Computational solutions for spatial transcriptomics

Iivari Kleino,1,⇑, Paulina Frolovaitė,1 Tomi Suomi,1 Laura L. Elo,1,⇑

1Turku Bioscience Centre, University of Turku and Åbo Akademi University Turku, Turku, Finland
2Institute of Biomedicine, University of Turku, Turku, Finland

A R T I C L E   I N F O

Article history:
Received 22 April 2022
Received in revised form 18 August 2022
Accepted 18 August 2022
Available online 1 September 2022

Keywords:
Spatial transcriptomics
Computational biology
Spatial data analysis frameworks
Single-cell analysis
Spatial deconvolution

A B S T R A C T

Transcriptome level expression data connected to the spatial organization of the cells and molecules would allow a comprehensive understanding of how gene expression is connected to the structure and function in the biological systems. The spatial transcriptomics platforms may soon provide such information. However, the current platforms still lack spatial resolution, capture only a fraction of the transcriptome heterogeneity, or lack the throughput for large scale studies. The strengths and weaknesses in current ST platforms and computational solutions need to be taken into account when planning spatial transcriptomics studies.

The basis of the computational ST analysis is the solutions developed for single-cell RNA-sequencing data, with advancements taking into account the spatial connectedness of the transcriptomes. The scRNA-seq tools are modified for spatial transcriptomics or new solutions like deep learning-based joint analysis of expression, spatial, and image data are developed to extract biological information in the spatially resolved transcriptomes. The computational ST analysis can reveal remarkable biological insights into spatial patterns of gene expression, cell signaling, and cell type variations in connection with cell type-specific signaling and organization in complex tissues.

This review covers the topics that help choosing the platform and computational solutions for spatial transcriptomics research. We focus on the currently available ST methods and platforms and their strengths and limitations. Of the computational solutions, we provide an overview of the analysis steps and tools used in the ST data analysis. The compatibility with the data types and the tools provided by the current ST analysis frameworks are summarized.

© 2022 The Authors. Published by Elsevier B.V. on behalf of Research Network of Computational and Structural Biotechnology. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

Contents

1. Introduction .......................................................................................................................... 4871
2. In situ spatial transcriptomics methods ................................................................................. 4872
  2.1. Strengths and limitations .............................................................................................. 4872

Abbreviations: AOI, area of illumination; Baysor, Bayesian Segmentation of Spatial Transcriptomics Data; BICCN, Brain Initiative Cell Census Network; BinSpect, Binary Spatial Extraction; BOLORAMIS, barcoded oligonucleotides ligated on RNA amplified for multiplexed and parallel in situ analyses; CCC, cell–cell communication; CCL, cell–cell interactions; CNV, copy-number variation; Dbit-Seq, Deterministic Barcoding in Tissue for spatial omics sequencing; DSP, digital spatial profiling; FA, factor analysis; FFPE, formalin-fixed, paraffin-embedded; FISH, fluorescence in situ hybridization; FISSSEQ, fluorescence in situ sequencing of RNA; FOV, Field of view; GRNs, gene regulation networks; GSEA, gene set enrichment analysis; GSVA, gene set variation analysis; HDST, high definition spatial transcriptomics; HMRF, hidden Markov random field; ICC, interaction changed genes; ISH, in situ hybridization; ISS, in situ sequencing; JSTA, Joint cell segmentation and cell type annotation; KNN, k-nearest neighbor; LCM, Laser Capture Microdissection; LCM-seq, laser capture microdissection coupled with RNA sequencing; LOH, loss of heterozygosity analysis; MC, Molecular Cartography; MERFISH, multiplexed error–robust FISH; NMF (NNMF), Non-negative matrix factorization; PCA, Principal Component Analysis; PIXEL-seq, Polony (or DNA cluster)-indexed library-sequencing; PE-lig, padlock ligation; QC, quality control; RNASeq, RNA sequencing; ROI, region of interest; SCENIC, Single-Cell Regulatory Network Inference and Clustering; scRNA-seq, single-cell RNA sequencing; scvi-tools, single-cell variational inference tools; seqFISH, sequential fluorescence in situ hybridization; sequ-smFISH, sequential single-molecule fluorescent in situ hybridization; SME, Spatial Morphological gene Expression normalization; smFISH, single molecule FISH; SPATA, SPAtial Transcriptomic Analysis; ST, Spatial Transcriptomics; ST Pipeline, Spatial Transcriptomics Pipeline; STARmap, spatially-resolved transcript ampcle readout mapping; t-SNE, t-distributed stochastic neighbor embedding; TIVA, Transcriptome in Vivo Analysis; TMA, tissue microarray; TME, tumor microenvironment; UMAP, Uniform Manifold Approximation and Projection for Dimension Reduction; UMI, unique molecular identifier; ZipSeq, zipcoded sequencing.

⇑ Corresponding author.

E-mail addresses: iivari.kleino@utu.fi (I. Kleino), laura.elo@utu.fi (L.L. Elo).

https://doi.org/10.1016/j.csbj.2022.08.043
2001-0370/© 2022 The Authors. Published by Elsevier B.V. on behalf of Research Network of Computational and Structural Biotechnology. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).
1. Introduction

To decipher the functions of the cellular systems and organs in multicellular organisms, one needs detailed information about the components and interactions between them at every scale level. In a homeostatic state, the cellular systems are in dynamic equilibrium. However, perturbations like immunological challenges (e.g., pathogens, autoimmune reactions), can flip the system into different states. While single-cell RNA-sequencing (scRNA-seq) provides a homeostatic state, the cellular systems are in dynamic equilibrium. However, perturbations like immunological challenges (e.g., pathogens, autoimmune reactions), can flip the system into different states. While single-cell RNA-sequencing (scRNA-seq) provides detailed information about the gene expression profiles and their heterogeneity in dissociated cells, spatial transcriptomics (ST) links the transcriptomes to their cellular locations providing spatial context. This information can be used to lay out a map of the possible connections between cells, factors affecting the cells (e.g., signaling molecules, available nutritional resources, or pathogens), and connections to the other systems or organs at the organism level. These spatial relationships can reveal how the different cell types and genetic programs are interconnected with each other and with the surrounding environment. Thus, various ST methods integrated with the other spatial and non-spatial methods are currently paving the way to a deep understanding of the structure and function of the living organisms at the system level.

Spatial transcriptomics derives from tissue in situ hybridization techniques detecting single mRNA species with DNA-oligo probes. In current highly multiplexed in situ hybridization ST methods, thousands of mRNA species are detected simultaneously by using clever probing strategies. Currently, the widest probe sets detect expressions of over 18 000 known protein-coding genes, and the mRNA capture and sequencing-based methods can detect different mRNA species in an untagged manner (including, but not limited to mRNAs, splice variants, lincRNAs, antisense RNAs, and structural RNAs). The transcriptomes can be analyzed from different sample types, including slices cut from freshly frozen and formalin-fixed, paraffin-embedded (FFPE) tissues. Each ST platform has its strengths and limitations, and therefore it is important to consider which platforms or their combinations are best for given research. The history of the spatial transcriptomics development is reviewed for example in Moses et al. 2022 [1] and Asp et al. [2].

The spatial transcriptomics methods can be categorized into three main approaches on the basis of how they capture and store spatial information: in situ methods, spatial DNA-barcoding, and regional selection spatial transcriptomics methods (Table 1). In situ methods contain both the spatial information and the identity of the transcripts in the acquired images of the samples labeled with fluorescent probes [3,4] or displaying in situ sequencing (ISS) signals [5,6]. These methods provide a location for each individual detected transcript at the single-molecule level, and the expression profiles at the subcellular, cellular, or regional levels are decoded from the collected image data (Fig. 1A).

