KTH ROYAL INSTITUTE
OF TECHNOLOGY

Doctoral Thesis in Biotechnology

Computational methods for analysis of spatial transcriptomics data

An exploration of the spatial gene expression landscape

ALMA ANDERSSON

Stockholm, Sweden 2022

Computational methods for analysis of spatial transcriptomics data

An exploration of the spatial gene expression landscape

ALMA ANDERSSON

Academic Dissertation which, with due permission of the KTH Royal Institute of Technology, is submitted for public defence for the Degree of Doctor of Philosophy on Friday the 18th of March 2022, at 10:00a.m. in Air&Fire, Tomtebodavägen 23A, Solna

Doctoral Thesis in Biotechnology
KTH Royal Institute of Technology
Stockholm, Sweden 2022

TRITA-CBH-FOU-2022:11
ISBN: 978-91-8040-142-5

Printed by: Universitetsservice US-AB, Sweden 2022

Abstract

Transcriptomics techniques, whether in the form of bulk, single cell/nuclei, or spatial methods have fueled a substantial expansion of our knowledge about the biological systems within and around us. In addition, the rate of innovation has accelerated over the last decade, resulting in a multitude of technological advances and new methods for generation of transcriptomics data. In 2009, isolating and characterizing the transcriptome of a single cell was seen as a major achievement, ten years later, in 2019, studies surveying a hundred thousand cells were commonplace. The field of spatial transcriptomics went through an equally transformative phase; from struggling with simultaneous characterization of a few targets, to seamlessly provide spatially resolved maps of the full transcriptome. Inevitably, we're approaching an inflection point where the generation of data is no longer the bottleneck, but rather its analysis. Alas, with standardized commercial products, high-quality spatial transcriptomics data can now be generated *en masse*. Hence, questions about data analysis have started to replace those of data generation. The work in this thesis seeks to address some of these emerging questions; the five articles it encompasses presents new methods for analysis of spatial transcriptomics data and examples of their application. Furthermore, it contains an introduction to current experimental and computational spatial transcriptomics techniques, as well as a section about data modeling.

In **Article I**, a probabilistic model for integration of single cell/nuclei and spatial transcriptomics data is presented. In short, the method allows for mixed signals – present in certain spatial transcriptomics platforms – to be decomposed into contributions from biologically relevant cell types or states derived from single cell/nuclei data. The model was implemented in code as a software, *stereoscope*, which is open source and publicly available. The same policy of open source and high transparency holds true for all software or code associated with this thesis. The *stereoscope* method has been used in several studies, one example being **Article II**, where we examined the spatial transcriptomics landscape of HER2-positive breast cancer patients. By integrating single cell and spatial transcriptomics data, several intriguing co-localization signals emerged. These signals allowed us to identify a signature for tertiary lymphoid structures and evidence of a trifold interaction involving: type I interferon signals, a T-cell subset, and a macrophage subset. However, the work also included other forms of explorative data analysis, such as unsupervised expression-based clustering. The clusters from this analysis, once annotated, exhibited high concordance with annotations provided by a pathologist and the tissue morphology. Taken together, this makes a compelling case for the use of spatial transcriptomics in the age of "digital pathology." Finally, we also derived "core signatures" from the expression-

based clusters, representing common expression profiles shared across the patients.

In **Article III**, we present a computational method, *sepal*, designed to identify genes with distinct spatial patterns, often referred to as *"spatially variable genes."* The method uses Fick's second law to simulate diffusion of transcripts in the tissue, measuring the time until convergence (a spatially uniform and homogeneous state). It then ranks the genes by their "diffusion time." The assumption being that genes exhibiting strong spatial patterns will take longer time to converge compared to genes with no pattern, thus relating the diffusion time to the degree of spatial structure.

Article IV constitutes a study of the mouse liver using spatial transcriptomics. As before, we employed *stereoscope* for the purpose of single cell integration, but realized more tailored computational tools – towards the specific tissue – were required to address certain questions. Thus, we developed two computational methods, one devoted to vein type identity prediction, the other enabling a change of data representation. In essence, to predict the vein identities, we first assembled spatially weighted composite expression profiles from – to the vein – neighboring observations. Then, a logistic classifier was trained using the composite profiles. Once the model was trained, it could be used to assign vein type identities to ambiguous or unannotated veins. In the second method, the two-dimensional spatial data was recast into a more informative one-dimensional representation by treating gene expression as a function of an observation's distance to its nearest vein structure.

The final work, **Article V**, expands the idea of recasting data into a more informative or helpful representation. More precisely, we present a method, *eggplant*, that allows the user to transfer spatial transcriptomics data from multiple sources to a common coordinate framework (CCF). Transfer of information to a CCF means spatial signals can be compared across conditions and time points, unlocking a plethora of valuable downstream analyses. For example, we perform spatiotemporal modeling of a synthetic system, and introduce the concept of "spatial arithmetics" to study local expression differences. With a growing corpus of spatial trancsriptomics data and ambitious international efforts like the Human Cell Atlas, we deem these sort of methods essential to leverage the data's full potential.

Sammanfattning

Transkriptomiktekniker, både i form av bulk, single cell/nuclei och spatiala metoder har tillåtit oss att utvidga vår kunskap om de biologiska system omkring likväl som inom oss. Under det senaste decenniet så har mängden innovationer inom området ökat på ett lavinartat sätt, och en uppsjö teknologiska avancemang har gjorts. Resultatet av detta är flertalet nya experimentella metoder. År 2009 så sågs isolering och karaktärisering av en enda cells transkriptom som ett stort framsteg, tio år senare (2019) så var studier med kartläggning av transkriptomet hos var och en av hundratusentals celler närmast osensationellt. Fältet som benämns spatial transcriptomics (sv. spatial transkriptomik) har genomgått en likvärdigt transformativ fas; det har gått från att kämpa med att uppskatta uttrycket av ett fåtal gener samtidigt till att kunna producera en spatial bild av samtliga gener i transkriptomet. Inte oväntat så närmar vi oss en inflektionspunkt där analys, istället för produktion av data, är den begränsande faktorn. Med standardiserade kommersiella produkter så kan högkvalitativ spatial transcriptomics data effektivt genereras i stor skala. Således har frågor kring analys av data börjat ersätta dem som berör dess framställning. Denna avhandling ämnar behandla vissa av dessa nya frågor; de fem artiklarna som den innefattar presenterar nya metoder för analys av spatial transcriptomics data samt exempel på deras applikationsområden. Avhandlingen ger även en överskådlig beskrivning av existerande metoder för produktion och analys av spatial transcriptomics data samt innehåller ett avsnitt om datamodellering.

I **Artikel I** så presenteras en probabilistisk modell för integration av single cell/nuclei och spatial transcriptomics data. Metoden möjliggör en dekomposering av de blandade signaler som är karaktäristiska för data från vissa spatial transcriptomics tekniker. Detta gör det möjligt att beskriva observationer utifrån deras sammansättning av biologiskt relevanta celltyper, definierade i single cell/nuclei data, istället för enbart genuttryck. Modellen implementerades även i kod som mjukvara och lanserades, med öppen källkod samt full tillgänglighet för allmänheten, under namnet *stereoscope*. Samma riktlinjer kring öppenhet och transparens gäller för all mjukvara och kod som är associerad med denna avhandling. Metoden, *stereoscope*, har använts i flertalet studier varav **Artikel II** är ett exempel. I detta arbete så undersökte vi det spatiala expressionslandskapet hos HER2-positiva bröstcancerpatienter. Genom att integra spatial och single cell data identifierade vi flertalet intressanta kolokaliseringssignaler. Från dessa signaler kunde vi definiera en signatur för tertiära lymfstrukturer samt se indikationer på en trevägsinteraktion mellan en interferon I signal, ett T-cell subset, och ett makrofag subset. Arbetet innefattade även ytterligare dataanalys, där vi nyttjade icke-vägledd (eng. unsupervised) klustring

av genexpressionsdatan. De resulterande klustrena, efter annotering, stämde väl överens med morfologin och annoteringar som tillhandahållits från en patolog. Sammantaget så bekräftar dessa resultat värdet i att använda spatial transcriptomics för "digital patologi". Slutligen, från genexpressionsklustren så kunde även "kärnsignaturer" identifieras, vilka representerar generella expressionsprofiler som delas av flertalet patienter.

I **Artikel III** så presenterar vi ytterligare en analysmetod, *sepal*, vilken är utvecklad för att identifiera gener med distinkta spatiala mönster, ofta refererade till som "spatialt variabla gener" (eng. spatially variable genes). Metoden använder först Ficks andra lag för att simulera diffusion av transkript i vävnaden, samtidigt som tiden till konvergens (ett spatialt homogent tillstånd) mäts. Sedan rankas varje gen baserat på dess "diffusionstid". Metoden bygger på antagandet att gener som uppvisar spatiala mönster generellt tar längre tid att konvergera jämfört med gener utan struktur.

Artikel IV redogör för en studie av muslevern genom användandet av spatial transcriptomics. Vi använde *stereoscope* med syfte att integrera single cell data även i detta projekt, men upplevde ett behov av mer skräddarsydda metoder för analys av den specifika vävnaden. Således introducerade vi två nya analysmetoder, en avsedd för predicering av venidentitet, den andra för att representera expressionsdatan på ett mer informativt sätt. För att predicera venidentiteter så skapade vi sammansatta och spatialt viktade genexpressionsprofiler baserat på observationer från respektive vens närliggande område. Därefter tränade vi en logistisk klassificerare med syfte att kunna identifiera huruvida en ven tillhörde klassen "centralven" eller "portalven" givet dess sammansatta genexpressionsprofil. Efter att modellen tränats så kunde den användas för att tillskriva oannoterade eller svårannoterade vener en av de två nämnda identiteterna. I den andra metoden så förflyttar vi tvådimensionell spatial transcriptomics data till en mer informativ endimensionell representation, detta genom att behandla genexpressionsuttrycket som en funktion av avståndet till en observations närmaste venstruktur.

I det sista arbetet, **Artikel V**, så vidareutvecklar vi idéen om att förflytta data till en mer informativ eller användbar representation. Mer exakt så presenterar vi en metod, *eggplant*, som tillåter användaren att projicera data från flertalet prover eller experiment till ett gemensamt koordinatsystem (eng. common coordinate framework, kort CCF). Genom att förflytta information till ett CCF så kan spatiala signaler jämföras mellan olika tillstånd och tidpunkter, vilket är nödvändigt för flertalet värdefulla sekundäranalyser. Exempel på sådana analyser i vår studie är: spatiotemporal modellering av ett syntetiskt system, och "spatial aritmetik" applicerad på experimentellt inhämtad vävnadsdata. Med en växande mängd av spatial transcriptomics data och ambitiösa internationella initiativ som "the Human Cell Atlas", så anser vi att liknande metoder är essentiella för att kunna nyttja datan till dess fulla potential.

Thesis Defense

The public defense of this thesis will take place on March 18th 2022. The event's location is: Air&Fire, Science for Life Laboratory, Tomtebodavägen 23A, Solna, Sweden (59°21'00.3"N 18°01'20.9"E). The attempted degree is Doctor of Philosophy (PhD) in Biotechnology.

- **Respondent**: M.Sc.Eng. Alma Andersson. Dept. of Gene Technology, KTH Royal Institute of Technology, Science for Life Laboratory, Solna, Sweden.
- **Chairman**: Docent Patrik Ståhl. Dept. of Gene Technology, KTH Royal Institute of Technology, Science for Life Laboratory, Solna, Sweden.
- **Faculty Opponent**: Dr. Omer Bayraktar. Wellcome Sanger Institute, Hinxton, Cambridge CB10 1SA, UK.

Evaluation committee

- **Member I**: Prof. Tuuli Lappalainen. Dept. of Gene Technology, KTH Royal Institute of Technology, Science for Life Laboratory, Solna, Sweden, New York Genome Center, NY, USA, Department of Systems Biology, Columbia University, NY, USA.
- **Member II**: Docent Marc Friedländer. Science for Life Laboratory, Department of Molecular Biosciences, The Wenner-Gren Institute, Stockholm University, Stockholm, Sweden.
- **Member III**: Dr. Åsa Björklund. Department of Cell and Molecular Biology, National Bioinformatics Infrastructure Sweden, Science for Life Laboratory Uppsala University, Uppsala, Sweden.

Respondent's supervisors

- **Main supervisor**: Prof. Joakim Lundeberg. Dept. of Gene Technology, KTH Royal Institute of Technology, Science for Life Laboratory, Solna, Sweden.
- **Co-supervisor:** Docent Olof Emanuelsson. Dept. of Gene Technology, KTH Royal Institute of Technology, Science for Life Laboratory, Solna, Sweden.

List of Publications

The work included in this thesis is listed in the section below. For comprehensiveness, an extended list of publications – which I have contributed to, but chosen to not include in the thesis – is also included.

0.1 Publications included in thesis

Article I: Andersson A, Bergenstråhle J, Asp M, Bergenstråhle L, Jurek A, Fernandez Navarro J, Lundeberg J **"Single-cell and spatial transcriptomics enables probabilistic inference of cell type topography"** *Communications Biology*, Oct 2020, https://doi.org/10.1038/s42003-020-01247-y

Article II: Andersson A, Larsson L, Stenbeck L, Salmén F, Ehinger A, Wu S, Al-Eryani G, Roden D, Swarbrick A, Borg Å, Frisén J, Engblom C, Lundeberg J, **"Spatial Deconvolution of HER2-positive Breast Tumors Reveals Novel Intercellular Relationships"** *Nature Communications*, Oct 2021, https://doi.org/10.1038/s41467-021-26271-2

Article III: Andersson A, Lundeberg J, **"sepal: identifying transcript profiles with spatial patterns by diffusion-based modeling"** *Oxford Bioinformatics*, Mar 2021, https://doi.org/10.1093/bioinformatics/btab164

Article IV: Hildebrandt F.F.A, Andersson A, Saarenpaa S, Larsson L, Van Hul N, Kanatani S, Masek J, Ellis E, Barragan A, Mollbrink A, Andersson E R, Lundeberg J, Ankarklev J, **"Spatial Transcriptomics to define transcriptional patterns of zonation and structural components in the liver"** *Nature Communications*, Dec 2021, https://doi.org/10.1038/s41467-021-27354-w

Article V: Andersson A, Andrusivová Ž, Li X, Sundström E, Lundeberg J, **"A Landmark-based Common Coordinate Framework for Spatial Transcriptomics Data"**, *bioRxiv* and *in review, Nature Methods*, Nov 2021, https://www.biorxiv.org/content/10.1101/2021.11.11.468178v1

0.2 Extended list of publications

0.2.1 Publications

- Gracia Villacampa E, Larsson L, Mirzazadeh R, Kvastad L, Andersson A, Mollbrink A, Kokaraki G, Monteil V, Schultz N, Appelberg K S, Montserrat N, Zhang H, Penninger J M, Miesbach W, Mirazimi Ali, Carlson J, Lundeberg J, **"Genome-wide spatial expression profiling in formalin-fixed tissues"** *Cell Genomics*, Dec 2021, https://doi.org/10.1016/j.xgen.2021.100065
- Bergenstråhle L, He B , Bergenstråhle J, Abalo X, Mirzazadeh R, Thrane K, Ji A, Andersson A , Larsson L, Stakenborg N, Boeckxstaens G, Khavari P, Zou J, Maaskola J, Lunderberg J **"Super-resolved spatial transcriptomics by deep data fusion"** *Nature Biotechnology*, Nov 2021, https://doi.org/10.1038/s41587-021-01075-3
- Bäckdahl J, Franzén L, Massier L, Li Q, Jalkanen J, Gao H, Andersson A, Bhalla N, Thorell A, Rydén M, L. Ståhl P, Mejhert N, **"Spatial mapping reveals human adipocyte subpopulations with distinct sensitivities to insulin"** *Cell Metabolism*, Aug 2021, https://doi.org/10.1016/j.cmet.2021.07.018
- Wu S.Z, Al-Eryani G, Roden D, Lee Junankar S, Harvey K, Kate, Andersson A, Thennavan A, Wang C, Torpy J.R, Bartonicek N, Wang T, Larsson L, Kaczorowski D, Weisenfeld N.I, Uytingco C.R, Chew J.G, Bent Z.W, Chan C, Gnanasambandapillai V, Dutertre C, Gluch L, Hui M.N, Beith J, Parker A, Robbins E, Segara D, Cooper C, Mak C, Chan B, Warrier S, Ginhoux F, Millar E, Powell J.E, Williams S.R, Liu X.S, O'Toole S, Lim E, Lundeberg J, Perou C.M, Swarbrick A, **"A single-cell and spatially resolved atlas of human breast cancers"** *Nature Genetics*, Sep 2021, https://doi.org/10.1038/s41588-021-00911-1
- He B, Bergenstråhle L, Stenbeck L, Abid A, Andersson A, Borg Å, Maaskola J, Lundeberg L, Zou J **"Integrating spatial gene expression and breast tumour morphology via deep learning"** *Nature Biomedical Engineering*, Jun 2020, https://doi.org/10.1038/s41551-020-0578-x
- Andersson A, Kasimova MA, Delemotte L, **"Exploring the Viral Channel KcvPBCV-1 Function via Computation,"** *The Journal of Membrane Biology*, Feb 2018., http://dx.doi.org/10.1007/s00232-018-0022-2

0.2.2 Preprints/Submissions

- Li X, Andrusivová Ž, Czarnewski P, Andersson A, Mattsson Langseth C, Gyllborg D, Braun E, Larsson L, Hu L, Alekseenko Z, Kopp Kallner H, Åkesson E, Nilsson M,

Linnarsson S, Lundeberg J, Sundström E, **"Decoding the development of the human spinal cord by multi-omics"** *in review, Nature Neuroscience*

- Erickson A, Berglund E, He M, Marklund M, Mirzazadeh R, Schultz N, Bergenstrahle L, Kvastad L, Andersson A, Bergenstrahle J, Larsson L, Shamikh A, Basmaci E, de Stahl T D, Rajakumar T, Thrane K, Ji A L, Khavari P A, Tarish F, Tanoglidi A, Maaskola J, Colling R, Mirtti T, Hamdy F C, Woodcock D J, Helleday T, Mills I G, Lamb A D, Lundeberg J, **"The spatial landscape of clonal somatic mutations in benign and malignant tissue"** *bioRxiv*, Jul 2021, `https://www.biorxiv.org/content/10.1101/2021.07.12.452018v1`

0.2.3 In Preparation

- **"Spatial transcriptions of antigen receptors maps lymphocyte clonality in human tissues"**, Author order not specified, *Estimated Submission: Q2 2022*

Acknowledgements

They say that we stand on the shoulders of giants, but I beg to differ. It is not the intellectual leaps of old scholars or pages of academic literature that have brought me here. Instead, it is the small everyday interactions that I've had with the people in my vicinity, each of them nudging me slightly in the right direction and giving me the momentum I needed to progress in work as well as life. Some of you have been with me from the very beginning, long before my PhD started, others I've just recently met, but all of you are very dear to me. Thus, this is my feeble attempt to express some of my gratitude – although words seldom do feelings justice.

Being "stuck in school" since I was six, I've had the pleasure to meet both some excellent teachers as well as some less impressive ones. About the latter, I shall speak no more, but the former group deserves some credit. To one of my first teachers, **Eva Zetterberg**, I want to say thank you for trying to bring order into chaos and showing how discipline is integral to success. **Bo Larsson** was my middle school teacher, you gave me free reigns to explore subjects of my own interest, this together with the support and encouragement you gave allowed me to grow, I appreciate that tremendously.

To my **Mom** and **Dad**, I can only say that I admire your patience. You've put up with me having lectures about Julius Caesar in the laundry room, executing chemical experiments in the kitchen, hogging the computer for days while writing useless programs, throwing a medal from the balcony because I didn't believe I deserved it (I still don't), and requesting books about quantum mechanics for Christmas. I definitely wasn't an *ordinary* child, but you always supported my interests and gave me the opportunity to pursue them fullheartedly, no matter how odd they were. You've let me prioritize schoolwork and my passions, as I've gotten older I realize what a privilege it is to have this freedom. In addition, you've always been there for me when I needed it the most. Knowing that you have my back gives me the courage to pursue my dreams. I've given you far too little credit for my achievements, but it's all because of you. I value you beyond everything else.

I also want to say thank you to my grandparents, I think you've been exposed to an almost equal amount of obscure lectures and odd interests as my parents. Sadly, both my grandfathers have passed, but their common denominator was an infectious enthusiasm for innovation and technology, something that indubitably rubbed off on me. My **Farmor** is kind, loving, and a force of nature, at age 80 she still clocks more steps a day than 90% of the population. All of you always seemed convinced that I could go far; this instilled massive confidence in me, and that's one of the most precious gifts you can give a child.

Thank you.

In 2017 I joined the Delemotte group at SciLifeLab and experienced what real research looks like, I loved every piece of it and my conviction of pursuing a research career solidified. During my time in the Delemotte group, I not only learned about Molecular Dynamics and enhanced sampling methods, but also met some amazing people. To my former PI and now friend, **Lucie Delemotte**, thank you for spotting my interest and taking a chance on a severely overexcited bachelor student, but also for the insightful advice and orienteering sessions. If I ever get my own group, it's my hope that I can be as good of a leader as you are.

My first exposure to the Spatial Transcriptomics (ST) group was in 2018, when Joseph Bergenstråhle a.k.a. **King Joey** asked me if I was interested in data analysis and then made me pursue a summer project in the group. Had it not been for you Joey, this thesis would look very different, and I'd lost out on meeting some remarkable people. Now, the time has come; lilla stycket is ready to fly, believe it or not.

When I began my PhD in 2019, **Žaneta** became my PhD-buddy, starting at the very same date. We kicked it off with a memorable visit to Stanford, Ben really was a fucker. Then, we had some amazing views on Skåla, experienced horizontal rain in Jotunheim, walked those pleasant snowscooter trails in Helags, froze our legs off at the car park in Sälen because we were to lazy to put up the tent, and also learnt that there are two Alby in Stockholm. You're a damn machine in the lab, but also one of the kindest and funniest people I've met. Thank you for all the laughs, you've added a lot of value to my time at Alpha-3.

Half a year into my PhD, the world got turned upside down by a global pandemic. Fortunately, our Alpha-3 adoptee/parasite **Franzi** lived just one floor above me, meaning we could go for morning runs and evening walks while working from home. Thus, instead of losing my mind because of solitude, I got a friend for life. Franzi, you have so much compassion and literally radiate kindness. All I can say is: stay excellent. Also, thank you for exposing my system to some proper food every now and then.

I've always loved a good adventure, and **Ludvig B** has been my faithful companion in many of them. From climbing mountains in Japan to raccoon spotting in California. I never quite understood how you put up with me constantly talking, being slightly paranoid about everything (e.g., mountain lions), and always lagging behind when running or climbing stairs. I'm convinced that I'll never meet a more humble or intelligent person than you; every time we talk I learn something new. Thank you for all the nice city-hikes, orienteering afternoons, latte art sessions, and taco Fridays – I will miss all of it dearly.

During the spring of 2021, I was the supervisor of **Julia**, then a master student. I'm still in awe of how well you tackled a project so far out of your comfort zone. Though, I'm even more impressed by the fact that you take ice-cold baths in mid-October and voluntarily agree to run with me at 6:am in the morning. You've brought a lot of perspective into my life this last year, and a conversation with you is never boring. I respect and applaud how you stay true to yourself, walk your own path, and are extremely open minded. You're a strong person, never forget that.

