

Spatially Resolved Gene Expression Analysis

Michaela Asp

Doctoral Thesis in Biotechnology
KTH Royal Institute of Technology
School of Engineering Sciences in Chemistry,
Biotechnology and Health
Department of Gene Technology
Science for Life Laboratory
Stockholm, Sweden 2018

© Michaela Asp
Stockholm 2018

KTH Royal Institute of Technology
School of Engineering Sciences in Chemistry, Biotechnology and Health
Department of Gene Technology

Science for Life Laboratory
SE-171 65 Solna
Sweden

Printed by Universitetsservice US-AB
Drottning Kristinasväg 53B
SE-100 44 Stockholm
Sweden

ISBN: 978-91-7729-965-3
TRITA-CBH-FOU-2018:43

Public defense of dissertation

This thesis will be defended **October 26th** at 10:00 in Gardaulan, Nobels väg 18, Folkhälsomyndigheten, Solna, for the degree of Doctor of Philosophy (PhD) in Biotechnology.

Respondent

Michaela Asp MSc. in Biotechnology
Department of Gene Technology, KTH Royal Institute of Technology,
Science for Life Laboratory, Stockholm, Sweden

Faculty opponent

Prof. Thierry Voet
Sanger Institute–EBI Single-Cell Genomics Centre, Wellcome Trust Sanger Institute,
Hinxton, UK
Department of Human Genetics, University of Leuven, Leuven, Belgium

Evaluation committee

Dr. Åsa Björklund
Department for Cell and Molecular Biology, National Bioinformatics Infrastructure Sweden,
Science for Life Laboratory, Uppsala University, Uppsala, Sweden

Assoc. Prof. Lars Feuk
Department of Immunology, Genetics and Pathology, Uppsala University, Uppsala, Sweden

Assoc. Prof. Marc Friedländer
Department of Molecular Biosciences, The Wenner-Gren Institute, Stockholm University,
Science for Life Laboratory, Stockholm, Sweden

Chairman

Prof. Stefan Ståhl
Division of Protein Technology, KTH Royal Institute of Technology,
AlbaNova University Center, Stockholm, Sweden

Respondent's main supervisor

Prof. Joakim Lundeberg
Department of Gene Technology, KTH Royal Institute of Technology,
Science for Life Laboratory, Stockholm, Sweden

Respondent's co-supervisor

Asst. Prof. Patrik L. Ståhl
Department of Gene Technology, KTH Royal Institute of Technology,
Science for Life Laboratory, Stockholm, Sweden

Abstract

Spatially resolved transcriptomics has greatly expanded our knowledge of complex multicellular biological systems. To date, several technologies have been developed that combine gene expression data with information about its spatial tissue context. There is as yet no single spatial method superior to all others, and the existing methods have jointly contributed to progress in this field of technology. Some challenges presented by existing protocols include having a limited number of targets, being labor extensive, being tissue-type dependent and having low throughput or limited resolution. Within the scope of this thesis, many aspects of these challenges have been taken into consideration, resulting in a detailed evaluation of a recently developed spatial transcriptome-wide method. This method, termed Spatial Transcriptomics (ST), enables the spatial location of gene activity to be preserved and visually links it to its histological position and anatomical context. Paper I describes all the details of the experimental protocol, which starts when intact tissue sections are placed on barcoded microarrays and finishes with high throughput sequencing. Here, spatially resolved transcriptome-wide data are obtained from both mouse olfactory bulb and breast cancer samples, demonstrating the broad tissue applicability and robustness of the approach. In Paper II, the ST technology is applied to samples of human adult heart, a tissue type that contains large proportions of fibrous tissue and thus makes RNA extraction substantially more challenging. New protocol strategies are optimized in order to generate spatially resolved transcriptome data from heart failure patients. This demonstrates the advantage of using the technology for the identification of lowly expressed biomarkers that have previously been seen to correlate with disease progression in patients suffering heart failure. Paper III shows that, although the ST technology has limited resolution compared to other techniques, it can be combined with single-cell RNA-sequencing and hence allow the spatial positions of individual cells to be recovered. The combined approach is applied to developing human heart tissue and reveals cellular heterogeneity of distinct compartments within the complete organ. Since the ST technology is based on the sequencing of mRNA tags, Paper IV describes a new version of the method, in which spatially resolved analysis of full-length transcripts is being developed. Exploring the spatial distribution of full-length transcripts in tissues enables further insights into alternative splicing and fusion transcripts and possible discoveries of new genes.

Keywords

RNA, RNA-sequencing, transcriptomics, spatial transcriptomics, single cells

Sammanfattning

Spatialt upplöst transkriptomik har markant breddat vår kunskap om komplexa multicellulära biologiska system. Idag har flertalet tekniker utvecklats med avsikt att länka genuttrycksdata med dess fysiska vävnadssammanhang. Även om det ännu inte finns en enda enskild teknik med överlägsna fördelar gentemot andra tekniker så har befintliga metoder gemensamt bidragit till framsteg av teknikutvecklingen. Befintliga protokoll omfattar dock fortfarande vissa utmaningar, så som ett limiterat antal gener, tungt laborativt arbete, beroende av vävnadstyp, storskalighet eller begränsad upplösning. Inom ramen för denna avhandling har aspekterna av dessa utmaningar beaktats och resulterat i en omfattande utvärdering av en nyligen utvecklad spatialt upplöst hel-transkriptom-omfattande metod. Denna metod går under namnet Spatial Transcriptomics (ST) och möjliggör en bevarad lokalisering av genaktivitet, vilken samtidigt kan kopplas visuellt till dess histologiska position och anatomiska sammanhang. Artikel I beskriver detaljerna kring det experimentella protokollet, vilket börjar med att placera intakta vävnadssnitt på avkodningsbara mikroarrayer och avslutas med storskalig sekvensering. Här visas spatialt upplöst hel-transkriptom data från såväl vävnad av musluktlob och bröstcancerprover, vilket visar på dess breda applicerbarhet och robusthet. Artikel II applicerar ST tekniken på adult humana hjärtprover, en vävnadstyp innehållande stora proportioner av fibrös vävnad vilket gör RNA extraheringen väsentligt mer utmanande. Nya strategier för protokollutförandet optimerades för att kunna generera spatialt upplöst transkriptomik data från patienter med hjärtsvikt. Här demonstreras fördelen med att använda tekniken för att detektera de lågt uttryckta biomarkörer som vanligtvis kan ses i sjukdomsutvecklingen av hjärtsvikt. Trots att ST tekniken har en begränsad upplösning jämfört med andra tekniker visar artikel III att metoden kan kombineras med RNA-sekvensering på enskilda celler, och på så sätt erhålla vävnadsspecifika positioner för enskilda celler. Detta kombinerade tillvägagångssätt appliceras på human hjärtutveckling och visar på cellulär heterogenitet inom de olika regionerna av organet. Eftersom ST tekniken är baserad på sekvensering av mRNA-taggar, beskriver artikel IV en ny version av metoden, där spatialt upplöst analys av full-längdstranskript utvecklas. Detta möjliggör ytterligare insikter om alternativ splicing, fusions transkript och upptäckten av nya gener.

Nyckelord

RNA, RNA-sekvensering, transkriptomik, spatialt upplöst transkriptomik, enskilda celler

List of publications

- I. Patrik L. Ståhl*, Fredrik Salmén*, Sanja Vickovic**, Anna Lundmark**, José Fernández Navarro, Jens Magnusson, Stefania Giacomello, **Michaela Asp**, Jakub O. Westholm, Mikael Huss, Annelie Mollbrink, Sten Linnarsson, Simone Codeluppi, Åke Borg, Fredrik Pontén, Paul Igor Costea, Pelin Sahlén, Jan Mulder, Olaf Bergmann, Joakim Lundeberg, Jonas Frisén. *Visualization and analysis of gene expression in tissue sections by spatial transcriptomics*. Science 353: 78-82 (2016). doi: 10.1126/science.aaf2403
- II. **Michaela Asp**, Fredrik Salmén*, Patrik L. Ståhl*, Sanja Vickovic*, Ulrika Felldin, Marie Löfling, José Fernandez Navarro, Jonas Maaskola, Maria J. Eriksson, Bengt Persson, Matthias Corbascio, Hans Persson, Cecilia Linde, Joakim Lundeberg. *Spatial detection of fetal marker genes expressed at low level in adult human heart tissue*. Sci. Rep. 7: 1-10 (2017). doi: 10.1038/s41598-017-13462-5
- III. **Michaela Asp**, Stefania Giacomello, Daniel Fürth*, Johan Reimegård*, Eva Wärdell*, Joaquin Custodio, Fredrik Salmén, Erik Sundström, Elisabet Åkesson, Patrik L. Ståhl, Magda Bienko, Agneta Månsson-Broberg, Christer Sylvén, Joakim Lundeberg. *An organ-wide gene expression atlas of the developing human heart*. Submitted.
- IV. **Michaela Asp**, Erik Borgström*, Alex Stuckey*, Joel Gruselius, Konstantin Carlberg, Zaneta Andrusivova, Fredrik Salmén, Max Käller, Patrik L. Ståhl, Joakim Lundeberg. *Spatial isoform profiling within individual tissue sections*. Manuscript.

* These authors contributed equally to the work.

**These authors contributed equally to the work.

*To my Grandmother on your 89th birthday
Till min Farmor på din 89 års dag*

Contents

INTRODUCTION	1
1. A spatial aspect of biology	1
1.1 The building blocks of an organism	1
1.2 The genetic code	1
1.3 The central dogma of molecular biology	2
1.4 The importance of spatial context in biology	3
2. Analyzing gene expression	5
2.1 Background	5
2.1.1 Northern blot	5
2.1.2 Polymerase Chain Reaction (PCR)	5
2.1.3 Quantitative real-time PCR (qPCR)	6
2.1.4 Microarrays	6
2.1.5 Sanger sequencing, EST, SAGE and CAGE	6
2.2 RNA sequencing (RNA-seq)	7
2.2.1 Massively parallel sequencing	7
2.2.2 Single molecule and long read sequencing	7
2.3 Single cell RNA sequencing (scRNA-seq)	8
2.3.1 First RNA-seq of a single cell	8
2.3.2 PCR amplification using the terminal transferase approach	8
2.3.3 PCR amplification using the template-switch approach	9
2.3.4 Linear amplification	9
2.3.5 Scaling up the multiplexity of scRNA-seq	9
2.3.6 Single cell combined approaches	10
3. Analyzing spatially resolved gene expression	11
3.1 Technologies based on microdissected gene expression	11
3.1.1 Laser capture microdissection (LCM)	11
3.1.2 Analyzing individual cryosections	11
3.1.3 RNA tomography (tomo-seq)	11
3.1.4 Transcriptome in vivo analysis (TIVA)	12
3.1.5 NICHE-seq	12
3.2 <i>In situ</i> hybridization technologies	13
3.2.1 Single-molecule RNA fluorescence in situ hybridization (smFISH)	13
3.2.2 Sequential hybridization (seqFISH)	13
3.2.3 Multiplexed error-robust FISH (MERFISH)	13
3.3 <i>In situ</i> sequencing technologies	15
3.3.1 In situ sequencing using padlock probes	15
3.3.2 Barcode in situ targeted sequencing (BaristaSeq)	16
3.3.3 Spatially-resolved Transcript Amplicon Readout Mapping (STARmap)	16
3.3.4 Fluorescent in situ RNA Sequencing (FISSEQ)	16
3.4 <i>In situ</i> capturing technologies	19
3.4.1 Spatial Transcriptomics (ST)	19
3.5 <i>In silico</i> reconstruction of spatial data	20
3.5.1 Using ISH as reference maps	20
3.5.2 Using ST as reference maps	20

PRESENT INVESTIGATION	22
Paper I – A spatially resolved transcriptome-wide approach to study whole tissue sections	22
Paper II – Detecting lowly expressed biomarkers in heart failure disease progression.....	24
Paper III – Tracing cellular heterogeneity within the developing human heart	25
Paper IV – Constructing spatially resolved full-length transcriptomes within whole tissue sections	26
Concluding remarks	28
Abbreviations	30
Acknowledgments.....	31
References	34

*Don't kid yourself that you're going to live again after you're dead; you're not.
Make the most of the one life you've got. Live it to the full.*

- Richard Dawkins

INTRODUCTION

1. A spatial aspect of biology

The word “spatial”, which originates from the Latin word *spatium*, meaning “space”, can be used when describing how objects relate to each other with regard to their relative positions. The spatial concept allows a biological system to be described as a global network where each element is influenced by its surrounding environment. Single cells in close proximity constitute tissues and individual organs are linked into organ systems. Studying biology therefore means that every piece of information has to be put into context in order to fully understand each element’s entire repertoire of mechanisms. Here follows an introduction to various biological units and their spatial relations to each other.

1.1 The building blocks of an organism

Organisms are life forms consisting of one or more cells: the basic biological unit of life¹. A unicellular organism, such as a member of the bacteria or archaea, is the simplest form of organism and absorbs its nutrition directly from the environment through its cellular membrane. Once it is ready to reproduce, it simply divides itself into two unicellular organisms. A multicellular organism, however, is a lot more complex. Animals, plants and fungi are examples of organisms that can contain many cells. For example, the number of cells in the human body has been estimated at ~37 trillion², all of which have defined roles within the biological system. Similar cells in multicellular organisms are grouped into tissues (muscle-, nervous-, connective tissue etc.) and different types of tissues together form organs (heart, brain, lungs etc.), which, in turn, make up whole organ systems (cardiovascular-, respiratory-, nervous system etc.) that constitute the entire organism. When it comes to the reproduction of a multicellular organism, the process is way more complex than just dividing itself (and far beyond the scope of this thesis).