Spatial barcoding (DNA-barcoding) is a collection of methods that use mRNA capturing barcoded DNA-oligos to incorporate the positional information in a DNA format with the transcript sequences (Fig. 1B-C). The methods are based on DNA-oligos with known barcoded capture regions that form either a solid surface spot array or bead array to capture the mRNAs diffused from the tissue sample [7–10]. Alternatively, the barcoded capture-oligos are injected on the sample from arrayed inlets in a custom chip [10,11]. The positional information as DNA-barcodes is attached to the copies of the transcripts in an enzymatic reaction on the mRNA capture sites. The spatial origins of transcripts are resolved from the transcript-spatial-barcode DNA-libraries by sequencing and mapping the transcripts to the spot-level spatial coordinates with the DNA-barcodes. The size and shape of the identifiable capture region depend on the used methodology and the smallest spots in high-density arrays capture transcripts at subcellular resolution (Fig. 1B). However, these are not yet generally available, and the spots in current commonly used spatial barcode arrays cover the area of more than one cell (Fig. 1C). The transcripts and cell profiles at spots can be visualized on the image of the sample as a sample/spot image (tissue image). Computational methods are then used to resolve spot or region-level cell compositions and single or subcellular level transcriptomes from the data.

Alternative to spot spatial barcoding are regional illumination and collection methods (Fig. 1D). Region selection barcoding is accomplished by the use of photoactivatable markers and regional photo-activation with laser illumination. In different approaches, the transcripts or cells on the target regions are tagged with barcode-oligos (TIVA tag, Zip-Seq) [12,13], which requires sequential labeling of each targeted region and, therefore, strongly limits the feasible number of samples. On the other hand, these methods
Spatial transcriptomics methods produce data at different spatial resolutions. The laser capture microdissection (LCM) has also been used to capture the individual cells or regions from the samples for LCM coupled RNA-sequencing (LCM-seq) [15,16]. Similar to GeoMx, transcripts from each collected cell or cell group are barcoded before sequencing or they can be processed independently as single sequencing libraries [17].

Already now, the insights gleaned with the developing ST platforms have shown the value of spatial transcriptomics in learning detailed information about the heterogeneity of the cells and inter-connectedness of the transcriptomes in multicellular systems. However, the availability of a myriad of platforms to produce large and complex omics datasets raise the requirement of advanced computational tools and skills needed to analyze the data. Careful experimental design and data analysis planning can streamline the production of high-quality findings. The recent reviews on spatial transcriptomics and multomics [1,27–32] cover the history, technology, and advances in the analysis in more detail; this review focuses on covering ways to help in choosing the suitable platform and analysis framework for spatial transcriptomics studies. We focus on the currently widely (and commercially) available platforms, their limitations, the available data analysis tools, and their suitability for the collection of different types of data. The state-of-the-art systems, not commercially available, are mentioned when appropriate and when expected to improve the methodology in the near future. The current computational ST analysis steps and frameworks are summarized, supplemented with optional standalone analysis options.

### 2. In situ spatial transcriptomics methods

#### 2.1. Strengths and limitations

The sequential single-molecule fluorescent in situ hybridization (sequ-smFISH) and the in situ sequencing (ISS) offer the highest spatial resolution among the ST methods by detecting individual transcript molecules in the samples at the optical resolution of the microscope system (Table 1, Fig. 1A). Of the ST methods, the sequ-smFISH methods have also the highest sensitivity, as they can detect transcripts with over 80% efficiency even when thousands of RNAs are targeted [3,6,33–35]. However, the in situ ST
methods are not quite yet detecting different mRNAs at transcriptome coverage level like the spatial barcoding ST methods based on polyA targeting mRNA capture.

A particular strength of in situ ST methods is their exceptional resolution and capture efficiency. The high capture efficiency allows reliable detection of even very lowly expressed transcripts. A targeted design of probe sets focusing on, for example, signaling pathways or transcription and genomic imprinting regulation has proven a great strategy with in situ ST. A prime example of this is a study of the early T cell progenitor development [36], in which Zhou et al. designed a probe set to analyze the expression of otherwise elusive transcription factors and master regulators expressed in progenitor T cells. In combination with a scRNA-seq, the study was able to provide insight into the synchronized and asynchronous developmental patterns in gene expressions during the early T cell development. The study also elegantly demonstrated the feasibility of sequential use of seqFISH, smFISH, and immunostaining on the same sample. However, it should be noted that the signals from transcripts degrade each FISH cycle [21] setting limits to the number of successive probing rounds and sets the preferential probing order to start from the weakest expressing transcripts. Indicating the power of integration of in situ ST and scRNA-seq methods, seqFISH was used to identify the exact location of scRNA-seq characterized cell types in mouse organogenesis [37].

The Brain Initiative Cell Census Network (BICCN), a collaborative effort to produce reference brain transcriptome atlases is showing in situ ST use in a very large-scale cell mapping project [38].

MERFISH and seqFISH have a single-molecule level resolution which allows analysis of subcellular transcript distributions at the single-cell level adding on top of single-cell transcriptome profiling. The spatial distribution can be used to detect single-cell level features like mRNA targeting, cell polarization, and localization of specific molecular complexes [4]. For functional single-cell transcriptome analysis, the nuclear to cytoplasmic ratio of immature and mature transcripts can be used to infer the regulatory states of genes at the single-cell level through the so-called RNA velocity analysis [33]. RNA velocity data can be used, for example, to construct detailed transcript regulatory profiles and to organize cells to pseudotime trajectories to analyze, for example, cell cycle progression [33].

Showing the use of sequential smFISH at the molecular complex level, Takei et al. resolved the functional architecture of cell nucleus in detail by adapting seqFISH+ method to co-detect localization of 3660 chromosomal loci with 17 functional chromosomal markers, and 70 selected mRNA species [39]. The study indicated a good general correlation between gene activity, localization to different functional nuclear zones, and association of functional chromosomal markers. However, at the single-cell level discordant localization of individual genes with active nuclear zones and the actual transcriptional activity of the genes suggested slower temporal dynamics for a chromosomal spatial organization than for the gene activity regulation.

A major limitation with the in situ ST methods is that the optical crowding poses a problem with very high-plexed probe sets as the detection efficiency degrades along with the number of the target mRNAs. This technological limitation is particularly noticeable with ISS methods as the molecular amplification used for signal enhancement increases the optical size of the targets [5,6,18]. Detection of transcriptome level number of targets would require resolving up to hundreds of thousands of transcripts per cell. The current state-of-the-art seqFISH+ with 60 pseudocolor probe set combined with a clever transcript encoding scheme, optimized sequential probing, and adoption of confocal microscope platform for detection allowed detection of over 30 000 individual transcripts with ~10 000 genes in cultured fibroblasts [4].

The number of simultaneously detected RNA targets in in situ ST is not limited only by the physical limitations set by the used microscopes. The sensitivity of smFISH is based on the number of in situ hybridization probes (ISH) targeting a given gene. For example, seqFISH+ needs the target RNA length to be at least 1 kb to accommodate the in situ hybridization probes for sensitive detection and encoding. This leaves out for example short mRNAs and many other short RNA species including miRNAs from the target repertoire. In MERFISH probes, the use of branched-DNA increased the detection sensitivity of MERFISH technology, and it can now detect even short 100–200 nucleotide long RNA targets with up to 60% efficiency and long RNAs with almost perfect efficiency [20]. This also allows the detection of a dramatically wider variety of RNA species including mRNA alternative splice variants. Remarkably, the increased sensitivity did not increase the optical size of the fluorescent pixel and with adoption of expansion microscope with more sensitive probes Xia et al. were able to measure the expression of around ~10 000 different RNAs simultaneously in cultured cells [33].

The ISS-based ST methods can detect very short RNA species, and the recently published new padlock ligation (PL-lig) based smFISH/ISS ST method BOLORAMIS can detect even miRNAs and single point mutations in cellular RNAs [18]. The sensitivity and specificity of BOLORAMIS and STARmap allow the detection of even different alternative splice forms of transcripts. However, the use of probes to cover transcripts in an untargeted manner is not feasible due to the cost of probes and without a technological leap in microscopy or probe design, the methodology is still limited by the optical crowding. The reverse transcription capture-based ISS methods [40,41] can detect RNA molecules in an untargeted manner but due to very low capture efficiency, they detect only a low number of transcripts per cell compared to other ST methods (Table 1).