Mengxiao, the floor's very own Inspector Gadget. If there's a tool, you have it – and you've probably read every review about it, twice. You always bring the cold hard truth, but also show immense care and consideration. Thank you for checking in on me whenever I'm gone, and really made an effort to cheer me up when my health hit the floor last year. The discussions we have are never boring and I value your opinion a lot. **Kim**, having you as a roomie at AGBT2020 was a blast, the *ugglestycke* still cracks me up. You're an endless source of positive energy, have the best stories, but also tend to share excellent life advice. **Ludlar**, the man who lived a thousands lives, or at least had a hundred jobs. Working with you is always a pleasure, you're competent as hell and never fail to deliver what you promise. You're also the bioinformatic glue that keeps this group together, something you get way too little appreciation for. **Camilla Engblom**, where should I start? If dictionaries were made of pictures, the entry for "badass" would be a photo of you. Not only are you like a premium version of Wikipedia for immunology, but you manage to keep a vibrant energy while juggling family and a gazillion projects. What do you take, and can I have some?

Sami, I wanted to make an inappropriate joke about Finns and knives here, but it didn't quite make the *cut*. Now, you might pretend to be grumpy and dark, but we all know that you're actually an extremely warm, attentive, and generous person. **Kostas**, having lunch with you is always a journey through uncharted territory, I never know quite what to expect. You never fail to lighten up the mood, no matter what; every workplace needs people like you. **Leire**, my fellow long-distance runner; it's been amazing to have someone to share my passion for running with. No matter what you do or where you go in the future, I'm convinced you'll have nothing but success. **Hailey**, Yo Bro! You might not share my obsession about running, but at least we tried a bike-run combo; what a complete failure that was – I was way too slow for you. I first met you as one of my students, and was thrilled to have you join the floor! Best of luck with everything, and don't overwork yourself! **Lovisa**, the queen of disgusting tissues, placenta and fat – the ultimate combo. It's really cool to see how you've immersed yourself into the bioinformatics analysis despite having a very different background, it reminds me to never stop learning. **Eva**, I'm sincerely impressed by how much biochemistry you know (and remember);

every damn enzyme or co-factor in the different reactions. You're also my hernia-twin, third of February will always have a special mark in my calendar! Thank you for all the nice mushroom hikes, even though all of my spots were completely useless. **Linnea**, the Celsius lady, always in the background taking a compromising photo or video. I bet you have more material on people than the NSA. Also, you were one of the few people who contributed with pictures for Brorsdag, I want to give you a special shout out for this.

Marco, I'm happy to see that you've finally embraced the concept of *förmiddag*, this is without a doubt my greatest legacy to the group. Your curiosity for the machine learning and programming topics makes me smile, I wish more people shared your enthusiasm. **Paulo**, the expert of so many subjects. You are the perfect teacher, always being driven and interested in the topic as well as in the students' learning. I believe the work you're doing for the HDCA and the group will be transformative. **Nayanika**, we went to the same masters program, even though it took me the better part of it to realize that such was the case (apparently, I'm a bit slow). Thank you for organizing an awesome kick-off. I've always wondered what it would feel like to sleep in a prison cell, it was awesome. **Markus**, you recently joined the group, but you've brought a lot of laughter to my last months at SciLife. Thank you for all the nice podcast recommendations and conversations about training, life, NFTs, and supplements. I'm fully convinced you'll smash this PhD! **Jörg**, I don't think I've met anyone with as many hobbies as you. Sorry for bullying you about the Swedish rug-flag, but I still think it's absurd that you bought it for that price. Despite this small flaw in judgment, you're funny as hell and the kind of person who always brings a good mood to the room, Sweden should be proud to have you as a citizen. **Irene** a.k.a. Kowalsky, you made a brief visit to our lab, but left a big imprint. I love your positive attitude and excitement. Also, finally, there's someone who's as addicted to exercise and outdoor activities as me. Now, I hate to admit it, but those Nuggati sandwiches and toddy were actually darn good; I'm a convert.

Annelie, fortunately you've been spared of having to take care of me in the wet-lab. Still, the work you do and the effort you put into this group's well-being is noticeable also outside of the wet-lab. I love how open and unpretentious your are – we could not wish for a better lab Mom. To **Reza**, the battery guy with a taste for travel, you are the embodiment of kindness and warmth. Thank you for being such a nice work-neighbor, always checking in on me and bringing encouraging words. Elsa is lucky to have you as a dad! **Maja**, I hate to break it to you, but your tea sucks. Echinacea and blueberry muffins, how can you drink that? Luckily your positive attitude and laugh make up for it tenfold, you really spread joy around you. **Chus**, you're a man of many talents, writing a book while also conducting high-quality science, hella impressive. **Stefania**, it's inspiring to see how hard you work and how much effort you invest in making good cutting-edge science. Thank you for the advice and care you've given. I have nothing but the highest

of confidence that you'll have a smashing academic career. **Pontus**, I still can't believe you managed to convince me to go in the ice cross course. Clinging on to sides for dear life, I felt more pathetic than in a long time. Thank you for sharing your mushroom spots with us, at least they, in contrast to mine, delivered. **José**, we had plenty of nice lunches during the weekends at SciLife. You always showed me respect and encouragement from the start, while also giving me some perspective on academia. You're a though guy, but have the warmest of hearts. **Linda**, the ever smiling supermom. I can barely take care of myself, so I have no idea how you managed to care of those small creatures (a.k.a. your children) while also completing a PhD. That's some next level shit, major kudos to you. **Mickan**, the beginning of my stay in the group just barely overlapped with the end of yours, but your attitude and positive vibe made quite an impression. Thank you for all the good hiking tips and laughs in the lunch room.

Pär, you were my neighbor when I was dispatched to the dark side; and you actually made it a lot brighter. Thank you for all the nice chats! **Simon**, we've shared a lot of conversations in the kitchen, with the topics ranging between everything from fire passages to the BALCO doping scandal. I've thoroughly enjoyed every single one of them. I wish you and your family the best, and if (when?) the Russians come, I know you'll be safe down in Småland. **Humam**, your journey is just crazy – I don't think anyone here deserves their PhD position more than you. I'm amazed by your positive attitude and kindness, despite having walked such a rough path, you're a real fighter. Good luck with the rest of the PhD, I'm sure you'll excel at it. **Patrik**, I've had the opportunity to act as a TA in several of your courses, and I've enjoyed it every time – it's clear that you really seek to deliver a good experience to everyone involved. The students at KTH are really fortunate to have such a passionate and skilled teacher.

Of course, the whole floor of **Alpha-3** deserves a thank you for providing such a nourishing environment. I feel blessed to have a workplace that I enjoy spending time at. Thus, to all of the remaining inhabitants of Alpha-3, if I haven't mentioned you specifically, please know that I still really appreciate your presence.

I also want to acknowledge some communities who's generosity with advice, time, and devotion I've taken advantage of more than once. This starts with **Wikipedia**, it rarely fails to answer my questions and is a remarkable source of continuously updated information – and only possible through the hard work of volunteers. Next, stackexchange.com hosting a set of math, stats, and programming subforums – without these sites I'd still be struggling with some of the equations presented in this thesis. Last but not least, the **PyTorch** community, who provides an excellent machine learning framework at zero cost, enabling people to effortlessly employ advanced models and play around with new ideas.

Finally, I want to thank my supervisor **Joakim Lundeberg**. I've never felt anything but trust and genuine support from you, already from day one. Despite being a successful PI you're far from content and still strive for the next big thing, constantly seeking to explore and innovate, I find that admirable. You've opened many doors for me and I am genuinely thankful for that. Up north – where I, apparently, have my origins – we're not very good at expressing our emotions; but you should know that I will miss being a part of the Lundeberg lab, and it's with mixed feelings that I leave the group. Also, don't worry, the *Andersson IT-helpdesk* remains open; although, it might be operating in a slightly different time zone.

Table of Contents

List of Figures

All models are wrong, but some are useful.
George Box

Chapter 1 :: Introduction

1.1 Research interests

In his foreword to the iconic novel "A brave new world," Aldeous Huxley states that: "*It is only by means of the sciences of life that the quality of life can be radically changed.*" This statement immediately resonated with me and feels just as true today as when I first read it several years ago. Mainly because I firmly believe that it is not until we **understand** the systems that constitute the fabric of life that we will be able to successfully use them to our advantage. Today, the "sciences of life" are rather referred to as the field of "life science," a collection of multiple disciplines ranging from mycology to quantum biology. It is within this hyper-diverse space of life science that I've spent the last years exploring a particular niche known as *transcriptomics*; which is the study of the transcriptional landscape in living organisms. I find this branch of life science particularly alluring because the associated technologies permit us to study the fundamental units of life – the cells – at an unprecedented level of detail and scale. While cells can be described as units, it's when they are assembled into larger structures, like organs or complete organisms, that we see their true power. In these complex structures, cells rarely operate in isolation, but rather as a part of the whole. Thus, to fully comprehend these systems, cells should ideally be studied in their native context. The particular flavor of transcriptomics prefixed with the term "spatial" seeks to do precisely this – study cells in their natural environment – by preserving the spatial relationship between the surveyed targets (e.g., cells or single RNA molecules). Spatial transcriptomics alone will not be sufficient to answer all the questions pertaining to life; such a statement borders to absurd. However, it is an instrumental tool in the collective effort to expand the boundaries of our knowledge.

If understanding is our primary goal, we must first accumulate knowledge, as this naturally precedes the former – we cannot understand what we do not know. Quite often, gathering of knowledge follows an iterative process: we propose an initial hypothesis, then develop the necessary tools to refute or confirm our beliefs, and update our models accordingly. These steps are repeated ad infinitum in a process we commonly refer to as *learning*. The effort required to keep the process alive doesn't remain constant through time, but tends to grow with the corpus of knowledge. While a single person might initially be able to contribute to all parts of the iterative cycle, most domains experience a distribution of labor; some individuals **develop** tools to test hypotheses, others **apply** these tools and interpret the results to either accept/reject a hypothesis or generate new ones. I consider myself a member of the first group; a developer of methods, models, and software for others to use in their pursuit of answers to questions relating to spatial transcriptomics.

My passion has always been in the abstract and theoretical, but I also desire purpose in the shape of a clear application of my work. At the time of my entry into the field, spatial transcriptomics offered a combination of all these components. It was vibrant, exciting, and in dire need of new tools to analyze the generated data. Thus, my ambition has been to design frameworks and mathematical methods that would allow the users to extract relevant information from their spatial transcriptomics data. As a consequence, my research interests have all been very general and pertained to fundamental aspects of

the data. Although broad in their character, these interests can be summarized in three main questions:

1. *How can spatial transcriptomics be leveraged to gain new biological insights?*
2. *How do we, mathematically and statistically, model spatial transcriptomics data in a befitting way?*
3. *What are meaningful metrics and representations of features derived from spatial transcripomics data?*

For (1), I want to clarify that I've never had any interest to pursue these insights myself. I pose this question because I want to understand what questions spatial transcriptomics data is capable of addressing – which is necessary to know before any form of method development can be initiated. In contrast, for (2), this is a topic I'm deeply invested in, as I believe it's imperative to contemplate such questions whenever we seek to model data. If we don't, we're at great risk of producing faulty models. Finally, (3) is relevant from an analysis perspective, as the choice of representation influences our conclusions. These three questions permeate the majority of my work, and I hope that presenting them in this way illuminates some of the themes that span across my projects.

1.2 Thesis themes and content

From the nature of my research interests, it should be fairly clear to the reader that the focus of this thesis will not be of a biological character. Instead, the content will touch upon ideas of data modeling, probabilistic inference, and analysis of count data. Nevertheless, since modeling of a system requires some basic understanding of its mechanisms, I've tried to provide the reader with sufficient information to: (i) put the methods into context, (ii) gauge whether our assumptions and method designs are justified and credible, and also (iii) understand why the research questions are relevant.

1.3 Thesis outline

This thesis is composed of a total of four chapters, all of which will be described in brief below. The first – and current – chapter is the **"Introduction,"** where I seek to introduce my interests, motivation, and attitude to the thesis work. The purpose of the second chapter, **"Background,"** is to provide the most basic and necessary information to put the presented work into context. The Background is not comprehensive, but it should at least familiarize the reader to some of the most essential concepts presented in the articles included in this thesis. It is written with a somewhat broad audience in mind, thus, anyone with exposure to high school biology and university-level math should be able to follow without much difficulty. The **"Epilogue"** accommodates some personal and highly subjective reflections on what has been and what I believe is yet to come. Finally, the **"Present Investigations"** chapter includes both the complete work and a short summary of the publications and manuscripts included in this thesis. While I've put a lot of effort into each chapter, the Epilogue lies closest to my heart and I'll encourage the curious reader to pay it a visit.

Chapter 2 :: Background

2.1 DNA, RNA, and protein - from information to action

As mentioned in the Introduction, biology is the stage rather than the act in this thesis, but certain aspects of biology are so integral to the spatial transcriptomics techniques that it merits them a brief description. The first concept I'd like to introduce, as is standard in most textbooks on molecular biology, is the *central dogma* together with its three main components: DNA, RNA, and protein molecules.

The central dogma essentially outlines the flow of information in the context of genetic material, thus describing how static information (DNA) can be used to produce a functional entity (protein), through an intermediary medium (RNA). DNA is a stable molecule that consists of two complementary sequences built from the four *nucleotides*: adenine (A), guanine (G), thymine (T), and cytosine (C). The process of "casting" information held by the DNA into RNA is referred to as *transcription*. When an amino acid chain (protein) is assembled according to the sequence specified by an RNA molecule, we call the action *translation*. These concepts will be elaborated on in more detail below, but are presented now for clarity, see Figure 2.1 for a schematic overview of the central dogma.

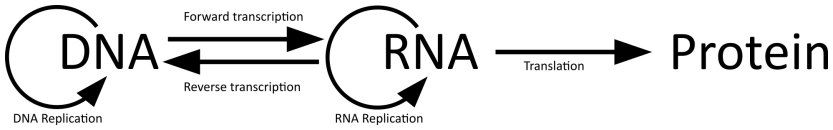

Figure 2.1: A schematic overview of the central dogma. The arrows indicate the direction of the informational flow.

The human genome consists of circa $3.2 \cdot 10^9$bp (base pairs, two paired nucleotides), or the double if both sets of chromosomes are counted. This number makes us blatantly unremarkable in terms of genome size.[1] For comparison, the Australian lungfish (*N.forsteri*) has an estimated genome size of $4.3 \cdot 10^{10}$bp,[2] the nematode *C.elegans*'s genome consists of approximately $1.0 \cdot 10^8$bp,[3] while $4.6 \cdot 10^6$bp make up the genetic material of the prokaryote *E.coli* (K12 strain).[4]

Out of the $3.2 \cdot 10^9$bp in the human genome, only 1-2% encode information that can be transcribed and then translated into a functional protein, meaning that 98-99% of our genetic material is "non-coding".[5, 6] Initially, upon this discovery, terms like "junk-DNA" were coined for the non-coding elements; which reflects the attitude towards them at that moment in time.[7] Later, as the non-coding regions have been more thoroughly studied, we've come to realize that far from all non-coding sequences are junk, some are even imperative for vital processes in our cells.[8, 9] While non-coding DNA might be foreign to individuals outside the scientific community, the genomic entity we call "gene" is a concept successfully disseminated to the broader public. This is in spite of the fact

that, in most eukaryotic cells, genes consist of a mixture of non-coding (e.g., promoter, introns, UTRs) and coding components (coding sequence).[10] Nevertheless, it is common to mainly associate genes with their coding part, and more specifically the protein they are destined to be translated into. It is often this "protein centric" view that tends to be conveyed to the non-scientific crowd, and also how genes will be treated throughout this thesis; even though it fails to do the complex nature of our genome full justice. Although somewhat dependent on the definition of what constitutes a gene, the estimated number of protein coding genes in humans is usually quoted to be in the magnitude of $2.0 \cdot 10^4$.[11]

In a normal (non-disease) state, the DNA is considered static in the sense that the information that it stores remains fixed; though, depending on the organism, its accessibility might vary over time – this is referred to as *epigenetic* changes. Here, accessibility means the ease by which the transcription complex can access particular regions of the DNA. The accessibility is regulated by how tightly the DNA is wound up around structural complexes built from a set of proteins known as *histones*. An *epigenetic* profile can be assigned to the DNA based on which regions that are more (*euchromatin*) or less (*heterochromatin*) exposed.

In contrast to the static DNA, RNA exhibits a more dynamic behavior where molecules are constantly transcribed and degraded within the cell.[12] RNA uses the same kind of fundamental constituents as DNA, four nucleotides, the only difference to DNA being that uracil (U) replaces the thymine (T) nucleotide. RNA exists in many different functional flavors, where the three main ones are: rRNA (r: ribosomal), tRNA (t: transfer), and mRNA (m: messenger). When genes are transcribed, the resulting product will be an immature mRNA molecule, that after certain processing – for example, splicing and polyadenylation – will reach a mature state. The processed mRNA molecules can then be translated into an amino acid sequence with the help of "translational machines" called *ribosomes*. The ribosomes are large complexes formed by rRNA and proteins. Just as the name indicates, an amino acid chain consists of multiple chemically (by a covalent peptide bond) linked amino acids. Amino acids are transported to the ribosomes by specific tRNA molecules which ensure that the amino acids are linked together in the correct order.[13] Perhaps somewhat unsurprisingly given the earlier protein centric statement, this thesis focuses exclusively on the cells' mRNA population.

In eukaryotes, there are 21 different amino acids that can be combined into an amino acid chain of any desired length. A majority of the amino acids (20) are immediately encoded by sequences in the DNA, but the 21st (selenocysteine) follows a slightly different regimen.[14] What might seem like a small pool of amino acids to sample from results in extreme diversity; if we were to construct a sequence of 10 amino acids, there would be approximately $1.66 \cdot 10^{13}$ alternative ways of doing this, a number that grows exponentially with the sequence length. Still, from the enormous space of potential amino acid sequences that we could synthesize, only a small subset is functional.[15] Rather than using the term "amino acid sequence," smaller chains (2-50 amino acids) are referred to as *peptides*, while the term *protein* is reserved for longer chains (> 50 amino acids).[1] In general, proteins tend to exhibit more intricate 3D structures than peptides. In fact, the structural configuration is so essential to a protein's function, that there exist specific helper pro-

[1] This 50 amino acid cutoff is taken from the *US Food and Drug Adminstration's* (FDA) policy document *"Definition of the Term 'Biological Product' "* (Docket ID. FDA-2018-N-2732). Other definitions exist, but are usually in the same order of magnitude.

teins (*chaperones*) with the sole purpose of making sure that an amino acid sequence is correctly *folded*.[16] Similar to mRNA molecules, the peptides and proteins can be subjected to modifications after their synthesis, referred to as *post-translational modifications* (PTMs). The PTMs often include the attachment of a second type of compound to the protein or peptide, with some examples being: phosphorylation (addition of a phosphate group), glycosylation (addition of sugar moieties), and lipidation (addition of lipids). The introduction of PTMs can have a wide range of effects on the proteins. PTMs could, for example, be added to improve stability or to activate/inactivate a protein.[17] Proteins (and to some extent peptides) are heavily involved in almost every aspect of cellular life, they orchestrate cell motility, provide structure, catalyze chemical reactions, cleanup waste, maintain the cell's environment, and much more. Proteins and peptides are not limited to *intra*cellular mechanisms, but also partake in *inter*cellular interaction and communication, as well as long-range signaling between organs; insulin being an example of the latter.

The man who introduced the notion of the central dogma in 1956, Francis Crick[2], was certain that the presumed DNA $\rightarrow$ RNA $\rightarrow$ Protein flow of information existed.[18] He also conceived a bidirectional flow of information between RNA and DNA. Chemically he saw no reason as to why RNA couldn't figure as a template for DNA, but was lacking both evidence and a clear application where it would be useful. This changed in the 1970's when Howard Temin and David Baltimore discovered and managed to isolate an enzyme (a catalytic protein) known as *reverse transcriptase* from members of the retrovirus family.[19] Reverse transcriptases (RTs) are capable of synthesizing DNA from an RNA template, meaning that they "reverse" the transcription process, hence the name. RTs are used by retroviruses to make a DNA copy of their genetic material (RNA-based) which then can be inserted into the infected host's DNA. The existence of RTs could easily be taken for another obscure oddity of the "virus world," but its discovery had a massive impact on the world of molecular biology, partially enabling a paradigm shift in the generation of transcriptomics data – this will be further expanded on in section 2.3.

In contrast to what have been presented so far, the central dogma also includes parts that don't cast information from one medium to another, but rather increase the amount of existing material. DNA *replication*, where a new copy of the DNA is made, is one such part. The key enzyme in DNA replication, synthesizing the new copy, is the *DNA polymerase*. Although not as common as DNA replication, RNA replication does occur, but seems to be a process exclusively present among viruses.[20] During RNA replication, a special kind of *RNA polymerase* is used.

The above description is an extremely simplified account of the different processes that comprise the central dogma, and while it does outline the main themes, it also neglects many of the more intricate aspects. For example, from this summary, it's easy to assume that the number of mRNA molecules found in a cell should correlate with the number of proteins associated with the transcribed gene; but this only holds true for some proteins – for others we may even observe an inverse relationship.[21, 22, 23] Such insights allude to the impact that regulatory mechanisms have on the molecular composition of the cell.

[2]Francis Crick is a complicated character. He indubitably made valuable contributions to the field of molecular biology. However, in later years several controversies concerning his behavior have surfaced. These include accusations of sexual harassment, clear evidence that he held pro-eugenics beliefs, and failure to properly credit other researchers. Such actions are, in my opinion, deplorable and *should* taint his reputation. Thus, while he deserves some credit, his less tasteful sides should also be highlighted.

Indeed, which pathways that are active in a cell have a huge influence over its identity. Two cells can share genomic profiles, but – depending on which processes that are active – behave and present as vastly different individuals. This phenomenon of different *cell types* or states is further examined in the next section.

2.2 Cells - the basic units of life

2.2.1 The essentials

Without DNA, RNA, and protein, there would be no life, but simply bringing these elements together will not create life. mRNA molecules need to interact with the ribosomes to be translated, the biochemical conditions (e.g., pH and osmolarity) must be correct for the proteins to fold and function correctly, and nucleotides must be present to enable transcription of the DNA.[13] To meet all of the necessary conditions, evolution has produced a closed compartment attuned to facilitate specific biochemical processes, this compartment is what we refer to as a *cell*. There's a huge variability in the architecture and composition of cells across different species, but there are also several shared features, of which some examples will be given below.

Perhaps the most essential feature of the cells is their protective barrier, the *cell membrane*, consisting of a lipid bilayer that separates the environment inside the cell from its surrounding. The viscous liquid found within the membrane is named the *cytosol*, and it envelops several intracellular pockets known as *organelles*.[24] Some of the organelles found in mammals are: mitochondria, the nucleus, the endoplasmic reticulum, the Golgi apparatus, and lysosomes. The mitochondria are colloquially described as – somewhat clichéd – the "powerhouses of the cell." The expression intends to convey how the electron transport chain – found in the mitochondrial membranes – produces most of a cell's "energy currency" ATP (adenosine triphosphate) in aerobic conditions.[25] The nucleus holds most of the genomic material (a small portion is also present in the mitochondria), meaning that transcription mainly occurs within the nucleus. In general, mRNA produced in the nucleus will eventually be transported through the nucleic membrane, and carried to parts of the endoplasmatic reticulum to which ribosomes are bound. As mentioned earlier, the ribosomes use the mRNA as a template and amino acids as ingredients to produce peptide or protein molecules. If the produced peptides/proteins are destined for the membrane or the extracellular region, they often pass through the Golgi apparatus, which packages the proteins in transport capsules (*vesicles*). In addition to acting as a trafficking hub, the Golgi apparatus sometimes performs post-translational modification. Lysosomes are involved in degradative processes, such as breaking down old cell parts. The cytosol is a crowded space filled with proteins, RNA molecules, organelles, metabolites, etc. Many of the cell's biochemical reactions occur in the cytosol, for example the ATP-generating *glycolysis*.[24] Having outlined some of the most basal aspects of cells, we'll next explore some of their finer nuances and attributes.

