperparameters were selected, we run 10 repetitions of spatial clustering of each sample separately and compared the predicted labels of the method against the annotation labels by computing their Adjusted Rand Index (ARI). For the joint benchmarking, we integrated all 9 test samples and performed 10 repetitions of joint clustering on the concatenated dataset. We compared the accordance of the predicted labels against the manual annotations by computing the ARI in two conditions: for each sample separately to assess the variability of the results between different samples; for all samples together, to assess the consistency of the labels across multiple samples. In both settings, spots with the unassigned label were included in the clustering but excluded from the evaluation. Additionally, running time and memory requirements were computed for 5 repetitions of joint clustering for each increasing number of samples, from 2 to 12.

*Hyperparameter tuning*

For each validation sample, through grid search, we run 5 repetitions for every combination of hyperparameters and selected the values associated with the maximum mean Adjusted Rand Index (ARI) across samples on the validation set.

The hyperparameter values candidates and the identified ones are the following:

BayesSpace

- N. highly variable genes [500, 1000, 2000, 5000]: 1000
- N. iterations [2000, 5000]: 5000
- N. principal components [5, 7, 10, 15, 20, 25, 30]: 15
- γ [1, 2, 3, 4]: 3

SEDR

- N. hidden features of last fully connected layer [10, 20, 30]: 10
- N. hidden features of last GCN layer [4, 8, 16]: 16
- N. principal components [100, 200, 300]: 100

STAGATE

- N. highly variable genes [1000, 2000, 5000]: 5000
- N. hidden features [256, 512, 1024]: 1024
- Latent size [5, 10, 30, 50]: 30

DR.SC

- Highly variable genes (HVGs) or spatially variable genes (SVGs): SVGs
- N. spatially variable genes [500, 1000, 2000, 5000]: 2000

CellCharter

- N. highly variable genes [500, 1000, 2000, 5000]: 5000
- Latent size [5, 10, 15]: 5
- N. neighborhood layers [2,3,4]: 4

*Joint benchmarking*

Joint clustering requires the integration of multiple samples into a single dataset. Given the presence of batch effect between patients, all methods for spatial clustering were preceded by a dimensionality reduction and batch effect removal step. CellCharter already uses scVI to perform dimensionality reduction of the spatial transcriptomics data, which was recently shown to be among the best data integration methods for transcriptomics data[68]. We used the same batch effect corrections for the same computational environment: Harmony[69] with default parameters for the R-based methods (BayesSpace); scVI with default parameters and early stopping for the Python-based methods (SEDR, STAGATE, and CellCharter). The R-based package DR.SC does not accept dimensionality-reduced data in input, as it implements its own dimensionality-reduction technique. Thus, we used DR.SC without batch effect removal to evaluate the ARI on the ground truth labels. However, we excluded it from the runtime and memory evaluation.

Spatial clustering of CODEX mouse spleen data

We applied CellCharter to a publicly-available dataset of mouse spleens imaged using the CODEX spatial proteomics technology[5]. The dataset is composed of 3 samples from healthy mice and 6 samples from mice with different stages of systemic lupus erythematosus (SLE). 30 proteins were measured in a total of over 700,000 cells. We used as cell type annotations the ones provided by the authors. First, we removed from the dataset the cells labeled as “dirt” and excluded MHCII from the markers because of inconsistent staining between the two conditions. As the last step of the preprocessing, we computed the network of spatial neighbors for each sample. Then, for each sample, we applied z-score normalization to the markers individually and used scArches[31] (v0.5.3) to perform dimensionality reduction using the trVAE model. We removed the last ReLU layer of the neural network to allow for continuous and real output values. The trVAE model was trained on the dataset using the mean squared error (MSE) loss, two hidden layers of size 128, no MMD, early stopping patience of 5 epochs, and a latent size of 10. We used the latent embeddings of all cells extracted from scArches as features and estimated the best number of clusters using our stability analysis with a k-neighborhood of 3, a range for the number of clusters between 2 and 20, and 10 repetitions. 11 and 4 were the numbers of clusters with the highest Fowlkes-Mallows Index and we chose 11 to have a more fine-grained view of the spatial architecture of the tissues. The model at 11 clusters with the highest marginal likelihood was used for the labeling of the cells, even though we didn’t find striking differences in the labeling between different runs, given the high Fowlkes-Mallows Index (0.78) at 11 clusters. Then, each cluster was labeled based on its cell type enrichment and the location of its cells in the tissue.