The in situ ST methods are limited also by the speed of the imaging. In both ISS and sequential smFISH, the transcripts are identified by sequentially imaging fluorescence signals using a microscope. Each cycle involves FISH probing or re-probing and imaging steps. The speed of imaging of the multiple FISH cycles with a high-resolution microscope is thus a bottleneck limiting the sample and detection area throughput with the in situ ST methods. The use of 60 pseudocolours for encoding transcripts in seqFISH+ lowered the needed FISH cycles to four (with the included error correction cycle) for theoretical detection of the expression of up to 24 000 genes [4]. This methodological advancement shortened the sample processing time to 1/8th (imaging and hybridization) from the original in seqFISH [34]. Commercial platforms with automated liquid handling and imaging and with the use of optimized probe sets for in situ ST are expected to improve the throughput and increase the availability of in situ ST for researchers.

### 2.2. Specific data processing steps with ST

Deriving cell-by-gene and cell-by-location matrices for ST analysis from the large set of raw in situ microscopy image data takes several computational steps. For example, in a 69-bit 10 000 gene MERFISH experiment, for each of 256 tiled fields of views (FOV), 23 rounds of three color fluorescent images were taken at six different focal z-planes with an additional single image at fiducial bead z-plane [33]. This is 111 872 single-channel microscope images to process into a form that can be used to extract 69-bit transcript codes and locations to construct cellular transcriptome profiles. This section covers the in situ ST-specific data analysis steps from the raw image data to the spatial cell type pattern analysis. The more advanced analysis steps commonly used with many ST dataset types are covered in Section 5.
2.2.1. Preprocessing and spot registration

The probed transcripts show in raw images as signal spots that, due to limitations in automated microscopy imaging and due to chromatic aberrations, are not necessarily in the same position (in register) in the sequentially imaged FOVs (Fig. 2A). Therefore, the images in each FOV are aligned using cross-correlation of the fiducial marker peak signals or nuclear stains (Fig. 2B) [46,42]. Usually, the signal aligning and enhancing steps also include chromatic aberration and illumination correction with microscope setup specific control images, image deconvolution, and background subtraction [43]. For a composite representation of the whole imaged sample, the single or multi-channel FOVs can be stitched together with a process guided by the overlaps in the FOV tiles [28,44].

In aligned images, each positive spot/pixel/voxel is a potential transcript, with a location in the pixel coordinate system and transcript identifier encoded by the levels of the sequential image channels. The transcript spots are called by finding the local maxima of the images and by selecting the values that are above a certain pixel threshold [45]. The barcodes are then extracted as per location signal strings for decoding (Fig. 2D) [3,42]. Details of the above steps and the barcode decoding with error correction and spot quality control (QC) heavily depend on the used microscope setup and in situ ST method (MERFISH, seqFISH or ISS). Readers interested in these details are referred to the procedures explained in the original research referenced in Table 1. The decoded transcripts are arranged into a gene-by-location matrix containing the identifiers and 2D or 3D location of every identified transcript in the used coordinate system.

2.2.2. Spatial segmentation

To construct single-cell or other spatial unit transcript profiles in the sample, segmentation and counting steps are performed to assign the detected transcripts into desired functional spatial units. Segmentation masks defining coordinates of the spatial units are done by tracing the shapes of the targeted features from the acquired images or algorithmically with the spatial distribution of the transcripts (Fig. 2C). Reflecting the challenges in the segmentation of different biological samples, several different computational methods and strategies have been developed for segmentation. These start from simple manual segmentation by marker thresholding and end with various automated machine learning utilizing strategies [43]. In a standard cell segmentation, the specific markers are used to divide the FOV area into nuclear, cytoplasmic, and empty regions. The individual cells are delineated by selecting the nuclei of each cell as a center and then propagating the cytoplasmic area around the nuclei. Different computational methods are employed to estimate the extent and shape of cell boundaries mathematically or by using marker signals or both. The various machine learning-based segmentation tools, like U-net [46], DeepCell [47], Mesmer [48], and CellPose [49], use ST images and pre- or post-stain marker data to predict spatial segmentation [50]. In cases where only transcript signals are present, the cell segmentation or segmentation-free cell transcriptome identification relies on algorithms that use spatial transcript distributions [51] or annotated transcript profiles are used in combination with the ST data to jointly segment and annotate the cell types, as in the recent methods SSAM [51], Baysor [52], and JSTA [53]. Finally, the identified transcripts within the spatial unit regions are counted based on the segmentation masks to construct the cell-by-gene and cell-by-location matrices for ST data analysis (Fig. 2D).

The image processing, segmentation, decoding, and counting steps can be performed by scripting with Python, R, or MatLab image processing packages or modules. The Starfish [54] and its fork SMART-Q with multomics capabilities collect python tools and scripts for building pipelines to process images and get the cell-by-gene and cell-by-location matrices from raw microscope images of different in situ ST methods. Other options offering useful functionalities for image processing, spot calling, and cell segmentation include PySpots [53,55], Cellpose [49], and FISH-quant [56] in python, and EBImage [57] and imager [58] in R. Of the complete ST data analysis frameworks Squidpy [59] has cell and nuclei segmentation capabilities available (Table 2).

3. Spatial barcoding methods

3.1. Strengths and limitations

Spatial barcoding ST methods are based on the limited diffusion of the target RNA molecules and their capture by the position-indicating barcoded DNA-oligos (Fig. 1B-C). Compared to in situ ST methods, spatial barcoding ST excels in availability, simplicity, and untargeted mRNA capture1. Nonetheless, it is behind in resolution and capture efficiency compared to many in situ ST methods (Table 1). Both the resolution and capture efficiency in spatial barcoding ST methods have increased gradually since early spot arrays [23] and custom microarrays [63], in which the center-to-center distance was in the 200 to 100-micron range. In these first-generation arrays, each “spot” captured transcripts from multiple cells, which complicated the cell and spatial data analysis.

---

1 New strategies are developed to widen the functional RNA species being captured and sequenced in scRNA-seq analysis. [60–62].
Recently, the spatial resolution in spatial barcoding arrays has increased to a level of single-cell size and beyond. In Slide-seq, this was obtained by micro ball-arrays [8,64] and, in DbiT-Seq, by chips with microchannels injecting the spatial barcode oligos [10]. The spatial barcoding reached the subcellular resolution with high-definition spatial transcriptomics (HDST) oligo arrays [9]. The highest resolution and the best capture efficiency have so far been achieved with Seq-Scope, Stereo-Seq, and Pixel-seq [7,24,25]. These recent high-resolution and high-density DNA-oligo spatial barcoding arrays not only increased the resolution to the submicron level, but the mRNA capture efficiency has also improved to a comparable level with the droplet-based scRNA-seq methods (Table 1). This allows robust untargeted detection of medium expression level genes at the single-cell level and even between the cellular compartments.

The HDST and Seq-Scope studies demonstrated that the high-resolution arrays can locate even rare cell types and resolve the gene expression differences at subcellular resolution, which makes for example the nuclear-to-cyttoplasmic type of RNA-velocity of analysis feasible for spatial barcode ST methods [9,25]. The preprints showcase the capabilities and limitations of Stereo-Seq with tumor leading-edge samples [65] and high-resolution spatial transcriptome atlases produced from regenerating axolotl brains and the developmental stages of the mouse, zebrafish, and fruit fly embryos [7,66–68]. Already these detailed transcriptome atlases, after being released into the public domain, could provide a rich resource of transcriptomic data to analyze the developmental process and brain regeneration of multicellular organisms. It will also be a great resource for testing and developing bioinformatics, data storage, and data handling methods for the analysis of such large spatial transcriptomics datasets.

