Visual analytics for spatially resolved omics data at single cell resolution: methods & applications

Somarakis, A.

1

Introduction

A disease is defined as “any harmful deviation from the normal structural or functional state of an organism”[1]. Organisms, though, are “republics of living elementary units ”[2].Therefore in 1839, Schwann stated that cells are the elementary parts of all tissue parts [2] defining them as “the fundamental units of living organisms” [3], thereby reducing the study of healthy and diseased tissue to the understanding of cellular functionality in the tissue.

To study cellular function, information regarding cellular properties is essential, including their molecular characteristics and position in the tissue. Hence, microscopic and molecular imaging systems are utilized to acquire information about cellular properties (e.g. protein abundance, RNA and DNA sequence) and simultaneously locate the position of the cell in the tissue. Until recently, it was either impossible to measure whole gene sequences or more than a couple of proteins at single-cell level [4]. Current ’spatial omics’ imaging modalities can measure tens of proteins [4, 5] to the whole human transcriptome [6, 7] at subcellular resolution. The detailed determination of the cellular properties at single cell resolution empowered researchers to precisely characterise the type of the cells in tissues and explore the spatial patterns they form. Both aspects are important towards the understanding of a cell’s role in tissue functionality, and to understand what goes wrong in diseased tissue.

The cell type’s heterogeneity provides information to the researchers about the complex biology in tissue , but only the exact determination of the spatial interactions completes the picture. For example, the presence of immune and cancer cells in a tissue sample exhibits an immune response to an existing tumour. Yet, the position of immune cells in relation to the tumour defines the current status of tumour as immune-excluded or immune-inhaled: this distinction is essential in treatment selection and outcome prediction in cancer. Especially now, with the current imaging modalities that offer much more information, the determination of the exact type of the immune and cancer cell interactions enables a deeper understanding of why tumours with similar status from different patients react differently to treatment. The analysis of the spatial cell type configuration in tissue is flanked by additional problems. The comparison of tissue samples originating from multiple subjects is prone to batch effects, as the different tissue preservation and acquisition protocols vary the range of acquired data among tissue samples. Also, the difficulty of boundary definition for complex cellular structures often deteriorates the quality of any finding and demands its “in situ” validation.

Most currently available methods [8, 9] are focused on the analysis of the data utilizing statistical hypothesis testing, requesting from the experts to choose in advance the specific patterns they want to test. This choice is further impeded by the novelty of the data, as the experts do not have a clear view of the main characteristics of the data and the patterns they can discover.A different approach which can incorporate gradually, in a data-driven way, the expert’s prior knowledge and cognitive skills would be more efficient. Visualizing the data, the expert can perceive its main characteristics in a more intelligible manner. In particular, Visual Analytics [10] combining visualization with data analysis techniques breaks down the identification of significant patterns to smaller tasks, where the expert can mobilize prior expertise and cognition to make more informative decisions throughout the identification process. Moreover, the expert can interactively, through the linked visual interfaces, not only explore in a stepwise manner the data, but also easier locate any finding

in the tissue and validate its quality and significance.

1.1. SPATIAL OMICS DATA

Even though the need for systematic exploration of cell contexts has been consolidated for years now [11, 12], the acquisition of images possible to capture in detail the molecular cellular required for high-resolution cell classification was not feasible. Only recently researchers managed to simultaneously identify multiple molecular properties, including DNA sequence[13], RNA transcripts[14, 15] or proteins[4, 16] alongside their spatial locations at cellular level. For example, recent state of the art acquisition methods [17] could measure simultaneously up to 10 proteins in each cell. With the methods presented in this thesis, expression information for more than 40 proteins is measured, enabling researchers to answer in much greater detail vital biological question, such as the identification of the constituent cell types in a tissue sample.

Each biological property is measured by a specific method, providing the expert the option to choose the one that is most suitable to answer a specific biological questions. Except for the biological relevance, the choice of the method is also influenced by the cost, the throughput speed and the characteristics of the tissue samples (e.g. part of the tissue, preservation method). In this thesis, applications were mostly focused on studies related to the immune system with Imaging Mass Cytometry [4]and multispectral immunofluorescence imaging [18], two relatively novel methods to image protein expression.

Imaging Mass Cytometry Data. Imaging Mass Cytometry data acquisition method [4] is a combination of mass cytometry [19], immunocytochemistry [20] (ICC) and immunohistochemistry [21] (IHC) techniques. More specifically, the acquisition starts with the labeling of tissue samples with antibodies conjugated to heavy metals following established IHC and ICC protocols. The selection of antibodies is crucial, as each of them binds to a specific tissue protein. Then, a laser ablates spot by spot every $1\mu m^2$ of the tissue. The ablated material is directed to a Mass Cytometer, where the amount of metals, and by extension of proteins, are measured for each spot. Each spot represents a pixel in the derived image. Up to now, more than 40 proteins can be measured simultaneously with Imaging Mass Cytometry.