1.2 The genetic code

Looking more closely into a cellular unit, nearly every one of those that constitute a human being contains a nucleus. The nucleus holds possibly the most famous biological molecule in the world – deoxyribonucleic acid (DNA). DNA is a long molecule, and if the content of all our cells together were stretched out it would reach to the sun and back more than 80 times^{2,3}. DNA contains the genetic code, or the “blueprint”, and carries all instructions regarding cellular functions. It also contains information determining our individual traits, such as eye color, whether odor can be

detected in the urine after eating asparagus and if we are susceptible to certain diseases – some of these traits are passed on to our offspring during reproduction. As an example, all blue-eyed people in the world share a common ancestor who lived ~10,000 years ago. Originally, we all had brown eyes but that single common ancestor carried a mutation affecting the production of melanin in the iris⁴, which resulted in a blue eye color.

Zooming in even further down to a DNA molecular unit, one can see that its structure is somewhat similar to a twisted ladder. The steps of the ladder are comprised of even smaller building blocks called nucleotides or bases. There are four types of DNA bases: adenine (A), thymine (T), guanine (G) and cytosine (C). It is the ordering of these bases (or the steps of the ladder) that makes up the personal blueprint for every individual. The presence of DNA has been known for over 100 years and it was isolated for the first time in 1869 by a Swiss chemist named Friedrich Miescher⁵. He named it “nuclein” as it was found inside the cells’ nuclei, and in 1944 it was shown that DNA is the hereditary material⁶. In 1953, Rosalind Franklin and Maurice Wilkins used X-ray data to demonstrate that DNA is assembled in a repeated helix structure. James Watson and Francis Crick got hold of the unpublished data (without Franklin's permission), and made the inference that the DNA molecule is composed of a double stranded helix, where A always pairs with T and C always with G. All these discoveries were published in the same issue of the journal *Nature*⁷⁻⁹. About 40 years later, in 1990, an international scientific research project set the goal of determining the order of all bases in the human genome, which has a total size of ~3 billion base pairs¹⁰. The project was named the Human Genome Project and as the project progressed, the private company Celera Genomics joined the race in 1998. Both the public and the private effort published their draft genomes in 2001^{11,12}.

1.3 The central dogma of molecular biology

Cells in our body contain almost the same blueprint¹³, so how come they do not all look the same or have the same function? The answer lies in the process of transferring genetic information from DNA into a final product within the cell, a process referred to as the central dogma of molecular biology. The central dogma, which was introduced by Francis Crick in 1958¹⁴, concerns three mechanisms performed by the cellular machinery. Existing DNA can be copied to produce new DNA (replication) or a ribonucleic acid (RNA) (transcription). The RNA molecule can possess direct activity and have, for example, a regulatory function within the cell. It can also act as an intermediate between DNA and the final product, a protein. RNA can be copied to make new RNA (replication) and used as the template for proteins (translation). Proteins, the main building blocks of the cell, cannot be replicated or

copied into DNA or RNA. The central dogma was refined in 1970 when it was discovered that RNA can be copied into DNA (reverse transcription)¹⁵.

The DNA molecule contains shorter segments referred to as “genes”. The definition of a gene varies within the scientific community but generally, one can explain the concept of a gene as a piece of DNA, which can be transcribed into RNA and has a function within the cell¹⁶. Different genes are turned on or off in different cells and genes that are active can be so at different levels. There are different “flavors” of RNA species, however, this thesis will focus mainly on the messenger RNA molecule (mRNA – a single-stranded RNA molecule that is later translated to proteins). As mRNAs are the intermediates between DNA and proteins, they can be used as indicators of a specific cell type. In 2016, an international effort called the Human Cell Atlas consortium set the goal of identifying all cell types in the human body¹. With the implementation of this large-scale project, the number of cell types identified is expected to increase significantly over the next few years. A cell’s complement of RNA molecules will hereafter be referred to as either its “transcriptome”, its “gene expression” or simply to as its “content of RNA molecules”.

1.4 The importance of spatial context in biology

Let us return to the spatial concept and how it describes biology. The molecular properties of individual cells within a multicellular organism can only be completely determined once we know their physical locations¹⁷. Cells within distinct tissue microenvironments express specific sets of genes, both influenced by, and influencing, the cells around them. This phenomenon governs, for example, the formation of gene expression gradients along the main embryonic body axes during different stages of development. These gradients direct the activation of the correct developmental gene programs, needed for the construction of specific organs^{18,19}. Moreover, cells located in the same organ are not always uniform, a state referred to as cellular heterogeneity which is always present to some degree in any cellular population. When it comes to tumor microenvironments, the cellular environments in which tumors exist, several sub-populations of cancer cells constituting the tumor can differ from each other completely in terms of both structural features and gene expression. Even down to a subcellular level, it is recognized that organelles and molecules are spatially localized in certain positions in order to carry out all necessary cellular functions. Proteins are spatially localized in certain structures²⁰ and similarly, DNA is found in both the nucleus and the mitochondria and, in plants, the plastids. Some cells, such as muscle fibers, even have multiple nuclei²¹ and the overall spatial distribution of DNA within the nucleus is arranged in a dynamic way^{22,23}. RNA can also have different localizations within the cell; for example is half

of the neuronal mRNAs that are present in rat hippocampus are enriched in the axons or dendrites compared to their levels in the cell body²⁴.

Earlier biological studies aiming to understand cellular heterogeneity within biological samples were carried out using histological methods. Tissue sections are first stained to highlight cellular structures or to tag particular molecules of interest. This is still done extensively today by, for example, pathologists wishing to distinguish diseased tissue from healthy and particularly to discriminate between cancer types²⁵. Because it is based on a subjective visual evaluation, there are occasions when pathologists disagree on their findings²⁶. An objective classification of cells within a sample could be achieved by combining the histological image with information about gene expression profiles. The work in this thesis concerns how we can connect gene expression profiles with morphological positions.

2. Analyzing gene expression

The entire set of an organism's DNA is called its genome. It contains both coding parts (genes) and non-coding parts. The coding parts give rise to RNA molecules, and the sum of these RNAs is referred to as its transcriptome. The human genome contains more than 20,000 distinct genes, expressed at different levels in different tissues. Determining the DNA sequence of a cell's genome tells us what the cell is capable of doing, whereas analyzing its RNA content instead gives us an indication of what the cell is doing at that particular point in time. A brief historical timeline of transcriptomics methods mentioned in this chapter is shown in Figure 1.

2.1 Background

A wide range of technologies for analyzing RNA molecules within a biological sample has preceded today's techniques. These technologies have typically been hybridization-, amplification- or sequencing-based approaches.

2.1.1 Northern blot

Early gene expression analysis was performed by means of the northern blot technology developed in 1977²⁷. RNA is first extracted from a tissue- or a cell culture sample, after which RNA molecules are separated by size using electrophoresis. Separated RNA molecules are then transferred to a blotting membrane, on which detection is achieved by hybridizing a labeled complementary DNA nucleotide sequence (a probe) corresponding to the sequence of a known expressed gene.

2.1.2 Polymerase Chain Reaction (PCR)

The polymerase chain reaction (PCR) was developed in the 1980s by Kary Mullis²⁸, who received the Nobel Prize for it in 1993. The technique is one of the most important scientific inventions and is still used extensively today in many protocols. It is sometimes referred to as "molecular photocopying"²⁹, where small pieces of DNA are amplified to produce a large number of copies. It constitutes a repeated series of heating cycles, where each cycle contains three separate steps: (i) Denaturing of double stranded DNA into single stranded DNA, (ii) Hybridization of specific primers (short DNA fragments), complementary to the flanking areas of the DNA region of interest, and (iii) Elongation, where the primers are extended over the DNA target by addition of free nucleotides by DNA polymerase. The extended primers generate new DNA target molecules, used as templates for the next amplification cycle. This means that theoretically the number of DNA molecules is doubled after each cycle and that the amplification rate is exponential. However, because some DNA fragments are

more accessible for amplification than others, the reaction efficiency is less than 100%³⁰⁻³².

2.1.3 Quantitative real-time PCR (qPCR)

The development of PCR technology subsequently enabled gene expression studies to be carried out by modifying the technique into a version called quantitative real-time PCR (qPCR)³³. In qPCR, the amplification of DNA molecules is monitored in real-time, and not solely at the end of the reaction. This is done by including fluorescent molecules in the PCR reaction mix (previously ethidium bromide was used but modern protocols often use SYBR Green^{34,35}), which then bind nonspecifically to double-stranded DNA. A more specific measurement can be achieved by using target-complementary probes attached to a fluorophore³⁵. As the fluorescent signal is proportional to the amount of DNA molecules it is possible to quantify the amount of molecules present in the beginning of the reaction.

2.1.4 Microarrays

Another method, which exceeds the capacity of qPCR in terms of multiplexing (the ability to process and analyzing multiple samples at once), is the microarray approach, which has been used extensively in gene expression studies since 1995³⁶. Here, clusters containing short complementary DNA nucleotide sequences (probes) corresponding to known gene sequences are attached to a solid-state surface. Extracted genetic material from a tissue sample is then labeled and applied to the microarray. If a gene (the target) is expressed, a hybridization-match between the array-probes and the target will occur, resulting in a signal. Relative abundances of target sequences can then be determined and quantitative measurements of gene expression within the sample can be made. However, microarrays suffer from three major limitations: (i) prior knowledge about gene sequences is required (ii) cross-hybridization can result in high background levels and (iii) the dynamic range of detection is limited, resulting in both background and saturated signals³⁷.

2.1.5 Sanger sequencing, EST, SAGE and CAGE

In 1977, Frederick Sanger and colleagues developed a sequencing method that could directly determine the base sequences of nucleic acids³⁸, for which he and Walter Gilbert received the Nobel Prize in 1980. It was the main technology used during the Human Genome Project and is still used today to validate sequencing data produced from newer technologies³⁹. The Sanger technology is able to read nucleotide sequences from cDNA and expressed-sequence-tags (EST) libraries (shorter subsequences of cDNA) without previous knowledge of the target sequence^{40,41}. However, EST sequencing suffers from being low-throughput, generally not quantitative, and costly for larger projects. Tag-based methods such as serial analysis

of gene expression (SAGE)⁴² and cap analysis of gene expression (CAGE)⁴³ were therefore developed to increase the throughput of sequencing data.

2.2 RNA sequencing (RNA-seq)

With the development of massively parallel sequencing technologies, one could process millions of DNA sequences in parallel, rather than 96 using Sanger⁴⁴. High-throughput technologies allow us to explore the entire transcriptome in a parallel fashion, an approach referred to as RNA-seq.

2.2.1 Massively parallel sequencing

The first commercial massively parallel sequencing approach was developed in 2005, when pyrosequencing⁴⁵ was multiplexed by immobilizing hundreds of thousands of unique DNA templates to nano-beads in a system named 454⁴⁶. Individual nucleotides are added one by one and each incorporation event is detected as light is emitted through a cascade of enzymatic reactions. The principle of nucleotide addition and detecting the incorporation of each nucleotide is called sequencing-by-synthesis (SBS). Ion Torrent technology⁴⁷ is another system utilizing SBS, but instead of detecting the incorporation event as a light signal, it registers the change in pH that occurs as a hydrogen ion is released at every base incorporation step. Another commercial massively parallel sequencing system released in 2007 is SOLiD (Sequencing by Oligo Ligation and Detection)⁴⁴. In contrast to SBS, consecutive rounds of ligation of fluorescently labeled oligonucleotides (oligos) are used to read out the sequences - a principle called sequencing-by-ligation (SBL). Many different high-throughput technologies have followed but undoubtedly the most successful one is the Solexa system, acquired by Illumina in 2007⁴⁸. The Illumina method is based on SBS, but uses fluorescently labeled nucleotides for detection.

2.2.2 Single molecule and long read sequencing

Existing high-throughput technologies have a sequencing read length of around 100-500 bases. Since most RNA molecules are much longer⁴⁹, shorter cDNA fragments have to be sequenced individually and pieced together computationally. In 2011, Pacific Biosciences (PacBio) released a commercially available platform able to perform single molecule sequencing on long DNA fragments⁵⁰. The technology is based on Single Molecule Real-Time (SMRT) sequencing, meaning that individual DNA sequences are recorded directly as fluorescently labeled nucleotides are incorporated. DNA molecules are attached to polymerases, which in turn are individually attached to the bottom of zeptoliter-sized wells, called zero-mode waveguides (ZMW)^{51,52}. Although the system has an average read length of about 15,000 bases, it does not yield the same throughput as the previous massively parallel sequencing technologies⁵³. For RNA studies, the PacBio Iso-Seq protocol has

been used to directly provide full-length information about human cDNA molecules, without the need to assemble them afterwards⁵⁴. Another technology that can be used for transcriptome studies is the Oxford Nanopore platform. This is based on immobilized protein nanopores on a polymer membrane⁵⁵. As an electrical potential is applied over the membrane, molecules are pushed through the pores and bases are detected when changes in the potential are registered. Generally, in practice the concept of sequencing RNA really means sequencing the cDNA (reverse transcribed RNA). The process of reverse transcription (RT) and additional amplification may introduce errors and biases. In 2017, Oxford Nanopore launched a new RNA sequencing solution making it possible to perform direct sequencing on full-length native RNA molecules^{56,57}.