2.2.2 Cell types and states

Akin to how the atomic elements listed in the periodic system figure as building blocks of our physical world, cells can be considered as the basic units of life. Indeed, just like the atom, a single cell possesses distinct properties and a certain amount of integrity. Cells can also be gathered in communities to form new entities (e.g., organs or tissues), analogous to the molecules that atoms comprise.[3] The simile between cells and atoms also captures the idea of there being different *cell types* (e.g., immune cells, neurons, muscle cells, etc.) that depending on how, w.r.t. ratios and organization, they are combined result in different outcomes. Variance and differences are present even among cells within the same cell types, less stark than between the types, but still present. Continuing with the atomistic narrative, one might think of these *cell states* as equivalent to the atoms' isotopes. Ideally, if we understand how the different cell types locate within their periodic system, we should be able to predict yet undiscovered cell types whose existence is implicated by the structure of the system; similar to how Dmitri Mendeleev predicted eka-aluminium (gallium), eka-silicon (germanium), and other elements *before* they were experimentally confirmed.[26] Although the "periodic system of cells" is an attractive and helpful representation, I want to acquaint the reader with a – to me – compelling and complementary alternative.

Around the same time (1957) as the central dogma was conceptualized by Crick, another scientist named Conrad Hal Waddington proposed a model designed to describe the cellular differentiation process during development.[27] Waddington postulated that the different cell types along the differentiation trajectory all could be positioned in an *energy landscape*. The stable cell types would correspond to *low energy states* while the more transient transitional states between them would have a higher energy. For clarity, just as in molecular systems, low energy states are favored and "desired" by the cells. The more differentiated (specialized) a cell type is, the lower its energy. This model also made Waddington describe the developmental process as *unidirectional*; a cell becoming more differentiated would be similar to a marble rolling down a hill, while a cell spontaneously dedifferentiating is equivalent to the marble rolling uphill by itself, i.e., nonsensical. Importantly, this model does not prohibit dedifferentation, but rather posits that without any external stimuli "pushing" the marble back up, we are unlikely to observe this in a natural setting. An example of such a stimuli being how the addition of Yamanaka factors to differentiated somatic cells can revert the differentiation process and convert them into stem cells (induced pluripotent stem cells, iPSCs).[28] While rare, unaided dedifferentation has been proven to occur in natural systems, especially in regenerative processes, thus invalidating Waddington's unidirectional assumption – but not nullifying the model at large.[29] In his research, Waddington related the energy of a cell's state to its epigenomic profile, but since then other modes of information – like the transcriptional state – have been used to map cell states to energy levels.[30, 31, 32] The *Waddington landscape* has also been employed to describe non-developmental processes. In fact, the landscape could be extended to harbor the complete set of states that a cell may reach, where "energy barriers" act as bulwarks separating one class of types from others.

[3] Comparing cells with atoms and the concept of a "periodic system of cells" is far from my own conception. I first encountered these ideas in a talk by *Aviv Regev*, and found the analogy very appealing as it aligned well with my chemistry background. To me, it offers a powerful way to convey the complexity of the cell type landscape in more familiar terms. Of course, I've adapted parts of it to my own liking, but I want to credit the source that made me aware of this perspective on cell types.

The model founded in Waddington's ideas treats cell types as points along a multidimensional continuum rather than discrete entities, as in the atom-like model; which may resonate better with someone from a biological background. The models are different, but not contradictory, and can each be helpful when pondering on questions about cell types and their relations. When attempting to model the transitional processes, the cell type space, and the effect of perturbations, I personally find it slightly more helpful to look at the system through Waddington's lense. The energy landscape model encompasses the intermediary states between the well-defined cell types, provides a clear explanation of cell type transitions, and can also easily capture hierarchical relationships between cell types. Furthermore, the energy landscape model casts concepts like cell states and transitions into a language that befits other quantitative fields like statistical mechanics, which have developed an extensive toolbox for analysis of similar systems.[33]

Here, a brief explanation, by an example, of what is meant by "relating a cell state to an energy level" will follow. Imagine a much simplified cell that only expresses two different genes (Gene 1 and Gene 2), the expression of each gene resides within the range of zero to one. The cell's state is exactly described by the tuple $(x, y) \in [0, 1] \times [0, 1]$, where x and y represents the expression levels of Gene 1 and Gene 2 respectively. Using the function $f(x, y)$, any given state (x, y) can be associated with a specific energy level. By applying f to every combination of possible cell states, a representation of the full energy landscape is obtained. The energy landscape can be modified by the introduction of a perturbation (g) to the system, changing the character of the mapping between cell state and energy level. A visual interpretation of this explanation is presented in Figure 2.2.

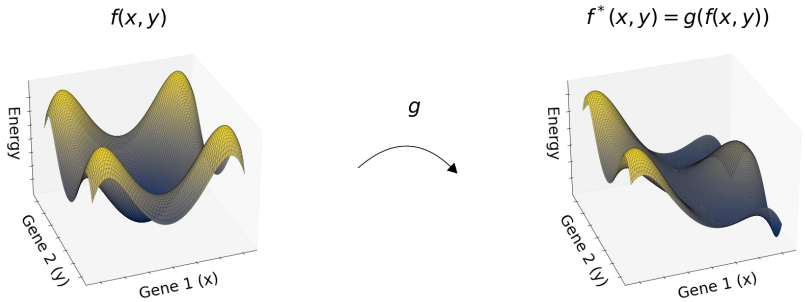

Figure 2.2: A simplified representation of the Waddington landscape. In a cell system only expressing two genes, Gene 1 (x) and Gene 2 (y), a function $f(x, y)$ can be used to map each cell state to an energy level. The image of f represents the energy landscape of the cell. By introducing some pertrubation g to the system, the energy landscape changes. Likewise, the map between state and energy level is updated according to $f^*(x, y) = g(f(x, y))$.

Assuming that the energy landscape contains all the states a cell could possibly occupy, disease and other aberrant states must also be represented. With cancer as an example, one might envision how a transition to the cancerous state requires the normal cells to climb an energy barrier. However, the introduction of a mutation (the hallmark of cancer) might suddenly disrupt the landscape, for example lowering the barrier or increasing the

energy of the normal state. The modifications to the energy landscape and disruption of relative energy between the healthy and disease state might then promote the migration from the former to the latter. In this case it's an *internal* (to the cell) process that modifies the energy landscape and permits the transition, but the alterant could just as well be *external*.[34, 35] For example, the cell's environment might aid in the reconfiguration of the energy landscape, creating new paths for the cell to travel along. Recently, it has been shown that cell-to-cell communication are critical for cell-fate decisions and how fluctuations in external signals have a large influence over the outcome of the differentiation process.[36] The implications being that one should not only study the cells' internal state, but also their environment and interactions. Experimental techniques that chart the molecular composition of cells without relating these signals to each other in the physical domain remain largely oblivious to the external factors that might exhibit influence over the cells. Thus, there's an obvious value to methods that enable the collection of molecular signals while also preserving their spatial context; enter spatial-omics techniques.

2.3 Spatial Transcriptomics - when context matters

The "-omics" suffix has become increasingly popular, and its use more liberal. A slew of exotic terms like "foodomics" or "museomics" have now been added to the set of more traditional fields like genomics and proteomics.[37, 38] Still, there's an apparent lack of consensus regarding the exact definition of omics. Every researcher, more or less, have his/her/their own interpretation of the term. In my opinion, the definition provided by Wikipedia (entry: "Omics", date: 2022-02-08) manages to convey the general ideas of the omics concept in a short and concise manner:

> "*Omics aims at the **collective** characterization and quantification of pools of biological molecules that translate into the structure, function, and dynamics of an organism or organisms.*"

I've chosen to highlight the "collective" part of this definition to emphasize how the term omics is strongly associated with, and often imply, studies where plenty of molecular signals are surveyed simultaneously. Omics-studies can be executed at different scales, ranging from sub-cellular resolution to whole organisms. The type of signals that an omics-field focuses on can often be deduced from its name: genomics → genomic material, transcriptomics → transcripts, and proteomics → proteins. The addition of "spatial" to an omics technique or field, indicates that the information collected to some extent preserves spatial relationships.

The work in this thesis belongs to the transcriptomics sphere, and more specifically, the niche of spatial transcriptomics. Thus, even though the other omics-fields offer a plethora of relevant insights, henceforth, they will only figure as peripheral elements of this thesis.

This "Spatial Transcriptomics" section will summarize the field both from an experi-

mental and a computational perspective. The subsections are written to give the reader an overview of its associated topic, for more detailed accounts I refer to any of the many reviews on the subject. However, before entering the realm of spatial transcriptomics, it's fit to first present *some* of the technical advancements that enabled the current-era spatial methods.

2.3.1 A prelude

The main objective in spatial transcriptomics is to elucidate both the position and identity of the transcripts present in the system of interest (e.g., a tissue sample). Multiple strategies exist for this purpose, but a majority of them rely on some integral concepts that will be, briefly, introduced here.

2.3.1.1 Sequencing-based identification

Having located or isolated a transcript, the subsequent task in transcriptomics studies is to determine its identity. One way of doing this is to *sequence* it, either *in situ* or *ex situ*. The basic premise of sequencing is, as the name indicates, to determine the nucleotide sequence of the target transcript.

One of the earliest and most successful sequencing approaches is *Sanger sequencing*, originally developed in the late 1970's for DNA sequencing by the (then to be) twofold Nobel laureate and eponym Frederick Sanger.[39] What follows here is a simplified explanation of the method. In Sanger sequencing, the target sequence (to be identified) is mixed with a pool of different components including: standard nucleotides, modified nucleotides, a DNA polymerase, and primers (to a single site). Often, the amount of target sequence is increased in an *amplification* step, which essentially makes several copies of the existing sequences. The primers will hybridize to a region of the target, allowing the polymerase to attach. Once attached, the polymerase can use the nucleotides to synthesize a, to the target, complementary strand. Under normal circumstances, the polymerase would extend the complementary strand until the end of the target is reached, incorporating the provided nucleotides in the nascent strand. What makes this setting deviant is the presence of the modified nucleotides, which will be inserted into the synthesized strand after a random number of unmodified nucleotides have been added. The modified nucleotides do not permit further extension, and thus terminates the elongation process. At the technique's inception, the modified nucleotides were identically labeled, requiring the reaction to be executed four times, one for each nucleotide type. Concurrent analysis of all four bases in a single reaction would later be enabled by technical advancements that allowed each nucleotide type to be colored by a uniquely colored fluorophore.

Assuming that the reagents are present in sufficient amounts, the elongation will terminate at each position in the target sequence, creating fragments of varying lengths. Next, the generated fragments will be separated by size, for example, using capillary gel electrophoresis. When using the refined approach with multiple fluorophores, the identity of each fragment is determined as it exits the size separation step by registering the fluorescent signal from its label. The result is a chromatogram that displays the final base identity as a function of sequence length, from which the original sequence easily can be reconstructed.[40] This solution is ingenious, but also very low throughput – it's

a tedious and cumbersome process that only determines the sequence of one target at a time. Furthermore, it assumes that the user knows or can guess at least some parts of the sequence prior to the experiment (to which the primer can bind). Thus, while useful, Sanger sequencing could never support the form of high throughput processing that the current omics-fields require. Despite the many drawbacks of Sanger sequencing, it leaves a huge heritage and represents a seminal point in the era of molecular biology. Indeed, Sanger sequencing fueled one of the largest collaborative science efforts in history: the human genome project.

Sanger sequencing belongs to the set of first generation sequencing techniques, which executed the sequencing task successfully but had several limitations (as mentioned above). Thus a series of improved second generation sequencing or *next generation sequencing* (NGS) techniques were invented in the mid to late 1990's, from which commercial products started to emerge around the mid-2000's.[41, 42] To illustrate the impact of these innovations: between 2005 and 2010, the per-base sequencing cost was halved every fifth month. This is more than three times faster than the microchip transistor capacity increase predicted by Moore's law.[43] One of the key characteristics of NGS techniques is how they allow massively parallel sequencing (MPS), meaning that several targets can be processed at once. The parallelization is often achieved by physical separation of the targets in a flow cell, which prevents signal cross-contamination in the sequencing processes. Of course, the various NGS platforms exploit different schemes for the sequencing, and they are all examples of impressive technical advancements from the first generation techniques. On a rudimentary level, the NGS methods use the same idea that the first generation methods introduced, being to somehow register the target sequence by either adding a single or multiple bases that elicit an informative signal (e.g., light emission) that can be registered and used to decipher the sequence. When working with RNA targets, they are typically converted to cDNA (complementary DNA) by reverse transcriptase enzymes before sequencing. The conversion step is implemented in the protocol because RNA molecules are unstable compared to DNA/cDNA and prone to degradation.[13] In addition to cDNA conversion, many NGS platforms also use sequencing adapters which are short oligonucleotide sequences that are ligated to the targets (usually after fragmentation). The adapters differ depending on the sequencing platform, but often contain sites for primers to attach to, a sequence to bind to a flow cell or equivalent, and potential sample indices if material from multiple samples are to be run together.[41]

In order to produce sufficiently strong signals, both first generation and NGS techniques require a certain amount of target material as input. This constitutes a challenge, because the required amount is often higher than what the source(s) can provide. Fortunately, another Nobel Prize awarded (1993) method – the polymerase chain reaction (PCR) – can be used to circumvent this seemingly insurmountable issue. PCR employs a protocol of repeated cycles where, in each cycle, complementary strands (DNA or cDNA) are denatured, primers bind (anneal) to the denatured strands, and new complementary strands are synthesized by a polymerase (extension) using the primed strands as templates.[44] Depending on which primers that are included among the reagents, PCR can be linear (primarily used in Sanger sequencing) or exponential (primarily used in NGS) in its amplification. While invaluable, PCR also comes at a cost, being that of unwanted *PCR bias*. It has repeatedly been observed how the PCR efficacy differs between targets, something that has serious consequences in quantitative studies where multiple targets are examined in parallel.[45, 46] With a difference in efficiency, the resulting pool of amplified mate-

rial does not necessarily preserve the proportional abundances between targets, and thus change their relative (as well as absolute) expression values. Nowadays, to mitigate PCR bias, a construct known as *"unique molecular identifier"'* (UMI) is usually attached to the targets before amplification. The UMI assigns a specific identity to each molecule, allowing all PCR products with the same UMI to be collapsed into a single observation after being sequenced.[47]

NGS techniques are fairly limited in the number of bases they can successfully sequence with high confidence; their ranges are usually in the magnitude of 20-400bp.[48] Hence, for most targets, only a fraction of their content is sequenced. For context and comparison, the average (human) mRNA molecule is 3392bp,[49] and Sanger sequencing can sequence up to 900bp.[50] Given the low numbers of sequenced bases in NGS techniques, they might seem to have been developed in vain, but clever computational strategies make them highly effective. By either: aligning the sequenced data to an existing reference genome/transcriptome, or assembling a reference *de novo*; the short sequences, known as *reads*, are sufficient to determine the origin of the target they represent. Naturally, if identification of isoforms or mutations is desired, the short sequence lengths become an issue, as they might not cover the variable regions.

Unsurprisingly, a third generation of sequencing techniques was introduced to not only provide high throughput, but also access to longer sequences, often operating at single molecule level. At the moment of writing, the set of third generation sequencing (TGS) techniques have not replaced the NGS techniques, like the latter did with those of the first generation; instead they co-exists and are often used as complementary tools. The TGS techniques are continuously being refined and improved, but they still have some disadvantages compared to NGS techniques, such as: relatively low accuracy, decreased throughput, and higher cost.[51]

In the forthcoming sections there will be examples of how sequencing techniques can be employed in a spatial context, but we'll first explore an alternative strategy for transcript identification and a non-spatial application of modern sequencing techniques.

2.3.1.2 Probe-based identification

Sequencing based strategies for transcript identification are excellent to use when there's an interest in the whole transcript population of a specimen, but this is not always the case. There are plenty of instances when only a subset of the transcripts is of relevance, and where this set is known *a priori* to the experiment. In such a scenario, sequencing of all transcripts easily becomes redundant and a more targeted approach is justified.

When charting the spatial position of a select set of transcripts is the primary objective, we can design complementary (to the target) *probes* – tagged with a marker (e.g., fluorophore) – that will hybridize with the target sequences in the specimen. The marker will then emit a signal that can be registered and used to map the spatial distribution of the target transcripts. Methods of this class, relying on hybridization between a probe and target, are known as *in situ* hybridization (ISH) techniques.[52] To be noted is how ISH methods give immediate information about the spatial location of the targets. Meanwhile, in sequencing-based methods, an extra encoding (e.g., by the inclusion of a barcode or

tag) step is often required to associate a spatial location with a target.

ISH methods have a long history, already in 1969 Pardue and Gall used rRNA probes (tagged with a heavy hydrogen isotope) to visualize an rRNA encoding gene in oocytes from the African clawed frog (*Xenopus laevis*).[53] Soon thereafter, fluorescent labels were introduced as alternatives to the radioisotopes. Circa 30 years after Pardue and Gall, in 1998, Femino et al. showed how fluorescent *in situ* hybridization (FISH) could be used to detect mRNA in tissue.[54]

Conceptually, probe-based identification might seem more straightforward than sequencing of transcripts to determine their identity; it's easy to execute *in situ* and relies on fewer chemical reactions. For very few, or a single, kind of target transcripts, the ISH methods are effective, but once multiplexing is attempted, issues start to emerge. With a limited number of fluorophores available, spectral overlap fast becomes a problem and prevents concurrent visualization of multiple targets (but, as we shall see, this can be circumvented). Targets can, of course, be visualized one-by-one, but this soon becomes ineffective and time consuming. Also, the more probes that are used, the more care and effort is required during the probe-design step to avoid unwanted probe-probe binding. Thus, one of the big challenges when using ISH for spatial transcriptomics is to increase throughput and multiplex capacity.

2.3.1.3 Single cell/nuclei RNA-sequencing

Despite techniques like ISH being present for more than 50 years, the spatial transcript**omics** field remained dormant until the mid-2010's when several high throughput techniques were launched. In fact, the current era of spatial transcriptomics techniques was preceded by the emergence and rise of another related field, being *single cell transcriptomics* often abbreviated as scRNA-seq. Figure 2.3 illustrates how the different fields exhibit similar growth trends, but with a slight translation to the right on the time axis for spatial transcriptomics. Single cell transcriptomics protocols, by design, tend to disrupt the spatial structure of a sample and eliminate the majority of spatial information, but it still pioneered many of the concepts used in analysis of spatial transcriptomics data. Furthermore, single cell and spatial transcriptomics data are often combined (computationally) by integrative means to obtain a more comprehensive understanding of the target specimen. Thus, despite its somewhat disparate nature, the single cell transcriptomics field has paved the way for spatial transcriptomics and played a crucial role in its development.

With the introduction of NGS techniques, transcriptomics studies – here referring to near transcriptome-wide surveys of a sample's expression landscape – became gradually more common. At this time, the *modus operandi* was to use *bulk* approaches, where the collected tissue specimen was treated as a singular observation. To clarify, with bulk methods, the tissue is processed experimentally and computationally, finally resulting in information about the abundance of the various transcripts across the *whole tissue*.[56] While informative, there's an evident issue with this approach: it lacks resolution. Tissues are inherently diverse, and rarely consist of a single cell type, but rather tend to host multiple cell communities. Still, with bulk-sequencing techniques, the origin of a transcript (i.e., which cell or cell type that produced it) remains unknown. Consequently, the bulk-methods convolve signals from different sources (cells), and mask nuances – such as multimodal

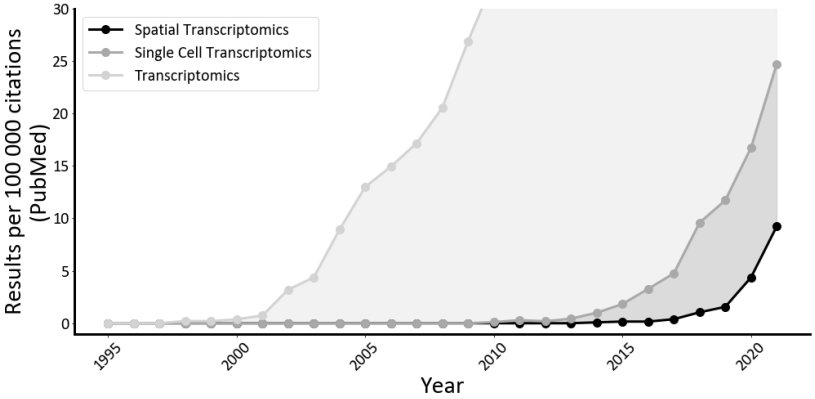

Figure 2.3: PubMed trends of different transcriptomics fields. The graph shows the (normalized) number of results per year for each of the terms: "Spatial Transcriptomics" (black), "Single Cell Transcriptomics" (dark gray), and "Transcriptomics" (light gray). The result is normalized to adjust for changes in general publication trends over time. The data was collected using the online service "PubMed by Year".[55]

gene expression distributions or contributions from rare cell types – that may exist among the cells. In spite of not being able to resolve individual cell profiles, bulk sequencing has proven tremendously valuable to advance our understanding in plenty of biological fields.[56] Though, for certain questions, single cell resolution is necessary, making bulk methods inadequate. The need for methods with higher cellular resolution was quickly identified and addressed. Already in 2009, Tang et al. showed how the transcriptional profile of a single cell could be charted.[57] With an existing proof of concept, it was just a matter of time before the concepts could be scaled to a larger set of cells, ushering in the era of single cell transcriptomics. In the years that followed, several strategies were proposed, resulting in an avalanche of new single cell transcriptomics methods. In 2012 Smart-seq was published by Ramsköld et al., where they outlined how full-length mRNA sequences could be retrieved for multiple single cells.[58] Two years later (2014), single-cell sequencing was selected as the "Method of the Year 2013" by Nature Methods.[59] Next, two seminal papers were published in 2015 and 2017 by Macosko et al. and Zheng et al. respectively, both describing how individual cells could be isolated in a single microfluidic droplet, where the necessary steps to tag (with a cell specific barcode) and amplify transcripts could be executed without intermixing of transcripts from different cells.[60, 61]. The experimental techniques quickly transitioned from the academic sphere to the industry, and today several commercial products for single cell sequencing exists. The launch of (relatively) easy and robust protocols for single cell sequencing accelerated the field, and plenty of studies have now used these techniques to address relevant biological questions.[62, 63, 64, 65, 66]

Like every experimental technique, single cell sequencing has its inherent biases. For example, certain cell types (e.g., adipocytes and neutrophils) are sensitive to the preparation steps and thus "burst" before entering a droplet, hence no data for these cell types are collected.[67, 68] Fortunately, a very similar but slightly different alternative to single cell sequencing exists, which sometimes (but not always) can capture the transcriptomics profiles from the sensitive cell types. This alternative technique is single *nuclei* RNA sequenc-

ing (sNuc-seq), which isolates each cell's nucleus and then sequence its mRNA content.[69] Together, single cell/nuclei sequencing act as some of the primary modalities in huge international research efforts such as the Human Cell Atlas and Tabula Sapiens.[70, 71] The single cell field in many ways laid the foundation for the spatial transcriptomics field, creating a precedent for what is expected from larger transcriptomics studies.