Multispectral immunofluorescence imaging Even though Imaging Mass Cytometry offers unprecedented detail in the measurement of protein abundance, its throughput is relatively low, limiting in practice the amount and size of tissue samples that can be scanned. On the other hand, a typical cohort study contains some hundreds to thousands of tissue samples. The Vectra imaging system based on multispectral immunofluorescence (IF) technology [18] can capture up to 7 proteins simultaneously, enabling researchers to conduct high throughput studies to the detriment of the amount of the measured proteins. Also, Vectra derived images offer higher resolution, depicting in every pixel a region of $0.25\mu m^2$ of the tissue sample.

As output, both modalities provide a stack of images, where each pixel of the image contains multiple scalar values at subcellular resolution. The development of our pipeline was based on, but is not limited, to the analysis of this type of data. In principle, most parts of our pipeline can be utilized for modalities that capture various properties (e.g. tissue transcriptomics) given that they provide their cell defining characteristics at cellular resolution.

1.2. PREPROCESSING OF MULTIPLEXED IMAGES

The transformation of pixel values to reliable cellular data is an open challenge, as it entails many and complex steps. Data normalization and cell segmentation are two of the most significant.

Data Normalization. Tissue preservation and handling methods influence the signal-to-noise ratio of the acquired images. A differentiation in the tissue fixation protocol or age among two samples can vary the range of their acquired values. Hence, normalization of the data and removal of noise during tissue acquisition are essential for the combination of images in cohorts for downstream analysis. Most of the time, an offset is expressed in the image values as an almost uniform disproportionate high value. A common procedure for the normalization of these data is the manual subtraction of the offset from the pixel values so as the expression pattern to match expert's prior knowledge. However, such an approach is prone to subjective observer errors and extremely time-consuming for large cohorts of images. On the other hand, automatic approaches [22] reduce processing time and inter-expert heterogeneity, but remove important biological variance as well. In Chapter 2.1, we propose a hybrid method to deal with the previously mentioned shortcomings for Imaging Mass Cytometry data.

Cell Segmentation. Cell segmentation is one of the most difficult problems in the analysis of histopathological images. Many methods have been utilized spanning from traditional unsupervised segmentation algorithms [23, 24], over fully-supervised deep learning methods [25], to semi-supervised approaches [8]. Usually, the unsupervised methods start with the nucleus segmentation followed by their dilation to identify the cellular borders. The deep learning methods extract features from multiple dimensions of the multiplexed images providing better accuracy, given that they are trained with a large amount of annotated images. The semi-supervised approaches are based on weakly annotated data and automated machine-learning methods [26], providing the best solution for data derived from recent imaging modalities with minimum available annotated images. However, these segmentation algorithms are developed for circular shaped cellular data (Figure 1.1a). For complex cellular structures, such as microglia cells (Figure 1.1b) where their branches have been detached from the cellular body during the image acquisition procedure, specific algorithms are needed. Current approaches are either focused on the identification of the cell's skeleton [27] or nucleus [28]. In Chapter 2.2, we propose an unsupervised method for the identification of microglia cells from high-dimensional multiplexed images.

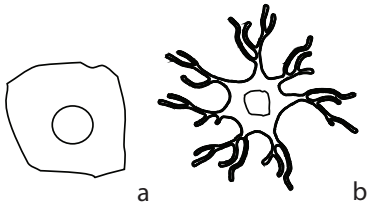

Figure 1.1: ***Cellular structures.*** *(a) A typical structure of a cell and (b) the structure of microglia cell, with the inner circural part (soma) and the many branches (processes).*

1

1.3. VISUAL ANALYTICS

Given the novelty of the aforementioned spatial cell phenotyping modalities, the experts do not always know in advance which patterns to expect in the data. The cellular composition of captured spatial omics images is often unknown, as the cellular characteristics that differentiate a patient and a healthy individual. When experts are neither aware of these characteristics, such as their main trends or outliers, nor the patterns they want to identify, experts should first perceive their data before starting affirming or rejecting a hypothesis. Exploratory Data Analysis [29] can play a significant role towards this direction.