2.3 Single cell RNA sequencing (scRNA-seq)

Initially RNA-seq only allowed for analyzing gene expression from bulk samples due to the large amount of RNA required for input. Whole pieces of tissue are homogenized and thousands of cells are analyzed simultaneously. This creates an average picture of gene expression within the bulk sample, potentially masking the products of genes that are expressed only by a minority of the cells. Measuring the gene activity of a single cell instead allows us to define gene expression profiles individually. However, there was one major challenge to be overcome in order to accomplish this; one single cell contains roughly 1-50 pg⁵⁸ of total RNA, of which only 1-5% (0.01-2.5 pg) comprise its ~300,000 mRNA molecules⁵⁹.

2.3.1 First RNA-seq of a single cell

In 2009, Tang et al.⁶⁰ improved existing single-cell RNA amplification protocols⁶¹⁻⁶³ and were able to apply their method to high-throughput RNA-seq and examine the whole transcriptome of one single cell. From then on, the number of methods available for analyzing single cells has exploded and today we are able to analyze hundreds of thousands of single cells in parallel.

2.3.2 PCR amplification using the terminal transferase approach

Most methods take advantage of the poly-A tail of eukaryotic mRNAs. By hybridizing a poly-T primer (a short single-stranded stretch of DNA nucleotides) to the poly-A tails, mRNA can be reverse transcribed to complementary DNA (cDNA). The cDNA then needs to be amplified, which can only be achieved when amplification adaptor sequences are added to the ends. The first adaptor can be incorporated together with the poly-T primer, while the second adaptor is introduced by different strategies. The Tang protocol enzymatically added another poly-A stretch to the cDNA using terminal deoxynucleotidyl transferase; a second poly-T primer containing the other

amplification adaptor can be hybridized to it. This approach was updated in 2013 to a more simplified protocol named Quartz-Seq⁶⁴. Here, the whole procedure was made more sensitive by completing all reaction steps in a single tube, without any need for purification. The protocol achieved higher multiplexing capacity in 2018⁶⁵, when a cell barcode (a sequence unique for one specific cell) and a molecular barcode (a sequence unique to one specific transcript) were introduced.

2.3.3 PCR amplification using the template-switch approach

Another approach to introducing the second amplification adaptor is to employ the template-switch mechanism⁶⁶. When the reverse transcriptase reaches the very end of the mRNA molecule, it adds a small additional number of nucleotides. The result is a short single-stranded overhanging sequence to which the second adaptor can hybridize. This has been applied in a couple of methods: in 2011 in STRT-seq⁶⁷ (Single-cell tagged reverse transcription) and in 2012 in Smart-seq⁴⁹. In STRT-seq, cDNA is fragmented and enriched for its 5' ends, while in Smart-Seq full-length information is retrieved enabling details about alternative RNA isoforms to be obtained. Both protocols have been further updated: STRT-seq introduced molecular barcodes⁶⁸ and Smart-seq2 improved its sensitivity and full-length coverage⁶⁹.

2.3.4 Linear amplification

The above-mentioned technologies are all based on *exponential* PCR amplification, which is prone to the risk of introducing sequence biases. One way to reduce amplification biases is by using *in vitro* transcription (IVT) amplification⁷⁰, a *linear* amplification-based approach. IVT was implemented in 2012 in a protocol termed CEL-Seq (Cell Expression by Linear amplification and Sequencing)⁷¹ with further optimization steps in 2016⁷². Up to 2014, protocols were able to analyze ~100 cells in parallel⁷³ but an automated IVT approach named MARS-Seq (massively parallel single-cell RNA-seq)⁷⁴ made it possible to analyze several thousands of cells in parallel.

2.3.5 Scaling up the multiplexity of scRNA-seq

To further increase the multiplexity capacity of scRNA-seq experiments up to tens of thousands⁷³ of cells, a method named CytoSeq⁷⁵, which was described in 2015, uses an approach that combines multiple shorter barcodes. This enabled the generation of a greater number of barcodes and thus increases the number of cells being analyzed. Shortly afterwards, two papers published in the same issue of the journal *Cell* introduced single cell droplet technology. The first of these, named InDrop (Index Droplets)⁷⁶, uses the previously described combinatorial indexing approach, while the second one, named Drop-seq⁷⁷, instead employs longer random barcodes. Recently, new multiplexing strategies have been developed and we are reaching a point where a number in the hundreds of thousands of cells can be analyzed

together⁷³. Here, cells are not isolated and barcoded individually. Instead, barcoding takes place inside the cell within pools of ~10-100 cells^{78,79}. Several rounds of pooling, mixing and re-pooling facilitates a decrease in the probability of two cells ending up with the same barcode.

2.3.6 Single cell combined approaches

Even though a snapshot of a cell's RNA content (which results from transcriptional bursts) can give us an indication of what cell type we are looking at, the ideal procedure would be to combine it with other omics technologies⁸⁰. Today, technologies combining single-cell transcriptomics with information such as the cell's genome^{81,82}, its proteome⁸³⁻⁸⁵ and its chromatin conformation²³ and accessibility²² have been developed.

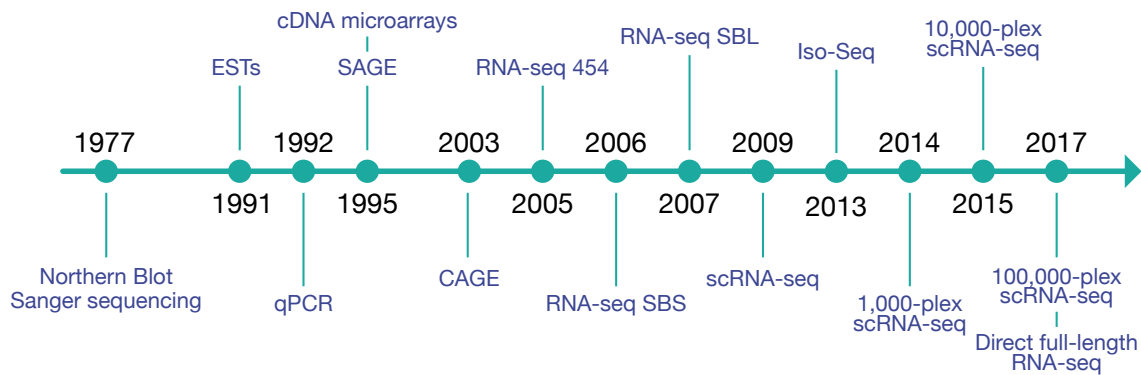


Figure 1. Brief historical timeline of transcriptomics.

3. Analyzing spatially resolved gene expression

There has been extensive analysis of gene expression in both bulk and single cell samples. However, understanding the complex network of biological functions within tissues and organs requires the spatial context of gene expression to be preserved. Today, several experimental and computational methods have been established which are designed to connect positional information with gene expression data. Characteristics of the experimental methods mentioned in this chapter are shown in Table 1.

3.1 Technologies based on microdissected gene expression

One way of capturing spatial gene expression information is simply by isolating regions of interest from within a sample. These regions can then be individually placed in test tubes for RNA extraction and subsequent gene expression profiling.

3.1.1 Laser capture microdissection (LCM)

Laser capture microdissection (LCM) is a technique in which a laser beam is used to cut out tissue regions identified under a microscope^{86,87}. In 2017, an extended version of the LCM protocol, named Geo-seq (geographical position sequencing)⁸⁸, combined LCM with scRNA-seq in order to profile the transcriptomes of tissue regions as small as ten single cells. Although LCM is a robust technology, it is very labor intensive, which limits the throughput of samples to be analyzed.

3.1.2 Analyzing individual cryosections

Another way of producing regionalized gene expression data is to slice tissue samples into thin cryosections and prepare individual sequencing libraries from each piece of tissue. In 2013 this transcriptome-wide approach was applied to *Drosophila* embryos, where cryosectioning and sequencing was performed along the anterior-posterior body axis, revealing both known and novel spatial gene expression patterns⁸⁹. High quantities of input RNA were needed for the preparation of sequencing libraries, a problem which was solved by adding large amounts of carrier RNA. This approach requires a large number of sequencing reads in order to obtain a sufficient number of the sequences originating from the individual tissue sections.

3.1.3 RNA tomography (tomo-seq)

A cryosectioning approach described in 2014, called RNA tomography (tomo-seq)⁹⁰, avoided the need of carrier RNA by linearly amplifying⁷¹ the cDNA of individual tissue sections. Here, the authors systematically cryosectioned three identical zebrafish embryos, each along one of the three main body axes: anterior-posterior, ventral-

dorsal and left-right. A 3D transcriptional profile of the embryo was then reconstructed computationally by overlapping RNA-seq information from all cryosections. This method is superior for both RNA quantification and spatial resolution compared to the earlier cryosectioning protocol⁸⁹. However, transcriptome-wide maps using tomo-seq can only be constructed using identical biological samples and cannot be applied to clinical samples.

3.1.4 Transcriptome in vivo analysis (TIVA)

In order to accurately explore gene expression in cells within their natural environment, the cells must be alive. One technology that enables spatially resolved transcriptomics of living cells in their natural environment is TIVA (transcriptome *in vivo* analysis)⁹¹, published in 2014. In the TIVA protocol, intact live tissue sections are exposed to TIVA tags (each tag is a photoactivatable mRNA capture molecule). These multifunctional tags are attached to a cell-penetrating peptide, allowing them to enter the cell cytosol. Next, TIVA tags can selectively be activated in cells of interest by laser photoactivation, after which the mRNA capture molecule can hybridize to mRNAs within the cell. TIVA tag-mRNA hybrids can then be purified and the captured mRNAs can be further analyzed by RNA-seq. Although it is currently the only method that can be applied to living tissue, its main limitation is its low throughput, as only a few single cells can be analyzed at a time. Furthermore, as TIVA is applied to live tissue, its use for the analysis of clinical samples is limited.

3.1.5 NICHE-seq

An alternative technology published in 2017, also using photoactivation, is NICHE-seq⁹². As the name implies, transcriptomes of individual cells within a specific niche (an area within a tissue that holds a specific microenvironment) are profiled. First, specific niches are visualized by intravenously transferring labeled landmark-cells (e.g. T and B cells) into transgenic mice expressing photoactivatable green fluorescent protein (GFP). Niches of interest are then photoactivated, followed by tissue dissociation, cell sorting of GFP⁺ activated cells and scRNA-seq. This technology is very high throughput and can analyze thousands of cells within a niche, but their exact positions within the photoactivated area are not known. As the current protocol depends on genetically engineered model organisms, it cannot be applied to human clinical samples.

3.2 *In situ* hybridization technologies

Instead of extracting RNA molecules from individual parts (or cells) within a tissue, one can visualize them directly in their original environment. This can be achieved by hybridizing a labeled probe complementary to the target of interest. An overview of *in situ* hybridization (ISH) technologies mentioned here is shown in Figure 2.

3.2.1 *Single-molecule RNA fluorescence in situ hybridization (smFISH)*

The ISH technique has existed since the 1960s⁹³ and has been used for visualizing gene expression from the early 1980s⁹⁴ onwards. Fluorescently labeled probes are individually hybridized to predefined RNA targets in order to visualize gene expression in fixed tissue. A further development of the fluorescence *in situ* hybridization (FISH) protocol uses instead many short probes to target different regions of the same transcript. This approach, named single-molecule RNA fluorescence *in situ* hybridization (smFISH), enables quantitative measurements of transcripts since one fluorescent spot indicates a single RNA molecule. Initially, a set of five probes (50 bases in length) each coupled to five fluorophores was used to target each transcript⁹⁵. However, labeling probes with a large set of fluorophores has two main problems: (i) They are difficult to synthesize and purify (ii) Fluorophores on the same probe can potentially interact with each other causing altered hybridization characteristics and self-quenching. An alternative smFISH method was developed in 2008, using instead a set of ~40 probes (20 bases in length), each coupled to a single fluorophore⁹⁶. Although smFISH has high sensitivity and subcellular spatial resolution, it can only target a few genes at a time.

3.2.2 *Sequential hybridization (seqFISH)*

A multiplex smFISH approach that uses sequential hybridizations (seqFISH)^{97,98} was developed in 2014. Individual transcripts are detected several times by serial rounds of hybridization, imaging and probe stripping. In every hybridization round, a pre-defined colored set of 24 encoding probes is used for each target. However, an increased number of hybridization rounds requires an increased number of smFISH probes, which makes seqFISH both expensive and time consuming.

3.2.3 *Multiplexed error-robust FISH (MERFISH)*

Another sequential hybridization method is multiplexed error-robust FISH (MERFISH)⁹⁹. Here, ~192 smFISH probes are hybridized to each RNA target, each probe carrying two flanking regions used for the readout. Fluorescent readout-probes are then hybridized to identify the targets in several rounds of hybridization, imaging, and signal extinguishing. Computational error-correction is performed after readout to account for imperfect hybridizations. Since the RNA targets only need to be hybridized with the smFISH probes once (a procedure taking ~10 hours), the

experimental time is reduced significantly. The following hybridization of the readout-probes takes ~15 minutes. In 2018, the MERFISH technique was combined with expansion microscopy¹⁰⁰ (an approach that physically enlarges tissues¹⁰¹) so as to increase the distance between overlapping RNA molecules and so increase the number of molecules that can be measured. Although no super-resolution microscopy equipment is required, the large volumes of image processing can be somewhat computationally demanding.