The high-density spatial barcoding arrays are a big developmental step towards a universal method in ST, even though the Seq-Scope, Stereo-seq, Pixel-seq, HDST, and DbiT-Seq are still behind in resolution and capture efficiency compared to the in situ ST methods. However, widespread adoption of high-density spatial barcoding arrays may be bottlenecked by the array production and array indexing as these require special techniques like customized use of Illumina sequencer [7], custom microchip production [10], or highly optimized arraying of oligos on to a polyacrylamide gel matrix [24]. Even though the state-of-the-art spatial barcoding methods are not yet widely available, they have shown that the fast pace in the development of ST is still going on strong and that by optimizing spatial barcoding, it is possible to acquire untargeted high gene coverage subcellular resolution data.

The latest high-density spatial barcode arrays are not yet generally available for researchers, and hence only pioneering studies have been published. Nonetheless, the spatial barcoding methods with lower resolution arrays have been available for some time and several studies have indicated their power for biological discoveries. The lower resolution spatial barcoding methods have been successfully used to dissect regional gene expression profiles in homeostatic tissues and developmental settings. For example, Slide-seq was used to produce an ST atlas of the mouse testis to decipher the complex organization of mammalian spermatogenesis at an unprecedented level [69]. The spatial transcriptomics with different pattern detection algorithms has allowed a detailed characterization of the normal gut functions and the functions in inflammatory disease [70,71]. The spatial transcriptomics has also revealed many novel features in various tumor samples [63,72–75]. These studies have, for example, revealed regional enrichment patterns for cancer cell subtypes and co-enrichments of cancer

Table 2

| Package       | Giotto | Seurat | STUtility | SPATA2 | Squidpy | scvi-tools | stLearn | GeoMx tools |
|---------------|--------|--------|-----------|--------|---------|------------|---------|-------------|
| Platforms     | R/Python | R | R | R | Python | Python | Python | R |
| Datacontainer | ig,mt,lc | ig,mt,lc | ig,mt,lc | ig,mt,lc | SPATA | Adt, img | Adt | gmx |
| ST data types | is,sb | is,sb | sb | sb | is,sb | is,sb | is,sb | is,sb |
| Spatial segmentation | Giotto | Seurat | Seurat | SPATA | | | |
| Nuclei count | | | | | | | | |
| QC and preprocessing | | | | | | | | |
| Descriptive statistics | | | | | | | | |
| Dimensionality reduction | | | | | | | | |
| Cell/spot clustering | | | | | | | | |
| Data visualizations | | | | | | | | |
| Factor analysis | | | | | | | | |
| Differential expression | | | | | | | | |
| Cell type annotation | | | | | | | | |
| Deconvolution | | | | | | | | |
| Reverse deconvolution | | | | | | | | |
| Cell type signature inference | | | | | | | | |
| Spatial representations | | | | | | | | |
| Genes with spatial patterns | | | | | | | | |
| Spatial domains | | | | | | | | |
| Cell neighborhood analysis | | | | | | | | |
| Neighbor dependent genes | | | | | | | | |
| Cell–cell interaction | | | | | | | | |
| Ligand-receptor analysis | | | | | | | | |
| Intergroup gene expression | | | | | | | | |
| GSEA and GSVA | | | | | | | | |
| CNV estimation | | | | | | | | |
| Spatial visualization | | | | | | | | |
| Interactive visualization | | | | | | | | |
| Interactive annotation | | | | | | | | |
| Image analysis | | | | | | | | |
| Features extraction images | | | | | | | | |
| Deep learning analysis | | | | | | | | |
| Cell trajectory analysis | | | | | | | | |

Note: Each dot represents single method for the task. Squidpy and scvi are build on top of Scanpy. Table 2 abbreviations see 2.
cells with different subtypes of non-malignant cells [63]. These regional gene expression patterns in turn help to identify candidate signaling pathways and mechanistic interactions between cancer, stromal, and immune cells and to identify prognostic markers [76,77].

3.2. Specific data processing steps with spatial barcoding ST

In spatial barcoding, the transcript and the location data are in a DNA format. The sequencing step converts this into a digital format that can be processed computationally. This section covers the spatial barcoding-specific data analysis steps from the raw sequence data to the spatial cell type pattern analysis. The more advanced analysis steps commonly used with many ST dataset types are covered in Section 5.

3.2.1. Preprocessing and location matrix generation

The preprocessing steps of the spatial barcoding raw sequence data are relatively straightforward and similar to scRNA-seq data. In spatial transcriptomics, instead of a gene to droplet barcode matrix, a gene to spot matrix is generated. The information to build the gene expression (transcript sequences) to spot position (spatial barcodes) matrix is in the paired-end sequencing reads. To identify the expressed genes, the sequences are aligned against an annotated reference genome using, for instance, STAR aligner [78] or Kallisto pseudo-aligner [79]. A second alignment step against a decoy genome can be used to filter out unwanted contaminating sequences [80]. Each sequence pair also contains the spatial barcode sequence and a unique molecular identifier (UMI), which is used to remove the PCR copies of the captured transcripts arising during sequencing library preparation. The expression levels of the genes are counted from the deduplicated, aligned, and barcode-associated reads, and a spatial barcode-by-gene matrix is generated [analogous to Fig. 2D]. STARsolo [81], bustools [79], ST Pipeline [80], Spaceranger count, and Slide-seq/drop-seq tools [69,82] are commonly used solutions to produce the spatial barcode-by-gene matrix.

The spatial barcode-by-location data connects each transcript to a location coordinate in the sample. The coordinates are used to construct spatial relationship graphs and grids for the spot or cell interconnection analyses, assign transcripts to cells after segmentation with subcellular resolution data, and for visualizations and joint analyses of the spots and different features on the associated tissue images. For instance, in the Visium platform, the barcode sequences and their positions in the spatial grid are fixed and the Spaceranger count creates a barcode-by-location matrix for spatial analyses. In spatial barcode methods with stochastic barcode spot locations, the spot barcode-by-location matrix is constructed in method-specific array sequencing and indexing step [7–9,24,25]. To visualize or jointly analyze transcripts with tissue images, the barcode coordinate system needs to be aligned with the one used with the tissue images. The Spaceranger count detects the spot positions (fiducial detection) from the bright-field image and creates data for the spot-image alignment. The spatial barcode-spot alignment with the images can also be done interactively [7,9] or by using custom scripts [8].

3.2.2. Estimation of the spot-wise cell type compositions

The spatial barcoding will locate the transcripts at multi- or subcellular spatial resolution. In multicellular resolution spatial barcoding, each spot can contain transcripts from multiple cells. The cell compositions and regional enrichment of different cell types and cell stages in the spots can be resolved with computational deconvolution, mapping, enrichment, and data-integration-based methods [83,84]. For instance, SPOTlight [85] uses non-negative matrix factorization (NNMF) and SpatialDecon [86] log-normal regression for deconvolution of the transcriptomics data, whereas Cell2Location is based on a Bayesian model [87] and Tangram is a deep learning framework to resolve cell types. The ST-framework Giotto (discussed in section 5.1.) uses enrichment-based options for cell type composition analysis [88] and Seurat anchor-based integration [89]. The spot compositions are reported as cell type proportions or probabilities of occurrence in the spot. The latest deconvolution methods can also report the estimated transcriptomes of the identified cell types. Most deconvolution methods use annotated reference transcriptome profiles derived from scRNA-seq or bulk RNA-seq datasets, making the accuracy and resolution of the cell type identification highly dependent on the compatibility of the used reference profiles with the cells in the target sample. However, the recently published conditional autoregressive-based deconvolution (CARD) [90] and the latent Dirichlet allocation based STDeconvolve [91] offer reference-free deconvolution methods, which is useful when optimal reference scRNA-seq profiles are not available.