Quantitative metrics can often provide a decent, but fragmented picture of data's characteristics, especially for the high-dimensional (i.e. multiple measured cellular properties) and spatial (i.e. cellular resolution) nature of such data. Anscombe's Quartet [31] is a typical example of the inability of quantitative metrics inability to provide the full spectrum of data characteristics. It contains four different two-dimensional datasets (Figure 1.2) with almost identical statistics; mean value for x and y, variance for x and y, correlation among x and y, linear regression line and coefficient. Their visual representations (Figure 1.2), though, illustrate four different datasets with discrete inherent characteristics, highlighting the definition given from Pickover and Tewksbury [32] for visualization, as the "*the art and science of making the unseen workings of nature visible*".

Therefore, we integrate visualization techniques in the exploratory analysis of our complex, novel data, as they are described in Chapters 3 and 4 to enable the expert "*to analyze data when they don't know exactly what questions they need to ask in advance*" [33].

Table 1.1: ***Anscombe's quartet*** *includes four datasets with identical statistics with 2 decimal points accuracy.*

	I		**II**		**III**		**IV**	
Observation	*x*	*y*	*x*	*y*	*x*	*y*	*x*	*y*
1	10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
2	8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
3	13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
4	9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
5	11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
6	14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
7	6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
8	4.0	4.26	4.0	3.10	4.0	5.39	19.0	12.50
9	12.0	10.84	12.0	9.13	12.0	8.15	8.0	5.56
10	7.0	4.82	7.0	7.26	7.0	6.42	8.0	7.91
11	5.0	5.68	5.0	4.74	5.0	5.73	8.0	6.89
Summary Statistics								
Mean	9	7.50	9	7.50	9	7.50	9	7.50
Variance	11	4.13	11	4.13	11	4.13	11	4.13
Correlation X, Y	0.82		0.82		0.82		0.82	
Linear Regression	*y = 3.00 + 0.500x*		*y = 3.00 + 0.500x*		*y = 3.00 + 0.500x*		*y = 3.00 + 0.500x*	

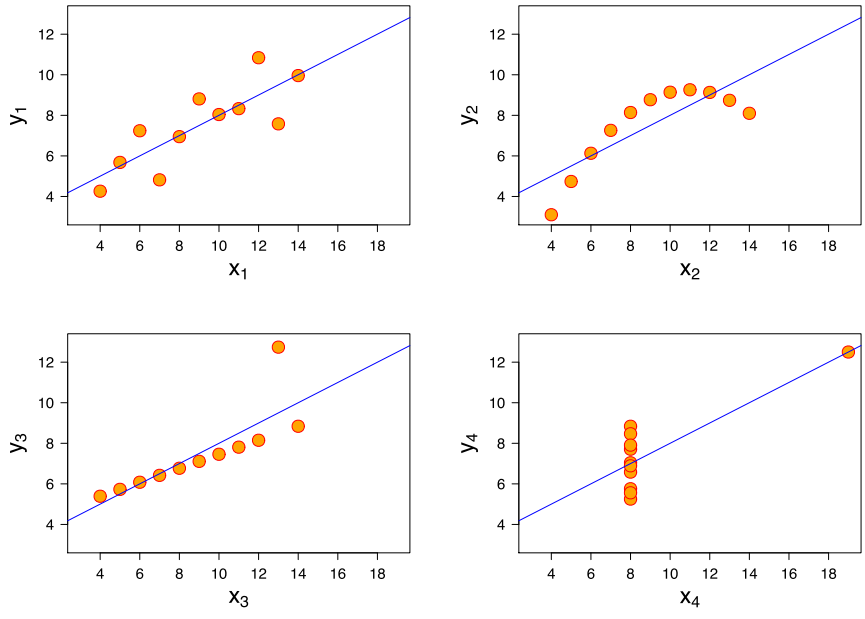

Figure 1.2: ***Anscombe's quartet*** *visual representation in the x-y space shows a significant difference from the above illustarted decriptive statistics. (Image Source: Wikipedia's Lemma [30])*

A significant factor of visualization efficiency is the expert that is performing the specific exploratory task. The identification of the number of clusters and outliers in a two-dimensional scatter plot is performed much more efficiently by a human than a computer. This can exemplify the functioning of visualization; a visual representation is designed to represent data in a way that can facilitate the expert to employ internal cognition and memory usage (i.e. points that are close together form a cluster, points that deviate from the group and are not close to another group are outliers) to gain a deeper understanding of the data (i.e. clusters) [33]. A vital part of the whole visualization process plays the interaction among the expert and computing system (Figure 1.3) throughout this cognitive process. This feedback mechanism is taking place through the interactivity among humans and computers, enabling the expert - on the fly - to explore multiple data attributes.