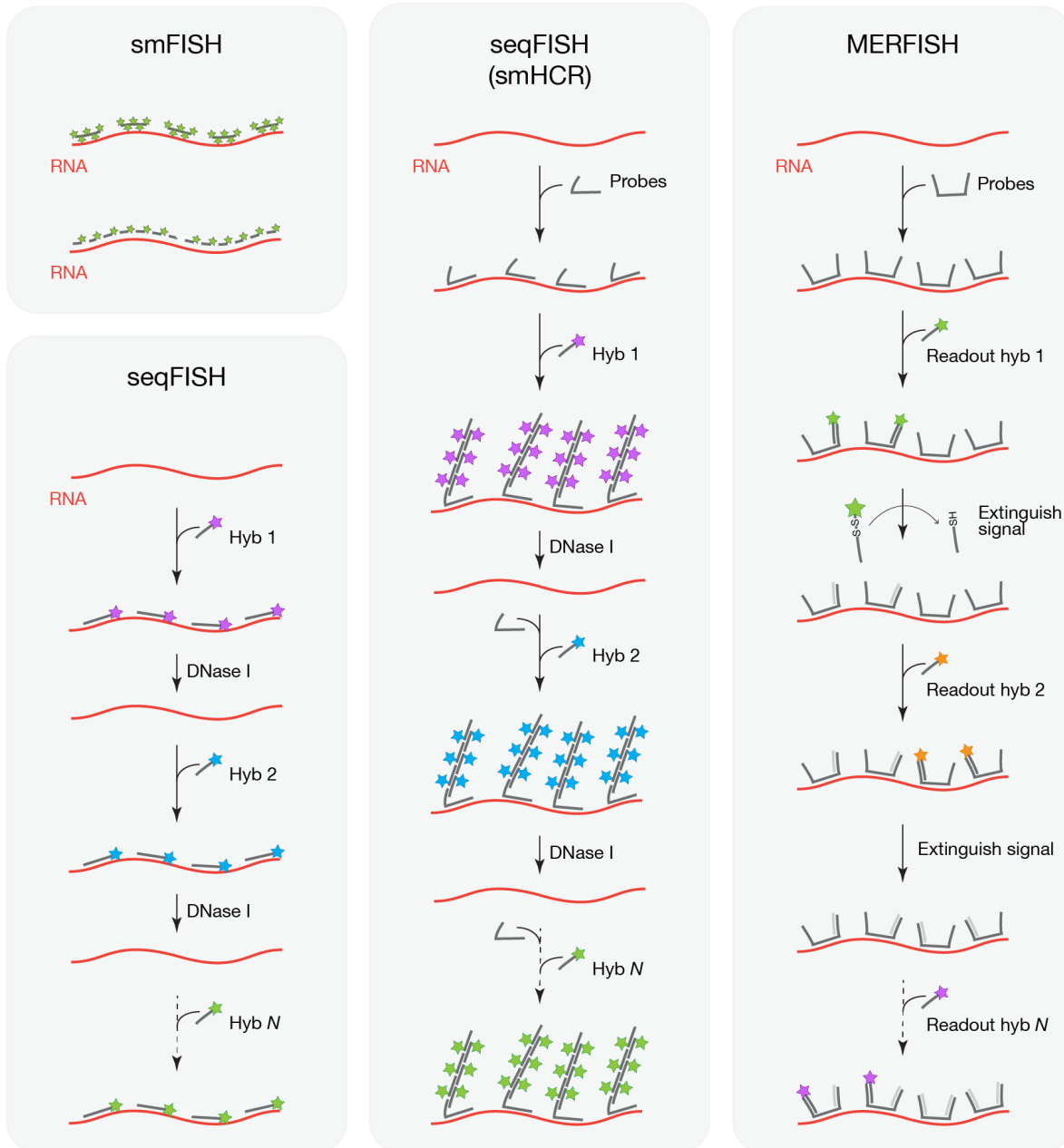


Figure 2. (*Figure appears on preceding page*)

Overview of ISH technologies. smFISH: In the original smFISH approach, a set of five probes labeled with five fluorophores each are hybridized to individual targets⁹⁵. Alternatively, a set of ~40 probes coupled to a single fluorophore is used⁹⁶. seqFISH: smFISH probes are hybridized consecutively to the target and stripped off by DNase I in between each round. The same probe sequences are used in different rounds of hybridization, but probes are connected to different fluorophores⁹⁷. seqFISH (smHCR): seqFISH in combination with single-molecule hybridization chain reaction (smHCR) to amplify the signal⁹⁸. MERFISH: A probe set with unique flanking read out regions is used for individual targets. Readout probes are hybridized sequentially in between rounds of signal extinguishing⁹⁹.

3.3 *In situ* sequencing technologies

Massively parallel sequencing can be performed directly on the RNA content of a cell while it remains in its tissue context. This is referred to as *in situ* sequencing (ISS) and several protocols have been developed. An overview of ISS technologies mentioned here is shown in Figure 3.

3.3.1 *In situ* sequencing using padlock probes

In 2013, the first *in situ* sequencing approach was published that used padlock probes to target known genes¹⁰². First, mRNA molecules within a tissue section are reverse transcribed to cDNAs, after which the original mRNAs are degraded in order to make the cDNAs accessible to the padlock probes. Padlock probes are single-stranded DNA molecules, in which each end contains regions complementary to a particular cDNA sequence. The probes can bind to the cDNA target either with a gap between the ends, or with the ends adjacent to each other. In the first case, the gap is filled by DNA polymerization and then, in both cases, the ends are ligated to create a circle of DNA. The targets are then amplified so that sequencing signals can be distinguished. This is done by amplifying the DNA circles, a process called rolling-circle amplification (RCA), resulting in micrometer-sized RCA products (RCPs) within the cells. Each RCP contains numerous repeats of the original padlock probe sequence. RCPs are then subjected to SBL, where either the gap-filled sequence, or a four-base-long barcode within the probe with adjacent ends, is decoded. By applying ISS to a breast cancer tissue section¹⁰², 31 targets were spatially localized. Although ISS with padlock probes facilitates subcellular resolution, the number of targets is limited due to the short sequencing read-length (of four bases) and the size of the RCPs.

3.3.2 Barcode *in situ* targeted sequencing (BaristaSeq)

A recently developed method called Barcode *in situ* targeted sequencing (BaristaSeq)¹⁰³ has an increased read-length of 15 bases. Instead of using SBL, it applies the Illumina sequencing chemistry (SBS), which has a higher signal-to-noise ratio than SBL. The Illumina chemistry requires repeated heat cycles, which is possible in BaristaSeq since RCPs are cross-linked using the same procedure as in FISSEQ¹⁰⁴ (described below).

3.3.3 Spatially-resolved Transcript Amplicon Readout Mapping (STARmap)

Another recently developed approach to ISS is Spatially-resolved Transcript Amplicon Readout Mapping (STARmap)¹⁰⁵. STARmap uses barcoded padlock probes that hybridize to the targets, but avoid the RT step by using a second primer, targeting the site next to the padlock probe. This duplex process is termed SNAIL (Specific Amplification of Nucleic Acids via Intramolecular Ligation). Both probes need to hybridize to the same target in order for the padlock probe to be circularized; it is then amplified by RCA to generate nanometer-sized single-stranded DNA products called nanoballs. Consequently, the SNAIL strategy reduces the noise that could otherwise occur from non-specific hybridization events when only a single padlock probe is used. Amine-modified bases are incorporated during the RCA, and used for embedding the nanoballs in a 3D scaffold by applying a specific hydrogel-tissue chemistry. The tissue-hydrogel complex is subsequently cleared of unbound proteins and lipids, enhancing the transparency of the tissue. Next, a modified SBL approach is applied to decode the five-base-long barcode. By using the hydrogel-tissue chemistry in combination with the modified sequencing protocol, more than 1,000 genes were targeted in mouse brain tissue sections.

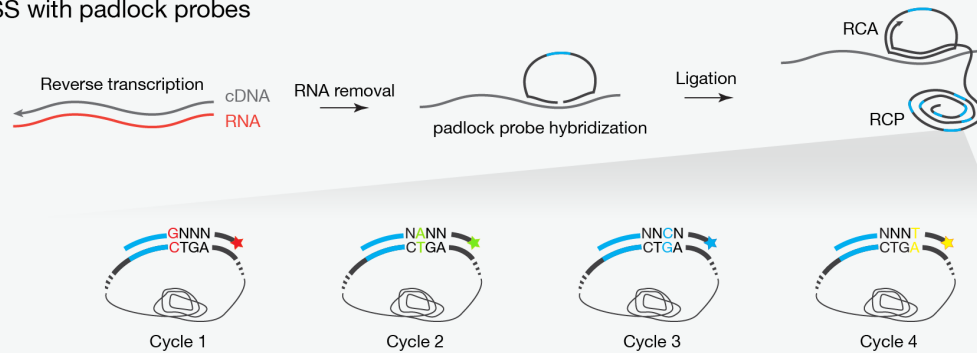
3.3.4 Fluorescent *in situ* RNA Sequencing (FISSEQ)

An untargeted ISS method named Fluorescent *in situ* RNA Sequencing (FISSEQ)¹⁰⁴ was published in 2014. This technology is similar to the padlock approach in that it also utilizes RCA to amplify the sequencing target and subsequent SBL. However, they differ from each other at several steps. First, in FISSEQ cDNA synthesis is performed using a mix of regular and modified amine-bases, together with tagged random-hexamer RT primers. Because of the incorporated amine-bases, cDNA is cross-linked to its cellular environment and thereafter circularized by ligation. Subsequent RCA creates single-stranded DNA nanoballs, whose positions are maintained via cross-linking to the cellular protein matrix. Lastly, sequencing is preformed using SBL with a read-length of 30 bases. However, sequencing every nanoball within a cell is problematic because of overcrowding; overlapping fluorescent nanoballs will be impossible to distinguish from each other. Instead, by extending the sequencing primers (which are complementary to the tag attached to the random-hexamer RT

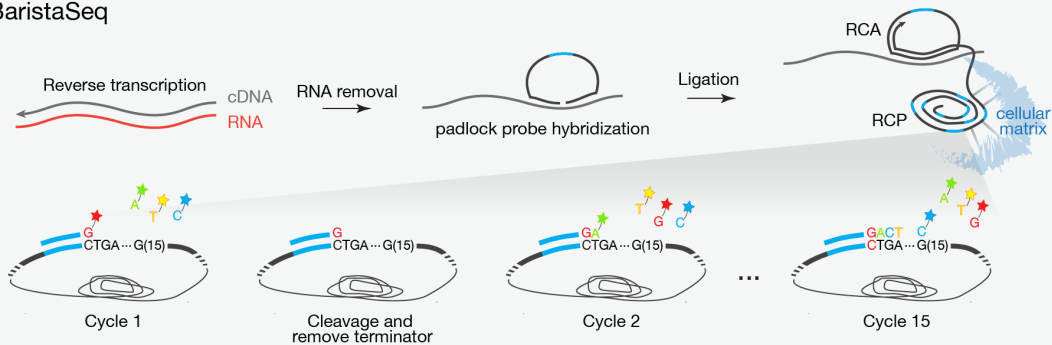
primers), nanoballs are randomly sampled so that only a subset is sequenced. By applying FISSEQ to cultured fibroblasts, more than 8,000 targets could be detected with subcellular resolution.

Although ISS technologies offer subcellular resolution, they all use micrometer- or nanometer-sized DNA balls to amplify the signal. Due to space limitation in the cell, the number of transcripts that can be detected is limited. All methods, except FISSEQ, are targeted approaches, and this further limits unbiased transcriptome-wide studies of intact tissue sections.