2.3.2 Experimental techniques

The field of spatial transcriptomics has already accumulated an impressive mass of experimental methods despite its young age. In 2021, Nature Methods granted spatial transcriptomics the same title (Method of the Year) as single cell sequencing received in 2014, which serves as an indicator of the increased interest in the associated techniques.[72] As is custom when a "new" field reaches beyond its origins and spreads to the broader public, plenty of reviews have been written on the topic, attempting to describe the different flavors of spatial transcriptomics.[73, 74, 75, 76] The aim of this section is somewhat more banausic: to provide the reader with a very primitive introduction of the different tactics used to generate spatial transcriptomics data. The interested reader is encouraged to visit any of the referenced reviews for a more comprehensive overview of the techniques.

With such an assorted set of spatial transcriptomics techniques, there's no obvious way to arrange them into discrete categories. One coarse grained way of classifying experimental methods is the binary split between sequencing-based and image-based methods. Here, I've decided to use a slightly more fine grained classification scheme, consisting of five different groups, inspired by the work of Asp et al. and Moses et al.[73, 74] The five categories are:

1. ROI selection techniques
2. *in situ* hybridization techniques
3. *in situ* sequencing techniques
4. *in silico* spatial reconstruction techniques
5. *in situ* capture techniques

More information about each category will be found in the subsections below.

2.3.2.1 ROI selection

The ROI selection techniques could be considered as the most "brute force" approaches to obtain spatial transcriptomics data. Their core principle is to excise a single or several regions of interest (ROIs) from a tissue specimen, separate them physically, and determine the transcriptional profile of the ROIs one-by-one. With this approach, the spatial resolution is fully determined by how fine ROIs one can select and separate from the rest of the tissue. Examples of manual tissue microdissection exist, but the application of lasers to cut out the ROIs followed by extraction – a procedure referred to as laser capture microdissection (LCM) – is generally preferred. In later years, refined versions of the ROI selection techniques have been developed, where the focus has shifted from physical

removal of the ROI to collecting the target transcripts within the ROI. To exemplify, in TIVA (transcriptome in vivo analysis tag), a photocleavable tag is introduced into live cells. Then, the cells of interest are exposed to light of the correct wavelength (405nm), which activates the tag and allows it to bind to the polyA-tail of the transcripts.[77] The tag contains a biotin molecule, making it easy to capture the hybridized tag-target complex with the help of streptavidin. Once captured, the mRNA can be released and sequenced, generating a transcriptional profile exclusively originating from the chosen cells. Another example of a more "modern" technique is the GeoMX Digital Spatial Profiling (DSP) platform from the company NanoString.[78] The GeoMX DSP platform uses a set of probes targeted towards a specific set of transcripts, and each of the probes has an oligonucleotide (oligo) tag encoding the probe identity. Once the probes have bound to the targets, the ROI is beamed with UV (ultra violet) light to release the oligos. The oligos are captured and registered as belonging to the examined ROI. In this case, the number of oligos is related to the number of target transcripts found within the tissue. As expected, the probe set limits the multiplexing capacity, and the first DSP platform didn't operate at the whole-transcriptome level, but the newer GeoMX Whole Transcriptome Atlas (WTA) platform supports panels in the magnitude of 18000 genes.[79] Other methods that will not be described in detail, but which belong to this category are: NICHE-seq, tomo-seq, and STRP-seq.[80, 81, 82]

2.3.2.2 *in situ* hybridization techniques

Instead of extracting the transcripts from an ROI and identifying them by sequencing their identity can be resolved *in situ* with the help of targeted probes (see section 2.3.1.2). The approach that Femino et al. used in 1998 utilized a set of at least five probes (50 nucleotides long) each designed for neighboring parts of the target transcript and labeled with five fluorophores.[54] The use of multiple shorter probes in the ISH context is referred to as *single molecule FISH* (smFISH). When interpreting the signal in smFISH experiments, the intensity is proportional to the number of bound target sequences. The large number of "fluorophores per probe" was required to provide a signal strong enough to be confidently detected. While successful, this approach was burdened with several issues such as self-quenching and difficulties with the synthesis of multilabeled probes.[83] In 2008, Raj et al. presented a new procedure for *in situ* identification of transcripts, which can be considered an upgrade of the 1998 approach. The new method utilized 48 or more probes (17-22 nucleotides long) binding to adjacent sequences, each with a single fluorophore, creating a sufficiently strong signal but without the drawbacks mentioned for the older method.[83] Despite its enhancements, the 2008 method still struggled with multiplexing, since only a few fluorophores are available and there are tens of thousands of potential targets in the transcriptome. Albeit not ideal for full-scale omics studies, smFISH and its derivatives are still seen as a "gold standard" for transcript identification, and often used to validate specific findings from more recent techniques.[84]

The true breakthrough of ISH-related methods came when the smFISH techniques were combined with *combinatorial barcoding*, which increased the multiplexing capacity manifold. With combinatorial barcoding, a target is no longer assigned a single color as its identifier but rather a given sequence of colors (representing a barcode). To give a simple example, if three probes are used and each is tagged with one of five different fluorophores,

transcripts of $5^3 = 125$ different genes could theoretically be targeted. In 2014, Lubeck et al. presented *sequential FISH* (seqFISH) which combined combinatorial barcoding with smFISH, allowing them to examine the spatial location of 12 different mRNA targets in 37 cells. The seqFISH technique uses several sequential rounds of probe hybridization and signal imaging, where each kind of target transcript is assigned a unique code.[85] While being able to multiplex, seqFISH is a fairly slow technique that have compatibility issues with highly autofluorescent tissue types. Notably, parts of the technology that seqFISH builds upon had already been presented by the same group two years earlier (2012).[86]

A caveat of relying on combinatorial barcodes is that that the techniques become sensitive to errors, if just a single misread is introduced in the imaging process, the identity of the target immediately becomes ambiguous. This can be partially remedied by adding redundant information and extra decoding rounds, but such actions come at the price of an increased number of probes and more hybridization rounds. Other *error correcting* schemes were introduced in techniques such as *single molecule hybridization chain reaction* (smHCR) and seqFISH+.[87, 88]

Efficient and a high level of error correction were key objectives in the method *multiplex error robust FISH*, or MERFISH for short, launched in 2016. Two different kinds of probes are used in MERFISH, *signal bearing* and *non-signal bearing*. First the non-signal bearing probes – with non-binding flanking regions – are bound to the target sequence. Next, in multiple rounds, signal bearing probes will bind to, and be released from, the flanking regions. Signal bearing probes will not bind to the flanking regions of each target in every round, meaning a binary response of non-bound and bound can be registered. From the set of binary patterns (barcodes) that can be generated (the cardinality is 2^N, where N is the number of rounds) a subset is selected with great care. The selected barcodes must all have a large discrepancy (Hamming distance ≥ 4), with the result that even if a round fails, a transcript's identity can still be determined.[52] Depending on the number of fluorophores used and the length of the barcodes, up to 10000 different target transcripts can be assessed.[89] The MERFISH method is currently available as a commercial platform, MERSCOPE™, distributed by the company Vizgen.

Other techniques belonging to this category are: split-FISH, osmFISH, EASI-FISH, RNAscope, and DNA Microscopy.[90, 91, 92, 93, 94] Since the targets tend to be single molecules, these methods are in general said to be operating at a *subcellular* resolution. With the help of computational means, transcripts are usually assigned to the cells present in the tissue, to create *spatial* single cell expression profiles; however, this is a non-trivial task that introduces an extra layer of effort into the protocol.

2.3.2.3 *in situ* sequencing techniques

This category could almost be considered a chimera of sequencing and probe-based techniques; in *in situ* sequencing (ISS), the transcripts' identities are determined by sequencing, but without removing them from the tissue. While the *in situ* aspect gives subcellular resolution, one of the main obstacles for ISS techniques is the limited space within the cell. The first example of ISS was presented in 2013 by Ke et al., the process begins with a conversion of mRNA to cDNA, then *padlock probes* are used to determine the cDNAs' identities.[95] These probes can, somewhat simplified, be described as circular sequences

being cut open, allowing the circle to unfold into a linear segment. The regions flanking the "cut", the ends in the linear sequence, are designed to bind to a specific region of their target. In the 2013 publication, two different strategies were proposed: (i) the probes would bind and form a gap between the flanking regions, which was to be filled in with nucleotides by a DNA polymerase followed by a ligation step; (ii) there would be no gap between the two binding flanks, just the cut, which could be closed by a single ligation step. In either strategy, the aim is to close the loop, making the linear probes circular. Once closed, to increase the signal strength, the padlock probes' sequences are amplified using rolling circle amplification (RCA). Finally, a procedure known as sequencing by ligation is applied to determine either the gap-filled sequence or a barcode sequence of the padlock probe. The ligation-only strategy has a higher sensitivy, but gap-filling allows for detection of transcript variants like isoforms or single-nucleotide variants (SNVs). In 2017 this particular ISS technique became commercialized by the company Cartana, which was later acquired by 10x Genomics in 2020. Following the acquisition, 10x Genomics later announced that the ISS technique would be launched in 2022 as the Xenium platform.

Other methods build upon similar principles as the 2013 method, both *barcode in situ targeted sequencing* (BaristaSeq) and *spatially resolved transcript amplicon readout mapping* (STARmap) use similar padlock probes but also introduce hydrogels or crosslinking with the cellular matrix to improve sequencing performance.[96, 97] All of the aforementioned ISS techniques are targeted in their character, i.e., the targets must be known prior to the experiment, but this is not an inherent feature of ISS techniques. Lee et al. presented *fluorescent in situ RNA sequencing* (FISSEQ), which uses random primers and matrix crosslinking to sequence transcripts from several thousand different genes.[98] The FISSEQ technology was commercialized in 2016 by ReadCoor Inc. and later acquired by 10x Genomics about the same time as Cartana.

2.3.2.4 *in silico* spatial reconstruction techniques

In most texts contrasting spatial transcriptomics and single cell sequencing methods, the reader can expect to encounter some version of the statement *"while single cell sequencing is informative, all spatial information is lost"*. In essence, this statement is correct, the dissociation step during sample preparation will indeed remove any immediate spatial information from the data. In contrast to, for example, the ISS techniques – where each observed signal is associated with a set of spatial coordinates – single cell (or nuclei) data will carry no such intelligence. Interestingly, the single cell data still possesses some inherent spatial information, encoded in the data itself. If properly distilled, this information can be utilized to *reconstruct* the spatial structure of a dissociated tissue; at least to some extent. The spatial reconstruction techniques aim to do exactly this, i.e., they collect non-spatial data and then use different computational algorithms to infer its original structure.

Already in 2015, two methods for spatial reconstruction were proposed by Achim et al. and Satija et al., both of them using reference maps of the biological system of interest.[99, 100] The latter (Satija et al.) was released in conjunction with the first version (v.1.0) of the now popular analysis framework *Seurat*. In the simpler version of the Seurat approach, a total of 47 landmark genes with distinct spatial expression are used to assign each cell to one of several bins in the spatial domain. For each of the 47 genes, a spatial bin is classified as "on" (expressed) or "off" (not expressed) based on ISH reference data of the

same structure. The on and off expression values are then inferred from single cell data, by using a bimodal mixture model. The expression of the landmark genes are modeled using a multivariate Gaussian distribution (with complete independence between the genes, i.e., a diagonal covariance matrix). The means and diagonal elements of the covariance matrix (the variances) are taken from the bimodal model. Next, for every cell, the posterior probability of it originating from a given bin is calculated and a probabilistic map across the bins is obtained. A cell can also be mapped to a specific bin, based on the spatial centroid of the probability map. To improve the performance of their model, they used a slightly more sophisticated model which relaxed the independence assumption, thus allowing landmark genes to covary.

Other methods subscribe to the idea of using a similarity measure between spatial transcriptomics and single cell data. For example, Peng et al. suggest that single cell data can be projected to a spatial domain using spatially variable genes and Spearman rank correlation.[101] More recently, deep learning methods have been proposed to solve the reconstruction task.[102]

In 2019, Nitzan et al. presented the method novoSpaRc for *de novo* spatial reconstruction, compatible with data where no reference is available, but also with the ability to leverage such information if accessible.[103] The key assumption in novoSpaRc is that cells neighboring in gene expression space likely reside at nearby positions in the physical space. Thus, the authors formulate an optimization task, cast in the form of an optimal transport problem, where a transition matrix that probabilisticly maps cells to spatial locations is the target variable. The transition matrix is optimized to minimize the discrepancy between distances in gene expression space and physical space between cells. If a reference is available, it may be incorporated into the inference process; the influence of the reference versus the observed data can be weighted according to the user's preference. The objective function also includes an entropy-based regularization term which favors less deterministic mapping of cells. The authors showcased their method by reconstructing several different tissues from both non-spatial (scRNA-seq) data and spatial data (Slide-seq) made non-spatial. By using originally spatial data, the reconstructions could be compared with the true spatial structure, thus being ideal for evaluation of the method.

Shortly after novoSpaRc was published, another method using an optimal transport problem was presented as SpaOTsc. Rather than using an entropy penalty term, SpaOTsc permits unbalanced transport.[104] Another method, CSOmap, uses ligand-receptor interactions and t-stochastic neighborhood embedding (t-SNE) to reconstruct the tissue.[105]

2.3.2.5 *in situ* capture techniques

One of the more appealing aspects of scRNA-seq is how it operates at a near full-transcriptome level and provides a fairly unbiased (compared to targeted methods) image of a cell's transcriptome. The *in situ* capture techniques attempts to leverage the same benefits of NGS by capturing transcripts *in situ* but sequencing them *ex situ*. To succeed, spatial information needs to be coupled with the transcript identities before they are removed from their capture location. Several research groups have approached the capture and coupling tasks from different angles, and the combination of strategies they prefer is mainly what sets the methods apart.

The *in situ* capture techniques entered the scene in 2016 when Ståhl et al. via their Science publication presented a method named *"Spatial Transcriptomics"*; to avoid confusion with the field itself, this will be exclusively referred to as ST.[106] The ST method uses a solid glass surface onto which oligonucleotide capture probes are printed in clusters (*spots*) arranged in an equidistant regular grid with circa 1000 nodes. The spots are positioned with a center-to-center distance of $200\mu m$ and have a diameter of $100\mu m$. Approximately $2.0 \cdot 10^8$ probes are found in each spot, all sharing the same *spatial barcode*. The purpose of the spatial barcode is to couple each probe with its physical position in the array. In addition to the spatial barcode, the probes contain: a cleavage site, amplification and sequencing handles, a UMI, and an mRNA capture region consisting of a polyT-sequence (complementary to the polyA-tail of eukaryotic mRNA). The tissue specimen to be examined is cryosectioned to thin slices (often $10\mu m$) and placed upon the ST array, to then be fixated and stained, followed by a brightfield imaging step of the whole tissue slice. Next, the tissue slice is permeabilized to release transcripts from the cell, which diffuse toward the glass surface and bind to the polyT-region of the capture probes. Once the transcripts are captured, reverse transcription is initiated. The reverse transcriptase uses the captured transcript as a template to extend the probe sequence. Finally, the probes are removed from the surface, using the cleavage site, and sequenced (after library preparation) with NGS. With the help of bioinformatic tools, the identity and capture location of every cDNA molecule is registered.

The first version of ST is commonly referred to as ST1K, distinguishing it from the updated ST2K version.[107] The difference between the two is how a denser sampling was implemented in the latter by hosting about 2000 spots, rather than 1000, in the same array area. ST was commercialized by the company Spatial Transcriptomics AB, from which 10x Genomics obtained the IP rights in late 2018. A year after the acquisition, in late 2019, 10x Genomics began shipping the product Visium. The Visium platform is a second upgrade of the ST technique, using approximately 5000 spots with a diameter of $55\mu m$ and a center-to-center distance of $100\mu m$ arranged according to the orange crate packing system.[108] With 10x Genomics being an established distributor of products in the genomics field, Visium spread fast and became one of the most widely used platforms for spatial transcriptomics studies in 2021.[74] None of the ST or Visium platforms have, so far, reached single cell resolution. As a consequence, transcripts captured at a given spot could originate from multiple cells, not all necessarily of the same cell type or state. Depending on the tissue and platform, the number of cells contributing with material to a spot could range from one to somewhere in the hundreds.[109] Furthermore, capturing transcripts by their polyA-sequence makes the technique fairly unbiased, as no prior target selection is required. Of course, this comes at the cost of only capturing polyadenylated transcripts, excluding specimens such as tRNA, rRNA, miRNA (micro RNA), snRNA (small nuclear RNA), piwi-interacting RNA (piRNA), etc. By only sequencing a small fraction of the captured mRNA, queries into isoform populations are very limited, although attempts to circumvent this have been made.[110]

Increased spatial resolution fast became the feature to optimize among developers of *in situ* capture methods. In 2019, two techniques with improved resolution compared to the ST1K and ST2K arrays were published close in time, *Slide-seqV1* and *high density spatial transcriptomics* (HDST).[111, 112] Both techniques used beads rather than printed probes, allowing them to achieve higher resolution, but also requiring them to do a *decod-*

ing step to determine the spatial position of each bead as this is not known beforehand. In Slide-seqV1 SOLiD sequencing is used to couple spatial barcodes with position, while HDST relies on FISH for decoding. The beads in Slide-seqV1 are $10\mu m$ in diameter, five times the size of the beads in HDST ($2\mu m$). A common consequence of increased resolution is reduced capture efficiency, the implication being that fewer transcripts per capture location are registered. Both Slide-seqV1 and HDST suffer from efficiency issues, although HDST reports worse performance than Slide-seqV1. In an attempt to improve the capture efficiency, Slide-seqV2 was launched in late 2020, presenting a platform with a more than 9x efficiency improvement from the first version and a performance that matched droplet-based single cell techniques, as well as outperforming Visium.[113]

DBiT-seq (2020), Seq-Scope (2021), PIXEL-seq (2021), Stereo-seq (2021) are examples of other techniques that also belong to the set of *in situ* capture methods, however, these will not be discussed in more detail here.[114, 115, 116, 117] Of all the listed methods in this category, Stereo-seq holds the most impressive resolution (spot diameter) of $220nm$, with an efficiency that – according to the authors' benchmark – is competitive with Visium's.

Working in a group so intertwined with the inception of ST and development of the *in situ* capture technologies, certain influence on my work is inevitable. Hence, many of the computational methods presented in this thesis have been developed for data produced by the ST or Visium platform, although frequently generalized to other techniques. With Visium's widespread use and the increasing amount of data generated from the platform, this preference is not completely unwarranted, but I still want to make the reader aware of it.[74]

2.3.3 Computational methods

This section serves to *introduce* the reader to some of the computational methods used for analysis of spatial transcriptomics data. Thus, I've intentionally chosen to keep the discussion at a superficial non-technical level. Instead, section 2.4 provides a more extensive account of concepts that relate to data modeling and analysis relevant to the work presented in this thesis.

2.3.3.1 Data character and content

All spatial transcriptomics techniques partly rely on computational methods to translate raw experimental data into a mature format that is informative, interpretable, and suitable for downstream analyses; one may even venture as far as to say that these methods are parts of the techniques themselves. The exact procedures for curation of data are dictated by the experimental platform, but tend to encompass one or more of the following actions: cell segmentation, assignment of transcripts to a donor cell, barcode demultiplexing, and mapping to a reference genome. This thesis focuses on analysis rather than processing of spatial transcriptomics data, hence, methods for pre-processing will not be discussed in any further detail here.

After the appropriate pre-processing steps have been applied to raw data, the product is usually an object that associates observed features of interest with spatial coordinates.

In capture-based techniques and those relying on microdissection (e.g., Visium, Slide-seq, HDST, and LCM), these features are gene expression vectors listing the number of unique transcripts from each gene observed at a given capture location or region; the expression vectors generally span over the whole transcriptome as *a priori* selection of targets is not required. For techniques based on *in situ* sequencing or hybridization (e.g., ISS, MERFISH, seqFISH, and FISSEQ), similar expression vectors can be assembled for each cell (though limited to the set of targets studied), alternatively one may only report the identity and spatial position of the target molecules without assigning them to a donor cell. In contrast to the latter set of techniques, the expression vectors of the sequencing-based methods can all be considered as *mixtures* with contributions from multiple cells, potentially heterogeneous w.r.t. cell types, a property with strong implications for the downstream analysis.

2.3.3.2 Drawing inspiration from the single cell sphere

Some statements regarding expression patterns can often be postulated by mere visual inspection of the curated data. However, more sophisticated methods to capture general trends and mine latent information from the data are frequently applied. Many of these methods are immediately borrowed from the single cell field, for example: clustering of datapoints (e.g., capture locations or cells), differential expression analysis between regions of interest or identified clusters, factor analysis, etc.[118] Consequently, several of the bioinformatic suites originally developed for single cell data analysis (e.g., Seurat and scanpy) have expanded their ecosystems to support handling of spatial data,[119, 120] but new tools specific to this particular kind of data are also available (e.g., STUtility, Giotto, stLearn and squidpy).[121, 122, 123, 124]

For the mixed data (e.g., Visium and Slide-seq), different schemes to decompose this into entities like factors or expression programs have been proposed; some cast this as a standard matrix factorization problem, others use probabilistic models with the aim to better account for different sources of variability in the data.[121, 125, 126] Attempts to delineate cell state dynamics using trajectory inference or velocities are staple elements in single cell studies, and application of these ideas to spatial data is alluring, but not without certain challenges. Adhering to the definition of RNA velocities presented by La Manno et al. (velocyto), information required to fit the dynamical model – abundance of spliced and unspliced reads – is hard to extract from spatial techniques, unless experiments are specifically designed for it.[127] To circumvent this issue when working with MERFISH data Xia et al. proposed a slightly modified model.[128] To order cells in pseudotime, they made a distinction between transcripts located in the nucleus and cytoplasm, instead of splice variants. Splicing information may be more accessible in the sequencing-based methods if TGS long-read sequencing is used, but even then, the characteristic mixed observations adds a layer of complexity to the velocity estimation; cells contributing to the same capture location do not necessarily populate similar positions in pseudotime.[110]

2.3.3.3 Taking context into consideration with spatially aware methods

Despite their broad applicability, single cell analysis methods discard information contained within the data, for example omitting notions of spatial coherence and correlation. Inevitably, development of *"spatially aware"* methods thus followed the surge in spatial data. One question that has been revisited at multiple occasions is how spatially variable genes can be identified. The strategy of Edsgard et al. (trendsceek) was to treat the expression data as a marked point process, then compute certain spatial summary statistics, and finally compare these statistics with those obtained from a set of null-distributions generated by "spatial shuffling" of expression values; this allowed genes with non-random spatial arrangement to be extracted.[129] Svensson et al. took a slightly different approach to find spatially variable genes with SpatialDE, using *Gaussian Processes* (GP) to model the (normalized) data.[130] For every gene, SpatialDE fits a full GP model with both a spatial and non-spatial term to explain the variance. The full model is then compared to a reduced model lacking the spatial term. Genes where the full model significantly outperforms the reduced one – accounting for the additional parameter – are considered spatially variable. More recently, the method SPARK, which uses a generalized linear spatial model (GLSM), was introduced by Sun et al.[131] In SPARK, observed expression values are taken as Poisson distributed, with the Poisson rate dependent on: certain explanatory variables (e.g., batch or replicate), random residual noise, and spatial correlation between signal locations. Here, the third term (spatial correlation) is described by a stationary GP. Analogously to SpatialDE, spatially variable genes are found by testing whether the term accounting for spatial correlations adds significant explanatory power to the model. In 2021 SpatialDE2 was released, which extends the SpatialDE model to a more general form. SpatialDE2 also supports GPU acceleration and spatial domain segmentation using HMRFs (Hidden Markov Random Fields).[132] An updated version of SPARK, named SPARK-X, was also released in 2021. SPARK-X is designed for computational efficiency and scaling to large spatial data sets, objectives achieved by implementing a non-parametric model.[133]

Arnol et al. also adopted the idea of using GPs to model spatial data in their method SVCA, this time extended to account for multiple types of effects (intrinsic, environmental, and cell-cell interactions) that might influence the gene expression.[134] From SVCA's design, different degrees of contribution to the observed variance in gene expression are attributed to the aforementioned effects. Both Walter et al. (FISHFactor) and Townes et al. (NSF) leverage the flexibility of GPs to achieve a form of *spatial factorization*. FISHFactor is tailored for single molecule data (e.g., from ISH or ISS techniques) while NSF is better suited for data produced using *in situ* capture methods.[135, 136] Ghazanfar et al. also explored means of examining spatial patterns and presented one solution with their method scHOT, although the gene-gene interplay rather than singular patterns was their focus.[137] In scHOT, weighted estimates of gene variance and gene-pair correlations are used to infer local interaction patterns within the spatial data, exposing genes that are differentially correlated across space.