However, it is impossible for the expert to explore raw cellular properties values or the total amount of spatial cellular combinations merely through interactive visualizations, due to their large amount. To that end, the utilization of automated data analysis techniques alongside interactive visualization systems is deemed necessary. The discipline that studies such systems is called Visual Analytics [10, 34]. In addition to that, there are numerous problems for which research questions can not be defined from the onset of their analysis. Hence, for such ill-defined problems automated solutions fall short Visual Analytics solutions can have substantial contribution by integrating analysis with hypothesis generation [35]. The rationale behind Visual Analytics, as illustrated in Figure 1.3, is based on the combination of the processing power of automated analysis techniques to deal with large amounts of data with the human's ability for analytical reasoning through interactive

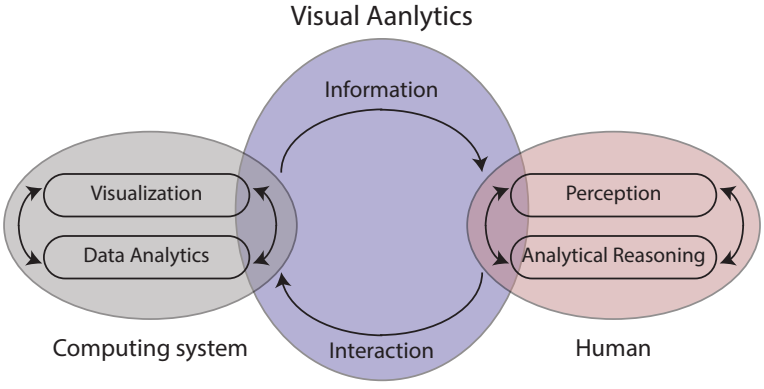

Figure 1.3: ***Visual Analytics process.*** *Starting from the creation of visualizations utilizing data analytics methods, provides insights to the expert. The expert, through her/his analytical reasoning, perceive the insights and interacts accordingly with the system to continue the data exploration.*

visualizations. The role of the experts (i.e., humans) in such a system is pivotal, as they interactively steer the data exploration [36].

Visual Analytics systems are being used in various scientific domains to facilitate the analysis of large amounts of complex data. Besides the exploration of tissue images, Visual Analytics is being used in Medicine for the exploration of public health data [37]. Moreover, Visual Analytics systems are being used for the exploration of the most efficient stock trading algorithm [38], exploration of high resolution remote sensing imagery for the support of precision agriculture [39] or the similarity exploration of texts [40].

Even though Visual Analytics is a valuable tool for the analysis of complex problems in large information spaces, it does not come without restrictions. Modern applications (e.g. more sensitive sensors, large scale scientific experiments) has resulted in the accumulation of enormous amounts of data. Following the increase of available data, algorithms have evolved to live up with the analysis requirements. Nonetheless, not all of these algorithms are appropriate to serve the main Visual Analytics tasks, namely the illustration of high-level abstractions synchronously with low-level details of the data [41]. Also, the combination of heterogeneous types of data is another limiting factor for many Visual Analytics applications. For scientific domains where multiple different data sources are combined is challenging for Visual Analytics approaches to create effective solutions. Furthermore, many restrictions are imposed from the involvement of the user in Visual Analytics approaches; the necessity for users to supervise 24/7 applications which include data streams and require their immediate response and supervision, or the required training in order to handle properly Visual Analytics tool and interpret any findings or results. Finally, a serious challenge which leads many times to the disuse of Visual Analytics tools is their limited compatibility with existing systems that the users utilize in their daily routine.

From the aforementioned restrictions and challenges it is evident that proper evaluation of the Visual Analytics systems should be performed. According to Keim et al. [10], Visual Analytics is both a science and a technique and thus should be evaluated based on

effectiveness, *efficiency* and *user satisfaction*. In other words, a Visual Analytics system should be evaluated based on its ability to enable users to fulfill the tasks that originally the system has been developed for, given a finite amount of resources. The evaluation includes the data, the tasks definition, the stakeholders and the Visual Analytics system itself. The evaluation of the data should take into account the heterogeneity and suitability of the data to answer the research questions. The defined tasks should be evaluated according to their complexity. The evaluation of the stakeholders should not be limited to the users, but also expanded to the the developers of the system. The evaluation of the Visual Analytics systems should include the graphical representations, the suitability and efficacy of the selected technologies. In general, through the evaluation process it is easier to deduct conclusions about precise well-defined problems (e.g. the efficacy of a visual analytics system to cluster the data from a two-dimensional scatter plot) than generic findings (e.g. the amount of saved time if we use one technique in favor of one other). Çöltekin et al. [42] in their research are trying to discover such generic conclusions recording and afterwards analysing the movement of the user's eyes. At the end of the day, proper evaluation of the Visual Analytics systems can enable the comparison of existing approaches and eventually lead to development of innovative methods and techniques.