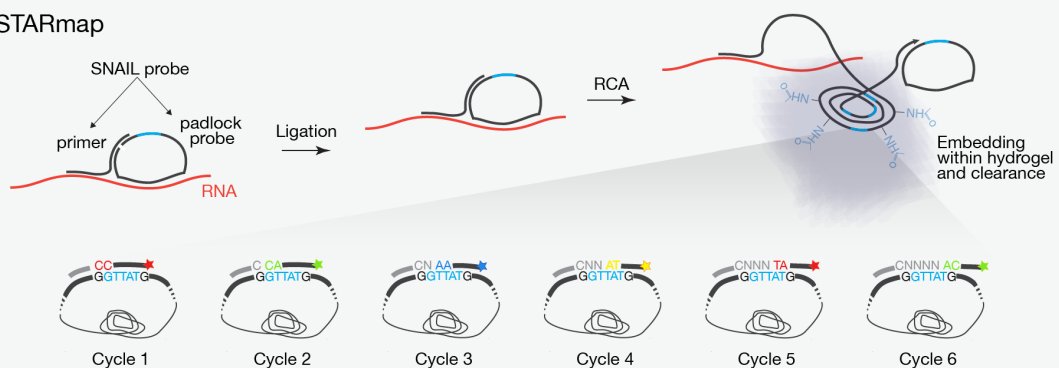
ISS with padlock probes



BaristaSeq



STARmap



FISSEQ

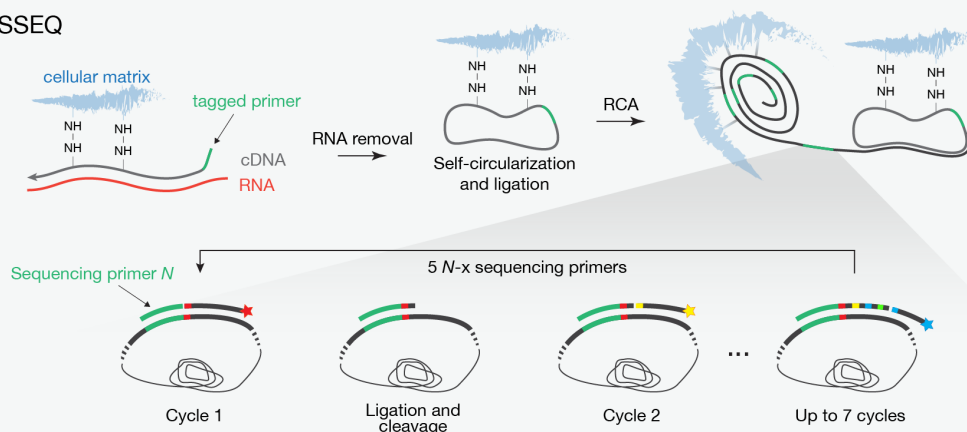


Figure 3. (*Figure appears on preceding page*)

Overview of ISS technologies. ISS with padlock probes: Padlock probe-based ISS using RCA and sequencing-by-ligation (the gapfilling approach is not shown here)¹⁰². BaristaSeq: Utilizes the padlock probe-approach but the RCP is cross-linked to the cellular matrix. Sequencing is performed using sequencing-by-synthesis¹⁰³. STARmap: A SNAIL probe complex binds directly to RNA. RCA is performed and the RCP is embedded within a hydrogel, which is thereafter cleared from unbound proteins and lipids. A modified version of sequencing-by-ligation is performed on the RCP¹⁰⁵. FISSEQ: cDNA is cross-linked to its cellular environment. RCA is performed on circulated cDNA and the RCP is again cross-linked to the cellular matrix. Sequencing is performed using sequencing-by-ligation¹⁰⁴.

3.4 *In situ* capturing technologies

The spatial techniques described so far have been based on either isolation of already-known tissue regions of interest, or *in situ* visualization of RNA molecules using hybridization or sequencing. Another approach is to capture and barcode whole transcriptomes *in situ*, then perform sequencing *ex situ*.

3.4.1 Spatial Transcriptomics (ST)

The Spatial Transcriptomics (ST)¹⁰⁶ technology, published in 2016, constitutes the fundamental base upon which this thesis is built. It is described in detail in the *Present Investigation* chapter. In short, a thin tissue section is placed on a microarray glass slide printed with 1,007 spots, where each spot contains ~200 million immobilized RT primers. Each primer within a spot bears a unique barcode sequence corresponding to its position within the array. Individual spots are 100 µm in diameter, with a center-to-center distance of 200 µm. After the tissue section has been placed on the microarray, it is fixed, stained, imaged and permeabilized. During the permeabilization process, mRNA molecules diffuse vertically down to the solid surface and hybridize locally to the RT primers. cDNA synthesis is then performed directly on the microarray, after which the tissue is removed and the cDNA-mRNA complexes are cleaved off. Subsequently, molecules are pooled into a single reaction, where amplification and library preparation take place. Massively parallel sequencing is then performed and the barcoded reads are superimposed back onto the tissue image. ST is applicable to a wide range of tissue types¹⁰⁶⁻¹¹² and allows for transcriptome-wide studies. However, the current spot size limits the spatial resolution to ~10-40 cells.

3.5 *In silico* reconstruction of spatial data

As well as the experimental approaches described in this chapter, there are also computational ways of constructing spatially resolved gene expression datasets. These technologies start from dissociated single cells and aim to computationally assign them to a spatial location by using a gene expression reference map.

3.5.1 Using ISH as reference maps

In 2015, two similar computational approaches using preexisting *in situ* hybridization (ISH) reference databases were published. The first (named Seurat) integrated scRNA-seq data into the spatial context of a zebrafish embryo¹¹³, while the other focused on the brain of the marine annelid *P. dumerilii*¹¹⁴. In both methods, reference maps are created by computationally dividing the tissue into smaller regions (bins). Each bin is given a score based on whether genes in the ISH database are present or absent in that particular region. Single cells are then placed on the tissue map by using their individual algorithms. The Seurat method applied to zebrafish traced back 851 single cells based on a reference map constructed of ISH data for 47 genes. In the *P. dumerilii* study, the authors used a larger reference set of 72 genes, mapping back 112 single cells. Another algorithm published in 2017 is DistMap¹¹⁵. Here, the research group used scRNA-seq data from over 10,000 single cells of dissociated *Drosophila* embryos to reconstruct a 6,000-cellular virtual embryo. By using a ISH reference set comprising 84 ISH genes, they discovered that most of the cells within the embryo have a unique transcriptional identity. All these computational methods are able to give single cells a relatively precise location within a tissue.

3.5.2 Using ST as reference maps

Using existing ISH databases to construct spatial reference maps limits analyses to those tissue types for which ISH atlases already exist. Accordingly this approach is not applicable to clinical samples. A way to overcome this limitation could be to construct a reference map based on transcriptomics-wide data, for example data obtained from ST technology. By combining scRNA-seq and ST on the same tissue, one could refine spatial maps of theoretically any type of tissue. This approach was applied to a tumor sample where one half was subjected to scRNA-seq, and the other half to ST¹¹⁶. Marker genes found in the scRNA-seq data were used to deconvolute the cell type compositions of different tissue regions. Another approach to this is described in Paper III in the *Present Investigation* chapter. Briefly, expression patterns found in the ST data are used to generate a spatial gene expression reference map, on which single cells are positioned based on gene profile matches.

Table 1. Comparison between spatially resolved transcriptomics technologies. Single-cell RNA-seq is not a spatial technique, but is shown here for reference.

Method	Spatial resolution	Approach	Detection efficiency
scRNA-seq	NA	Transcriptome-wide	5-40% ¹¹⁷
LCM	Cellular	Targeted or Transcriptome-wide	NA
Geo-seq	10 cells	Transcriptome-wide	NA
tomo-seq	Anatomical features	Transcriptome-wide	NA
TIVA	Cellular	Transcriptome-wide	NA
NICHE-seq	Cellular	Transcriptome-wide	NA
smFISH	Subcellular	Targeted	Nearly 100% ¹¹⁸
seqFISH	Subcellular	Targeted	84% ⁹⁸
MERFISH	Subcellular	Targeted	80% ⁹⁹
ISS using barcoded padlock probes	Subcellular	Targeted	30% ¹¹⁹
BaristaSeq	Subcellular	Targeted	30% ¹⁰³
STARmap	Subcellular	Targeted	No less than scRNA-seq ¹⁰⁵
FISSEQ	Subcellular	Transcriptome-wide	<0.005% ¹²⁰
Spatial Transcriptomics	Anatomical features	Transcriptome-wide	6.9% ¹⁰⁶

PRESENT INVESTIGATION

This thesis is based on four research papers, all of which address the field of spatial transcriptomics. **Paper I** is the landmark paper describing a novel technology for spatially resolved transcriptomics, while **Paper II** describes the application of the technology to adult human cardiac tissue. In **Paper III**, the technology is applied to developing human heart tissue in combination with scRNA-seq to facilitate increased spatial resolution. Finally, **Paper IV** extends the technology described in **Paper I**, making it also applicable to full-length gene expression analyses.

Paper I – A spatially resolved transcriptome-wide approach to study whole tissue sections

Spatially resolved transcriptomics provides us with new insights into the molecular diversity of complex multicellular organisms. Several approaches have been established in order to preserve gene expression information together with its tissue localization. However, existing challenges for many spatial technologies include the extent of existing knowledge about the targets, the labor-intensive nature of the methods or the fact that they are not applicable to clinical samples. This paper describes a method whereby whole intact tissue sections can be studied in a spatial whole-transcriptome manner. Firstly, fresh frozen tissue is cryosectioned into a thin nearly-unicellular layer and placed on a barcoded microarray. The microarray contains six 6,200 x 6,600 μm areas, each covering 1,007 individually printed spots with a diameter of 100 μm . Each spot contains a cluster of ~ 200 million immobilized poly-T oligos. All oligos within the same spot carry a unique positional barcode sequence, specifying the x and y coordinate of the spot in the array. The tissue section is then fixed and imaged, and subsequently subjected to permeabilization. During the permeabilization step, mRNA molecules will diffuse down to the microarray surface and hybridize to the oligos underneath it. Reverse transcription then takes place directly on the surface. Next, the tissue section is removed and the mRNA-cDNA hybrids are detached from the microarray for library preparation and high-throughput sequencing. Barcoded sequencing reads can then be overlaid on the histological image of the tissue. Hence, individual transcriptomes obtained from each spot will provide spatial gene expression profiles. The technology was applied to mouse olfactory bulb tissue and the resulting spatial gene expression profiles correlated well with the ISH based Allen Brain Atlas. Additionally, the technology showed a detection efficiency of $\sim 6.9\%$ when compared to smFISH. One key step in the protocol is the permeabilization. If the tissue is permeabilized too much, mRNA molecules will diffuse horizontally and their spatial contexts will be lost. On the other

hand, if the tissue is permeabilized too little, mRNAs may not be extracted from the cells at all. In order to investigate the amount of transcript diffusion, microarrays uniformly coated with non-barcoded poly-T oligos were used together with a modified reverse transcription mixture containing Cy3-labeled nucleotides. The protocol was performed as described above up to tissue removal. The intensity of signals generated from the Cy3-cDNA footprint was compared with the histological staining of the tissue, revealing a diffusion distance of $\sim 1.7 \mu\text{m}$. Lastly, the technology is applied to explore tumor heterogeneity within human breast cancer biopsies, showing its applicability to clinical samples. Although this technology provides a less labor-intensive, transcriptome-wide approach applicable to a wide range of samples, it does not have single-cell resolution. Furthermore, the experimental procedure of the current protocol restricts analyses of full-length transcripts.

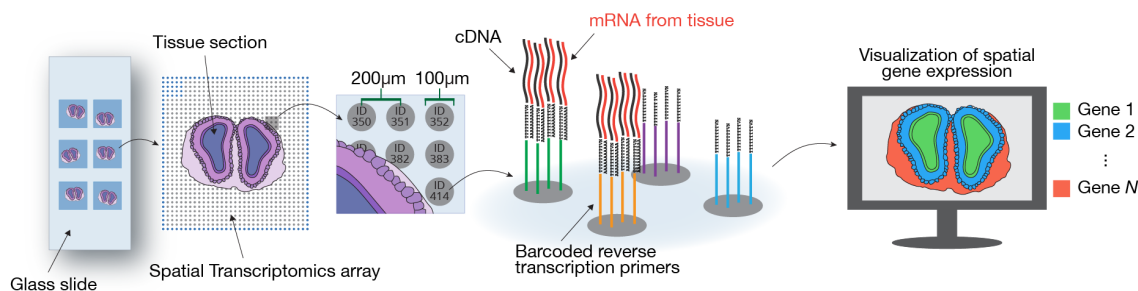


Figure 4. Overview of the Spatial Transcriptomics technology.

Paper II – Detecting lowly expressed biomarkers in heart failure disease progression

Studying human adult heart biopsies has previously only been carried out using microarrays or bulk RNA-seq approaches. However, this generates an average profile of gene expression across the sample and the signal from lowly expressed genes related to subpopulations of cells can be lost. Identifying weakly expressed genes can be of importance when investigating heart failure progression within clinical samples. Earlier studies have shown that genes normally only expressed during the fetal stage are reactivated during heart failure pathogenesis, and thus can be used as biomarkers for disease progression. Here, the technology described in **Paper I** (ST) is applied to heart failure clinical samples. The paper represents a pilot technical study in which adult heart biopsies from three individual patients are investigated. This was performed in order to explore tissue handling, permeabilization procedures and the sensitivity of the method before applying it to future larger clinical studies. One of the main challenges was to retain sufficient amounts of high quality RNA. Adult heart tissue contains a large proportion of fibrous tissue and the cellular density is low, meaning that RNA extraction can be problematic. Initially, the experimental protocol was modified in order to make adult cardiac RNA accessible. The modifications included both a stronger permeabilization treatment as well as a harsher procedure for tissue removal. Investigating the expression of six potential biomarkers across all samples revealed that the ST method provides a higher sensitivity than that which can be achieved through bulk data. Although the study shows the advantage of using ST technology for detecting biomarkers expressed at low levels, it covers only a limited set of patients and therefore no conclusions about HF disease progression can be made.

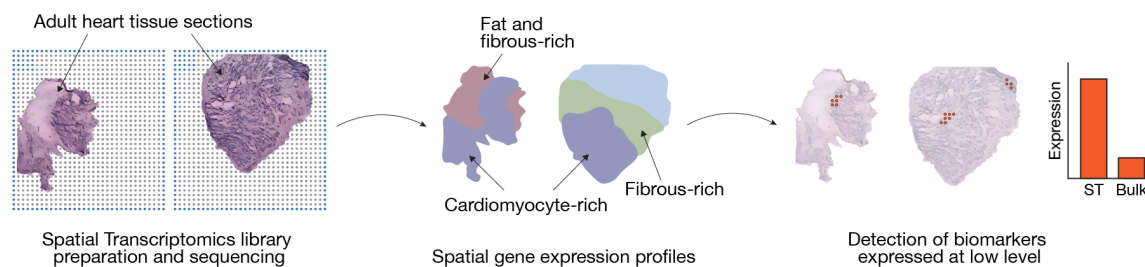


Figure 5. Applying Spatial Transcriptomics technology to human adult heart tissue.