Spatial expression data can easily be described using concepts borrowed from the field of graph theory; with every source of a signal (e.g., capture location or cell) interpreted as a node in a graph and edges connecting nodes to their neighbors (defined by physical distance). This interpretation was conceptualized by Zhu et al. who used HMRFs to extract spatial domains from ISS data.[138] Domain membership of each cell was treated as

a latent (hidden) variable to be inferred under the assumptions that: (i) members of the same domain should have similar gene expression profiles, and (ii) neighboring cells likely belong to the same domain. HMRFs have also been applied to infer spatial organization of tumor clones – defined by a set of CNAs (copy number aberrations) – in ST data. The referenced method, STARCH, assumes that nearby capture-locations are inclined to share genetic profiles, and takes advantage of the presumed spatial correlations to strengthen CNA-signals.[139]

2.3.3.4 Combining spatial and single cell-or nuclei data

Plenty of attention has been given to the area of data integration, perhaps most intensively how information from single cell/nuclei RNA-seq studies could be transferred to data derived from spatial experiments. Thus, single cell/nuclei data – where no mixing of cells occurs – has repeatedly been used to deconvolve the expression profiles found in ST, Visium, and Slide-seq; producing estimates of cell type abundance at the capture locations. Examples of some deconvolution strategies together with their release dates are:

- NMFReg – March 2019, a method that decomposes the single cell data into signatures by non-negative matrix factorization (NMF), then assigns a signature to each cell type, and finally uses a non-negative least-squares approach to estimate the loadings (contributions) of respective signature to the observed expression data at the capture locations.[111]
- *stereoscope* – December 2019 (preprint), a probabilistic method presented in **Article I** which assumes that both single cell/nuclei and spatial data follows a negative binomial distribution. Importantly, the expression of a gene within a cell is conditioned on the cell's identity (cell type). In the spatial data, for every capture location, *stereoscope* tries to find the combination (proportions) of cell types that – based on the cell type specific parameters estimated from the single cell data – best explains the observed expression values. This is evaluated by computing the data likelihood given the compound distribution generated by combining cell type specific distributions according to the proportions. *stereoscope* is implemented in PyTorch, which supports GPU acceleration. The model does not use the mean-dispersion parameterization of the negative binomial, but the "rate and success probability" parametrization.
- MIA (Multimodal Intersection Analysis) – January 2020, where representative cell type gene sets derived from single cell data (e.g., marker genes) are compared to the gene expression in the spatial observations. The statistical significance (based on a hypergeometric test) of the overlap is then used to assess enrichment and depletion of types within the specific region or observation.[140]
- cell2location – November 2020 (preprint), akin to *stereoscope*, models both single cell/nuclei and spatial data as negative binomial distributed, but uses a Bayesian framework to infer latent parameters that can be related to cell type abundance.[141] In contrast to *stereoscope* and RCTD (see below), which employ penalized maximum likelihood approach for parameter inference, cell2location relies on variational inference, which reportedly makes it superior w.r.t. computational time. In the implementation, the more common mean-dispersion parametrization is used for the

negative binomial distribution. cell2location also supports two ways of estimating the single cell/nuclei signatures used to guide the decomposition of spatial data into contributions from cell types: (i) one highly efficient strategy where the mean for each cell type is computed immediately from the observations, and (ii) a more computationally demanding approach where regularized negative binomial regression is used to account for batch differences.

- RCTD – February 2021, is a probabilistic method developed for Slide-seq data but generalized to all *in situ* capture platforms with mixed signals.[142] In contrast to *stereoscope* and cell2location, RCTD models the UMI count data using a hierarchical Poisson-lognormal mixture. The choice of a Poisson-lognormal mixture model for UMI count data is not as common as the Gamma-Poisson mixture (effectively producing a negative binomial model), but definitely not unheard of, see section 2.4.2. RCTD extensively models, and tries to account for, platform effects that otherwise could mask relevant biological signals or confound the result. In addition, it supports different "inference modes" relating to the expected number of cell types that are present at each capture location. Having decomposed the spatial data, RCTD – by conditioning on cell type – provides means to identify spatially variable genes where the variation is not driven by the spatial cell type distribution.

- Tangram – October 2021, does not assume that the gene expression follows a specific distribution, but rather aims to find an optimal mapping between cells from single cell/nuclei data to the spatial observations, though still using a probabilistic strategy.[143] Tangram attempts to simultaneously maximize agreement between cell densities and the expression profiles, comparing the observed spatial data to the mapped single cell/nuclei data. To measure agreement they use Kullback-Leibler divergence (KLD) between the densities and cosine distance for the expression profiles. If cell numbers are unknown, the KLD term can be excluded. An optional entropy regularizer can be utilized to make the cell assignments more localized. Data-driven cell filtering is also included as an optional feature of the model. The filtering only keeps the "best" cells, where fitness is learnt from the data. Tangram is platform-agnostic and compatible with a majority of the spatial transcriptomics techniques. For those spatial techniques where only a few targets are surveyed, Tangram has a module for imputation of gene expression.

Projection of single cell/nuclei annotation labels to spatial data are relevant even if deconvolution is not necessary. This is exemplified by Qian et al. who, in conjunction with presenting a probabilistic method for the task (pciSeq), showed how it enabled finer cell type calling in mouse brain ISS data.[144] pciSeq treats the data as realizations of a Gamma-Poisson mixture (i.e., a negative binomial distribution). Probabilities of each cell belonging to a specific cell type are obtained from pciSeq after approximating the posterior over cell types. The Seurat suite also offers single cell data integration by embedding the two data types in a joint space where "anchors" (similar objects from respective type) are identified and used to transfer labels.[145]

2.3.3.5 Methods with influences from deep learning

Other modalities than sequencing data can of course be integrated with spatial data. For example, He et al. illustrated how morphological cues contained in images of a tissue can

be leveraged to predict gene expression in ST data once a mapping (here, a neural network) between the two is established.[146] With their deep generative model Bergenstråhle et al. further explored how tissue images enable computational enhancement of the gene expression resolution, referred to as inferred *super-resolution*. In the model, image and gene expression are considered as generated by a latent tissue state. The posterior over the latent state and a set of other hidden variables, given the image and expression, is approximated by variational inference where the variational parameters of the latent state are encoded by a convolutional neural network (a.k.a. recognition network) that takes the image as an input.[147] The network architecture has some resemblance with U-Net,[148] allowing information to be shared across different levels of granularity. After training the model, expression levels at all pixels in the image can be inferred, thus, increasing the resolution significantly. This model also uses a negative binomial distribution to describe the expression data, while the image intensities are assumed to be sampled from a Gaussian distribution.

More methods relying on deep learning have started to emerge in the later years, likely due to a combination of increased access to data and a growing interest in the field. For example, SpaGCN uses a graph convolutional network (GCN) to integrate gene expression, histology, and spatial information to identify spatial domains within the samples. The framework CoSTA relies on the more traditional convolutional networks to find families of genes with similar spatial patterns.[149] Most likely, integration of spatial data with one or more modalities (e.g., protein, genome, metabolome or epigenome data) will emerge as a topic of intensive research. Given its success in other fields and versatility, deep learning will likely become an indispensable tool for advancements in the domain of computational biology. For my personal thoughts on the topics of multimodal integration and deep learning, I refer the reader to the Epilogue (section: “What I predict”).

2.4 Modeling gene expression - a mathematical perspective

Despite us being surrounded by the same systems, the language we use to describe them tends to vary profoundly. A medical doctor might use schematic images to outline the intricate anatomy of the heart while an engineer finds it more appealing to describe the flow of blood using Navier–Stokes equation. Neither approach is wrong, they each serve a particular purposes and highlights aspects of the system most relevant to their respective user. When modeling gene expression or transcriptomics data, my language of preference has been statistics. Biological systems are inherently noisy and uncertain, which makes the non-deterministic nature of statistics a perfect fit for them. This section will introduce some basic concepts relating to modeling of gene expression and spatial data.

2.4.1 The basics - setting the scene

Sequencing-based data, like that originating from scRNA-seq or *in situ* capture spatial transcriptomics methods, is usually presented as *count* data. The name stems from the fact that we count the number of observations from each feature. Since we can’t observe “a fraction of an observation”, the count data always consist of non-negative integer values. The canonical way to represent count data is by tabulation, or through a matrix, listing observations along one dimension and features along the other. While *raw*

count data exclusively contains non-negative integer values, it's commonplace to – when working with gene expression data – *transform* these values (e.g., to reduce variance or account for certain biases), resulting in values that can span the whole real domain.[150] Several distributions could theoretically be used to model gene expression data, perhaps the most natural options being: the binomial distribution, the multinomial distribution, the Poisson distribution, the negative binomial distribution, the normal distribution, and non-parametric distributions. Below, each of these will be briefly described. After the distributions have been introduced, a discussion regarding their aptitude for modeling gene expression will follow.

2.4.1.1 The binomial distribution

The binomial distribution is a discrete univariate distribution with two parameters, the success probability (p) and the number of independent experiments (n). If a stochasitc variable X is distributed according to a binomial distribution, one may consider X as representing the number of successes in n independent trials, where the probability of success in each trial is equal to p. Formally we have:

$$X \sim Bin(n,p) \rightarrow P(X=x) = \binom{n}{x} p^x (1-p)^{(n-x)} \tag{2.1}$$

For the expected value and variance, the following relationships hold true:

$$\mathbb{E}[X] = np, \quad Var[X] = np(1-p) \tag{2.2}$$

For large n, the binomial can be approximated with a normal distribution (see section 2.4.1.5) according to:

$$Bin(n,p) \overset{n \mapsto \infty}{\sim} \mathcal{N}(np, np(1-p)) \tag{2.3}$$

If n is large while p is small, the binomial can be approximated with a Poisson distribution (see section 2.4.1.3) according to:

$$Bin(n,p) \overset{n \mapsto \infty, p \mapsto 0}{\sim} Poi(np) \tag{2.4}$$

2.4.1.2 The multinomial distribution

The multinomial distribution is the multivariate version of the binomial distribution. While the binomial distribution models repeated trials with a binary outcome, the trials in the multinomial can take any of k different outcomes. The multinomial is given as:

$$\boldsymbol{X} \sim Mul(n, \boldsymbol{p}) \rightarrow P(\boldsymbol{X} = \{x_1, \cdots, x_k\}) = \frac{n!}{x_1! \cdots x_k!} \prod_i p_i^{x_i}, \quad \sum_i p_i = 1 \tag{2.5}$$

The expected value and variance for an outcome i over all n trials are similar to those of the binomial distribution. The covariance between outcomes is always negative. The following identities hold true for the multinomial:

$$\mathbb{E}[X_i] = np_i, \quad Var[X_i] = np_i(1-p_i), \quad Cov[X_i, X_j] = -np_i p_j \tag{2.6}$$

2.4.1.3 The Poisson distribution

The Poisson distribution, just like the binomial, is a univariate discrete distribution. The Poisson has a single parameter, the *rate* or *mean* parameter λ. The distribution is designed to describe the number of independent events within a set unit of time or space. For a Poisson distributed variable X, the following holds true:

$$X \sim Poi(\lambda) \rightarrow P(X = x) = \frac{\lambda^x e^{-\lambda}}{x!} \tag{2.7}$$

The mean and variance are given as:

$$\mathbb{E}[X] = \lambda, \quad Var[X] = \lambda \tag{2.8}$$

As is evident from the above expression, there's a clear dependency between the mean and the variance: a one unit increase in the mean will result in a one unit increase in the variance, i.e., they are equal.

Superficially, the Poisson and binomial distributions share many features, but there are some key differences. The binomial models discrete occurances over a discrete set of trials (domain), while the domain is continuous for the Poisson. There's a limited number of attempts (n) where a success could occur in the binomial, each with a probability p. In contrast, the Poisson permits an infinite number of attempts over an interval of time or given space, but all with a minuscule chance of success – as manifested in the binomial to Poisson approximation.

2.4.1.4 The negative binomial distribution

The negative binomial is yet another univariate discrete distribution, that relates to the binomial, but with some important alterations. The reader will notice how the negative binomial is discussed in greater detail than many of the other distributions, this is because it is one of the most common choices of distributions when attempting to model count data; this preference will be explained in later sections.

The negative binomial, rather than modeling the number of successes given n trials and a success probability p, models the number of failures before the r:th success. Other interpretations exist as well, for example where one measures the number of successes before a given number of failures. Using the first alternative, the negative binomial can be described as:

$$X \sim \text{NegBin}(p, r) \rightarrow P(X = x) = \binom{r + x - 1}{x}(1 - p)^x p^r \tag{2.9}$$

Where x represents the number of failures. With this parametrization, the expected value and variance become:

$$\mathbb{E}[X] = \frac{pr}{1 - p}, \quad Var[X] = \frac{pr}{(1 - p)^2} \tag{2.10}$$

Using the gamma function (Γ) the negative binomial can be extended to support all non-negative real values of r, the expression being:

$$P(X = x) = \frac{\Gamma(x + r)}{\Gamma(x + 1)\Gamma(r)}(1 - p)^x p^r \tag{2.11}$$

Popular suites for probablistic modeling like PyTorch and Tensorflow tend to replace the success probability p with an *odds* parameter (o). When logged, as is common during optimization, the odds parameter is mapped to the whole real domain, and no constraints have to be imposed during optimization. Using the odds parameter, the expression becomes:

$$P(X = x) = \frac{\Gamma(x + r)}{\Gamma(x + 1)\Gamma(r)} \frac{o^x}{(o + 1)^{r+x}}, \quad o = \frac{p}{(1 - p)} \tag{2.12}$$

In regression problems, another parametrization is usually preferred, using the *mean* (μ) and *dispersion* (ϕ), defined as below:

$$\mu = \frac{pr}{(1 - p)}, \quad \phi = r \tag{2.13}$$

With the resulting form:

$$P(X = x) = \frac{\Gamma(x + \phi)}{\Gamma(x + 1)\Gamma(\phi)} \Big(\frac{\mu}{\phi + \mu}\Big)^x \Big(\frac{\phi}{\phi + \mu}\Big)^\phi \tag{2.14}$$

Where the mean and variance can be described as:

$$\mathbb{E}[X] = \mu, \quad Var[X] = \mu + \frac{\mu^2}{\phi} \tag{2.15}$$

From these expressions of the mean and variance it becomes clear that the negative binomial can accommodate *overdispersion*, where the variance is larger than the mean. In this thesis, the notation NegBin2(μ, ϕ) is used to indicate usage of the mean and dispersion parametrization.

Interestingly, the negative binomial relates to the Poisson distribution, as the latter represents a limiting case of the former when $\phi \mapsto \infty$. Hence, the negative binomial can be considered a generalization of the Poisson, where ϕ controls the amount of additional spread compared to a Poisson distribution with the same mean. A brief proof of this statement for the case $\phi \in \mathbb{N}$ can be seen in Appendix A.1. However, there is an even deeper relation between the two, the negative binomial can be seen as an hierarchical Gamma-Poission mixture, where the rate parameter in the Poisson is distributed according to a Gamma distribution. That is, if:

$$\begin{aligned} X &\sim Poi(\lambda) \\ \lambda &\sim Gamma(\phi, \phi/\mu) \end{aligned} \tag{2.16}$$

Then, the marginal distribution of X is:

$$X \sim \text{NegBin2}(\mu, \phi) \tag{2.17}$$

For a proof of this relationship, see Appendix A.2.

2.4.1.5 Uni and multivariate Gaussian

The univariate Gaussian differs from the previously discussed distributions, as it models continuous variables in the whole real domain. The univariate Gaussian is not suitable for modeling of raw counts, but could be a candidate distribution if the counts have been normalized. If a variable X follows a univariate Gaussian distribution, then:

$$X \sim \mathcal{N}(\mu, \sigma) \rightarrow P(X = x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp(-\frac{1}{2}\Big[\frac{x-\mu}{\sigma}\Big]^2) \tag{2.18}$$

With the expected value and variance:

$$\mathbb{E}[X] = \mu, \quad Var[X] = \sigma^2 \tag{2.19}$$

The multivariate case has a similar form, but uses vectors rather than scalars:

$$\boldsymbol{X} \sim \text{MVN}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \rightarrow P(\boldsymbol{X} = \boldsymbol{x}) = \frac{1}{(2\pi|\boldsymbol{\Sigma}|)^{1/2}} \exp(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})) \tag{2.20}$$

With the expected value and variance for an outcome i and covariance between outcome i and j being:

$$\mathbb{E}[X_i] = \mu_i, \quad Var[X_i] = \sigma_i^2, \quad Cov[X_i, X_j] = \sigma_i\sigma_j \tag{2.21}$$

2.4.1.6 Non-parametric distributions

With non-parametric distributions there are no explicit assumptions about their shape and character, instead they adapt to the data. Non-parametric distributions aim to find a shape of the distribution that matches the observed data without introducing too much volatility, and they can be constructed in several ways of varying complexity. To illustrate the concept, one of the most common methods for estimation of probability density functions is described below, more specifically *kernel density estimation* (KDE). For simplicity, we'll study the univariate case.

In KDE, the observed data is assumed to be independent and identically distributed. To estimate the shape of the distribution, the following estimator is used:

$$\hat{f}(x) = \frac{1}{n}\sum_{i}^{n} K(x, x_i) \tag{2.22}$$

Where $K(.,.)$ is a *kernel*. There exist several kernel functions, but a popular choice is the Gaussian kernel (K_{Gauss}):

$$K_{Gauss}(x, x') = \frac{1}{\sqrt{2\pi\sigma^2}} \exp(-\frac{1}{2}\Big[\frac{x-x'}{\sigma}\Big]^2) \tag{2.23}$$

2.4.2 Model construction

Here, the process of measuring transcripts within a given cell or at a certain location using sequencing-based techniques will be examined from a statistical perspective. Note, in this discussion we assume that the observations consist of UMI counts and not raw reads.

Here, I've opted to use the general term "source" to describe the origin of the transcripts; however, this is interchangeable with for example: a cell, a Visium spot, or a Slide-seq bead.

At the source, there will be a large pool of transcripts of varying identity (i.e., transcribed from different genes), this represents the source's total transcript load. Ideally we'd like to characterize all transcripts in the pool, but most techniques will only allow the source to be represented by a small fraction of the larger population. For example, it's estimated that a single mammalian cell holds circa $2 \cdot 10^5$ transcripts, while the number of UMI counts used to represent cells when using scRNA-seq protocols tends to be in the range of $1 \cdot 10^3 - 1 \cdot 10^4$ UMI counts.[151, 60, 61] The lower value from the experimental methods can be explained by several factors, such as the efficacy of both the capture medium and downstream biochemical reactions (e.g., reverse transcription).[152]

Which transcripts in the pool that will be successfully captured and processed is determined by a random process. In the simplest of models, we can assume that the probability of a certain type of transcript being captured solely depends on its relative abundance. This model is "simple" in the sense that *"all transcripts are treated as equals,"*, meaning that the presence of potential bias in the measurement process is ignored. However, such bias is known to exist in real systems, for example, transcripts locate to different positions and organelles in the cell, which may impact their chance of successfully being measured. Also, certain transcripts are more sensitive to the processing compared to others, and thus have a lower chance of "surviving".[153] Although not exhibiting the highest degree of verisimilitude, the simple model offers valuable insights and will suffice to build an understanding of the underlying processes dictating the character of the observed gene expression. Though, the simple model will not be fully spared from critique in this thesis; after its introduction, there will follow a commentary on some of its flaws.

2.4.2.1 The simple model

In a distilled version of the measurement process, one may consider it as a case of "picking transcripts" from the large pool of total transcripts and placing them in a "UMI count bucket," which is used to represent the source. Importantly, this means that we do *not* replace the transcripts removed from the pool. This scenario is best described using the *multivariate hypergeometric distribution*, as it gives the probability of k successes in n trials – without replacement – where there are multiple different outcomes. Though, given how only a small fraction of the transcripts are sampled, one may consider the probability of selecting a certain type of transcript as unaffected by removal of the corresponding transcript from the pool. With fixed success probabilities, one may treat the process as if replacement does occur, making the *multinomial distribution* a valid choice of distribution to describe the measurement process.

Let x_{ij} symbolize the total counts of transcript type j in source i, and $t_i = \sum_j x_{ij}$ the total transcript load in source i. Note, hereafter, the term *gene* will be used instead of "transcript type." For the UMI counts, let y_{ij} be the analog to x_{ij} and n_i to t_i. Using the multinomial model, we then have:

$$\boldsymbol{y}_i = \{y_{i1}, \ldots, y_{1M}\} \sim Mul(n_i, \boldsymbol{\pi}_i), \quad \boldsymbol{\pi}_i = \{\pi_{i1}, \ldots, \pi_{iM}\}, \quad \sum_j \pi_{ij} = 1 \tag{2.24}$$

Where $\pi_{ij} = x_{ij}/t_i$. Thus, π_{ij} represents the relative expression of gene j in source i. Note that the π_i values are unknown, i.e., they cannot be immediately determined from the UMI count data. Assuming that n_i is a random variable following a Poisson distribution with rate λ_i, the probability of observing a given expression vector $\boldsymbol{y}_i$ is given as:

$$\begin{aligned} P(\boldsymbol{y} = \boldsymbol{y}_i) &= Mul(\boldsymbol{y}_i|n_i, \boldsymbol{\pi}_i) Poi(n_i|\lambda_i) = \\ &= \frac{n_i!}{\prod_j y_{ij}!} \prod_j \pi_j^{y_{ij}} \cdot \frac{\lambda_i^{n_i} e^{-\lambda_i}}{n_i!} = \\ &= \Big(\prod_j \frac{(\lambda_i \pi_j)^{y_{ij}}}{y_{ij}!}\Big) e^{-\lambda_i} = \Big(\prod_j \frac{(\lambda_i \pi_j)^{y_{ij}}}{y_{ij}!}\Big) e^{-\lambda_i \cdot 1} = \Big(\prod_j \frac{(\lambda_i \pi_j)^{y_{ij}}}{y_{ij}!}\Big) e^{-\lambda_i \cdot \sum_j \pi_{ij}} = \\ &= \Big(\prod_j \frac{(\lambda_i \pi_j)^{y_{ij}}}{y_{ij}!}\Big) \Big(\prod_j e^{-\lambda_i \pi_{ij}}\Big) = \prod_j \frac{(\lambda_i \pi_j)^{y_{ij}}}{y_{ij}!} e^{-\lambda_i \pi_{ij}} = \\ &= \prod_j Poi(y_{ij}|\lambda_i \pi_{ij}) \end{aligned} \tag{2.25}$$

The final expression in Eq. 2.25 represents the joint probability of several independent Poisson distributions (one for each gene). The rate parameter for the j:th Poisson distribution being $\lambda_i \pi_{ij}$. With only one observation of the total UMI counts in a source (n_i), $\lambda_i = n_i$ becomes the natural choice.[142] Thus, rather than looking at the joint expression vector $\boldsymbol{y}_i$ for cell i, we can also operate with the individual observations y_{ij}. Hence, the measurement step – capturing the transcripts expressed in a source – can be modeled with a Poisson distribution. Still, this model only encompass parts of the process generating the observed UMI counts; the biological variability remains – so far – unaccounted for.