*Comparison between CellCharter and STAGATE*

We compared the running time and label assignment on the CODEX mouse spleen dataset between the two best-performing methods according to our benchmark: CellCharter and STAGATE. For STAGATE, given the large number of cells included in the CODEX mouse spleen dataset, it was not possible to fit the whole dataset into the GPU memory. We relied on the batch training strategy of dividing each sample into different subgraphs according to the x and y coordinates and using a subgraph as a batch in the training process. We split every sample into 24 subgraphs, 4 subgraphs based on the x coordinate and 6 subgraphs based on the y coordinate. Given 9 samples, the split resulted in 216 subgraphs. We trained STAGATE using a single hidden layer with default parameters: size = 512, latent size = 10,

learning rate = 0.001, weight decay = 0.0001, number of epochs = 1000, and performed clustering at 11 clusters. On the other hand, CellCharter didn't require any sample splitting. Finally, we compared the running time and label assignment of STAGATE against a single run of CellCharter. For assessing the quality of the labels, since no ground truth of the anatomical area of the cells is available, we relied on the pathologist annotations[5] for a visual comparison.

*Cluster neighborhood enrichment of CODEX mouse spleen data*

We computed the cluster neighborhood enrichment for each condition separately, removing the intra-cluster links to highlight only interactions between different clusters. We then performed differential neighborhood enrichment between healthy and systemic lupus samples. We run 1000 permutations to estimate the significance using a threshold for the p-value of 0.01. All significant pairs of clusters remained significant upon Benjamini-Hochberg correction with an adjusted p-value threshold of 0.05.

*Shape characterization of CODEX mouse spleen data*

For the CODEX dataset, for each spatial cluster, we obtained the cluster components of cells belonging to it as the connected components with a size greater than 250 cells. Then, for each cluster component, we computed the alpha shape with a starting value of alpha equal to 2000 pixels and a minimum hole area ratio $a$ of 0.1.

<u>Spatial clustering of CosMx NSCLC data</u>

We applied CellCharter to a publicly-available dataset of NSCLC samples from the Nanostring image-based spatial transcriptomics CosMx technology[41]. The dataset is composed of 8 samples from 4 lung adenocarcinoma and 1 lung squamous cell carcinoma patient. The 3 slides from patient LUAD-5 were obtained from serial sections, while the 2 slides from patient LUAD-9 were obtained from non-serial sections. The dataset comprises 960 genes measured on more than 750,000 cells. We filtered out genes expressed in fewer than 3 cells and cells with fewer than 3 genes expressed, performed CPM and $\log_2$ normalization. Then, we run dimensionality reduction and batch effect removal using scVI (v1.6.2). We used the default parameters (one hidden layer of size 128, latent size 10, early stopping with patience of 45 epochs), batch effect removal based on the sample, and using the patient as a categorical covariate. We used the latent embeddings of all cells extracted from scVI as features, a k-neighborhood of 3, and estimated the best candidates for the number of clusters using our stability analysis with 10 repetitions and a range for the number of clusters from 2 to 25. It returned 3, 8, and 20 as the best candidates and we chose 20 for downstream analysis.

*Cluster and cell type neighborhood enrichment of CosMx NSCLC data*

Cluster neighborhood enrichment between CellCharter's 20 spatial clusters was run for all 8 samples together. On the other hand, cell type neighborhood enrichment was run for each sample separately. In both cases, we removed intra-cluster (or intra-cell type) links to highlight only interactions between different clusters (or cell types).

*Differential expression analysis of CosMx NSCLC data*

Using Seurat[70], we log-normalized the counts of the original unnormalized data using a scale factor equal to 10000. Then we used MAST[71] to determine the differentially expressed genes between the tumor cells of the spatial clusters C0 and C12. We selected the genes with a $\log_2$

fold-change higher than 0.5 and a p-value lower than 0.05. We performed gene set enrichment analysis of the differentially expressed genes using the GSEA tool[72] (v.4.2.3). Given that the CosMx NSCLC dataset contains gene expression measurements for 960 genes, to avoid a possible bias because of the different gene universe, we used the pre-ranked version, which runs the gene set enrichment analysis against the full list of 960 genes ranked from the highest to the lowest fold-change. We run GSEA against the GO:BP (Gene Ontology Biological Process) and the CP:KEGG (Canonical Pathways KEGG) gene set databases at version 7.5.1.