1.4. CONTRIBUTION & OUTLINE

The main contribution of our work is the development of an end-to-end exploratory data analysis methodology , (Figure 1.4) addressing the main challenges researchers face during the analysis of highly multiplexed tissue images at cellular level, in close cooperation with the clinical researchers from Leiden University Medical Centre (LUMC). A birds-eye overview of our pipeline is described in our previous work [43]. More specifically, our pipeline covers the preprocessing of the data (Figure 1.4a), including the removal of the biological irrelevant intra-cohort heterogeneity and the segmentation of complex cellular structures, preparing the data for the main task of the analysis, the exploration of spatial cellular patterns. Then, we enable the researchers to visually explore the data, through two interactive and data-driven Visual Analytics frameworks; ImaCytE [44] (Figure 1.4b), for the identification of spatial cellular patterns in tissue samples and SpaCeCo [45] (Figure 1.4c), for the comparison of distinct clinical cohorts of patients. In particular:

- A workflow focused on the normalization of cohorts of images derived from Imaging Mass Cytometry which entails the semi-supervised classification of pixels, as either actual signal or background and the assignment of a cell's expression value, as the fraction of actual signal pixels that belong to its segmented area.

- An automated algorithm for the segmentation of microglia, a complex cellular structure with a main circular part and many branches, based on its nucleus and membrane expression markers utilizing an advanced level-set-based approach [46].

- ImaCytE; A Visual Analytics framework, which was developed for the end-to-end, in depth, analysis of individual images derived from Imaging Mass Cytometry, but can be applied to any spatial omics modality offering cellular resolution. The main novelty of our framework is not only the interactive identification of cohesive cell types, but through motifs the stratification of the cell types into subgroups with

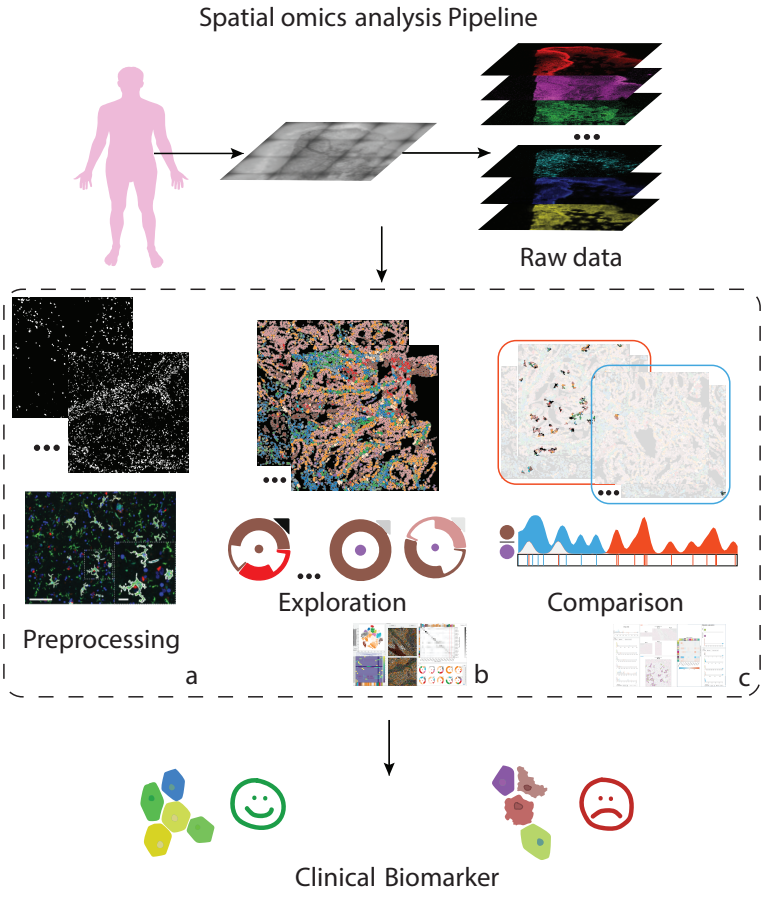

Figure 1.4: *An end-to-end pipeline for the analysis of highly multiplexed tissue images at cellular resolution, covering from (a) the preprocessing over (b) the exploration to (c) the comparison of the data.*

unique microenvironment characteristics and their inspection interactively in the tissue.

- SpaCeCo; A Visual Analytics tool, for the comparison of cohorts with labelled cellular images, based on both their cell types and the spatial patterns they form. Moreover, the tool enables in every step of the comparison the identification of outliers in each cohort and location of any finding in the tissue.