To elaborate and exemplify, we'll examine Figure 2.4, where the variance of a gene's expression – across a set of cells annotated as Memory B-cells – is plotted as a function of its mean. Upon studying the figure, it's evident how the variance is not linearly related to the mean. If the UMI counts were distributed according to the Poisson measurement model presented above, a one-to-one linear relationship would be expected between the mean and variance, but the data clearly shows signs of overdispersion.

The presence of this overdispersion can be attributed to biological variance; despite all cells belonging to the same cell type, they are individual entities. As a consequence, the product $n_i \pi_{ij}$ should also be considered a stochastic variable. A valid expression model can be obtained by assuming $\nu_{ij} = n_i \pi_{ij} \sim Gamma(a, b)$. This model assumes that $\boldsymbol{\pi}_i$ follows a Dirichlet distribution generated by sampling values from a Gamma distribution followed by normalization, see Appendix A.3 for a more elaborate explanation. A complete model of the UMI counts is formed by combining the expression and measurement model accordingly:

$$\begin{aligned} y_{ij}|n_i, \pi_{ij} &\sim Poi(n_i \pi_{ij}) = Poi(\nu_{ij}) \\ \nu_{ij} &\sim Gamma(\phi_{ij}, \phi_{ij}/\mu_{ij}) \end{aligned} \tag{2.26}$$

Which is a Gamma-Poisson mixture and thus equates the marginal distribution of y_{ij} to a negative binomial distribution with mean μ and dispersion ϕ, given as:

$$y_{ij} \sim \text{NegBin2}(\mu_{ij}, \phi_{ij}) \tag{2.27}$$

Hence, using a negative binomial distribution to model UMI count data, as many methods do,[155, 156, 150, 157] can be considered as somewhat theoretically justified. Revisiting Figure 2.4, it's clear how an assumed quadratic relationship (which the negative binomial induce) between the mean and variance outperforms a linear relationship associated with the Poisson. Having advocated for the Gamma-Poisson mixture model, it should be said that other combinations of measurement and expression models have also been

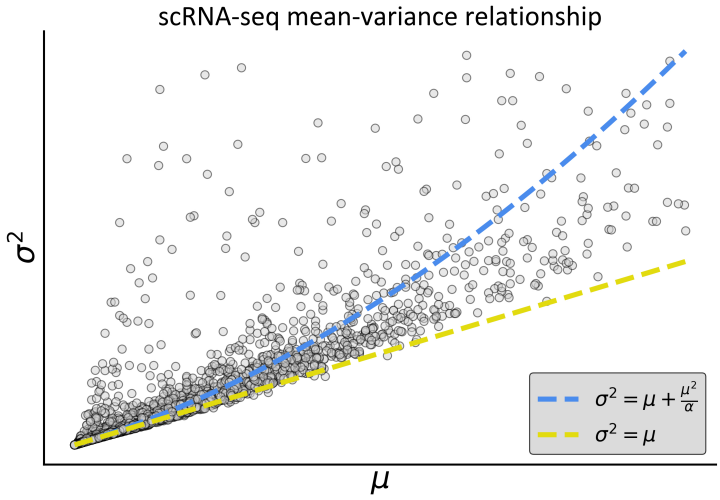

Figure 2.4: Mean-variance relationship in UMI count data. Each data point represents a gene, where the the position on the x and y-axis correspond to the mean (μ) respectively variance (σ^2) taken across all observations in the data set. The data set consists of all cells annotated as "Memory B-cells" in the publication by Wu et al.[154] Two curves representing different relationships between the mean and variance are visualized in the graph: a linear (yellow dashed line) and a quadratic (blue dashed line). The quadratic curve was fitted using least squares.

used; for example, replacing the Gamma distribution with a lognormal.[142, 158] Still, the Gamma-Poisson is an attractive option, as the integrals in the expression for the marginal distribution have a analytical solution represented by the negative binomial.

2.4.2.2 Flaws of the simple model

As mentioned in the previous section, the simple model assumes unbiased sampling of transcripts from the pool during measurement; a relatively unlikely scenario. If bias is present, and not able to be accounted for, the inferred values $\tilde{\pi}_{ij}$ does not estimate the true expressivity (π_{ij}) of a gene j within a source i, but rather its biased counterpart. That is, $\tilde{\pi}_{ij} \hat{=} b_{ij}\pi_{ij}$, where b_{ij} is a bias term.[153] If the bias is source independent and only depends on the gene ($b_{ij} = b_j$), differential gene expression analysis can still be conducted with the condition that ratios are used. Ratios are required because the bias terms cancel out, while they would remain if differences in absolute values were used instead:

$$\tilde{\pi}_{ij}/\tilde{\pi}_{i'j} = \pi_{ij}b_j/\pi_{i'j}b_j = \pi_{ij}/\pi_{i'j} \tag{2.28}$$

Complications arise when the bias is source dependent ($b_{ij} \neq b_{i'j}$), and requires a more complex model together with further assumptions. Using the simpler model despite bias being present will likely result in an overestimation of the cell-to-cell expression variation.[153]

Additionally, using the Gamma distribution for the expression model might be too restrictive. For example, Sarkar and Stephens showed that a non-parametric model was

favored over the Gamma model for 20-69% of genes across several data sets. Another issue arise if transcripts from certain genes are present in very low amounts, as this invalidates the assumption that the sampling of transcripts (without replacement) can be approximated by a process that assumes replacement (i.e., the multinomial).[153]

Despite the aforementioned flaws, the negative binomial model is still popular to use when modeling UMI count data. With a parametric model, interpretable parameters can be inferred and it's convenient to use in regression models or optimization frameworks.

2.4.2.3 Modeling transformed data

So far the discussion has pertained to modeling of *raw* UMI counts, but another approach is to *transform* the data prior to model construction. It's common practice to apply scaling (often relating to a source's library size) and log-transformation to UMI count data, as illustrated in Eq. 2.29.

$$y_{ij}^{\text{trans}} = \log(\frac{y_{ij}^{\text{obs}}}{\sum_k y_{ik}^{\text{obs}}} + c) \tag{2.29}$$

Where c represents a pseudocount to prevent zero values. Log-transformations are prevalent in statistical analysis because it tends to have a variance stabilizing effect and – sometimes – makes the data more normally distributed.[159] Both of these properties, homogeneous variance and normal distribution, are often requirements for statistical tests and models. However, concerns about log-transforming count data have been raised and even advised against.[160] In their method SpatialDE, Svensson et al. used *Anscombe's* transformation instead of the log-transform; which is better adapted for data following a negative binomial distribution.[130] The reason Svensson et al. does not immediately model the data as negative binomial is because they use a Gaussian Process to capture spatial relationships, where the observed data is assumed to follow a multivariate normal distribution. For more details on the Gaussian Process, see Methods of **Article V**.

2.4.2.4 A word on zero inflation

Single cell and spatial transcriptomics data is often referred to as being "sparse," which aims to convey the large abundance of zeros in the count matrices used to represent experimental data. Due to the stochastic nature of the sampling in the measurement process, a certain part of the target population will remain unobserved. Thus, zero-observations, or sparse data, are not inherently disruptive to any analysis. They only become problematic if present in *excess*, i.e., if they are more abundant than expected from the sampling process. Excessive zeros would imply the presence of dropouts caused by technical issues. Such issues could have several negative consequences, for example: estimates of differential gene expression might be convoluted, and true biological variation masked. The phenomenon of excessive zeros is usually referred to as *zero inflation*, see Eq. 2.30 for a mathematical definition.

$$P(x) = \begin{cases} (1-\pi)f(x;\theta) + \pi & \text{if } x = 0 \\ (1-\pi)f(x;\theta) & \text{else} \end{cases} \tag{2.30}$$

f is the pdf/pmf of the underlying (non zero-inflated) distribution, θ the parameters of f, and π the probability of extra zeros. The debate about whether sequencing data exhibits zero inflation or not has split the single cell field into two different camps: those who model data as zero inflated, and those who oppose this practice.

There exist several methods where the idea of zero inflation has been incorporated into their core architecture, two examples being: ZIFA and ZINB-WaVE.[161, 162] In ZINB-WaVE the authors claim superiority to the standard negative binomial model, and support this by comparing their method to others relying on a standard negative binomial model. The other camp argues that the observed zeros aren't excessive and their presence can be explained with an appropriate statistical model. In his correspondence letter (to Nature Biotechnology) titled *"Droplet scRNA-seq is not zero-inflated,"* Svensson takes a clear stance in the question while also presenting compelling evidence in favor of his opinion. Svensson shows how observed frequencies of zero counts agree with those expected from a negative binomial distribution, and how deviations from this can be explained by biological variance.[163] Similarly, Sarkar and Stephens calls for the abolishment of the term "dropout" and state that there's a lack of evidence for the need of zero-inflated models, but acknowledge that there might be a use of it given the right circumstances. When comparing different combinations of measurement and expression models, they evaluate the performance of a point-Gamma expression model (rendering a zero-inflated negative binomial distribution), but only find a small percentage (2-16%) of genes that have even weak support for this observation model.[153]

Personally, without having investigated the issue thoroughly myself, I consider the arguments presented by the opponents to zero-inflated models stronger than those who are in favor of them; especially because the proponents have failed to produce a robust explanation of the mechanism that would produce the supposedly excessive zeros. The question of zero inflation has not been as widely discussed in the context of spatial transcriptomics techniques, but I see no reason to assume, by default, that expression of a gene would be zero-inflated among these methods. Nevertheless, issues with permeabilization or other technical aspects could motivate its application to a select set of genes.

Chapter 3 :: Epilogue

Science is a precise art where guesses and hypotheticals are often frowned upon; when comments of this kind are made, they are usually confined to the last few sentences in the discussion section of our publications. Review articles are the rare exception, in these the authors sometimes make predictions about the future, though they take caution to not be overtly visionary. However, as a young scientist without much prestige, you're unlikely to be given an outlet where your thoughts and ideas can be voiced – but the Doctoral thesis presents one such platform. Therefore, I'm much inclined to take this opportunity and share some personal reflections. These include what I've learnt so far, a few predictions about what I expect, and what I hope to see in the future. The reader may do as he/she/they pleases and skip this part if it's deemed too unscientific, but if some degree of speculation and subjectivity is accepted, it might offer an interesting read.

3.1 What I've learnt

The *"end of history illusion"* is a phenomenon in psychology where individuals agree that up until the current point in time they've experienced continuous and significant growth, but believe that, henceforth, they will not change by any considerable amount. This illusion is persistent across all ages, and repeatedly proven to be incorrect. We humans are malleable and never seem to solidify. No matter where in life we are, we continue to develop, change, and grow.

I was convinced that I'd learn a lot during my PhD, scientifically – but would I be affected on a personal level? Most likely not. Despite me being aware of the aforementioned illusion, I was impermeable to the idea that this experience would leave much of an imprint on me. I guess that this is at its best described as arrogance and at its worst as stupidity.

Starting my PhD on the 12:th of June 2019, I've spent exactly 1010 days – or 2 years, 9 months, and 6 days – pursuing my degree. This time has been nothing short of transformative. Agreeably, approximately three years is not a huge amount of time, but these years have been densely packed with new experiences, encounters, and impressions. I've acquired many new skills, but I also leave this era of my life as a very different person than the one who entered it. Below follows a curated list of insights that I've collected over the course of my PhD, relating to science as well as personal topics.

- **A high level of complexity does not equal a high level of success**. Among computational methods, it's rarely the most advanced methods that surface as the most popular ones. If you desire spread and impact, study the field, seek questions that are frequently being asked but rarely answered; then tailor your method towards this. Never develop a method and *then* invent a question for it to address.
- **Develop for you audience, not yourself**. If you're capable of formulating a statistical or mathematical model, and then implement it in code, you are likely

more proficient in these areas than the average user of your tool. Therefore, if you want people to use your software, make the interface intuitive and provide a layman's explanation of how it works. Good documentation with loads of examples is key to success. If possible, integrate your method into already existing frameworks, this makes it easy for users to explore without having to learn a new syntax. From my experience, methods that are easy to operate are often favored over less user friendly ones, even though the latter might have much better performance.

- **Listen to people when they complain**. If someone expresses that they are struggling with something, they are likely not alone. Embrace the opportunity and be the one to deliver the solution. This is one of the easiest ways to identify areas where you can make a useful contribution.

- **Seek diversity and honor others' expertise**. The best collaborations are those where the people involved have complementary strengths and show mutual respect for each other's skills. There's a difference to being proud of your expertise and being arrogant about it. A project thrives when the members don't consider their own contribution more (or less) important than anyone else's, but acknowledge that everyone is essential for the process to move forward.

- **Time spent planning is often doubly rewarded.** I'm addicted to fast progress, but have learnt that a short pause can save plenty of time. Making informed design choices, and not just blindly throwing yourself at the first idea, almost always results in a more pleasant and faster overall process. A quick fix for the situation at hand might seem tempting, but adapting general solutions usually pays off in the end.

- **Garbage data will give you garbage results**. You wouldn't pick up a roadkill, cook it, and then expect it to taste like a dish from a Michelin star restaurant. The same should hold true for data; one needs to have reasonable expectations about what information that can be derived from it. There's a difference between a bioinformat*ician* and a mag*ician*, the latter can turn nothing into something, the former cannot. Sometimes, the data is just not good enough to answer certain questions, if such is the case, there are only two reasonable options: (i) ask a different question, or (ii) generate new data.

- **Don't bring nuclear weapons to a gun fight.** Sometimes enthusiasm and excitement about new powerful methods makes us blind to the fact that the problem at hand likely could be solved with simpler means. For some questions, a simple regression model will do just as fine – and possibly even better – than a fancy deep learning model. It's easy to be caught up in the storm of buzz words, but take some time to contemplate what level of complexity your problem actually requires.

- **Aim to be the dumbest person in the room**. The best way to grow is to position yourself in an environment where people are more skilled than yourself, it accelerates learning and forces you to be alert. Comfort is truly the enemy of improvement.

- **Don't set yourself on fire to keep others warm**. I believe one should always strive to help others when we can, but at some point, it can also become problematic. If you *consistently* are the one who does the extra work, covers for others, and stays late – then you're not helping, you're being taken advantage of. We're all familiar with the airplane safety instructions telling us to put on our own masks before

helping someone else, this is equally applicable to the workplace. If you want to have a positive impact on the people around you, the most important thing is that you feel good about your own situation.

- **Never compromise on health**. In January 2021 I experienced something close to a physical collapse, my body simply quit on me. I could barely walk for two months, and for six more months, every day of my life felt like a living hell – I did not enjoy living. Every morning, I put on an alarm that counted down the hours that I had left to be awake and aware of my situation. Still, when night came, I barely slept. Instead, I woke up multiple times having issues breathing or in a state of complete sleep paralysis. A combination of bad nutrition, an extreme (according to some people) amount of exercise, and working ten to twelve hours a day (including weekends) put me in a state of severe exhaustion. It was not until I became a prisoner of my own body that I realized how much my previous freedom meant to me. It's hard realizing that you're not an exception, but just as human as everyone else. However, in the end, this realization is healthy. If there's one thing I will bring with me from these years, it's that *nothing* is worth sacrificing one's well-being or health for.

- **Perspective is everything.** There's a quote from the, truly awful, series *"Pirates of the Caribbean"* that reads: *"The problem is not the problem. The problem is your attitude about the problem."* Even though I cringe just by thinking about Captain Jack Sparrow, these words have stayed with me. I've experienced first hand how you can't plan every aspect of life. Unexpected things can, and will, happen. Our attitude determines how we experience these events, whether it becomes a tragedy or a lesson. I've tried to adopt more of a "gratitude mindset"; instead of being frustrated when things don't go my way, I try to celebrate what has gone right so far. This attitude is not always easy to maintain, and one is of course allowed to feel anger, but it's a feeling that becomes toxic if we let it linger for too long. Implementing this mindset have made me a much happier individual and helped me through some really dark times.

3.2 What I predict

In 2016, when the Spatial Transcriptomics (ST) technique was published, I had just finished the second year of my bachelor and was yet to hear the term "transcriptomics". Thus, I'm acutely aware of the fact that I belong to the younger generation of the transcriptomics field, and do not have the same experience as many of my peers. Still, having worked somewhat intensively in the niche of computational method development for spatial transcriptomics, I have a few predictions about the future, which I'll take the freedom to share here.

- **Deep learning methods will become staple goods**. Although I'm fascinated by deep learning (DL) methods, none of my works have so far exploited the power of these architectures – mainly because I've felt as if the questions I had could be addressed with simpler methods, or because the data wasn't there. However, the trend of access to more data, increasingly sophisticated and user friendly frameworks – paired with the development of new kinds of models – makes me certain that DL

will revolutionize the single cell and spatial transcriptomics fields, just as it has many other aspects of our life. Currently, a lot of the DL-based methods simply apply existing general models (e.g., taken from the natural language processing field) to a problem in the transcriptomics sphere. However, I believe we'll migrate from this approach towards using *bespoke models*, where prior information about the biological systems are integrated into the model architecture. In the very near future, methods leveraging graph convolutional networks (GCNs) and their aptitude for irregular data will likely become a popular element in many methods for analysis of spatial transcriptomics data.

- **Emergence of perturbation studies.** The majority of publications and projects that include spatial transcriptomics data have so far been observational. A sample is collected, analyzed, and relevant observations presented. At some rare occasions, samples representing case and control exist, but usually with limited meta data and no control over confounding variables. While interesting, this setup mainly permits exploratory data analysis (EDA), but does not lend itself well to infer causal relationships. To go beyond mere associations or correlations, an intervention or *perturbation* of the system is necessary. Thus, I'm certain that it's just a question of time until techniques to the likes of Perturb-seq are combined with spatial assays. With the introduction of perturbations, we'll be able to deduce how gene expression impacts spatial structure, and potentially also the reciprocal relationships. With access to such data, *causal inference* will likely become an essential tool for modeling and understanding causative effects. This is something I'm genuinely excited about.

- **Preference of generative models.** Many of the models we currently employ are of a discriminative nature, but I anticipate a shift towards *generative models*. Discriminative models assumes some functional form of the posterior, in contrast, generative models learns the joint probability distribution over all variables. Generative models are more susceptible to incorporation of prior information about the systems being studied, and better at representing causal relationships. Thus, they neatly tie together the two previous statement about a need for bespoke models and causal links.

- **Challenges of multimodal analysis.** To me, the trend in technology development can best be summarized with the Pokémon slogan: *"Gotta Catch 'Em All."* The transcriptome, epigenome, proteome, and metabolome – we want them all, at the same time, from the same cell. Alas, 10x Genomics already have an assay where RNA-seq and ATAC-seq data from the same cell can be obtained, as well as an second assay where spatial RNA-seq information and protein abundance are collected simultaneously. Except for increased ability to resolve cell types and states, very few examples where multimodal data is superior to unimodal data have so far been presented, but there's no lack of ideas.

  One of the commonly mentioned aspirations is to learn relationships between the different modalities, which can be used to *predict* one modality from another, for example, deducing protein levels from gene expression. Here, I will take a somewhat controversial and conservative stance by stating that: prediction of one modality from another will prove to be more challenging than many expect. I base this statement on two concepts: *temporal delays* and *missing information*. I'll elaborate on both these issues below.

Changes to one part of the central dogma usually don't manifest immediately in other parts, some form of delay tends to be present. Thus, data ($\boldsymbol{x}_t$) collected from one modality at time t isn't necessarily informative about the feature values ($\boldsymbol{y}_t$) of a different modality at the same time point. Instead – due to the lag – $\boldsymbol{x}_t$ relates to the values ($\boldsymbol{y}_{t'}$) at a later point t'. This discrepancy causes an issue in learning, because the two modalities are related according to:

$$\boldsymbol{y}_{t'} = f(\boldsymbol{x}_t) \tag{3.1}$$

However, in most multimodal assays, we observe ($\boldsymbol{x}_t, \boldsymbol{y}_t$), meaning the data required to learn f is not available. Potentially, $\boldsymbol{y}_{t'}$ could be inferred from $\boldsymbol{y}_t$ by learning a second map g such that $\boldsymbol{y}_{t'} = g(\boldsymbol{y}_t)$. Then f can be learnt by first transforming $\boldsymbol{y}_t$ through g. Now, to find g, the derivative $\partial \boldsymbol{y}_t / \partial t$ must likely be deduced. To estimate this derivative, at least one more data point close in time (w.r.t. protein turnover timescales) is required. Unfortunately, experimental assays only capture a single snapshot of the system at a particular time. As a consequence, estimation of such derivatives is usually infeasible. The dilemma described above is what I refer to as temporal delay.

Next, I'll address the second caveat, that of missing data. The path from one modality to another often involves several steps and regulatory mechanisms, not exclusively relying on elements of the observed modality. Thus, Eq. 3.1 should be updated to:

$$\boldsymbol{y}_{t'} = f(\boldsymbol{x}_t, \boldsymbol{u}_t) \tag{3.2}$$

Where $\boldsymbol{u}_t$ represent entities with an influence over the regulatory mechanisms (e.g., enzyme levels or metabolic concentrations). Note that it's possible that u_t and y_t overlaps. Assuming Eq. 3.2 is true, data must also be collected on $\boldsymbol{u}_t$ for predictions about $\boldsymbol{y}_{t'}$ to be made, solely relying on $\boldsymbol{x}_t$ is not sufficient. Thus, $\boldsymbol{x}_t$ does not contain all the information needed to predict $\boldsymbol{y}_t$. Of course, if $t \approx t'$ and $f(\boldsymbol{x}_t, \boldsymbol{u}_t) \approx f(\boldsymbol{x}_t)$, the problem is reduced to a much simpler one. Still, when such is not the case, we should accept that the prediction task is challenging. I definitely don't think it's beyond our capabilities, but while I expect methods for *integration* of different data modalities to emerge soon after the experimental technologies, general methods to model intermodal relationships will take more time to mature.

- **The group before the individual.** As mentioned in section 2.2.2, both internal and external factors influence a cell's state. In my opinion, there's still a need for general methods that tries to model how the local environment of a cell affects its behavior. Conditional models for gene expression already exist, one example being those that condition on cell type, often resulting in sets of marker genes or gene signatures. These models could be expanded to also include conditioning on the local environment of a cell, for example, the proportion of different cell types in its neighborhood. Such models add a new, interconnected, layer of information to our understanding of how cells operate in biological systems. Indeed, early attempts to construct models of this kind have already been made (e.g., node-centric expression modeling by the Theis Lab), and I dare to predict an abundance of them in a couple of years from now.

3.3 What I hope

Having outlined the lessons I've learnt and my predictions for the future, only one thing remains: listing some of the thing I hope for, but am less certain of.

- **Revised educational programs.** In genomics, almost every new technological method is accompanied by a suite of computational tools to analyze the data. Ever more frequently, high impact journals publish purely computational methods designed to unveil previously occluded insights that only emerge by clever modeling of the data. Thus, it's evident that computational expertise is just as important to advance life science as biological and technical knowledge. If further proof is needed, in 2021, SciLifeLab and the Wallenberg National Program announced several DDLS (data driven life science) fellowships, acknowledging the importance of computational competence. Still, the essential skills needed in computational biology, such as: statistics, mathematics, probability theory, modeling, and programming, are severely underrepresented in many of the biotechnology programs at Swedish universities. We need to step up our game if we want maintain our status within the life sciences as an innovative and leading nation, and remain competitive with international institutions like the Broad or the Wellcome Trust Sanger Institute. The foundation must be laid early on, educating PhD students is not good enough, computational biology tracks should be instituted already at the Master level and potentially even seep into the bachelor programs. I sincerely hope that the educational programs will be updated, to also prepare students – with an interest – for the challenges a computational biologist faces.