Cancer cell signature scoring

We visualized the spatial distribution of tumor cells in LUAD-9 expressing different levels of specific signatures. Signatures were constructed by taking the most significant GO:BP and CP:KEGG gene sets obtained from the differential expression analysis between clusters C0 and C12. Additionally, we included signatures of 41 meta-programs determined from scRNA-seq data spanning over 24 tumor types[58] (**Suppl. Table 3)** and we constructed the signature "cell proliferation" by merging the 4 cell cycle-related signatures. Given a gene set, we computed its score for each cell using scanpy's *tl.score_genes* function[73] (v1.9.1). Given the 960 genes measured by the CosMx technology, signatures had a relatively small size (**Suppl. Table 3**). To overcome the noise caused by the small size of the signatures, we smoothed the signature score of each cell by taking the average over the 50 nearest cancer cell neighbors.

Cancer cell signature score dependence on neutrophil distance

We assessed the level of the hypoxia signature score of tumor cells at increasing distance from the neutrophils. To achieve that, we computed the $L$-hop neighbors of all neutrophils of patients LUAD-9 and LUAD-12, with $L$ from 2 to 40. For each $L$-hop adjacency matrix, we selected only the links between neutrophils and tumor cells and used CellCharter's neighborhood aggregation on the hypoxia score to obtain the mean hypoxia score of the tumor cells at distance $L$ from the neutrophils. We executed the procedure for all neutrophils, for the neutrophils in the spatial cluster C11, and for the neutrophils not in C11.

Association between hypoxia and neutrophil infiltration

Bulk gene expression datasets of lung adenocarcinoma primary tumors were downloaded from public repositories reported in the respective studies[11,74–81]. Neutrophil infiltration score was computed with ConsensusTME[82] implemented in the corresponding R package (v0.0.1.9000, parameters: cancer = "LUAD", statMethod = "ssgsea"). The hypoxia signature score for these bulk samples was computed using singscore[83] v1.14.0 with default parameters.

## DATA AVAILABILITY

The raw datasets used are publicly available from their original authors[5,34,41]. The .h5ad files of the processed and dimensionality-reduced data are available at https://github.com/CSOgroup/cellcharter_analyses.

## CODE AVAILABILITY

CellCharter is released as an open-source Python library on Github at https://github.com/CSOgroup/cellcharter. Code to reproduce the analyses is available at https://github.com/CSOgroup/cellcharter_analyses.

## REFERENCES


1. Chappell, L., Russell, A. J. C. & Voet, T. Single-Cell (Multi)omics Technologies. *Annu. Rev. Genomics Hum. Genet.* **19**, 15–41 (2018).
2. Marx, V. Method of the Year: spatially resolved transcriptomics. *Nat. Methods* **18**, 9–14 (2021).
3. Moffitt, J. R., Lundberg, E. & Heyn, H. The emerging landscape of spatial profiling technologies. *Nat. Rev. Genet.* **23**, 741–759 (2022).
4. Moses, L. & Pachter, L. Museum of spatial transcriptomics. *Nat. Methods* **19**, 534–546 (2022).
5. Goltsev, Y. *et al.* Deep Profiling of Mouse Splenic Architecture with CODEX Multiplexed Imaging. *Cell* **174**, 968-981.e15 (2018).
6. Ståhl, P. L. *et al.* Visualization and analysis of gene expression in tissue sections by spatial transcriptomics. *Science* **353**, 78–82 (2016).
7. Callaway, E. M. *et al.* A multimodal cell census and atlas of the mammalian primary motor cortex. *Nature* **598**, 86–102 (2021).
8. Bassiouni, R., Gibbs, L. D., Craig, D. W., Carpten, J. D. & McEachron, T. A. Applicability of spatial transcriptional profiling to cancer research. *Mol. Cell* **81**, 1631–1639 (2021).
9. Hunter, M. V., Moncada, R., Weiss, J. M., Yanai, I. & White, R. M. Spatially resolved transcriptomics reveals the architecture of the tumor-microenvironment interface. *Nat. Commun.* **12**, 6278 (2021).
10. Keren, L. *et al.* A Structured Tumor-Immune Microenvironment in Triple Negative Breast Cancer Revealed by Multiplexed Ion Beam Imaging. *Cell* **174**, 1373-1387.e19 (2018).
11. Tavernari, D. *et al.* Nongenetic Evolution Drives Lung Adenocarcinoma Spatial Heterogeneity and Progression. *Cancer Discov.* **11**, 1490–1507 (2021).