3.2.3. Spatial segmentation

The objective of segmentation is to construct masks to assign the detected transcripts to individual cells, nuclei, or larger objects. The segmentation with spatial barcoding data is essentially the same process as with in situ based ST methods and is based on the stained image data of the target tissue in combination with the actual gene signals. Usually, ST data have associated image data with hematoxylin and eosin (H&E) staining, indicating the location of nuclei and selected structural features. In some cases, fluorescent staining for different cellular markers is available to guide the segmentation. Despite intense development, the segmentation of tissues with densely packed cells is still one of the most challenging steps in the analysis of spatial data. The decisions for segmentation strategy and methods depend heavily on the sample type, the platform used, and the available data to guide the segmentation. Several methods varying from manual assignment to statistical, supervised, semi-supervised, and unsupervised machine learning applications have been developed for the task with other data types (see in situ ST data segmentation). So far, the high-resolution spatial barcoding transcriptomics has not been widely available, and a simple grid segmentation has been used in some of the studies [7,65]. We anticipate the development of deep learning models for feature segmentation and cell type identification with the large spatial barcoding ST datasets.

4. Regional selection spatial transcriptomics methods

Laser capture microdissection coupled with RNA-sequencing (LCM-seq) and digital spatial profiling (DSP) are methods that use flexible regions of interest (ROI) binning in spatial transcriptomics [14–17]. Each selection bin contains transcripts from one ROI and they are barcoded for backtracking of the location in the sample. After sequencing and processing, the bins are used similarly to the spots in spatial barcoding arrays, except for that the regions can vary in size, shape, and can be discontinuous (Fig. 1D). Depending on the researchers’ choices, each bin can contain transcripts from one or more cells and the ROIs can be selected, for example, to be homogenous in size and shape, cell type, cell marker, or consist of a functional region in the sample [14,92].

4.1. Strengths and limitations

Recently the DSP, under the commercial name GeoMx, has gained popularity as a nearly whole transcriptome level ST method with flexible ROI selection that works also with FFPE samples. In
GeoMx, the samples are hybridized with analyte probes detecting RNA or proteins, which upon laser illumination release their oligonucleotide tags with the identifiers of the target transcripts or proteins. The oligonucleotide tags are collected after each round of illumination and enzymatically barcoded for the collection bin identification. In LCM-seq the transcripts collected to each bin are barcoded by using barcoded DNA-oligos in reverse transcription reaction. In both methods, a DNA library containing the pooled bins is sequenced and the transcript count-to-bin matrix is resolved.

In GeoMx up to four fluorescent markers can be used to guide the sampling or ROIs can be selected based on the physical properties of the sample like distance from the feature of interest and/or by following the shape of the biological unit of interest. In LCM-seq the choice of selection markers depends on the used LCM-seq instrument. The strategies for ROI selection/sampling vary and are decided by the researcher based on prior knowledge and suitability for the research setting [14,92]. The flexible sampling allows the development of complex sampling strategies including for example selection of cell type-specific ROIs for profiling and then use of these in deconvolution of the mixed cell ROIs in the same sample. The available data analysis options are limited by the sampling strategy, hence the markers and the cell pooling to bins at the sampling phase should be carefully thought out before the experimental phase [92].

The spatial resolution of the GeoMx and LCM-seq data is determined by the sampling plan. The ROIs can be as small as single cells; however, the throughput in GeoMx and LCM-seq does not allow single-cell analysis on a scale comparable to in situ or spatial barcoding ST methods. GeoMx is also limited by the reliable quantification of transcripts, which in DSP requires 20 to 300 cells per ROI [14] and is thus below the current other ST methods. The transcript detection efficiency in LCM-seq depends on the sequencing library generation method and can be at the highest on the level with tube-based scRNA-seq. GeoMx does not offer untargeted mRNA capture like LCM-seq and spatial barcoding ST. The largest currently available pre-designed probe sets for GeoMx can detect over 18,000 protein-coding genes with a possibility for customization and co-detection of a number of protein targets. An advantage of the GeoMx probing protocol is that it is minimally destructive, and the tissue slides may be used for other applications like H&E staining or immunohistochemistry for retrieving additional information [14]. The GeoMx can shine in the analysis of sample cohorts in which only a limited number of biological compartments or cell types need to be sampled to answer the research questions. For example, GeoMx DSP results have revealed spatial heterogeneity in gene profiles in the host response to SARS-CoV-2 infection in lung samples [93]. DSP has also been used to profile diabetic foot ulcers [94], detect alterations in diabetic kidneys [95], and analyze intra-tumor and inter-tumor heterogeneity in various tumor types [96–99]. With the aid of deconvolution, GeoMx data was used to estimate cell level heterogeneity of tumor-infiltrating lymphocytes in different tumor microenvironments [86]. LCM-seq showed its usability in a study where neurons were collected from different regions of the brainstem and spinal cord for the identification of genes protecting neurons from spinal muscular atrophy associated cell death [100].

4.2. Specific data processing steps with regional selection ST

The cells or regions for analysis are usually pooled at the ROI selection phase in GeoMx and LCM-seq either due to detection efficiency or throughput limitations. Unless carefully sampled, the ROIs will have cells in different stages or consist of heterogeneous cell types. Hence, the analysis depends heavily on the choices taken at ROI selection. In its simplest form, the transcriptomic profiles of GeoMx and LCM-seq ROIs are compared to each other to identify differentially expressed genes. GeoMx proprietary analysis software offers tools for basic analysis. For custom analysis, the GeoMx data can be converted from proprietary file formats to the more accessible spatial data format with the GeoMx tools package in R, which offers the capability for filtering ROIs and probes with quality control parameters associated with sequencing, alignment, and negative control probes included in the pre-designed probe sets. The GeoMx tools package also includes tools for normalization, dimensionality reduction, clustering, differential gene expression analysis, and options for visualization, such as UMAP, t-SNE, and volcano plots. Like spatial barcoding ST data, the GeoMx and LCM-seq multicell ROIs can be analyzed for cell compositions with deconvolution tools, such as the SpatialDecom R package for constrained log-normal regression-based deconvolution of GeoMx and other ST data [86]. For cell type prediction SpatialDecom uses provided cell type signatures or custom ones that can be inferred from RNAseq and scRNA-seq datasets. It can also form new cell profiles “on the fly” from pure cell ROIs and use tumor cell-type-specific ROIs to identify confusing genes in the target cell profiles to aid deconvolution of non-tumor cell types. To enhance estimates, SpatialDecom uses background counts of decoy probes and can also use nuclei counts to estimate total cell counts. For neighborhood analysis, the SpatialDecom includes tools to model the gene expression profiles and up and downregulation of genes in ROIs by using the estimated cell counts with a so-called reverse deconvolution method. Additional computational analyses can be done with custom scripts or other available ST-tools after the conversion of data to a suitable format. However, advanced analysis using spatial relationships (see Section 5) can be performed with the GeoMx and LCM-seq spatial datasets only if the spatial coordinates are resolved for the ROIs.

5. Advanced solutions in the analysis of spatial transcriptomics data

The common objectives in spatial transcriptomics data analysis include identification, quantitation, and annotation of the spatial patterns (domains) at multicellular, cellular, and molecular levels, and the statistical analysis of these quantitated features at different length scales (Fig. 3). Many processing and analysis steps with the spatially resolved single-cell transcriptomes are identical to common scRNA-seq analysis steps and the same tools can be used taking into account the method-specific confounding factors and limitations in the ST data. The following sections cover the commonly used data analysis steps and some advanced computational solutions to analyze biological phenomena in spatial data. We focus on methods specialized for ST data analysis. While a detailed description of all the methods mentioned here is beyond the scope of this review, more detailed descriptions of the ST data analysis methods can be found in the original research articles or the recent reviews [27–29,83,101,102].