- **Increased diversity.** If there's one thing I'm not stoked about, it's gender quotation and female-exclusive events; to me they have an opposite effect of their intended purpose. These actions belittle women's competence and give the impression that we need extra help or special rules to succeed. However, women are clearly underrepresented in the computational field; at many hackathons or meetings, I've found myself – as a woman – in a very small minority, and am often assumed to be someone representing the wet-lab side. I'm not upset by this, and have never been met with anything but respect when correcting people, but I don't think it has to be like this. Girls and young women should be equally encouraged to purse STEM subjects as their male counterparts, and all of us – me included – should probably revise or abolish some of our stereotypes. So, I dearly hope for a future where the computational fields become more diverse and inclusive. Of course, diversity extends beyond gender, the same arguments can – and should – be made about ethnicity, age, religion, sexual identity, etc. Being a white woman living in Sweden, I fully acknowledge my privileges, and that my encounters with prejudice are probably dwarfed by those from other – less fortunate – groups. Still, I can only speak of my own experiences and observations.

- **Breaking the limit**. My third, and final, wish for the future is to pass the qualifying time for the Boston Marathon. To then – of course – complete the race.

Bibliography

[1] Rachel L Goldfeder, Dennis P Wall, Muin J Khoury, John P A Ioannidis, and Euan A Ashley. Human genome sequencing at the population scale: A primer on high-throughput DNA sequencing and analysis. *American Journal of Epidemiology*, 186(8):1000–1009, May 2017.

[2] Axel Meyer, Siegfried Schloissnig, Paolo Franchini, Kang Du, Joost M. Woltering, Iker Irisarri, Wai Yee Wong, Sergej Nowoshilow, Susanne Kneitz, Akane Kawaguchi, Andrej Fabrizius, Peiwen Xiong, Corentin Dechaud, Herman P. Spaink, Jean-Nicolas Volff, Oleg Simakov, Thorsten Burmester, Elly M. Tanaka, and Manfred Schartl. Giant lungfish genome elucidates the conquest of land by vertebrates. *Nature*, 590(7845):284–289, January 2021.

[3] R. Waterston and J. Sulston. The genome of caenorhabditis elegans. *Proceedings of the National Academy of Sciences*, 92(24):10836–10840, November 1995.

[4] Frederick R. Blattner, Guy Plunkett, Craig A. Bloch, Nicole T. Perna, Valerie Burland, Monica Riley, Julio Collado-Vides, Jeremy D. Glasner, Christopher K. Rode, George F. Mayhew, Jason Gregor, Nelson Wayne Davis, Heather A. Kirkpatrick, Michael A. Goeden, Debra J. Rose, Bob Mau, and Ying Shao. The complete genome sequence of escherichia coli k-12. *Science*, 277(5331):1453–1462, September 1997.

[5] Ran Elkon and Reuven Agami. Characterization of noncoding regulatory DNA in the human genome. *Nature Biotechnology*, 35(8):732–746, August 2017.

[6] The ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature*, 489(7414):57–74, September 2012.

[7] Hyunmin Lee, Zhaolei Zhang, and Henry M. Krause. Long noncoding RNAs and repetitive elements: Junk or intimate evolutionary partners? *Trends in Genetics*, 35(12):892–902, December 2019.

[8] Brian S. Gloss and Marcel E. Dinger. Realizing the significance of noncoding functionality in clinical genomics. *Experimental & Molecular Medicine*, 50(8):1–8, August 2018.

[9] Luisa Statello, Chun-Jie Guo, Ling-Ling Chen, and Maite Huarte. Gene regulation by long non-coding RNAs and its biological functions. *Nature Reviews Molecular Cell Biology*, 22(2):96–118, December 2020.

[10] Degeng Wang. "molecular gene": Interpretation in the right context. *Biology & Philosophy*, 20(2-3):453–464, March 2005.

[11] Cassandra Willyard. New human gene tally reignites debate. *Nature*, 558(7710):354–355, June 2018.

[12] Fivos Borbolis and Popi Syntichaki. Cytoplasmic mRNA turnover and ageing. *Mechanisms of Ageing and Development*, 152:32–42, December 2015.

[13] Jeremy Berg. *Biochemistry.* Macmillan inernational W.H. Freeman and Company, New York, 2019.

[14] Henderson James Cleaves. Selenocysteine. In *Encyclopedia of Astrobiology*, pages 1495–1496. Springer Berlin Heidelberg, 2011.

[15] Niels Gregersen, Peter Bross, Søren Vang, and Jane H. Christensen. Protein misfolding and human disease. *Annual Review of Genomics and Human Genetics*, 7(1):103–124, September 2006.

[16] Anthony L. Fink. Chaperone-mediated protein folding. *Physiological Reviews*, 79(2):425–449, April 1999.

[17] Christopher Walsh. *Posttranslational modification of proteins : expanding nature's inventory.* Roberts and Co. Publishers, Englewood, Colo, 2006.

[18] Matthew Cobb. 60 years ago, francis crick changed the logic of biology. *PLOS Biology*, 15(9):e2003243, September 2017.

[19] John M. Coffin and Hung Fan. The discovery of reverse transcriptase. *Annual Review of Virology*, 3(1):29–51, September 2016.

[20] Michael M. C. Lai. RNA replication without RNA-dependent RNA polymerase: Surprises from hepatitis delta virus. *Journal of Virology*, 79(13):7951–7958, jul 2005.

[21] Christopher Buccitelli and Matthias Selbach. mRNAs, proteins and the emerging principles of gene expression control. *Nature Reviews Genetics*, 21(10):630–644, July 2020.

[22] Pelin Akan, Andrey Alexeyenko, Paul Costea, Lilia Hedberg, Beata Solnestam, Sverker Lundin, Jimmie Hällman, Emma Lundberg, Mathias Uhlén, and Joakim Lundeberg. Comprehensive analysis of the genome transcriptome and proteome landscapes of three tumor cell lines. *Genome Medicine*, 4(11):86, 2012.

[23] Emma Lundberg, Linn Fagerberg, Daniel Klevebring, Ivan Matic, Tamar Geiger, Juergen Cox, Cajsa Älgenäs, Joakim Lundeberg, Matthias Mann, and Mathias Uhlen. Defining the transcriptome and proteome in three functionally different human cell lines. *Molecular Systems Biology*, 6(1):450, January 2010.

[24] Bruce Alberts. *Essential cell biology.* Garland Science, New York, 2015.

[25] P F Chinnery. Mitochondria. *Journal of Neurology, Neurosurgery & Psychiatry*, 74(9):1188–1199, September 2003.

[26] Philip J. Stewart. Mendeleev's predictions: success and failure. *Foundations of Chemistry*, 21(1):3–9, April 2018.

[27] C.H. Waddington and H. Kacser. *The Strategy of the Genes: A Discussion of Some Aspects of Theoretical Biology.* Allen & Unwin, 1957.

[28] Kazutoshi Takahashi and Shinya Yamanaka. Induction of pluripotent stem cells from mouse embryonic and adult fibroblast cultures by defined factors. *Cell*, 126(4):663–676, August 2006.

[29] Yongchang Yao and Chunming Wang. Dedifferentiation: inspiration for devising engineering strategies for regenerative medicine. *npj Regenerative Medicine*, 5(1), July 2020.

[30] Geoffrey Schiebinger, Jian Shu, Marcin Tabaka, Brian Cleary, Vidya Subramanian, Aryeh Solomon, Joshua Gould, Siyan Liu, Stacie Lin, Peter Berube, Lia Lee, Jenny Chen, Justin Brumbaugh, Philippe Rigollet, Konrad Hochedlinger, Rudolf Jaenisch, Aviv Regev, and Eric S. Lander. Optimal-transport analysis of single-cell gene expression identifies developmental trajectories in reprogramming. *Cell*, 176(4):928–943.e22, February 2019.

[31] Jiajun Zhang, Qing Nie, and Tianshou Zhou. Revealing dynamic mechanisms of cell fate decisions from single-cell transcriptomic data. *Frontiers in Genetics*, 10, December 2019.

[32] J. Wang, K. Zhang, L. Xu, and E. Wang. Quantifying the waddington landscape and biological paths for development and differentiation. *Proceedings of the National Academy of Sciences*, 108(20):8257–8262, May 2011.

[33] Andrew E. Teschendorff and Andrew P. Feinberg. Statistical mechanics meets single-cell biology. *Nature Reviews Genetics*, 22(7):459–476, April 2021.

[34] Jian Xu, Xinfeng Liu, Jieli Chen, Alex Zacharek, Xu Cui, Smita Savant-Bhonsale, Michael Chopp, and Zhenguo Liu. Cell–cell interaction promotes rat marrow stromal cell differentiation into endothelial cell via activation of TACE/TNF-alpha signaling. *Cell Transplantation*, 19(1):43–53, January 2010.

[35] Silvia S. Chen, Wendy Fitzgerald, Joshua Zimmerberg, Hynda K. Kleinman, and Leonid Margolis. Cell-cell and cell-extracellular matrix interactions regulate embryonic stem cell differentiation. *STEM CELLS*, 25(3):553–561, March 2007.

[36] Megan K. Rommelfanger and Adam L. MacLean. A single-cell resolved cell-cell communication model explains lineage commitment in hematopoiesis. *Development*, 148(24), December 2021.

[37] Francesco Capozzi and Alessandra Bordoni. Foodomics: a new comprehensive approach to food and nutrition. *Genes & Nutrition*, 8(1):1–4, August 2012.

[38] Elsa Call, Christoph Mayer, Victoria Twort, Lars Dietz, Niklas Wahlberg, and Marianne Espeland. Museomics: Phylogenomics of the moth family epicopeiidae (lepidoptera) using target enrichment. *Insect Systematics and Diversity*, 5(2), March 2021.

[39] F. Sanger, S. Nicklen, and A. R. Coulson. DNA sequencing with chain-terminating inhibitors. *Proceedings of the National Academy of Sciences*, 74(12):5463–5467, December 1977.

[40] Lilian T. C. França, Emanuel Carrilho, and Tarso B. L. Kist. A review of DNA sequencing techniques. *Quarterly Reviews of Biophysics*, 35(2):169–200, May 2002.

[41] Sara Goodwin, John D. McPherson, John D, and W. Richard McCombie. Coming of age: ten years of next-generation sequencing technologies. *Nature Reviews Genetics*, 17(6):333–351, Jun 2016.

[42] James M. Heather and Benjamin Chain. The sequence of sequencers: The history of sequencing DNA. *Genomics*, 107(1):1–8, January 2016.

[43] Lincoln D Stein. The case for cloud computing in genome informatics. *Genome Biology*, 11(5):207, 2010.

[44] Lilit Garibyan and Nidhi Avashia. Polymerase chain reaction. *Journal of Investigative Dermatology*, 133(3):1–4, March 2013.

[45] Silvia G. Acinas, Ramahi Sarma-Rupavtarm, Vanja Klepac-Ceraj, and Martin F. Polz. PCR-induced sequence artifacts and bias: Insights from comparison of two 16s rRNA clone libraries constructed from the same sample. *Applied and Environmental Microbiology*, 71(12):8966–8969, December 2005.

[46] Adam Roberts, Cole Trapnell, Julie Donaghey, John L Rinn, and Lior Pachter. Improving RNA-seq expression estimates by correcting for fragment bias. *Genome Biology*, 12(3):R22, 2011.

[47] Johnny A. Sena, Giulia Galotto, Nico P. Devitt, Melanie C. Connick, Jennifer L. Jacobi, Pooja E. Umale, Luis Vidali, and Callum J. Bell. Unique molecular identifiers reveal a novel sequencing artefact with implications for RNA-seq based gene expression analysis. *Scientific Reports*, 8(1), September 2018.

[48] Shawn E. Levy and Richard M. Myers. Advancements in next-generation sequencing. *Annual Review of Genomics and Human Genetics*, 17(1):95–115, August 2016.

[49] Allison Piovesan, Maria Caracausi, Francesca Antonaros, Maria Chiara Pelleri, and Lorenza Vitale. GeneBase 1.1: a tool to summarize data from NCBI gene datasets and its application to an update of human gene statistics. *Database*, 2016:baw153, 2016.

[50] Olena Morozova and Marco A. Marra. Applications of next-generation sequencing technologies in functional genomics. *Genomics*, 92(5):255–264, November 2008.

[51] Shanika L. Amarasinghe, Shian Su, Xueyi Dong, Luke Zappia, Matthew E. Ritchie, and Quentin Gouil. Opportunities and challenges in long-read sequencing data analysis. *Genome Biology*, 21(1), February 2020.

[52] Jeffrey R. Moffitt, Junjie Hao, Guiping Wang, Kok Hao Chen, Hazen P. Babcock, and Xiaowei Zhuang. High-throughput single-cell gene-expression profiling with multiplexed error-robust fluorescence in situ hybridization. *Proceedings of the National Academy of Sciences*, 113(39):11046–11051, September 2016.

[53] M. L. Pardue and J. G. Gall. MOLECULAR HYBRIDIZATION OF RADIOACTIVE DNA TO THE DNA OF CYTOLOGICAL PREPARATIONS. *Proceedings of the National Academy of Sciences*, 64(2):600–604, October 1969.

[54] Andrea M. Femino, Fredric S. Fay, Kevin Fogarty, and Robert H. Singer. Visualization of single RNA transcripts in situ. *Science*, 280(5363):585–590, April 1998.

[55] Sperr E. Pubmed by year, 2016.

[56] Rory Stark, Marta Grzelak, and James Hadfield. RNA sequencing: the teenage years. *Nature Reviews Genetics*, 20(11):631–656, July 2019.

[57] Fuchou Tang, Catalin Barbacioru, Yangzhou Wang, Ellen Nordman, Clarence Lee, Nanlan Xu, Xiaohui Wang, John Bodeau, Brian B Tuch, Asim Siddiqui, Kaiqin Lao, and M Azim Surani. mRNA-seq whole-transcriptome analysis of a single cell. *Nature Methods*, 6(5):377–382, April 2009.

[58] Daniel Ramsköld, Shujun Luo, Yu-Chieh Wang, Robin Li, Qiaolin Deng, Omid R Faridani, Gregory A Daniels, Irina Khrebtukova, Jeanne F Loring, Louise C Laurent, Gary P Schroth, and Rickard Sandberg. Full-length mRNA-seq from single-cell levels of RNA and individual circulating tumor cells. *Nature Biotechnology*, 30(8):777–782, July 2012.

[59] Method of the year 2013. *Nature Methods*, 11(1):1–1, December 2013.

[60] Evan Z. Macosko, Anindita Basu, Rahul Satija, James Nemesh, Karthik Shekhar, Melissa Goldman, Itay Tirosh, Allison R. Bialas, Nolan Kamitaki, Emily M. Martersteck, John J. Trombetta, David A. Weitz, Joshua R. Sanes, Alex K. Shalek, Aviv Regev, and Steven A. McCarroll. Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. *Cell*, 161(5):1202–1214, May 2015.

[61] Grace X. Y. Zheng, Jessica M. Terry, Phillip Belgrader, Paul Ryvkin, Zachary W. Bent, Ryan Wilson, Solongo B. Ziraldo, Tobias D. Wheeler, Geoff P. McDermott, Junjie Zhu, Mark T. Gregory, Joe Shuga, Luz Montesclaros, Jason G. Underwood, Donald A. Masquelier, Stefanie Y. Nishimura, Michael Schnall-Levin, Paul W. Wyatt, Christopher M. Hindson, Rajiv Bharadwaj, Alexander Wong, Kevin D. Ness, Lan W. Beppu, H. Joachim Deeg, Christopher McFarland, Keith R. Loeb, William J. Valente, Nolan G. Ericson, Emily A. Stevens, Jerald P. Radich, Tarjei S. Mikkelsen, Benjamin J. Hindson, and Jason H. Bielas. Massively parallel digital transcriptional profiling of single cells. *Nature Communications*, 8(1), January 2017.

[62] Richard C. V. Tyser, Elmir Mahammadov, Shota Nakanoh, Ludovic Vallier, Antonio Scialdone, and Shankar Srinivas. Single-cell transcriptomic characterization of a gastrulating human embryo. *Nature*, 600(7888):285–289, November 2021.

[63] Cole Trapnell, Davide Cacchiarelli, Jonna Grimsby, Prapti Pokharel, Shuqiang Li, Michael Morse, Niall J Lennon, Kenneth J Livak, Tarjei S Mikkelsen, and John L Rinn. The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. *Nature Biotechnology*, 32(4):381–386, March 2014.

[64] Xin Zou, Ke Chen, Jiawei Zou, Peiyi Han, Jie Hao, and Zeguang Han. Single-cell RNA-seq data analysis on the receptor ACE2 expression reveals the potential risk of different human organs vulnerable to 2019-nCoV infection. *Frontiers of Medicine*, 14(2):185–192, March 2020.

[65] Anoop P. Patel, Itay Tirosh, John J. Trombetta, Alex K. Shalek, Shawn M. Gillespie, Hiroaki Wakimoto, Daniel P. Cahill, Brian V. Nahed, William T. Curry, Robert L. Martuza, David N. Louis, Orit Rozenblatt-Rosen, Mario L. Suvà, Aviv Regev, and Bradley E. Bernstein. Single-cell RNA-seq highlights intratumoral heterogeneity in primary glioblastoma. *Science*, 344(6190):1396–1401, June 2014.

[66] Amit Zeisel, Ana B. Muñoz-Manchado, Simone Codeluppi, Peter Lönnerberg, Gioele La Manno, Anna Juréus, Sueli Marques, Hermany Munguba, Liqun He, Christer Betsholtz, Charlotte Rolny, Gonçalo Castelo-Branco, Jens Hjerling-Leffler,

and Sten Linnarsson. Cell types in the mouse cortex and hippocampus revealed by single-cell RNA-seq. *Science*, 347(6226):1138–1142, March 2015.

[67] Jesper Bäckdahl, Lovisa Franzén, Lucas Massier, Qian Li, Jutta Jalkanen, Hui Gao, Alma Andersson, Nayanika Bhalla, Anders Thorell, Mikael Rydén, Patrik L. Ståhl, and Niklas Mejhert. Spatial mapping reveals human adipocyte subpopulations with distinct sensitivities to insulin. *Cell Metabolism*, 33(9):1869–1882.e6, September 2021.

[68] 10x Genomics. Q/a: Can i process neutrophils (or other granulocytes) using 10x single cell applications?, 2020.

[69] Gabrielle J. Benitez and Kosaku Shinoda. Isolation of adipose tissue nuclei for single-cell genomic applications. *Journal of Visualized Experiments*, (160), June 2020.

[70] Aviv Regev, Sarah A Teichmann, Eric S Lander, Ido Amit, Christophe Benoist, Ewan Birney, Bernd Bodenmiller, Peter Campbell, Piero Carninci, Menna Clatworthy, Hans Clevers, Bart Deplancke, Ian Dunham, James Eberwine, Roland Eils, Wolfgang Enard, Andrew Farmer, Lars Fugger, Berthold Göttgens, Nir Hacohen, Muzlifah Haniffa, Martin Hemberg, Seung Kim, Paul Klenerman, Arnold Kriegstein, Ed Lein, Sten Linnarsson, Emma Lundberg, Joakim Lundeberg, Partha Majumder, John C Marioni, Miriam Merad, Musa Mhlanga, Martijn Nawijn, Mihai Netea, Garry Nolan, Dana Pe'er, Anthony Phillipakis, Chris P Ponting, Stephen Quake, Wolf Reik, Orit Rozenblatt-Rosen, Joshua Sanes, Rahul Satija, Ton N Schumacher, Alex Shalek, Ehud Shapiro, Padmanee Sharma, Jay W Shin, Oliver Stegle, Michael Stratton, Michael J T Stubbington, Fabian J Theis, Matthias Uhlen, Alexander van Oudenaarden, Allon Wagner, Fiona Watt, Jonathan Weissman, Barbara Wold, Ramnik Xavier, and Nir Yosef and. The human cell atlas. *eLife*, 6, December 2017.

[71] and Stephen R Quake. The tabula sapiens: a multiple organ single cell transcriptomic atlas of humans. *bioRxiv*, July 2021.

[72] Method of the year 2020: spatially resolved transcriptomics. *Nature Methods*, 18(1):1–1, January 2021.

[73] Michaela Asp, Joseph Bergenstråhle, and Joakim Lundeberg. Spatially resolved transcriptomes—next generation tools for tissue exploration. *BioEssays*, 42(10):1900221, May 2020.

[74] Lambda Moses and Lior Pachter. Museum of spatial transcriptomics. *bioRxiv*, May 2021.

[75] Lisa N. Waylen, Hieu T. Nim, Luciano G. Martelotto, and Mirana Ramialison. From whole-mount to single-cell spatial assessment of gene expression in 3d. *Communications Biology*, 3(1), October 2020.

[76] Anjali Rao, Dalia Barkley, Gustavo S. França, and Itai Yanai. Exploring tissue architecture using spatial transcriptomics. *Nature*, 596(7871):211–220, August 2021.

[77] Ditte Lovatt, Brittani K Ruble, Jaehee Lee, Hannah Dueck, Tae Kyung Kim, Stephen Fisher, Chantal Francis, Jennifer M Spaethling, John A Wolf, M Sean Grady, Alexandra V Ulyanova, Sean B Yeldell, Julianne C Griepenburg, Peter T Buckley, Junhyong Kim, Jai-Yoon Sul, Ivan J Dmochowski, and James Eberwine.

Transcriptome in vivo analysis (TIVA) of spatially defined single cells in live tissue. *Nature Methods*, 11(2):190–196, January 2014.

[78] Christopher R. Merritt, Giang T. Ong, Sarah E. Church, Kristi Barker, Patrick Danaher, Gary Geiss, Margaret Hoang, Jaemyeong Jung, Yan Liang, Jill McKay-Fleisch, Karen Nguyen, Zach Norgaard, Kristina Sorg, Isaac Sprague, Charles Warren, Sarah Warren, Philippa J. Webster, Zoey Zhou, Daniel R. Zollinger, Dwayne L. Dunaway, Gordon B. Mills, and Joseph M. Beechem. Multiplex digital spatial profiling of proteins and RNA in fixed tissue. *Nature Biotechnology*, 38(5):586–599, May 2020.

[79] Kenny Roberts, Alexander Aivazidis, Vitalii Kleshchevnikov, Tong Li, Robin Fropf, Michael Rhodes, Joseph M. Beechem, Martin Hemberg, and Omer Ali Bayraktar. Transcriptome-wide spatial RNA profiling maps the cellular architecture of the developing human neocortex. *bioRxiv*, March 2021.

[80] Chiara Medaglia, Amir Giladi, Liat Stoler-Barak, Marco De Giovanni, Tomer Meir Salame, Adi Biram, Eyal David, Hanjie Li, Matteo Iannacone, Ziv Shulman, and Ido Amit. Spatial reconstruction of immune niches by combining photoactivatable reporters and scRNA-seq. *Science*, 358(6370):1622–1626, December 2017.

[81] F. Kruse, J.P. Junker, A. van Oudenaarden, and J. Bakkers. Tomo-seq: A method to obtain genome-wide expression data with spatial resolution. In *Methods in Cell Biology*, pages 299–307. Elsevier, 2016.