Paper III – Tracing cellular heterogeneity within the developing human heart

Little is known about the process of human cardiac morphogenesis as a whole. The early structure of the primitive heart tube will eventually transform into a fully functional adult heart. During this process, various stem-cell populations furnish the heart with its entire repertoire of cell types, a procedure not yet completely understood. In this paper, the ST technology described in **Paper I** is applied to developing human heart tissue in combination with scRNA-seq to investigate gene expression programs during human heart development. Initially, two heart tissue samples of the same age (~7 weeks-old) were processed either through the ST protocol or by scRNA-seq. Spatial gene expression profiles were linked to distinct anatomical structures within the heart and one of the regions, called the outflow tract, deviated particularly strongly from the rest. Based on gene profile matches between cardiac regions (ST data) and single cells (scRNA-seq data), cellular heterogeneity across the regions could be tracked. By using the combined approach, the outflow tract region could be characterized as having a relatively higher cellular diversity compared to the other regions. Furthermore, two additional heart samples from an earlier (~5 weeks-old) and a later (~9 weeks-old) developmental time point were processed by the ST protocol. Comparisons between the three time points showed low temporal differences in gene expression (from week 5 to week 9) in contrast to the differences seen within the heart. Here, a framework is set up that combines scRNA-seq data with ST technology in order to trace single cells to a location *in situ*. This paper is a first step towards a complete organ-wide gene expression atlas of the developing human heart. However, additional samples would provide a more detailed map of cardiac structures and facilitate more in-depth investigations regarding cell-to-cell interactions.

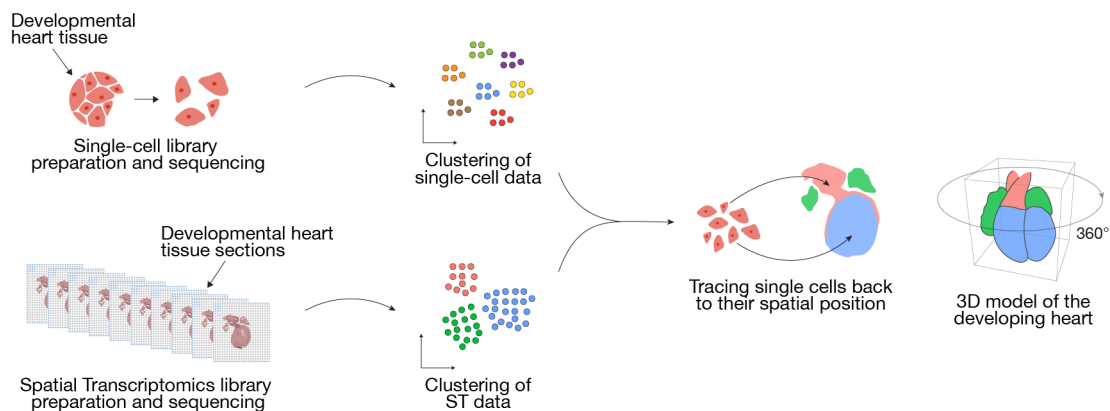


Figure 6. Combining scRNA-seq and Spatial Transcriptomics.

Paper IV – Constructing spatially resolved full-length transcriptomes within whole tissue sections

In the protocol described in **Paper I**, tissue sections are placed on a barcoded microarray and fixed with formaldehyde, a chemical agent responsible for creating covalent cross-links between the mRNAs and their surrounding molecules. The reverse transcription is therefore restricted by the site of a cross-link, resulting in short tags of mRNA data. This approach is very valuable for quantitative gene expression studies, but it does not provide full-length transcript information. More detailed investigations into alternative splicing, fusion transcripts and mutations are therefore largely precluded. This paper describes an alternative protocol using ST barcoded microarrays to study complete transcript structures within individual tissue sections. The protocol differs from the standard procedure in the following steps; (i) fixation, (ii) reverse transcription, (iii), amplification and (iv) library preparation. Firstly, the tissue is fixed with methanol to keep entire RNA molecules intact and accessible to the reverse transcriptase. Secondly, the reverse transcription reaction is performed with an enzyme possessing terminal transferase activity, resulting in the addition of a couple of non-templated nucleotides to the 3' end of the cDNA. A template-switch oligo (designed with an amplification handle) can then hybridize to the extended part of the cDNA, and the RT enzyme can continue to extend over the template-switch oligo. The cDNA constructs will consequently contain amplification handles on both sides and can be subsequently amplified by PCR. Next, the diluted cDNA library is placed in 10x Genomics droplets and a second barcode is incorporated into the cDNA molecules. This procedure generates shorter barcoded cDNA fragments suitable for massively parallel sequencing. Droplet barcodes and spatial barcodes are then linked computationally to reconstruct full-length transcripts. This paper demonstrates that spatially barcoded libraries can be combined with the 10x Genomics droplet station to identify spatial positions of full-length cDNAs within a tissue context. However, future studies based on redesigned barcoded surface-probes and/or the droplet barcodes would provide a more seamless system for spatial isoform profiling.

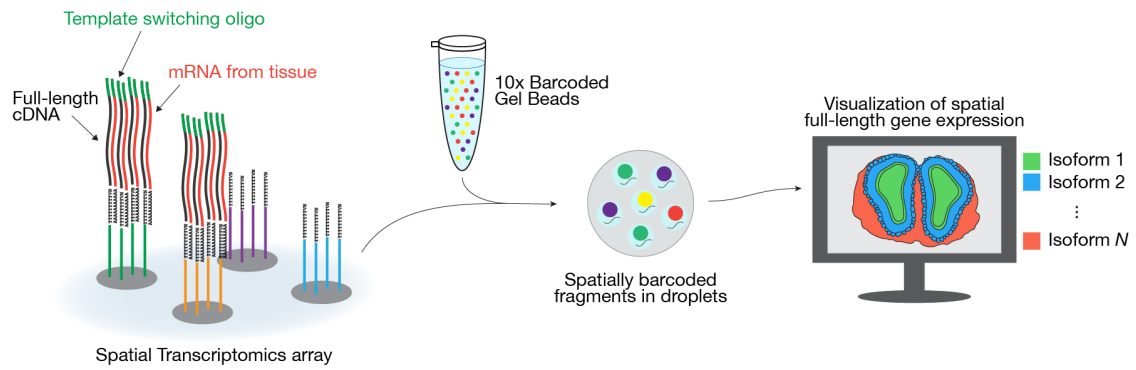


Figure 7. Spatially resolved full-length transcriptomics.

Concluding remarks

More than ten years ago, massively parallel sequencing and the introduction of RNA-seq revolutionized the field of transcriptomics. Shortly after, it was even possible to analyze the individual transcriptome of a single cell. Going from analyzing bulk samples to single cells has truly opened our minds to the cellular diversity in, e.g., complex tissue systems and developmental processes. By adding a spatial dimension to the analysis of these heterogeneous tissue structures, we are beginning to understand biological networks and interactions within the system.

Today, there is a wide range of spatial technologies but none of them is superior in every respect to the others. Dissection approaches, such as Geo-seq and TIVA, can be combined with scRNA-seq to retain complete transcriptome-wide information. TIVA can also be applied to living tissue, a unique feature no other spatial technique possesses. However, both these methods suffer from low throughput when it comes to analyzing a number of regions/cells in parallel. NICHE-seq and tomo-seq, on the other hand, offer high throughput, but are not applicable to clinical samples. The transcriptome-wide ST technology can be applied to a variety of tissue types, including clinical samples, but it does not achieve single cell resolution. *In situ* approaches, in contrast, have exceptional spatial resolution, reaching down even to the subcellular level, but are restricted regarding the fact that targeted genes need to be defined *a priori*. An exception is FISSEQ, but this method has a lower detection efficiency compared to the others.

A major obstacle for many spatial technologies today is the efficiency of detection. In order to increase the efficiency, RNA losses throughout the workflow need to be minimized. This can be achieved by using minimum number of experimental steps and ideally without performing any wash steps. However, RT may be the most efficiency-limiting step, in which not all transcripts are converted to cDNA. Future protocols will likely have workflows with substantially fewer steps and many will probably exclude the RT step by directly targeting native RNA.

Further implementation of spatial techniques in the clinical setting would benefit from protocols compatible with formaldehyde fixed-paraffin embedded tissue. Although possible by some present-day methods, it is more laborious and it is still challenging to achieve sufficient quality. Whether or not spatial techniques will be routinely used in future applications within clinics, they could certainly be a powerful screening tool for the detection of disease biomarkers.

For a technology to be widely accessible, it may be beneficial either to commercialize the method in the form of a kit or a service, or to make the protocol procedure cost efficient and compatible with regular lab equipment. LCM, ST and padlock-probe based ISS studies have all been published outside their lab of creation. Which one of the existing spatial methods will reach the widest future community of scientists, either in its current form or with complementary improvements, remains to be seen.

Beyond analyzing spatially resolved gene expression lies the challenge of analyzing spatially resolved multiomics. This integration of transcriptomics, genomics, proteomics and certainly many other types of omics could advance our knowledge of cell heterogeneity and interactions between cells. Existing protocols for combined single cell omics could surely be applied to spatial techniques using dissection approaches. In terms of analyzing individual tissue sections, using technologies such as ISH, ISS and ST, it would currently be possible to process consecutive tissue sections in order to integrate transcriptomics and proteomics for example. Another future prospect, beyond spatially resolved transcriptomics, would be to monitor real-time temporal changes in gene expression in living tissue, a possibility that may not be too far away considering the technological progression made during the last decade alone.

One thing is for sure; teamwork is fundamental to technological development and scientific discoveries, and it is often most effective when it brings together people across disciplinary divides. One excellent example is the global Human Cell Atlas initiative, which is gathering the efforts of many researchers to profile all the cell types in the human body. By the time it is completed, this international, open access project will have produced an enormous library of descriptions of at least 10 billion cells¹. As this project proceeds, new technologies and collaborations will advance in a rapid manner. These are technologies that will produce a tremendous amount of data, in need of analysis and interpretation with new, standardized procedures. The field of spatially resolved transcriptomics is undoubtedly only beginning to show us what is possible.

Abbreviations

A	Adenine
BaristaSeq	Barcode in situ targeted sequencing
C	Cytosine
CAGE	Cap analysis of gene expression
cDNA	Complementary DNA
CEL-Seq	Cell expression by linear amplification and sequencing
DNA	Deoxyribonucleic acid
EST	Expressed-sequence-tags
FISH	Fluorescence <i>in situ</i> hybridization
FISSEQ	Fluorescent <i>in situ</i> RNA Sequencing
G	Guanine
Geo-seq	Geographical position sequencing
GFP	Green fluorescent protein
InDrop	Index droplets
ISH	<i>In situ</i> hybridization
ISS	<i>In situ</i> sequencing
IVT	<i>In vitro</i> transcription
LCM	Laser capture microdissection
MARS-Seq	Massively parallel single-cell RNA-sequencing
MERFISH	Multiplexed error-robust FISH
mRNA	messenger RNA
PacBio	Pacific Biosciences
PCR	Polymerase chain reaction
qPCR	Quantitative real-time PCR
RCA	Rolling-circle amplification
RCP	Rolling-circle amplification products
RNA	Ribonucleic acid
RNA-seq	RNA sequencing
RT	Reverse transcription
SAGE	Serial analysis of gene expression
SBL	Sequencing-by-ligation
SBS	Sequencing-by-synthesis
scRNA-seq	Single cell RNA sequencing
seqFISH	Sequential hybridization
smFISH	Single-molecule RNA fluorescence <i>in situ</i> hybridization
smHCR	Single-molecule hybridization chain reaction
SMRT	Single molecule real-time
SNAIL	Specific Amplification of Nucleic Acids via Intramolecular Ligation
SOLiD	Sequencing by oligo ligation and detection
ST	Spatial Transcriptomics
STARmap	Spatially-resolved transcript amplicon readout mapping
STRT-seq	Single-cell tagged reverse transcription
T	Thymine
TIVA	Transcriptome in vivo analysis
Tomo-seq	RNA tomography
ZMW	Zero-mode waveguides

Acknowledgments

I started my PhD in August 2013 by setting up two goals: (i) developing my ability to critically analyze and manage challenges and (ii) improving my presentation skills and not shy away from opportunities to speak about my research. I did not know then that this journey would give me so much more than just that. During my PhD years, I have not just achieved the goals that I initially set up, but also developed professionally and personally on so many levels. All of this would not have been possible without a bunch of amazing, intelligent, driven and inspiring people I have had the fortune of having around me:

Joakim! My favorite professor! ;) Thank you for giving me the opportunity to do my PhD in your research group. And Thank you for supporting me in so many conferences and courses around the world, making it possible for me to build my own solid foundation of networks. It has been such an experience and I have certainly learnt a lot during the process. During times of difficulties, I have very much appreciated the chocolate placed on my desk. My co-supervisor **Patrik** – Thank you for all scientific advices and great small talks. You are truly one of my greatest role models! My unofficial co-supervisor, and also very dear friend, **Stefania** – a true source of inspiration! Thank you for both guiding and pushing me scientifically, and for many memorable adventures: New York, Malta, Cyprus, the Swedish mountains... (and no, there is no possibility to flush a Swedish outdoor toilet (“dass”)). To **Afshin** – Thank you for your openness and for always having time for scientific discussions, and for critically reviewing the content of this thesis. I also want to highlight **Anders, Kim, Joseph** and **Kostas** here for fantastic feedback on the early versions of this thesis.