The dissertation continues with Chapter 2, which presents an overview of the preprocessing part of our pipeline, including the data normalization and the segmentation of complex cellular structures. Chapter 3 and 4 form the core part of our analysis pipeline and explain in depth the concepts and functionality of our two Visual Analytics frameworks; ImaCytE and SpaCeCo. Chapter 5 outlines a study in Alzheimer's disease, where major

parts of our pipeline have been utilized for discovery research. Finally, Chapter 6 summarizes this work and reflects on the lessons learnt during this research and upon directions for future work.

References

[1] D. G. Scarpelli and W. Burrows, "Disease's definition." https://www.britannica.com/science/disease, 1976–98. [Online; Accessed: 2021-07-14].

[2] P. Mazzarello, "A unifying concept: the history of cell theory," *Nature Cell Biology*, vol. 1, no. 1, pp. E13–E15, 1999.

[3] A. Regev, S. A. Teichmann, E. S. Lander, I. Amit, C. Benoist, E. Birney, B. Bodenmiller, P. Campbell, P. Carninci, M. Clatworthy, *et al.*, "Science forum: the human cell atlas," *Elife*, vol. 6, p. e27041, 2017.

[4] C. Giesen, H. A. Wang, D. Schapiro, N. Zivanovic, A. Jacobs, B. Hattendorf, P. J. Schüffler, D. Grolimund, J. M. Buhmann, S. Brandt, Z. Varga, P. J. Wild, D. Günther, and B. Bodenmiller, "Highly multiplexed imaging of tumor tissues with subcellular resolution by mass cytometry," *Nature Methods*, vol. 11, no. 4, pp. 417–422, 2014.

[5] Y. Goltsev, N. Samusik, J. Kennedy-Darling, S. Bhate, M. Hale, G. Vazquez, S. Black, and G. P. Nolan, "Deep profiling of mouse splenic architecture with codex multiplexed imaging," *Cell*, vol. 174, no. 4, pp. 968–981, 2018.

[6] R. Ke, M. Mignardi, A. Pacureanu, J. Svedlund, J. Botling, C. Wählby, and M. Nilsson, "In situ sequencing for rna analysis in preserved tissue and cells," *Nature Methods*, vol. 10, no. 9, pp. 857–860, 2013.

[7] J. H. Lee, E. R. Daugharthy, J. Scheiman, R. Kalhor, T. C. Ferrante, R. Terry, B. M. Turczyk, J. L. Yang, H. S. Lee, J. Aach, K. Zhang, and G. M. Church, "Fluorescent in situ sequencing (FISSEQ) of rna for gene expression profiling in intact cells and tissues," *Nature Protocols*, vol. 10, no. 3, pp. 442–458, 2015.

[8] D. Schapiro, H. W. Jackson, S. Raghuraman, J. R. Fischer, V. R. T. Zanotelli, D. Schulz, C. Giesen, R. Catena, Z. Varga, and B. Bodenmiller, "histoCAT: analysis of cell phenotypes and interactions in multiplex image cytometry data," *Nat Methods*, vol. 14, no. 9, pp. 873–876, 2017.

[9] Y. Yuan, H. Failmezger, O. M. Rueda, H. Raza Ali, S. Gräf, S. F. Chin, R. F. Schwarz, C. Curtis, M. J. Dunning, H. Bardwell, N. Johnson, S. Doyle, G. Turashvili, E. Provenzano, S. Aparicio, C. Caldas, and F. Markowetz, "Quantitative image analysis of cellular heterogeneity in breast tumors complements genomic profiling," *Science Translational Medicine*, vol. 4, no. 157, pp. 157ra143–157ra143, 2012.

[10] D. Keim, J. Kohlhammer, G. Ellis, and F. Mansmann, "Mastering the information age: solving problems with visual analytics," Goslar: Eurographics Association, 2010.

[11] J. Eberwine, H. Yeh, K. Miyashiro, Y. Cao, S. Nair, R. Finnell, M. Zettel, and P. Coleman, "Analysis of gene expression in single live neurons," *Proceedings of the National Academy of Sciences*, vol. 89, no. 7, pp. 3010–3014, 1992.

[12] R. N. Van Gelder, M. E. von Zastrow, A. Yool, W. C. Dement, J. D. Barchas, and J. H. Eberwine, "Amplified rna synthesized from limited quantities of heterogeneous cdna," *Proceedings of the National Academy of Sciences*, vol. 87, no. 5, pp. 1663–1667, 1990.

[13] A. McKenna, G. M. Findlay, J. A. Gagnon, M. S. Horwitz, A. F. Schier, and J. Shendure, "Whole-organism lineage tracing by combinatorial and cumulative genome editing," *Science*, vol. 353, no. 6298, 2016.