12. Karras, P. *et al.* A cellular hierarchy in melanoma uncouples growth and metastasis. *Nature* **610**, 190–198 (2022).

13. Lomakin, A. *et al.* Spatial genomics maps the structure, nature and evolution of cancer clones. *Nature* **611**, 594–602 (2022).

14. Lundberg, E. & Borner, G. H. H. Spatial proteomics: a powerful discovery tool for cell biology. *Nat. Rev. Mol. Cell Biol.* **20**, 285–302 (2019).

15. Tian, L., Chen, F. & Macosko, E. Z. The expanding vistas of spatial transcriptomics. *Nat. Biotechnol.* 1–10 (2022) doi:10.1038/s41587-022-01448-2.

16. Rao, A., Barkley, D., França, G. S. & Yanai, I. Exploring tissue architecture using spatial transcriptomics. *Nature* **596**, 211–220 (2021).

17. Liu, Y. *et al.* High-Spatial-Resolution Multi-Omics Sequencing via Deterministic Barcoding in Tissue. *Cell* **183**, 1665-1681.e18 (2020).

18. Vickovic, S. *et al.* High-definition spatial transcriptomics for in situ tissue profiling. *Nat. Methods* **16**, 987–990 (2019).

19. Cho, C.-S. *et al.* Microscopic examination of spatial transcriptome using Seq-Scope. *Cell* **184**, 3559-3572.e22 (2021).

20. Chen, A. *et al.* Spatiotemporal transcriptomic atlas of mouse organogenesis using DNA nanoball-patterned arrays. *Cell* **185**, 1777-1792.e21 (2022).

21. Fu, X. *et al.* Polony gels enable amplifiable DNA stamping and spatial transcriptomics of chronic pain. *Cell* **185**, 4621-4633.e17 (2022).

22. Chatzis, S. P. & Tsechpenakis, G. The Infinite Hidden Markov Random Field Model. *IEEE Trans. Neural Netw.* **21**, 1004–1014 (2010).

23. Scarselli, F., Gori, M., Tsoi, A. C., Hagenbuchner, M. & Monfardini, G. The Graph Neural Network Model. *IEEE Trans. Neural Netw.* **20**, 61–80 (2009).

24. Zhao, E. *et al.* Spatial transcriptomics at subspot resolution with BayesSpace. *Nat. Biotechnol.* **39**, 1375–1384 (2021).

25. Liu, W. *et al.* Joint dimension reduction and clustering analysis of single-cell RNA-seq and spatial transcriptomics data. *Nucleic Acids Res.* gkac219 (2022) doi:10.1093/nar/gkac219.

26. Kipf, T. N. & Welling, M. Semi-Supervised Classification with Graph Convolutional Networks. Preprint at https://doi.org/10.48550/arXiv.1609.02907 (2017).

27. Fu, H. *et al.* Unsupervised Spatially Embedded Deep Representation of Spatial Transcriptomics. 2021.06.15.448542 Preprint at https://doi.org/10.1101/2021.06.15.448542 (2021).

28. Dong, K. & Zhang, S. Deciphering spatial domains from spatially resolved transcriptomics with an adaptive graph attention auto-encoder. *Nat. Commun.* **13**, 1739 (2022).

29. Kingma, D. P. & Welling, M. Auto-Encoding Variational Bayes. Preprint at https://doi.org/10.48550/arXiv.1312.6114 (2014).

30. Lopez, R., Regier, J., Cole, M. B., Jordan, M. I. & Yosef, N. Deep generative modeling for single-cell transcriptomics. *Nat. Methods* **15**, 1053–1058 (2018).

31. Lotfollahi, M. *et al.* Mapping single-cell data to reference atlases by transfer learning. *Nat. Biotechnol.* **40**, 121–130 (2022).

32. Fowlkes, E. B. & Mallows, C. L. A Method for Comparing Two Hierarchical Clusterings. *J. Am. Stat. Assoc.* **78**, 553–569 (1983).

33. Edelsbrunner, H., Kirkpatrick, D. & Seidel, R. On the shape of a set of points in the plane. *IEEE Trans. Inf. Theory* **29**, 551–559 (1983).

34. Maynard, K. R. *et al.* Transcriptome-scale spatial gene expression in the human dorsolateral prefrontal cortex. *Nat. Neurosci.* **24**, 425–436 (2021).