5.1. Available ST data analysis frameworks

Common scripting languages, such as R and Python, and various open-source data analysis and visualization packages can be used to build ST data analysis and visualization pipelines. Recently, some consolidated analysis solutions have been released which enable efficient and reproducible ST data analysis without extensive programming and scripting work. Different computational frameworks combine variable sets of commonly used and new scRNA-seq and ST data analysis tools chosen to work together. For a quick reference, the general capabilities in the analysis frameworks Giotto [88], Seurat [89,103], STUtility [104], SPATA2 [105],

I. Kleino, P. Frolova, T. Suomi et al. Computational and Structural Biotechnology Journal 20 (2022) 4870–4884
learning-based analysis and spatial trajectory and pseudotime anal-
ysis framework using Python with a focus on integrative deep
higher-level machine learning packages. stLearn is another ST anal-
ysis in addition to statistical analysis, can leverage the powerful Python
ers in 3D. As modular Python frameworks2, Squidpy and scvi-tools,
izes in analysis and visualization of sequential spatial sample lay-
ng packages in R. In addition to general ST analysis, STUtility special-
yses. SPATA2 is focused on trajectory and pseudotime analyses and
integrates tools for gene set enrichment and variance analysis (GSEA
and GSVA) and to estimate genomic copy number variation (CNV)
ased on transcriptome profiling along the genomic locations [109].

5.2. Preprocessing and quality control steps

The data imported into the ST data analysis frameworks (Giotto,
Squidpy, Seurat, STUtility, SPATA2, scvi-tools, stLearn, and GeoMx)
at a minimum consists of cell/spot-by-gene matrix and cell/spot-
by-location matrix, and optional associated image data of the sam-
ple tissue (Table 2). All the ST frameworks include the common
preprocessing tools for normalization, filtering, and dimensional
duction used to ensure the quality and consistency of the ST data
[83,110]. There is no single solution for all datasets as the data may
contain different levels of variation from the confounding sources.
The usual standard preprocessing steps are the same as used for
scRNA-seq data, including normalization across data points, scal-
ing, filtering of bad quality cells or spots, and low abundance genes
based on the total number of detected molecules [59,104,106,111].
The scRNA-seq specific methods like STransformation [112],
Snorm [113], or SCRAN [114] can also be used to remove unwanted variations. A recent preview of the stLearn package
showcased a deep learning-aided normalization method specialized
for spatial data, which takes into account image data and the
neighboring spatial data to adjust the gene expression values
[107]. Usually, data is treated also with different dimensionality
reducing algorithms for easing computationally heavy analysis
tasks like clustering and for visualization of the variation of the
data in low dimensional space.

5.3. Finding patterns in spatial transcriptomics data

Cells and molecules are spatially organized to optimize the
structure and function of the biological units they are part of
(Fig. 3A). The main goal of ST analysis is to identify spatial patterns
and domains at different scales in the ST data. In many tissues and
tumors, the specific cell types or cell stages form distinct spatial
omo- or heterotypic cell and gene expression patterns, which
be used to identify the functional units or cell niches. Strategies
starting from cell types or from gene expression patterns are used
to detect spatial patterns and identify spatially distinct domains
and niches.

5.3.1. Finding cell type patterns from ST data

A simple strategy to get an overview of the cellular patterns is
to cluster and annotate the cells or spots based on their transcrip-
tomes and then visualize the spatial variation on tissue overlay
plots (spatial plots). For transcriptome-based cell type profiling,
the available ST data analysis frameworks offer dimensionality
duction and clustering methods and options for visualizations as
dimensionally reduced data plots and spatial plots (Fig. 3B). In
addition to the widely used clustering alternatives for scRNA-seq
data, such as K-means, hierarchical, Louvain, and Seurat has a
pecific sample clustering algorithm for ST data based on modular-
nity tuning in the KNN (k-nearest neighbor) distance graph and
STUtility has implemented NNMF dimensionality reduction that
s to detect spatial patterns and identify spatially distinct domains
and niches.

---

2 Recently a scverse consortium started maintaining the Python ST-frameworks
Scanpy, Squidpy, and scvi-tools scverse.org].
tion spatial barcoding data, the deconvolution tools discussed in Section 3.2.3. provide probabilistic cell type compositions of the spots that can be used to divide the sample into cellular regions and determine cell neighborhood composition. (Fig. 3B).

For cell type annotation of single-cell resolution ST data, the same strategies and tools can be used as with the scRNA-seq data. Typically, the available ST data analysis frameworks have one or more differential gene expression analysis tools for marker gene profiling and manual marker-based cell type annotation. For instance, scvi-tools has two automated cell type annotation methods, and the option to use the scArches [120] transfer learning models, in addition to dataset integration, to annotate new data sets even with different modality reference data. Standalone reference-based annotation tools like SingleLe [121], CHETAH [122], and scGate [123] in R, and Cello [124] in Python can be included in the SA analysis solutions for cell type annotation. Additionally, methods are available that use both the tissue image and ST data for the annotation. For example, the deep learning-based solution SpaCell [125] uses tissue image and ST data jointly for cell type and disease stage annotation. The identified cell types or unannotated cell clusters can be visually and statistically analyzed for their spatial organization by plotting on the sample image and then comparing to, for example, available tissue atlases or a pathologist’s tissue annotation of the corresponding H&E stained tissues image.

Regulon-based analysis could provide further cell stage clustering by identifying coregulated genes and transcription factor networks (regulons) underlying the functional cell states. Currently, none of the ST data analysis frameworks incorporate such gene regulatory network analysis by default. For scRNA-seq data, for instance, SCENIC (Single-Cell Regulatory Network Inference and Clustering) and pySCENIC are tools to detect regulons [126,127]. SCENIC uses the cellular regulon activity patterns to cluster the cells.

5.3.2. Detecting spatial gene expression patterns

At the molecular level, genes may show spatial expression patterns which reflect their cell type-specific intrinsic programs and extrinsic effects on the local cell community or tissue microdomain (Fig. 3C). Cell molecular and organellar subcellular distribution data can be used jointly with spatial transcriptomics data to provide more detailed information about gene-phenotype associations in cells (Fig. 3H) [28,29]. RNA velocity analysis based on exon–intron ratios in transcripts [128] or inferred from nuclear to cytoplasmic gene expression ratios can be used to order cells to pseudotime trajectories for cell fate analysis [33].

Many ST data analysis frameworks implement methods developed for the identification of genes showing spatial expression patterns with the aid of the spatial representations of the ST data (Table 2) [29,88]. A simple form is a spatial grid in which the average expression of genes in each grid box area is measured or calculated. Many spatial barcode methods naturally produce data in such a form as the spots are spatially organized in a grid with a fixed size, relative locations, and the number of neighboring spots, and each spot integrates gene expressions from the covered areas. For high-resolution ST data, the grid box dimensions can be tuned to computationally integrate expressions from subcellular to multicellular-sized regions. A common approach is to construct a spatial network with nodes/vertices representing the cells and links/edges representing the connections between the cells, where the edges can have associated weights derived from their physical distance or other connection metrics (Fig. 3D).

For detecting genes with spatial trends in their expression, Giotto implements their Binary Spatial Extraction (BinSpect) method, which finds genes with spatially coherent expression patterns [88], as well as the previously published tools SpatialDE [129], SPARK [130], and trendsceek [131]. SpatialDE3 uses Gaussian process regression to identify genes with spatial expression patterns, SPARK identifies trending expressions using generalized spatial linear models, whereas trendsceek uses nonparametric marked point processes to detect spatial trends in gene expression. Seurat has its own implementation similar to trendsceek, SPATA2 implements SPARK, Squidpy and STUtility use autocorrelation-based detection of genes showing patterned expression, whereas stLearn uses factor analysis based on tissue microenvironment detection to detect genes correlated with the identified factors. Additionally, standalone solutions exist, such as MERINGUE [133], which uses an autocorrelation-based analysis method, and Hotspot [134], which uses a graph-based approach to identify the most informative genes.

In addition to detecting genes showing spatially distinct expression patterns, another typical goal of ST data analysis is to detect spatial domains with coherent gene expression. These may differ from the patterns detected based on the cell type analysis [135]. In the available ST data analysis frameworks, various computational solutions are used to specify the spatial domains by summarizing the detected genes with distinct coherent gene expression patterns into metagenes. For instance, SPARK and SpatialDE use clustering to form metagenes from the detected spatial genes. Additionally, Giotto implements a hidden Markov random field (HMRF) method [135] to detect spatial domains, which uses the relationship information in the spatial graphs to compare the gene expression of each cell to its neighborhood (domain) to search for coherent gene expression patterns and to assign cells into a predefined number of spatial domains. The standalone solution MERINGUE uses cross-correlations for grouping the identified expression patterns into spatial domains and for spatially informed cell clustering [133].