To all my **co-authors** and collaborators, especially to: **Matthias** – for providing me the experience of taking part of several heart surgeries. **Eva W** – for your great knowledge of handling tissue samples and for devoting a lot of your time to the developmental heart project. **Christer** – All PhD students should have their own Christer! It has been, and still is, a true pleasure to work with you. **Johan R** – for your calm and patience despite my one thousands questions about SIMCA and Git. For not cancelling a telephone meeting even though a bunch of screaming kids are hanging around your legs.

To my Dear **Uppsala team**, where it all begun in 2011. Thank you **Ulf** and **Inger** for introducing me to the world of sequencing technologies. To “flocken”: **Linnéa** – for the best mentorship, **Ida** – my science-partner-in-crime and devoted travel-partner,

Sanna – the best mud-runner ever to team up with. All three of you: Thank you for true friendship reaching far beyond the border of science. To **Adam** – the guy who kills in networking! Thanks for giving me the first introduction to bioinformatics and for being a great colleague and friend. **Lars, Jocke** and **Ammar** – Always so much fun to hang out with you guys! Both scientifically and non-scientifically... So many great moments at ASHG in San Diego!

Many thanks to former and present people at the premium Alfa 3 floor: **Bahram, Christian, Nemo, Pär L, Max, Carsten, Francesco, Aleksandra, Florian, Mikael H, Jacke, Verena, Beata, Kicki, Guillermo, 2x Mattias O, Remi, Rapolas, Emanuela, Simon, Mengxiao, Magdalena, Fanny, Lina, Chuan, Anna K, Tobias, Sailendra, Luisa, Niyaz, Erik, Anna, Sára** and **Alma**. **Sanja** – for introducing me early on in the lab and for many pub nights. **Per K** – for giving great feedback after several DNA club presentations. **Phil** – for hiking and climbing gear discussions as well as where to travel next. **Joel** – for uncountable numbers of concerts, pub sessions, fun run sessions... **Conny** and **Olov** – for being my best ski-competitors! **Sverker** – for the best laugh and for many late night walks through the city. **Yue** – for your happy spirit and energy! **Johannes** – for great co-organization of numerous Wild-Orsa-Ski-BiggestDalahorse-Appleglögg events deep down in the darkest of Dalarna. **Benji** – for switching my work in the lab with your bioinformatics during a period. We both were completely lost and the project ended up in nothing, although I learnt how to manage the terminal! And don't worry, I promise never ever to become “mogen”. **Fredrik** – for always giving me answers to my restricted number of 5-questions-per-day questions. Thanks for teaching me everything you know, except shooting, which you are embarrassingly bad at. **Anders** – In all honesty, Thanks for being a true friend.

To inspiring PIs and researchers at SciLifeLab: **Pelin, Anders A** and **Peter S. Marc F** and **Wenjing** for educating me in miRNAs. **Paul H** for fun times during Gene Technology courses.

To the amazing **Spatial Transcriptomics** group: **Kim** – Ma Sistá! An amazing woman seeing the good side of all people. For your caring and openness! I <3 U. **Joseph** – for keeping us updated on the latest happenings in the business world. **Linnéa** – for your constant smile and endless laughs around the lab. **Emelie** – An adorable and hard core Värmländska. **Annelie** – Our very beloved lab-mammy! Lab-mummy? **Linda** – for your never-ending positivity, energy and fighting spirit. **Jose** – My very Spanish friend always up for a relió! And no, I don't have a cat... **Kostas** – for always being helpful and supportive, both in and outside the lab. **Maja** – the positive girl always up for a Mackmyra event! **Ludvig L** – for nerdy hiking-trip talks. **Sami, Eva** and **Zaneta** – the new stars of the ST team. Thank you also **Jonas M, Ludvig B, Britta** and **Sidra**.

You have all contributed to the amazing atmosphere. Thank you **David** – Sort of actually part of ST although not officially. My perfect yin and yang matching partner in the lab as well as organizing poker tournaments.

To former premium'ers who left significant footprints in the SciLifeLab building: **Mario** – My awesome friend whom I share the same hair style with. No one can throw together better mega-, giga- and peta- parties like you and spreading the need of a 4th of July celebration around SciLifeLab. Move back to Sweden. Now. **Sebbe** – My partner throughout KTH, Uppsala and SciLifeLab. Where is the next adventure taking us? Thank you for being one of my very best friends and for all the hula hula in the lab.

To all my supporting and amazing **friends** outside academia: Now you know what I have been doing the past 5 years ;)

And finally, to my very beloved and wonderful Family:

Mamma, Pappa, Millan, Josh and Wille – Thank you for always supporting me in whatever reasonable and unreasonable decisions I make in life. To **André** – Thank you for feeding me with an infinite amount of love. And food. Jag älskar er.

Thank you all! Tack alla!

References

1. Regev, A. Science Forum: The Human Cell Atlas. *Elife* **6**:e27041, (2017).
2. Bianconi, E. *et al.* An estimation of the number of cells in the human body. *Ann. Hum. Biol.* **40**, 463–471 (2013).
3. NASA Science. Retrieved July 26, 2018, from <https://solarsystem.nasa.gov/planets/earth/overview/>
4. Eiberg, H. *et al.* Blue eye color in humans may be caused by a perfectly associated founder mutation in a regulatory element located within the HERC2 gene inhibiting OCA2 expression. *Hum. Genet.* **123**, 177–187 (2008).
5. Dahm, R. Friedrich Miescher and the discovery of DNA. *Dev. Biol.* **278**, 274–288 (2005).
6. Avery, O. T., MacLeod, C. M. & McCarty, M. Studies on the chemical nature of the substance inducing transformation of pneumococcal types: Induction of transformation by a desoxyribonucleic acid fraction isolated from pneumococcus type III. *J. Exp. Med.* **79**, 137–158 (1944).
7. Franklin, R. E. & Gosling, R. G. Molecular Configuration in Sodium Thymonucleate. *Nature* **171**, 740 (1953).
8. Wilkins, M. H. F., Stokes, A. R. & Wilson, H. R. Molecular Structure of Nucleic Acids: Molecular Structure of Deoxypentose Nucleic Acids. *Nature* **171**, 738 (1953).
9. Watson, J. D. & Crick, F. H. C. Molecular structure of nucleic acids: A structure for deoxyribose nucleic acid. *Nature* **171**, 737–738 (1953).
10. GRCh38.p12. Retrieved August 9, 2018, from https://www.ncbi.nlm.nih.gov/assembly/GCF_000001405.38
11. International Human Genome Sequencing Consortium. Initial sequencing and analysis of the human genome. *Nature* **409**, 860–921 (2001).
12. Venter, J. C. *et al.* The Sequence of the Human Genome. *Science*. **291**, 1304–1351 (2001).
13. Macaulay, I. C. & Voet, T. Single Cell Genomics: Advances and Future Perspectives. *PLOS Genet.* **10**, e1004126 (2014).
14. Crick, F. On Protein Synthesis. *Symp. Soc. Exp. Biol. Number XII Biol. Replication Macromol. Cambridge Univ. Press.* 138–163 (1958).
15. Crick, F. Central dogma of molecular biology. *Nature* **227**, 561–563 (1970).
16. Genetics Home Reference. in *National Institutes of Health. U.S. National Library of Medicine. Department of Health & Human Services* 11–12 (2018).
17. Moor, A. E. & Itzkovitz, S. Spatial transcriptomics: paving the way for tissue-level systems biology. *Curr. Opin. Biotechnol.* **46**, 126–133 (2017).
18. Ramel, M.-C. & Hill, C. S. The ventral to dorsal BMP activity gradient in the early zebrafish

- p>embryo is determined by graded expression of BMP ligands.
- Dev. Biol.*
- 378**
- , 170–182 (2013).
19. Reeves, G. T. *et al.* Dorsal-ventral gene expression in the *Drosophila* embryo reflects the dynamics and precision of the dorsal nuclear gradient. *Dev. Cell* **22**, 544–557 (2012).
 20. Thul, P. J. *et al.* A subcellular map of the human proteome. *Science*. **356**, eaal3321 (2017).
 21. Bruusgaard, J. C., Liestøl, K., Ekmark, M., Kollstad, K. & Gundersen, K. Number and spatial distribution of nuclei in the muscle fibres of normal mice studied in vivo. *J. Physiol.* **551**, 467–478 (2003).
 22. Cusanovich, D. A. *et al.* Multiplex single-cell profiling of chromatin accessibility by combinatorial cellular indexing. *Science*. **348**, 910–914 (2015).
 23. Ramani, V. *et al.* Massively multiplex single-cell Hi-C. *Nat. Methods* **14**, 263–266 (2017).
 24. Buxbaum, A. R., Haimovich, G. & Singer, R. H. In the right place at the right time: visualizing and understanding mRNA localization. *Nat. Rev. Mol. Cell Biol.* **16**, 95–109 (2015).
 25. Crosetto, N., Bienko, M. & van Oudenaarden, A. Spatially resolved transcriptomics and beyond. *Nat. Genet.* **16**, 57–66 (2015).
 26. Elmore, J. G. *et al.* Diagnostic concordance among pathologists interpreting breast biopsy specimens. *JAMA* **313**, 1122–1132 (2015).
 27. Alwine, J. C., Kemp, D. J. & Stark, G. R. Method for detection of specific RNAs in agarose gels by transfer to diazobenzyloxymethyl-paper and hybridization with DNA probes. *Proc. Natl. Acad. Sci. U. S. A.* **74**, 5350–5354 (1977).
 28. Mullis, K. Faloona, F. Scharf, S. Saiki, R. Specific Enzymatic Amplification of DNA in Vitro: The Polymerase Chain Reaction. *Cold Spring Harb. Symp. Quant. Biol.* **51**, 263–273 (1986).
 29. National Human Genome Research Institute. Retrieved August 9, 2018, from <https://www.genome.gov/10000207/polymerase-chain-reaction-pcr-fact-sheet/>
 30. Taylor, S., Wakem, M., Dijkman, G., Alsarraj, M. & Nguyen, M. A practical approach to RT-qPCR—Publishing data that conform to the MIQE guidelines. *Methods* **50**, S1–S5 (2010).
 31. Aird, D. *et al.* Analyzing and minimizing PCR amplification bias in Illumina sequencing libraries. *Genome Biol.* **12**, (2011).
 32. Kanagawa, T. Bias and artifacts in multitemplate polymerase chain reactions (PCR). *J. Biosci. Bioeng.* **96**, 317–323 (2003).
 33. Higuchi, R., Dollinger, G., Sean Walsh, P. & Griffith, R. Simultaneous amplification and detection of specific DNA sequences. *Biotechnol.* **10**, 413–417 (1992).
 34. Morrison, T., Weis, J. & Wittwer, C. Quantification of low-copy transcripts by continuous SYBR Green I monitoring during amplification. *Biotechniques* **24**, 954–962 (1998).
 35. Huggett, J. & Bustin, S. A. Standardisation and reporting for nucleic acid quantification. *Accredit. Qual. Assur.* **16**, 399 (2011).
 36. Schena, M., Shalon, D., Davis, R. W. & Brown, P. O. Quantitative monitoring of gene expression

- patterns with a complementary DNA microarray. *Science*. **270**, 467–470 (1995).
37. Okoniewski, M. & Miller, C. J. Hybridization interactions between probesets in short oligo microarrays lead to spurious correlations. *BMC Bioinformatics* **7**, 276 (2006).
 38. Sanger, F., Nicklen, S. & Coulson, A. R. DNA sequencing with chain-terminating inhibitors. *Proc. Natl. Acad. Sci. U. S. A.* **74**, 5463–5467 (1977).
 39. Hagemann, I. S. in *Clinical Genomics* (eds. Kulkarni, S. & Pfeifer, J. B. T.-C. G.) 3–19 (Academic Press, 2015). doi:<https://doi.org/10.1016/B978-0-12-404748-8.00001-0>
 40. Adams, M. D. *et al.* Complementary DNA sequencing: expressed sequence tags and human genome project. *Science*. **252**, 1651–1656 (1991).
 41. Boguski, M. S., Tolstoshev, C. M. & Bassett, D. E. Gene discovery in dbEST. *Science*. **265**, 1993–1994 (1994).
 42. Velculescu, V. E., Zhang, L., Vogelstein, B. & Kinzler, K. W. Serial Analysis of Gene Expression. *Science*. **270**, 484–487 (1995).
 43. Shiraki, T. *et al.* Cap analysis gene expression for high-throughput analysis of transcriptional starting point and identification of promoter usage. *Proc. Natl. Acad. Sci.* **100**, 15776–15781 (2003).
 44. Mardis, E. R. The impact of next-generation sequencing technology on genetics. *Trends Genet.* **24**, 133–141 (2008).
 45. Ronaghi, M., Uhlén, M. & Nyrén, P. A Sequencing Method Based on Real-Time Pyrophosphate. *Science*. **281**, 363–365 (1998).
 46. Margulies, M. *et al.* Genome sequencing in microfabricated high-density picolitre reactors. *Nature* **437**, 376–380 (2005).
 47. Rothberg, J. M. *et al.* An integrated semiconductor device enabling non-optical genome sequencing. *Nature* **475**, 348–352 (2011).
 48. Illumina. History of sequencing by synthesis. Retrieved August 18, 2018, from <https://emea.illumina.com/science/technology/next-generation-sequencing/illumina-sequencing-history.html?langsel=/it/>
 49. Ramsköld, D. *et al.* Full-Length mRNA-Seq from single cell levels of RNA and individual circulating tumor cells. *Nat. Biotechnol.* **30**, 777–782 (2012).
 50. genomeweb. Retrieved August 18, 2018, from <https://www.genomeweb.com/sequencing/pacbio-ships-first-two-commercial-systems-order-backlog-grows-44#.W3ghgS2B3-Z>
 51. Levene, M. J. *et al.* Zero-Mode Waveguides for Single-Molecule Analysis at High Concentrations. *Science*. **299**, 682–686 (2003).
 52. Eid, J. *et al.* Real-Time DNA Sequencing from Single Polymerase Molecules. *Science*. **323**, 133–138 (2009).
 53. PacBio. SMRT Sequencing: Read Lengths. Retrieved August 18, 2018, from <https://www.pacb.com/smrt-science/smrt-sequencing/read-lengths/>