35. Ramiscal, R. R. & Vinuesa, C. G. T-cell subsets in the germinal center. *Immunol. Rev.* **252**, 146–155 (2013).

36. Pusztaszeri, M. P., Seelentag, W. & Bosman, F. T. Immunohistochemical Expression of Endothelial Markers CD31, CD34, von Willebrand Factor, and Fli-1 in Normal Human Tissues. *J. Histochem. Cytochem.* **54**, 385–395 (2006).

37. Greaves, M. & Maley, C. C. Clonal evolution in cancer. *Nature* **481**, 306–313 (2012).

38. Yuan, S., Norgard, R. J. & Stanger, B. Z. Cellular Plasticity in Cancer. *Cancer Discov.* **9**, 837–851 (2019).

39. Swanton, C. Intratumor Heterogeneity: Evolution through Space and Time. *Cancer Res.* **72**, 4875–4882 (2012).

40. Black, J. R. M. & McGranahan, N. Genetic and non-genetic clonal diversity in cancer evolution. *Nat. Rev. Cancer* **21**, 379–392 (2021).

41. He, S. *et al.* High-plex Multiomic Analysis in FFPE at Subcellular Level by Spatial Molecular Imaging. 2021.11.03.467020 Preprint at https://doi.org/10.1101/2021.11.03.467020 (2022).

42. Wang, Y. *et al.* N-myc downstream regulated gene 1(NDRG1) promotes the stem-like properties of lung cancer cells through stabilized c-Myc. *Cancer Lett.* **401**, 53–62 (2017).

43. Ma, J., Gao, Q., Zeng, S. & Shen, H. Knockdown of NDRG1 promote epithelial–mesenchymal transition of colorectal cancer via NF-κB signaling. *J. Surg. Oncol.* **114**, 520–527 (2016).

44. Zhu, H. & Zhang, S. Hypoxia inducible factor-1α/vascular endothelial growth factor signaling activation correlates with response to radiotherapy and its inhibition reduces hypoxia-induced angiogenesis in lung cancer. *J. Cell. Biochem.* **119**, 7707–7718 (2018).

45. Rajarathnam, K., Schnoor, M., Richardson, R. M. & Rajagopal, S. How do chemokines navigate neutrophils to the target site: Dissecting the structural mechanisms and signaling pathways. *Cell. Signal.* **54**, 69–80 (2019).

46. Behrens, C. *et al.* EZH2 Protein Expression Associates with the Early Pathogenesis, Tumor Progression, and Prognosis of Non–Small Cell Lung Carcinoma. *Clin. Cancer Res.* **19**, 6556–6565 (2013).

47. Blaisdell, A. *et al.* Neutrophils Oppose Uterine Epithelial Carcinogenesis via Debridement of Hypoxic Tumor Cells. *Cancer Cell* **28**, 785–799 (2015).

48. Yee, P. P. *et al.* Neutrophil-induced ferroptosis promotes tumor necrosis in glioblastoma progression. *Nat. Commun.* **11**, 5424 (2020).

49. Su, H. *et al.* Identification of hub genes associated with neutrophils infiltration in colorectal cancer. *J. Cell. Mol. Med.* **25**, 3371–3380 (2021).

50. scverse. https://scverse.org/.

51. Deng, Y. *et al.* Spatial-CUT&Tag: Spatially resolved chromatin modification profiling at the cellular level. *Science* **375**, 681–686 (2022).

52. Zhao, T. *et al.* Spatial genomics enables multi-modal study of clonal heterogeneity in tissues. *Nature* **601**, 85–91 (2022).

53. Deng, Y. *et al.* Spatial profiling of chromatin accessibility in mouse and human tissues. *Nature* **609**, 375–383 (2022).

54. Vickovic, S. *et al.* SM-Omics is an automated platform for high-throughput spatial multi-omics. *Nat. Commun.* **13**, 795 (2022).

55. Ben-Chetrit, N. *et al.* Integration of whole transcriptome spatial profiling with protein markers. *Nat. Biotechnol.* 1–6 (2023) doi:10.1038/s41587-022-01536-3.

56. Chen, R. J. *et al.* Scaling Vision Transformers to Gigapixel Images via Hierarchical Self-Supervised Learning. in 16144–16155 (2022).

57. Sikkema, L. *et al.* An integrated cell atlas of the human lung in health and disease. 2022.03.10.483747 Preprint at https://doi.org/10.1101/2022.03.10.483747 (2022).