The statistically detected spatial domains of coherent gene expression can then be corroborated by comparing them to domains identified from the corresponding image data, for instance, by manual annotation or deep learning-based feature detection [136]. The joint analysis of molecular and image data is leveraged in recently developed methods, including stLearn [107], SpaGCN [137], and SpaCell [125]. stLearn uses a two-step clustering method in which the clustered expression data is further clustered with the spatial information to find finer spatial domains [107]. SpaGCN integrates the image RGB pixel data along with the spatial coordinates to tune spatial expression graph weights and then uses graph convolution and iterative clustering to find spatial domains and spatial genes and metagenes with differential expression [137]. SpaCell uses deep learning-based analysis of the images of H&E stained samples jointly with the matched measured spatial molecular data to classify the cell types and disease stages on the ST samples.

5.3.3. Cellular neighborhoods, interactions, and communication

Understanding the compositions of cell neighborhoods and molecular level cell–cell interactions (CCI) and cell–cell communication (CCC) in tissues and organs will open the path for building comprehensive models of biological systems. Also, detection of critical molecular components of cellular connectedness would allow identification of crucial molecular pathways causing pathologies and potential ways to treat these [138]. The molecular level cell connectedness analysis has progressed with giant steps recently with the availability of dissociated scRNA-seq datasets and a number of computational data analysis tools [138,139]. Now, with spatially aware transcriptome data, modeling and inference of molecular interconnectedness of cells at different length scales is becoming reality.

5 GPU enhanced SpatialDE2 has been launched [132].
Spatial cell neighborhood analysis detects preferred cell–cell adjacencies and cell community compositions that can suggest homo- and heterotypic attractive or repulsive connections between different cell types (Fig. 3E). The analysis is based on statistical enrichment of cell–cell interactions. Of the available ST data analysis frameworks, only Giotto and Squidpy offer cell neighborhood analysis by default. Statistical analysis in both is based on the enrichment method introduced in HistoCat [140], where the the occurrences of cell–cell interactions are compared to the randomized baseline in the data.

Spatial information also allows the prediction of cell adhesion and cell signaling ligand-receptor pairs mediating cell connections in tissues at auto-, juxta-, para-, and endocrine length scales (Fig. 3F-G). The basis for statistical analysis is a cell–cell relationship representation and a list of known ligand-receptor and cell-adhesion ligand-receptor pairs, including also molecular compositions, as the receptors and ligands may be homo- or hetero-multimeric in nature. The ligand-receptor list can be constructed from protein interaction datasets. However, standalone tools like CellphoneDB [141,142], NicheNet [143], and CellChat [144] usually provide manually curated ligand-receptor lists for the analysis, and collections like FANTOM5 [145] and Omnipath [146] have larger ligand-receptor datasets available. Typically, the cell-communication scores are calculated for each ligand-receptor pair in all relevant pairs of cell types. The highest-scoring pairs are the most likely ones mediating connections between the particular cell types.

For instance, Giotto implements a spatially-aware method that calculates cell-communication scores associated with the cell type neighborhoods, whereas stLearn and Squidpy have their own implementations of the enrichment test of CellphoneDB [141,142], which is a method commonly used to identify important ligand-receptor pairs in scRNA-seq data analysis. Two recent standalone methods, tensor composition-based Tensor-cell2cell [147], and random forest-based MISTy [148], aim to identify significant ligand-receptor pairs simultaneously at multiple length scales (Fig. 3F). Also, spatially-aware versions of the CellphoneDB [149] and NicheNet [143], and several additional standalone spatially aware ligand-receptor analysis tools based on different statistical and machine learning methods have been released and are described in recent reviews [138,139].

5.3.4. Interactive visualization

All the ST data analysis frameworks provide multiple options for production of graphs and spatial feature visualizations of the performed data analysis. Most of the frameworks offer also some level of interactive visualization for exploratory data browsing and visual inspection of the results along with the analysis. Giotto enables exploratory viewing and selection of subsets of data for reanalysis and SPATA2 has interactive tool for manual spatial trajectory drawing and annotation. Squidpy has its own image data container type and connects to Napari, a Python-based GPU accelerated image analysis software, for advanced data visualizations and image-based analysis. Squidpy allows the use of machine learning packages for feature extraction from the image data (H&E and fluorescent staining), including cell and nuclei segmentation used in subcellular spatial domain computations. Seurat and stLearn have interactive viewing options with “click and select” types of gene selection and visualization adjustments. STUtility provides tools for image trimming and for automatic and manual image alignment of consecutive sample sections of spatial barcoding data for the construction of a rotating 3D stack of the spot images. At the moment scvi-tools does not offer interactive spatial visualization integrated into the framework, however the use of anndata data container offers workaround to import of analysis results to other software for interactive visualization. Various spatial transcriptomics visualization options are reviewed in [83].

5.4. Summary and Outlook

Both the methodological and computational solutions for spatial transcriptomics are evolving at a fast pace, providing coverage to larger sample areas at higher resolution thus producing spatial datasets with increasing complexity and size. New advanced commercial spatial transcriptomics platforms allowing wide adoption of spatial transcriptics in biological and medical research fields have been promised to be launched by the end of the year 2022. The newest ST methodologies allow the collection of nearly complete transcriptomic datasets with subcellular resolution at the organ and systems level. The collected datasets will contain a wealth of biological information about the spatial relationships of the cells that can be used to supplement, interpret, and find causal relationships for connecting the genetic programs in multicellular systems.

Methods to mine these vast datasets by combining statistical modeling, supervised and unsupervised machine learning methods, and data integration techniques with the current knowledge of signaling, metabolic and biochemical pathways, and gene regulation networks holds promise to identify complete sets of components taking part in biological processes and to modeling dynamic cellular functions in interconnected multicellular systems.

Computationally, already the size of terabytes of raw data per sample holding spatial information of hundreds of millions individual transcripts in the forthcoming datasets [22] will pose a challenge for use of many current computational analysis methods with current computational resources. On the other hand, these new datasets open an unprecedented opportunity for the development of new computational methods for mining and interpreting the vast amounts of biological and biomedical data.

The exploratory visualization and sharing of the data and analysis results in interactive formats with collaborators and in the public domain are also becoming an important part of the computational analysis of complex biological datasets. The option to share the spatial data as standalone datasets with the possibility for interactive visualization with dedicated viewers is currently not easily attainable with the available ST data analysis platforms. To provide maximum benefit for all communities, scientific and other, the vast information in spatial datasets could be provided also in interactive visualization formats.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

Prof. Elo reports grants from the European Research Council ERC (677943), European Union’s Horizon 2020 research and innovation programme (955321), Academy of Finland (310561, 314443, 329278, 335434, 335611 and 341342), and Sigrid Juselius Foundation during the conduct of the study. Our research is also supported by Biocenter Finland, and ELIXIR Finland. The figures were created with BioRender.com.

References

[1] Moses L, Pachter L. Museum of spatial transcriptomics. Nature Methods 2022;19:534–46. https://doi.org/10.1038/s41592-022-01409-2.
[2] Asp M, Bergensstråhle J, Lundberg J. Spatially Resolved Transcriptomes-Next Generation Tools for Tissue Exploration. BioEssays 2020;42:1–6. https://doi.org/10.1002/bies.201900221

[3] Chen KH, Boettiger AN, Moffitt JR, Wang S, Zhuang X. Spatially resolved, highly multiplexed RNA profiling in single cells. Science 2015;348:1360–3. https://doi.org/10.1126/science.aaf229.