54. Sharon, D., Tilgner, H., Grubert, F. & Snyder, M. A single-molecule long-read survey of the human transcriptome. *Nat. Biotechnol.* **31**, 1009–1014 (2013).
55. Clarke, J. *et al.* Continuous base identification for single-molecule nanopore DNA sequencing. *Nat. Nanotechnol.* **4**, 265–270 (2009).
56. Oxford Nanopore. Company history. Retrieved August 18, 2018, from <https://nanoporetech.com/about-us/history>
57. Smith, A. M., Jain, M., Mulroney, L., Garalde, D. R. & Akeson, M. Reading canonical and modified nucleotides in 16S ribosomal RNA using nanopore direct RNA sequencing. *bioRxiv [PREPRINT]* (2017). doi:<https://doi.org/10.1101/132274>
58. Boon, W. C. *et al.* Increasing cDNA Yields from Single-cell Quantities of mRNA in Standard Laboratory Reverse Transcriptase Reactions using Acoustic Microstreaming. *J. Vis. Exp.* **53**, 3144 (2011).
59. Velculescu, V. E. *et al.* Analysis of human transcriptomes. *Nat. Genet.* **23**, 387–388 (1999).
60. Tang, F. *et al.* mRNA-Seq whole-transcriptome analysis of a single cell. *Nat. Methods* **6**, 377–382 (2009).
61. Tietjen, I. *et al.* Single-Cell Transcriptional Analysis of Neuronal Progenitors. *Neuron* **38**, 161–175 (2003).
62. Kurimoto, K. *et al.* An improved single-cell cDNA amplification method for efficient high-density oligonucleotide microarray analysis. *Nucleic Acids Res.* **34**, e42 (2006).
63. Kurimoto, K., Yabuta, Y., Ohinata, Y. & Saitou, M. Global single-cell cDNA amplification to provide a template for representative high-density oligonucleotide microarray analysis. *Nat. Protoc.* **2**, 739–752 (2007).
64. Sasagawa, Y. *et al.* Quartz-Seq: a highly reproducible and sensitive single-cell RNA sequencing method, reveals non-genetic gene-expression heterogeneity. *Genome Biol.* **14**, 3097 (2013).
65. Sasagawa, Y. *et al.* Quartz-Seq2: a high-throughput single-cell RNA-sequencing method that effectively uses limited sequence reads. *Genome Biol.* **19**, 29 (2018).
66. Schmidt, W. M. & Mueller, M. W. CapSelect: a highly sensitive method for 5' CAP-dependent enrichment of full-length cDNA in PCR-mediated analysis of mRNAs. *Nucleic Acids Res.* **27**, e31 (1999).
67. Islam, S. *et al.* Characterization of the single-cell transcriptional landscape by highly multiplex RNA-seq. *Genome Res.* **21**, 1160–1167 (2011).
68. Islam, S. *et al.* Quantitative single-cell RNA-seq with unique molecular identifiers. *Nat. Methods* **11**, 163–166 (2014).
69. Picelli, S. *et al.* Smart-seq2 for sensitive full-length transcriptome profiling in single cells. *Nature methods* **10**, 1096–1098 (2013).
70. Eberwine, J. *et al.* Analysis of gene expression in single live neurons. *Proc. Natl. Acad. Sci.* **89**, 3010–3014 (1992).

71. Hashimshony, T., Wagner, F., Sher, N. & Yanai, I. CEL-Seq: single-cell RNA-seq by multiplexed linear amplification. *Cell Rep.* **2**, 666–673 (2012).
72. Hashimshony, T. *et al.* CEL-Seq2: Sensitive highly-multiplexed single-cell RNA-Seq. *Genome Biol.* **17**, 77 (2016).
73. Svensson, V., Vento-Tormo, R. & Teichmann, S. A. Exponential scaling of single-cell RNA-seq in the past decade. *Nat. Protoc.* **13**, 599–604 (2018).
74. Jaitin, D. A. *et al.* Massively parallel single-cell RNA-seq for marker-free decomposition of tissues into cell types. *Science.* **343**, 776–779 (2014).
75. Fan, H. C., Fu, G. K. & Fodor, S. P. A. Combinatorial labeling of single cells for gene expression cytometry. *Science.* **347**, 1258367 (2015).
76. Klein, A. M. *et al.* Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells. *Cell* **161**, 1187–1201 (2015).
77. Macosko, E. Z. *et al.* Highly Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets. *Cell* **161**, 1202–1214 (2015).
78. Cao, J. *et al.* Comprehensive single-cell transcriptional profiling of a multicellular organism. *Science.* **357**, 661–667 (2017).
79. Rosenberg, A. B. *et al.* Single-cell profiling of the developing mouse brain and spinal cord with split-pool barcoding. *Science.* **360**, 176–182 (2018).
80. Macaulay, I. C., Ponting, C. P. & Voet, T. Single-Cell Multiomics: Multiple Measurements from Single Cells. *Trends in Genetics* **33**, 155–168 (2017).
81. Macaulay, I. C. *et al.* G&T-seq : parallel sequencing of single- cell genomes and transcriptomes. *Nat. Methods* **12**, 519–522 (2015).
82. Dey, S. S., Kester, L., Spanjaard, B., Bienko, M. & van Oudenaarden, A. Integrated genome and transcriptome sequencing of the same cell. *Nat. Biotechnol.* **33**, 285–289 (2015).
83. Darmanis, S. *et al.* Simultaneous Multiplexed Measurement of RNA and Proteins in Single Cells. *Cell Rep.* **14**, 380–389 (2016).
84. Stoeckius, M. *et al.* Simultaneous epitope and transcriptome measurement in single cells. *Nat. Methods* **14**, 865–868 (2017).
85. Frei, A. P. *et al.* Highly multiplexed simultaneous detection of RNAs and proteins in single cells. *Nat. Methods* **13**, 269–275 (2016).
86. Emmert-Buck, M. R. *et al.* Laser Capture Microdissection. *Science.* **274**, 998–1001 (1996).
87. Simone, N. L., Bonner, R. F., Gillespie, J. W., Emmert-Buck, M. R. & Liotta, L. A. Laser-capture microdissection: opening the microscopic frontier to molecular analysis. *Trends Genet.* **14**, 272–276 (1998).
88. Chen, J. *et al.* Spatial transcriptomic analysis of cryosectioned tissue samples with Geo-seq. *Nat. Protoc.* **12**, 566–580 (2017).

89. Combs, P. A. & Eisen, M. B. Sequencing mRNA from Cryo-Sliced *Drosophila* Embryos to Determine Genome-Wide Spatial Patterns of Gene Expression. *PLoS One* **8**, 2–8 (2013).
90. Junker, J. P. *et al.* Genome-wide RNA Tomography in the Zebrafish Embryo. *Cell* **159**, 662–675 (2014).
91. Lovatt, D. *et al.* Transcriptome in vivo analysis (TIVA) of spatially defined single cells in live tissue. *Nat. Methods* **11**, 190–6 (2014).
92. Medaglia, C. *et al.* Spatial reconstruction of immune niches by combining photoactivatable reporters and scRNA-seq. *Science*. **358**, 1622–1626 (2017).
93. Gall, J. G. & Pardue, M. Lou. Formation and detection of RNA-DNA hybrid molecules in cytological preparations. *Proc. Natl. Acad. Sci.* **63**, 378–383 (1969).
94. Singer, R. H. & Ward, D. C. Actin gene expression visualized in chicken muscle tissue culture by using in situ hybridization with a biotinated nucleotide analog. *Proc. Natl. Acad. Sci.* **79**, 7331–7335 (1982).
95. Femino, A. M., Fay, F. S., Fogarty, K. & Singer, R. H. Visualization of Single RNA Transcripts in Situ. *Science*. **280**, 585–590 (1998).
96. Raj, A., van den Bogaard, P., Rifkin, S. A., van Oudenaarden, A. & Tyagi, S. Imaging individual mRNA molecules using multiple singly labeled probes. *Nat. Methods* **5**, 877–879 (2008).
97. Lubeck, E., Coskun, A. F., Zhiyentayev, T., Ahmad, M. & Cai, L. Single-cell in situ RNA profiling by sequential hybridization. *Nat. Methods* **11**, 360–361 (2014).
98. Shah, S., Lubeck, E., Zhou, W. & Cai, L. In Situ Transcription Profiling of Single Cells Reveals Spatial Organization of Cells in the Mouse Hippocampus. *Neuron* **92**, 342–357 (2016).
99. Chen, K. H., Boettiger, A. N., Moffitt, J. R., Wang, S. & Zhuang, X. Spatially resolved, highly multiplexed RNA profiling in single cells. *Science*. **348**, 1360–1363 (2015).
100. Wang, G., Moffitt, J. R. & Zhuang, X. Multiplexed imaging of high-density libraries of RNAs with MERFISH and expansion microscopy. *Sci. Rep.* **8**, 1–13 (2018).
101. Chen, F. *et al.* Nanoscale imaging of RNA with expansion microscopy. *Nat. Methods* **13**, 679–684 (2016).
102. Ke, R. *et al.* In situ sequencing for RNA analysis in preserved tissue and cells. *Nat. Methods* **10**, 857–860 (2013).
103. Chen, X., Sun, Y. C., Church, G. M., Lee, J. H. & Zador, A. M. Efficient in situ barcode sequencing using padlock probe-based BaristaSeq. *Nucleic Acids Res.* **46**, e22 (2018).
104. Lee, J. H. *et al.* Highly Multiplexed Subcellular RNA Sequencing in Situ. *Science*. **343**, 1360–1363 (2014).
105. Wang, X. *et al.* Three-dimensional intact-tissue sequencing of single-cell transcriptional states. *Science*. eaat5691 (2018). doi:10.1126/science.aat5691
106. Ståhl, P. L. *et al.* Visualization and analysis of gene expression in tissue sections by spatial transcriptomics. *Science*. **353**, 78–82 (2016).

107. Salmén, F. *et al.* Multidimensional transcriptomics provides detailed information about immune cell distribution and identity in HER2 + breast tumors. *bioRxiv [PREPRINT]* (2018). doi:<https://doi.org/10.1101/358937>
108. Berglund, E. *et al.* Spatial maps of prostate cancer transcriptomes reveal an unexplored landscape of heterogeneity. *Nat. Commun.* **9**, 1–13 (2018).
109. Giacomello, S. *et al.* Spatially resolved transcriptome profiling in model plant species. *Nat. Plants* **3**, 1–11 (2017).
110. Lundmark, A. *et al.* Gene expression profiling of periodontitis-affected gingival tissue by spatial transcriptomics. *Sci. Rep.* **8**, (2018).
111. Asp, M. *et al.* Spatial detection of fetal marker genes expressed at low level in adult human heart tissue. *Sci. Rep.* **7**, (2017).
112. Maniatis, S. *et al.* Spatiotemporal dynamics of molecular pathology in amyotrophic lateral sclerosis. *bioRxiv [PREPRINT]* (2018). doi:<https://doi.org/10.1101/389270>
113. Satija, R., Farrell, J. A., Gennert, D., Schier, A. F. & Regev, A. Spatial reconstruction of single-cell gene expression data. *Nat. Biotechnol.* **33**, 495–502 (2015).
114. Achim, K. *et al.* High-throughput spatial mapping of single-cell RNA-seq data to tissue of origin. *Nat. Biotechnol.* **33**, 503–509 (2015).
115. Karaïskos, N. *et al.* The *Drosophila* embryo at single-cell transcriptome resolution. *Science*. **358**, 194–199 (2017).
116. Moncada, R. *et al.* Building a tumor atlas: integrating single-cell RNA-Seq data with spatial transcriptomics in pancreatic ductal adenocarcinoma. *bioRxiv [PREPRINT]* (2018). doi:<https://doi.org/10.1101/254375>
117. Grün, D. & Van Oudenaarden, A. Design and Analysis of Single-Cell Sequencing Experiments. *Cell* **163**, 799–810 (2015).
118. Lein, E., Borm, L. E. & Linnarsson, S. The promise of spatial transcriptomics for neuroscience in the era of molecular cell typing. *Science*. **358**, 64–69 (2017).
119. Larsson, C., Grundberg, I., Söderberg, O. & Nilsson, M. In situ detection and genotyping of individual mRNA molecules. *Nat. Methods* **7**, 395–397 (2010).
120. Lee, J. H. Quantitative approaches for investigating the spatial context of gene expression. *Wiley Interdiscip. Rev. Syst. Biol. Med.* **9**, e1369 (2017).