[14] K. H. Chen, A. N. Boettiger, J. R. Moffitt, S. Wang, and X. Zhuang, "Spatially resolved, highly multiplexed rna profiling in single cells," *Science*, vol. 348, no. 6233, 2015.

[15] S. Shah, E. Lubeck, W. Zhou, and L. Cai, "In situ transcription profiling of single cells reveals spatial organization of cells in the mouse hippocampus," *Neuron*, vol. 92, no. 2, pp. 342–357, 2016.

[16] M. Angelo, S. C. Bendall, R. Finck, M. B. Hale, C. Hitzman, A. D. Borowsky, R. M. Levenson, J. B. Lowe, S. D. Liu, S. Zhao, *et al.*, "Multiplexed ion beam imaging of human breast tumors," *Nature medicine*, vol. 20, no. 4, pp. 436–442, 2014.

[17] M. J. Gerdes, C. J. Sevinsky, A. Sood, S. Adak, M. O. Bello, A. Bordwell, A. Can, A. Corwin, S. Dinn, R. J. Filkins, *et al.*, "Highly multiplexed single-cell analysis of formalin-fixed, paraffin-embedded cancer tissue," *Proceedings of the National Academy of Sciences*, vol. 110, no. 29, pp. 11982–11987, 2013.

[18] M. E. Ijsselsteijn, T. P. Brouwer, Z. Abdulrahman, E. Reidy, A. Ramalheiro, A. M. Heeren, A. Vahrmeijer, E. S. Jordanova, and N. F. de Miranda, "Cancer immunophenotyping by seven-colour multispectral imaging without tyramide signal amplification," *The Journal of Pathology: Clinical Research*, vol. 5, pp. 3–11, 2019.

[19] D. R. Bandura, V. I. Baranov, O. I. Ornatsky, A. Antonov, R. Kinach, X. Lou, S. Pavlov, S. Vorobiev, J. E. Dick, and S. D. Tanner, "Mass cytometry: Technique for real time single cell multitarget immunoassay based on inductively coupled plasma time-of-flight mass spectrometry," *Analytical Chemistry*, vol. 81, no. 16, pp. 6813–6822, 2009.

[20] M. Bibbo and D. Wilbur, *Comprehensive Cytopathology E-Book*. Elsevier Health Sciences, 2014.

[21] J. Seuma, J. Bunch, A. Cox, C. McLeod, J. Bell, and C. Murray, "Combination of immunohistochemistry and laser ablation icp mass spectrometry for imaging of cancer biomarkers," *Proteomics*, vol. 8, no. 18, pp. 3775–3784, 2008.

[22] J.-P. Fortin, N. Cullen, Y. I. Sheline, W. D. Taylor, I. Aselcioglu, P. A. Cook, P. Adams, C. Cooper, M. Fava, P. J. McGrath, *et al.*, "Harmonization of cortical thickness measurements across scanners and sites," *Neuroimage*, vol. 167, pp. 104–120, 2018.

[23] S. Beucher, "Use of watersheds in contour detection," in *Proceedings of the International Workshop on Image Processing*, CCETT, 1979.

[24] S. Osher and J. A. Sethian, "Fronts propagating with curvature-dependent speed: Algorithms based on hamilton-jacobi formulations," *Journal of Computational Physics*, vol. 79, no. 1, pp. 12–49, 1988.

[25] N. F. Greenwald, G. Miller, E. Moen, A. Kong, A. Kagel, C. C. Fullaway, B. J. McIntosh, K. Leow, M. S. Schwartz, T. Dougherty, C. Pavelchek, S. Cui, I. Camplisson, O. Bar-Tal, J. Singh, M. Fong, G. Chaudhry, Z. Abraham, J. Moseley, S. Warshawsky, E. Soon, S. Greenbaum, T. Risom, T. Hollmann, L. Keren, W. Graf, M. Angelo, and D. Van Valen, "Whole-cell segmentation of tissue images with human-level performance using large-scale data annotation and deep learning," *bioRxiv*, 2021.

[26] S. Berg, D. Kutra, T. Kroeger, C. N. Straehle, B. X. Kausler, C. Haubold, M. Schiegg, J. Ales, T. Beier, M. Rudy, *et al.*, "Ilastik: interactive machine learning for (bio) image analysis," *Nature Methods*, vol. 16, no. 12, pp. 1226–1232, 2019.

[27] M. Abdolhoseini, M. G. Kluge, F. R. Walker, and S. J. Johnson, "Segmentation, tracing, and quantification of microglial cells from 3d image stacks," *Scientific reports*, vol. 9, no. 1, pp. 1–10, 2019.