[4] Eng CHL, Lawson M, Zhu Q, Dries R, Koutenova N, Takei Y, et al. Transcriptome-scale super-resolved imaging in tissue by RNA seqFISH+. Nature 2019;568:235–9. https://doi.org/10.1038/s41586-019-1049-y

[5] Lee JH, Daugharthy ER, Scheiman J, Kalhor R, Ferrante TC, Terry R, et al. Fluorescent in situ sequencing (FISSEQ) for gene expression profiling in intact cells and tissues. Nature Protocols 2015;10:442–58. https://doi.org/10.1038/nprot.2014.191

[6] Eng CHL, Lawson M, Zhu Q, Dries R, Koutenova N, Takei Y, et al. Transcriptome-scale super-resolved imaging in tissue by RNA seqFISH+. Nature 2019;568:235–9. https://doi.org/10.1038/s41586-019-1049-y

[7] Vibko S, Eraslan G, Salmén F, Flugshammer J, Stenbeck L, Schapiro D, et al. High-definition spatial transcriptomics for in situ tissue profiling. Nature Methods 2019;16:987–90. https://doi.org/10.1038/s41592-019-0548-y

[8] Liao J, Liu Y, Qin X, Enninful A, Bai Z, Deng Y, et al. Spatial multi-omics sensitive spatial transcriptomics at near-cellular resolution with Slide-seqV2. Nature Methods 2021;19:313–9. https://doi.org/10.1038/s41592-020-0972-6

[9] Xia C, Fan J, Emanuel G, Hao J, Zhuang X. Spatial transcriptome profiling by MERFISH reveals subcellular RNA compartmentalization and cell cycle-dependent gene expression. Proceedings of the National Academy of Sciences of the United States of America 2019;116:19409–9. https://doi.org/10.1073/pnas.1912459116

[10] Shah S, Lubek J, Zhou W, Lai Y, Jennings G, et al. High-throughput single-cell gene expression profiling with spatially resolved transcriptional mapping. Cell 2018;174:363–376.e16. https://doi.org/10.1016/j.cell.2018.05.035

[11] Nonaka I, Iba S, Nishimura M, Uchimura K, Hasegawa M, et al. Single-cell RNA-seqFISH accurately detects transcripts in single cells and reveals robust spatial organization in the Hippocampus. Neuron 2019;102:355–68.e16. https://doi.org/10.1016/j.neuron.2019.07.008.

[12] Nakamura S, Nishimoto M, Ono Y, Koyama H, Seno M, et al. Single cell sequencing for fixed tissue via DBiT-seq. STAR Protocols 2021;2:.https://doi.org/10.1002/stpr.88

[13] Takai Y, Kunz J, Keyse C, O’Donnell L, Sáez-Fernández A, et al. Integrated spatial and single-cell transcriptomic data elucidates mouse organogenesis. Nature Biotechnology 2022;40:74–85. https://doi.org/10.1038/s41587-021-01006-2

[14] Callaway EM, Dong HW, Ecker JR, Hawrylycz MJ, Huang ZJ, Lein ES, et al. A high-resolution spatial transcriptome of the mouse brain. Nature 2021;598:86–102. https://doi.org/10.1038/s41586-021-03950-0

[15] Huang KH, Eichorst JP, McGinnis CS, Patterson DM, Chow ED, Kersten K, et al. Single-cell transcriptomics reveals regulatory Gene Expression Dynamics Leading to Lineage Commitment in Early T Cell Development. Cell Systems 2019;9:321–37. https://doi.org/10.1016/j.cels.2019.09.008.

[16] loholt T, Ghazanfar S, Missarov A, Koutenova N, Pierson N, Griffiths JA, et al. Integration of spatial and single-cell transcriptomic data elucidates mouse organogenesis. Nature Biotechnology 2022;40:74–85. https://doi.org/10.1038/s41587-021-01006-2

[17] Takei Y, Yen J, Zhong S, Olkkainen N, Pierson N, White J, et al. Integrated spatial genomics reveals global architecture of single nuclei. Nature 2021;590:344–50. https://doi.org/10.1038/s41586-021-02126-z

[18] Ke R, Mignardi M, Funicelli A, Svedlund J, Botling J, Wåhby C, et al. In situ sequencing for RNA analysis in preserved tissue and cells. Nature Methods 2013;10:857–60.https://doi.org/10.1038/nmeth.2563

[19] Shah S, Takei Y, Zhou W, Lubek J, Yen J, Eng CHL, et al. Dynamic and Spatial Genomics of the Nascent Transcriptome by Intron-seqFISH. Cell 2018;174:363–376.e16. https://doi.org/10.1016/j.cell.2018.05.015.

[20] Mittorf JR, Hao J, Wang G, Chen KH, Babcock HP, Zhuang X. High-throughput single-cell geneexpression profiling with multiplexed error-robust fluorescence in situ hybridization. Proceedings of the National Academy of Sciences of the United States of America 2016;113:11046–51. https://doi.org/10.1073/pnas.1612826113.

[21] Watson ER, Taherian Fard A, Marbach D, Computational Methods for Single-Cell Imaging and Omics Data Integration. Frontiers in Molecular Biosciences 2022;8:1–25. https://doi.org/10.3389/fmolb.2021.768106

[22] Eng CHL, Shah S, Thomasse J, Cai L. Profiling the transcriptome with RNA SPOTs. Nature Methods 2017;14:1153–5. https://doi.org/10.1038/nmeth.4505

[23] Zhou Q, Shah S, Dries R, Cai L, Yuan GC. Identification of spatially associated subpopulations by combining scRNAseq and sequential fluorescence in situ hybridization data. Nature Biotechnology 2018;36:1183–90. https://doi.org/10.1038/nbt.4265.

[24] Falk T, Mai D, Benisch R, Çiçek O, Abdelkadir A, Marrakchi Y, et al. U-Net: deep learning for cell counting, detection, and morphometry. Nature Methods 2019;16:67–70. https://doi.org/10.1038/s41592-018-0263-1

[25] Van Valen DA, Kudo T, Lane KM, Macklin DN, Quach NT, DeFelice MM, et al. Deep Learning Automates the Quantitative Analysis of Individual Cells in Live-Cell Imaging Experiments. PLoS Computational Biology 2016;12:.

[26] Greenwald NF, Miller G, Moen E, Kong A, Kajal A, Dougherty T, et al. Whole-cell segmentation of tissue images with human-level performance using large-scale data annotation and deep learning. Nature Biotechnology 2021. https://doi.org/10.1038/s41587-021-00755-9.

[27] Stringer C, Wang T, Michaelos M, Pachittari M. Cellpose: a general algorithm for cellular segmentation. Nature Methods 2021;18:100–6. https://doi.org/10.1038/s41592-020-0914-5.

[28] Scherthaner WD, Beerenwinkel P, Bu, Ching HH, Rocheleau J. Leveraging multimodal microscopy to optimize deep learning models for cell segmentation. APL Bioeng 2021;5:. https://doi.org/10.1063/5.100532.
signalling in human. Nature Communications 2015;6:7866. https://doi.org/10.1038/ncomms8866.

[146] Türei D, Valdeolivas A, Gul L, Palacio-Escat N, Klein M, Ivanova O, et al. Integrated intra- and intercellular signaling knowledge for multicellular omics analysis. Molecular Systems Biology 2021;17:1–6. https://doi.org/10.15252/msb.20209923.

[147] Armingol E, Baghdassarian HM, Martino C, Perez-Lopez A, Knight R, Lewis NE. Context-aware deconvolution of cell-cell communication with Tensor-cell2cell. bioRxiv 2021:2021.09.20.461129.

[148] Tanevski J, Flores ROR, Gabor A, Schapiro D, Saez-Rodriguez J. Explainable multiview framework for dissecting spatial relationships from highly multiplexed data. Genome Biol 2022;23:97. https://doi.org/10.1186/s13059-022-02663-5.

[149] Garcia-Alonso L, Handfield LF, Roberts K, Nikolakopoulou K, Fernando RC, Gardner L, et al. Mapping the temporal and spatial dynamics of the human endometrium in vivo and in vitro. Nature Genetics 2021;53:1698–711. https://doi.org/10.1038/s41588-021-00972-2.