58. Gavish, A. *et al.* The transcriptional hallmarks of intra-tumor heterogeneity across a thousand tumors. 2021.12.19.473368 Preprint at https://doi.org/10.1101/2021.12.19.473368 (2021).

59. Bruni, D., Angell, H. K. & Galon, J. The immune contexture and Immunoscore in cancer prognosis and therapeutic efficacy. *Nat. Rev. Cancer* **20**, 662–680 (2020).

60. Howard, R., Kanetsky, P. A. & Egan, K. M. Exploring the prognostic value of the neutrophil-to-lymphocyte ratio in cancer. *Sci. Rep.* **9**, 19673 (2019).

61. Palla, G. *et al.* Squidpy: a scalable framework for spatial omics analysis. *Nat. Methods* **19**, 171–178 (2022).

62. Hagberg, A., Swart, P. & S Chult, D. *Exploring network structure, dynamics, and function using networkx*. https://www.osti.gov/biblio/960616 (2008).

63. Borchert, O. PyCave. https://github.com/borchero/pycave.

64. Shapely. https://github.com/shapely/shapely.

65. Zhang, T. Y. & Suen, C. Y. A fast parallel algorithm for thinning digital patterns. *Commun. ACM* **27**, 236–239 (1984).

66. Walt, S. van der *et al.* scikit-image: image processing in Python. *PeerJ* **2**, e453 (2014).

67. sknw. https://github.com/Image-Py/sknw.

68. Luecken, M. D. *et al.* Benchmarking atlas-level data integration in single-cell genomics. *Nat. Methods* **19**, 41–50 (2022).

69. Korsunsky, I. *et al.* Fast, sensitive and accurate integration of single-cell data with Harmony. *Nat. Methods* **16**, 1289–1296 (2019).

70. Satija, R., Farrell, J. A., Gennert, D., Schier, A. F. & Regev, A. Spatial reconstruction of single-cell gene expression data. *Nat. Biotechnol.* **33**, 495–502 (2015).

71. Finak, G. *et al.* MAST: a flexible statistical framework for assessing transcriptional changes and characterizing heterogeneity in single-cell RNA sequencing data. *Genome Biol.* **16**, 278 (2015).

72. Subramanian, A. *et al.* Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci.* **102**, 15545–15550 (2005).

73. Wolf, F. A., Angerer, P. & Theis, F. J. SCANPY: large-scale single-cell gene expression data analysis. *Genome Biol.* **19**, 15 (2018).

74. Collisson, E. A. *et al.* Comprehensive molecular profiling of lung adenocarcinoma. *Nature* **511**, 543–550 (2014).

75. Shedden, K. *et al.* Gene expression–based survival prediction in lung adenocarcinoma: a multi-site, blinded validation study. *Nat. Med.* **14**, 822–827 (2008).

76. Schabath, M. B. *et al.* Differential association of STK11 and TP53 with KRAS mutation-associated gene expression, proliferation and immune surveillance in lung adenocarcinoma. *Oncogene* **35**, 3209–3216 (2016).

77. Okayama, H. *et al.* Identification of Genes Upregulated in ALK-Positive and EGFR/KRAS/ALK-Negative Lung Adenocarcinomas. *Cancer Res.* **72**, 100–111 (2012).

78. Der, S. D. *et al.* Validation of a Histology-Independent Prognostic Gene Signature for Early-Stage, Non–Small-Cell Lung Cancer Including Stage IA Patients. *J. Thorac. Oncol.* **9**, 59–64 (2014).

79. Chen, J. *et al.* Genomic landscape of lung adenocarcinoma in East Asians. *Nat. Genet.* **52**, 177–186 (2020).

80. Mezheyeuski, A. *et al.* Multispectral imaging for quantitative and compartment-specific immune infiltrates reveals distinct immune profiles that classify lung cancer patients. *J. Pathol.* **244**, 421–431 (2018).

81. Ding, L. *et al.* Somatic mutations affect key pathways in lung adenocarcinoma. *Nature* **455**, 1069–1075 (2008).

82. Jiménez-Sánchez, A., Cast, O. & Miller, M. L. Comprehensive Benchmarking and Integration of Tumor Microenvironment Cell Estimation Methods. *Cancer Res.* **79**, 6238–6246 (2019).

83. Foroutan, M. *et al.* Single sample scoring of molecular phenotypes. *BMC Bioinformatics* **19**, 404 (2018).