[28] PerkinElmer, "inForm Advanced Image Analysis Software - For Accurately Quantifying Biomarkers in Tissue Sections." https://tinyurl.com/y2ckm848, 2012. [Online; Accessed: 2021-07-14].

[29] J. W. Tukey, "The Future of Data Analysis," *The Annals of Mathematical Statistics*, vol. 33, no. 1, pp. 1 – 67, 1962.

[30] Wikipedia contributors, "Anscombe's quartet — Wikipedia, the free encyclopedia." https://en.wikipedia.org/wiki/Anscombe [Online; Accessed 2021-06-10].

[31] F. J. Anscombe, "Graphs in statistical analysis," *The American Statistician*, vol. 27, no. 1, pp. 17–21, 1973.

[32] C. A. Pickover and S. K. Tewksbury, *Frontiers of scientific visualization*. John Wiley & Sons, Inc., 1994.

[33] T. Munzner, *Visualization analysis and design*. A K Peters/CRC press, 2014.

[34] A. Endert, M. S. Hossain, N. Ramakrishnan, C. North, P. Fiaux, and C. Andrews, "The human is the loop: new directions for visual analytics," *Journal of intelligent information systems*, vol. 43, no. 3, pp. 411–435, 2014.

[35] D. A. Keim, F. Mansmann, and J. Thomas, "Visual analytics: How much visualization and how much analytics?," *SIGKDD Explor. Newsl.*, vol. 11, p. 5–8, may 2010.

[36] R. Raidou, *Visual analytics for digital radiotherapy : towards a comprehensible pipeline*. PhD thesis, Technische Universiteit Eindhoven, 2017.

[37] B. Preim, P. Klemm, H. Hauser, K. Hegenscheid, S. Oeltze, K. Toennies, and H. Völzke, "Visual analytics of image-centric cohort studies in epidemiology," in *Visualization in Medicine and Life Sciences III. Mathematics and Visualization.* (L. Linsen, B. Hamann, and H. C. Hege, eds.), pp. 221–248, Springer, 2016.

[38] K. Tsang, H. Li, F. Lam, Y. Mu, Y. Wang, and H. Qu, "Tradao: A visual analytics system for trading algorithm optimization," in *Proceedings of the 32nd IEEE Visualization Conference (IEEE VIS)*, pp. 61–65, 2020.

[39] M. P. Wachowiak, D. F. Walters, J. M. Kovacs, R. Wachowiak-Smolíková, and A. L. James, "Visual analytics and remote sensing imagery to support community-based research for precision agriculture in emerging areas," *Computers and Electronics in Agriculture*, vol. 143, pp. 149–164, 2017.

[40] A. Abdul-Rahman, G. Roe, M. Olsen, C. Gladstone, R. Whaling, N. Cronk, R. Morrissey, and M. Chen, "Constructive visual analytics for text similarity detection," in *Computer Graphics Forum*, vol. 36, pp. 237–248, Wiley Online Library, 2017.

[41] D. A. Keim, F. Mansmann, J. Schneidewind, and H. Ziegler, "Challenges in visual data analysis," in *Tenth International Conference on Information Visualisation (IV'06)*, pp. 9–16, IEEE, 2006.

[42] A. Çöltekin, S. Fabrikant, and M. Lacayo, "Exploring the efficiency of users' visual analytics strategies based on sequence analysis of eye movement recordings," *International Journal of Geographical Information Science*, vol. 24, no. 10, pp. 1559–1575, 2010.

[43] A. Somarakis, M. E. Ijsselsteijn, B. Kenkhuis, V. v. Unen, S. J. Luk, F. Koning, L. v. d. Weerd, N. F. C. C. d. Miranda, B. P. F. Lelieveldt, and T. Höllt, "Visual Analysis of Tissue Images at Cellular Level," in *EuroVis 2021 - Dirk Bartz Prize* (S. Oeltze-Jafra and R. G. Raidou, eds.), The Eurographics Association, 2021.

[44] A. Somarakis and T. Höllt, "ImaCytE open source software." https://www.doi.org/10.5281/zenodo.3345951, 2019. [Online; Accessed: 2021-07-14].

[45] A. Somarakis and T. Höllt, "SpaCeCo open source software." https://www.doi.org/10.5281/zenodo.3885814, 2020. [Online; Accessed: 2021-07-14].

[46] O. Dzyubachyk, W. A. Van Cappellen, J. Essers, W. J. Niessen, and E. Meijering, "Advanced level-set-based cell tracking in time-lapse fluorescence microscopy," *IEEE Transactions on Medical Imaging*, vol. 29, no. 3, pp. 852–867, 2010.