[82] Halima Hannah Schede, Christian G. Schneider, Johanna Stergiadou, Lars E. Borm, Anurag Ranjak, Tracy M. Yamawaki, Fabrice P. A. David, Peter Lönnerberg, Maria Antonietta Tosches, Simone Codeluppi, and Gioele La Manno. Spatial tissue profiling by imaging-free molecular tomography. *Nature Biotechnology*, 39(8):968–977, April 2021.

[83] Arjun Raj, Patrick van den Bogaard, Scott A Rifkin, Alexander van Oudenaarden, and Sanjay Tyagi. Imaging individual mRNA molecules using multiple singly labeled probes. *Nature Methods*, 5(10):877–879, September 2008.

[84] Chee-Huat Linus Eng, Sheel Shah, Julian Thomassie, and Long Cai. Profiling the transcriptome with RNA SPOTs. *Nature Methods*, 14(12):1153–1155, November 2017.

[85] Eric Lubeck, Ahmet F Coskun, Timur Zhiyentayev, Mubhij Ahmad, and Long Cai. Single-cell in situ RNA profiling by sequential hybridization. *Nature Methods*, 11(4):360–361, March 2014.

[86] Eric Lubeck and Long Cai. Single-cell systems biology by super-resolution imaging and combinatorial labeling. *Nature Methods*, 9(7):743–748, June 2012.

[87] Sheel Shah, Eric Lubeck, Maayan Schwarzkopf, Ting fang He, Alon Greenbaum, Chang ho Sohn, Antti Lignell, Harry M. T. Choi, Viviana Gradinaru, Niles A. Pierce, and Long Cai. Single-molecule RNA detection at depth via hybridization chain reaction and tissue hydrogel embedding and clearing. *Development*, January 2016.

[88] Chee-Huat Linus Eng, Michael Lawson, Qian Zhu, Ruben Dries, Noushin Koulena, Yodai Takei, Jina Yun, Christopher Cronin, Christoph Karp, Guo-Cheng Yuan, and Long Cai. Transcriptome-scale super-resolved imaging in tissues by RNA seqFISH+. *Nature*, 568(7751):235–239, March 2019.

[89] Chenglong Xia, Jean Fan, George Emanuel, Junjie Hao, and Xiaowei Zhuang. Spatial transcriptome profiling by MERFISH reveals subcellular RNA compartmentalization and cell cycle-dependent gene expression. *Proceedings of the National Academy of Sciences*, 116(39):19490–19499, September 2019.

[90] Jolene Jie Lin Goh, Nigel Chou, Wan Yi Seow, Norbert Ha, Chung Pui Paul Cheng, Yun-Ching Chang, Ziqing Winston Zhao, and Kok Hao Chen. Highly specific multiplexed RNA imaging in tissues with split-FISH. *Nature Methods*, 17(7):689–693, June 2020.

[91] Simone Codeluppi, Lars E. Borm, Amit Zeisel, Gioele La Manno, Josina A. van Lunteren, Camilla I. Svensson, and Sten Linnarsson. Spatial organization of the somatosensory cortex revealed by osmFISH. *Nature Methods*, 15(11):932–935, October 2018.

[92] Yuhan Wang, Mark Eddison, Greg Fleishman, Martin Weigert, Shengjin Xu, Tim Wang, Konrad Rokicki, Cristian Goina, Fredrick E. Henry, Andrew L. Lemire, Uwe Schmidt, Hui Yang, Karel Svoboda, Eugene W. Myers, Stephan Saalfeld, Wyatt Korff, Scott M. Sternson, and Paul W. Tillberg. EASI-FISH for thick tissue defines lateral hypothalamus spatio-molecular organization. *Cell*, 184(26):6361–6377.e24, December 2021.

[93] Fay Wang, John Flanagan, Nan Su, Li-Chong Wang, Son Bui, Allissa Nielson, Xingyong Wu, Hong-Thuy Vo, Xiao-Jun Ma, and Yuling Luo. RNAscope. *The Journal of Molecular Diagnostics*, 14(1):22–29, January 2012.

[94] Joshua A. Weinstein, Aviv Regev, and Feng Zhang. DNA microscopy: Optics-free spatio-genetic imaging by a stand-alone chemical reaction. *Cell*, 178(1):229–241.e16, June 2019.

[95] Rongqin Ke, Marco Mignardi, Alexandra Pacureanu, Jessica Svedlund, Johan Botling, Carolina Wählby, and Mats Nilsson. In situ sequencing for RNA analysis in preserved tissue and cells. *Nature Methods*, 10(9):857–860, July 2013.

[96] Xiaoyin Chen, Yu-Chi Sun, George M Church, Je Hyuk Lee, and Anthony M Zador. Efficient in situ barcode sequencing using padlock probe-based BaristaSeq. *Nucleic Acids Research*, 46(4):e22–e22, November 2017.

[97] Xiao Wang, William E. Allen, Matthew A. Wright, Emily L. Sylwestrak, Nikolay Samusik, Sam Vesuna, Kathryn Evans, Cindy Liu, Charu Ramakrishnan, Jia Liu, Garry P. Nolan, Felice-Alessio Bava, and Karl Deisseroth. Three-dimensional intact-tissue sequencing of single-cell transcriptional states. *Science*, 361(6400), July 2018.

[98] Je Hyuk Lee, Evan R. Daugharthy, Jonathan Scheiman, Reza Kalhor, Joyce L. Yang, Thomas C. Ferrante, Richard Terry, Sauveur S. F. Jeanty, Chao Li, Ryoji Amamoto, Derek T. Peters, Brian M. Turczyk, Adam H. Marblestone, Samuel A. Inverso, Amy Bernard, Prashant Mali, Xavier Rios, John Aach, and George M. Church.

Highly multiplexed subcellular RNA sequencing in situ. *Science*, 343(6177):1360–1363, March 2014.

[99] Kaia Achim, Jean-Baptiste Pettit, Luis R Saraiva, Daria Gavriouchkina, Tomas Larsson, Detlev Arendt, and John C Marioni. High-throughput spatial mapping of single-cell RNA-seq data to tissue of origin. *Nature Biotechnology*, 33(5):503–509, April 2015.

[100] Rahul Satija, Jeffrey A Farrell, David Gennert, Alexander F Schier, and Aviv Regev. Spatial reconstruction of single-cell gene expression data. *Nature Biotechnology*, 33:495–502, 2015.

[101] Guangdun Peng, Shengbao Suo, Jun Chen, Weiyang Chen, Chang Liu, Fang Yu, Ran Wang, Shirui Chen, Na Sun, Guizhong Cui, Lu Song, Patrick P.L. Tam, Jing-Dong J. Han, and Naihe Jing. Spatial transcriptome for the molecular annotation of lineage fates and cell identity in mid-gastrula mouse embryo. *Developmental Cell*, 36(6):681–697, March 2016.

[102] Floyd Maseda, Zixuan Cang, and Qing Nie. DEEPsc: A deep learning-based map connecting single-cell transcriptomics and spatial imaging data. *Frontiers in Genetics*, 12, March 2021.

[103] Mor Nitzan, Nikos Karaiskos, Nir Friedman, and Nikolaus Rajewsky. Gene expression cartography. *Nature*, 576(7785):132–137, November 2019.

[104] Zixuan Cang and Qing Nie. Inferring spatial and signaling relationships between cells from single cell transcriptomic data. *Nature Communications*, 11(1), April 2020.

[105] Xianwen Ren, Guojie Zhong, Qiming Zhang, Lei Zhang, Yujie Sun, and Zemin Zhang. Reconstruction of cell spatial organization from single-cell RNA sequencing data based on ligand-receptor mediated self-assembly. *Cell Research*, 30(9):763–778, June 2020.

[106] 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, and Jonas Frisén. Visualization and analysis of gene expression in tissue sections by spatial transcriptomics. *Science*, 353(6294):78–82, July 2016.

[107] Emelie Berglund, Sami Saarenpää, Anders Jemt, Joel Gruselius, Ludvig Larsson, Ludvig Bergenstråhle, Joakim Lundeberg, and Stefania Giacomello. Automation of spatial transcriptomics library preparation to enable rapid and robust insights into spatial organization of tissues. *BMC Genomics*, 21(1), April 2020.

[108] 10x Genomics. Visium product information.

[109] Manuel Saiselet, Joël Rodrigues-Vitória, Adrien Tourneur, Ligia Craciun, Alex Spinette, Denis Larsimont, Guy Andry, Joakim Lundeberg, Carine Maenhaut, and Vincent Detours. Transcriptional output, cell-type densities, and normalization in spatial transcriptomics. *Journal of Molecular Cell Biology*, 12(11):906–908, June 2020.

[110] Kevin Lebrigand, Joseph Bergenstråhle, Kim Thrane, Annelie Mollbrink, Konstantinos Meletis, Pascal Barbry, Rainer Waldmann, and Joakim Lundeberg. The spatial landscape of gene expression isoforms in tissue sections. *bioRxiv*, August 2020.

[111] Samuel G. Rodriques, Robert R. Stickels, Aleksandrina Goeva, Carly A. Martin, Evan Murray, Charles R. Vanderburg, Joshua Welch, Linlin M. Chen, Fei Chen, and Evan Z. Macosko. Slide-seq: A scalable technology for measuring genome-wide expression at high spatial resolution. *Science*, 363(6434):1463–1467, March 2019.

[112] Sanja Vickovic, Gökcen Eraslan, Fredrik Salmén, Johanna Klughammer, Linnea Stenbeck, Denis Schapiro, Tarmo Äijö, Richard Bonneau, Ludvig Bergenstråhle, José Fernandéz Navarro, Joshua Gould, Gabriel K. Griffin, Åke Borg, Mostafa Ronaghi, Jonas Frisén, Joakim Lundeberg, Aviv Regev, and Patrik L. Ståhl. High-definition spatial transcriptomics for in situ tissue profiling. *Nature Methods*, 16(10):987–990, September 2019.

[113] Robert R. Stickels, Evan Murray, Pawan Kumar, Jilong Li, Jamie L. Marshall, Daniela J. Di Bella, Paola Arlotta, Evan Z. Macosko, and Fei Chen. Highly sensitive spatial transcriptomics at near-cellular resolution with slide-seqV2. *Nature Biotechnology*, 39(3):313–319, December 2020.

[114] Yang Liu, Mingyu Yang, Yanxiang Deng, Graham Su, Archibald Enninful, Cindy C. Guo, Toma Tebaldi, Di Zhang, Dongjoo Kim, Zhiliang Bai, Eileen Norris, Alisia Pan, Jiatong Li, Yang Xiao, Stephanie Halene, and Rong Fan. High-spatial-resolution multi-omics sequencing via deterministic barcoding in tissue. *Cell*, 183(6):1665–1681.e18, December 2020.

[115] Chun-Seok Cho, Jingyue Xi, Yichen Si, Sung-Rye Park, Jer-En Hsu, Myungjin Kim, Goo Jun, Hyun Min Kang, and Jun Hee Lee. Microscopic examination of spatial transcriptome using seq-scope. *Cell*, 184(13):3559–3572.e22, June 2021.

[116] Xiaonan Fu, Li Sun, Jane Y. Chen, Runze Dong, Yiing Lin, Richard D. Palmiter, Shin Lin, and Liangcai Gu. Continuous polony gels for tissue mapping with high resolution and RNA capture efficiency. *bioRxiv*, March 2021.

[117] Ao Chen, Sha Liao, Mengnan Cheng, Kailong Ma, Liang Wu, Yiwei Lai, Xiaojie Qiu, Jin Yang, Wenjiao Li, Jiangshan Xu, Shijie Hao, Xin Wang, Huifang Lu, Xi Chen, Xing Liu, Xin Huang, Feng Lin, Zhao Li, Yan Hong, Defeng Fu, Yujia Jiang, Jian Peng, Shuai Liu, Mengzhe Shen, Chuanyu Liu, Quanshui Li, Yue Yuan, Huiwen Zheng, Zhifeng Wang, Zhaohui Wang, Xin Huang, Haitao Xiang, Lei Han, Baoming Qin, Pengcheng Guo, Pura Muñoz-Cánoves, Jean Paul Thiery, Qingfeng Wu, Fuxiang Zhao, Mei Li, Haoyan Kuang, Junhou Hui, Ou Wang, Haorong Lu, Bo Wang, Shiping Liu, Ming Ni, Wenwei Zhang, Feng Mu, Ye Yin, Huanming Yang, Michael Lisby, Richard J. Cornall, Jan Mulder, Mathias Uhlen, Miguel A. Esteban, Yuxiang Li, Longqi Liu, Xun Xu, and Jian Wang. Spatiotemporal transcriptomic atlas of mouse organogenesis using DNA nanoball patterned arrays. *bioRxiv*, January 2021.

[118] Ludvig Larsson, Jonas Frisén, and Joakim Lundeberg. Spatially resolved transcriptomics adds a new dimension to genomics. *Nature Methods*, 18(1):15–18, January 2021.

[119] Satija Lab. Analysis, visualization, and integration of spatial datasets with seurat, 2021-08-30.

[120] scanpy team. scanpy 1.5.0 release notes, 2020-05-15.

[121] Joseph Bergenstråhle, Ludvig Larsson, and Joakim Lundeberg. Seamless integration of image and molecular analysis for spatial transcriptomics workflows. *BMC Genomics*, 21(1), July 2020.

[122] Ruben Dries, Qian Zhu, Rui Dong, Chee-Huat Linus Eng, Huipeng Li, Kan Liu, Yuntian Fu, Tianxiao Zhao, Arpan Sarkar, Feng Bao, Rani E. George, Nico Pierson, Long Cai, and Guo-Cheng Yuan. Giotto: a toolbox for integrative analysis and visualization of spatial expression data. *Genome Biology*, 22(1), March 2021.

[123] Duy Pham, Xiao Tan, Jun Xu, Laura F. Grice, Pui Yeng Lam, Arti Raghubar, Jana Vukovic, Marc J. Ruitenberg, and Quan Nguyen. stLearn: integrating spatial location, tissue morphology and gene expression to find cell types, cell-cell interactions and spatial trajectories within undissociated tissues. *bioRxiv*, May 2020.

[124] Giovanni Palla, Hannah Spitzer, Michal Klein, David Fischer, Anna Christina Schaar, Louis Benedikt Kuemmerle, Sergei Rybakov, Ignacio L. Ibarra, Olle Holmberg, Isaac Virshup, Mohammad Lotfollahi, Sabrina Richter, and Fabian J. Theis. Squidpy: a scalable framework for spatial single cell analysis. *bioRxiv*, February 2021.

[125] Brendan F. Miller, Feiyang Huang, Lyla Atta, Arpan Sahoo, and Jean Fan. Reference-free cell-type deconvolution of multi-cellular pixel-resolution spatially resolved transcriptomics data. *bioRxiv*, June 2021.

[126] Jonas Maaskola, Ludvig Bergenstråhle, Aleksandra Jurek, José Fernández Navarro, Jens Lagergren, and Joakim Lundeberg. Charting tissue expression anatomy by spatial transcriptome decomposition. *bioRxiv*, July 2018.

[127] Gioele La Manno, Ruslan Soldatov, Amit Zeisel, Emelie Braun, Hannah Hochgerner, Viktor Petukhov, Katja Lidschreiber, Maria E. Kastriti, Peter Lönnerberg, Alessandro Furlan, Jean Fan, Lars E. Borm, Zehua Liu, David van Bruggen, Jimin Guo, Xiaoling He, Roger Barker, Erik Sundström, Gonçalo Castelo-Branco, Patrick Cramer, Igor Adameyko, Sten Linnarsson, and Peter V. Kharchenko. RNA velocity of single cells. *Nature*, 560(7719):494–498, August 2018.

[128] Chenglong Xia, Jean Fan, George Emanuel, Junjie Hao, and Xiaowei Zhuang. Spatial transcriptome profiling by MERFISH reveals subcellular RNA compartmentalization and cell cycle-dependent gene expression. *Proceedings of the National Academy of Sciences*, 116(39):19490–19499, September 2019.

[129] Daniel Edsgärd, Per Johnsson, and Rickard Sandberg. Identification of spatial expression trends in single-cell gene expression data. *Nature Methods*, 15(5):339–342, March 2018.

[130] Valentine Svensson, Sarah A Teichmann, and Oliver Stegle. SpatialDE: identification of spatially variable genes. *Nature Methods*, 15(5):343–346, March 2018.

[131] Shiquan Sun, Jiaqiang Zhu, and Xiang Zhou. Statistical analysis of spatial expression patterns for spatially resolved transcriptomic studies. *Nature Methods*, 17(2):193–200, January 2020.

[132] Ilia Kats, Roser Vento-Tormo, and Oliver Stegle. SpatialDE2: Fast and localized variance component analysis of spatial transcriptomics. *bioRxiv*, October 2021.

[133] Jiaqiang Zhu, Shiquan Sun, and Xiang Zhou. SPARK-x: non-parametric modeling enables scalable and robust detection of spatial expression patterns for large spatial transcriptomic studies. *Genome Biology*, 22(1), June 2021.

[134] Damien Arnol, Denis Schapiro, Bernd Bodenmiller, Julio Saez-Rodriguez, and Oliver Stegle. Modeling cell-cell interactions from spatial molecular data with spatial variance component analysis. *Cell Reports*, 29(1):202–211.e6, October 2019.

[135] Florin Walter, Oliver Stegle, and Britta Velten. FISHFactor: A probabilistic factor model for spatial transcriptomics data with subcellular resolution. *bioRxiv*, November 2021.

[136] F. William Townes and Barbara E. Engelhardt. Nonnegative spatial factorization. *arXiv*, 2021.

[137] Shila Ghazanfar, Yingxin Lin, Xianbin Su, David Ming Lin, Ellis Patrick, Ze-Guang Han, John C. Marioni, and Jean Yee Hwa Yang. Investigating higher-order interactions in single-cell data with scHOT. *Nature Methods*, 17(8):799–806, July 2020.

[138] Qian Zhu, Sheel Shah, Ruben Dries, Long Cai, and Guo-Cheng Yuan. Identification of spatially associated subpopulations by combining scRNAseq and sequential fluorescence in situ hybridization data. *Nature Biotechnology*, 36(12):1183–1190, October 2018.

[139] Rebecca Elyanow, Ron Zeira, Max Land, and Benjamin J Raphael. STARCH: copy number and clone inference from spatial transcriptomics data. *Physical Biology*, 18(3):035001, March 2021.

[140] Reuben Moncada, Dalia Barkley, Florian Wagner, Marta Chiodin, Joseph C. Devlin, Maayan Baron, Cristina H. Hajdu, Diane M. Simeone, and Itai Yanai. Integrating microarray-based spatial transcriptomics and single-cell RNA-seq reveals tissue architecture in pancreatic ductal adenocarcinomas. *Nature Biotechnology*, 38(3):333–342, January 2020.

[141] Vitalii Kleshchevnikov, Artem Shmatko, Emma Dann, Alexander Aivazidis, Hamish W King, Tong Li, Artem Lomakin, Veronika Kedlian, Mika Sarkin Jain, Jun Sung Park, Lauma Ramona, Elizabeth Tuck, Anna Arutyunyan, Roser Vento-Tormo, Moritz Gerstung, Louisa James, Oliver Stegle, and Omer Ali Bayraktar. Comprehensive mapping of tissue cell architecture via integrated single cell and spatial transcriptomics. *bioRxiv*, November 2020.

[142] Dylan M. Cable, Evan Murray, Luli S. Zou, Aleksandrina Goeva, Evan Z. Macosko, Fei Chen, and Rafael A. Irizarry. Robust decomposition of cell type mixtures in spatial transcriptomics. *Nature Biotechnology*, February 2021.

[143] Tommaso Biancalani, Gabriele Scalia, Lorenzo Buffoni, Raghav Avasthi, Ziqing Lu, Aman Sanger, Neriman Tokcan, Charles R. Vanderburg, Åsa Segerstolpe, Meng Zhang, Inbal Avraham-Davidi, Sanja Vickovic, Mor Nitzan, Sai Ma, Ayshwarya Subramanian, Michal Lipinski, Jason Buenrostro, Nik Bear Brown, Duccio Fanelli,

Xiaowei Zhuang, Evan Z. Macosko, and Aviv Regev. Deep learning and alignment of spatially resolved single-cell transcriptomes with tangram. *Nature Methods*, 18(11):1352–1362, October 2021.

[144] Xiaoyan Qian, Kenneth D. Harris, Thomas Hauling, Dimitris Nicoloutsopoulos, Ana B. Muñoz-Manchado, Nathan Skene, Jens Hjerling-Leffler, and Mats Nilsson. Probabilistic cell typing enables fine mapping of closely related cell types in situ. *Nature Methods*, 17(1):101–106, November 2019.

[145] Tim Stuart, Andrew Butler, Paul Hoffman, Christoph Hafemeister, Efthymia Papalexi, William M Mauck III, Yuhan Hao, Marlon Stoeckius, Peter Smibert, and Rahul Satija. Comprehensive integration of single-cell data. *Cell*, 177:1888–1902, 2019.

[146] Bryan He, Ludvig Bergenstråhle, Linnea Stenbeck, Abubakar Abid, Alma Andersson, Åke Borg, Jonas Maaskola, Joakim Lundeberg, and James Zou. Integrating spatial gene expression and breast tumour morphology via deep learning. *Nature Biomedical Engineering*, 4(8):827–834, June 2020.

[147] Ludvig Bergenstråhle, Bryan He, Joseph Bergenstråhle, Xesús Abalo, Reza Mirzazadeh, Kim Thrane, Andrew L. Ji, Alma Andersson, Ludvig Larsson, Nathalie Stakenborg, Guy Boeckxstaens, Paul Khavari, James Zou, Joakim Lundeberg, and Jonas Maaskola. Super-resolved spatial transcriptomics by deep data fusion. *Nature Biotechnology*, November 2021.

[148] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. *CoRR*, abs/1505.04597, 2015.

[149] Yang Xu and Rachel Patton McCord. CoSTA: unsupervised convolutional neural network learning for spatial transcriptomics analysis. *BMC Bioinformatics*, 22(1), August 2021.

[150] Christoph Hafemeister and Rahul Satija. Normalization and variance stabilization of single-cell RNA-seq data using regularized negative binomial regression. *Genome Biology*, 20(1), December 2019.

[151] Ron Milo. *Cell biology by the numbers*. Garland Science, Taylor & Francis Group, New York, NY, 2016.

[152] Ashraful Haque, Jessica Engel, Sarah A. Teichmann, and Tapio Lönnberg. A practical guide to single-cell RNA-sequencing for biomedical research and clinical applications. *Genome Medicine*, 9(1), August 2017.

[153] Abhishek Sarkar and Matthew Stephens. Separating measurement and expression models clarifies confusion in single-cell RNA sequencing analysis. *Nature Genetics*, 53(6):770–777, May 2021.

[154] Sunny Z. Wu, Ghamdan Al-Eryani, Daniel Lee Roden, Simon Junankar, Kate Harvey, Alma Andersson, Aatish Thennavan, Chenfei Wang, James R. Torpy, Nenad Bartonicek, Taopeng Wang, Ludvig Larsson, Dominik Kaczorowski, Neil I. Weisenfeld, Cedric R. Uytingco, Jennifer G. Chew, Zachary W. Bent, Chia-Ling Chan, Vikkitharan Gnanasambandapillai, Charles-Antoine Dutertre, Laurence Gluch, Mun N. Hui, Jane Beith, Andrew Parker, Elizabeth Robbins, Davendra

Segara, Caroline Cooper, Cindy Mak, Belinda Chan, Sanjay Warrier, Florent Ginhoux, Ewan Millar, Joseph E. Powell, Stephen R. Williams, X. Shirley Liu, Sandra O'Toole, Elgene Lim, Joakim Lundeberg, Charles M. Perou, and Alexander Swarbrick. A single-cell and spatially resolved atlas of human breast cancers. *Nature Genetics*, 53(9):1334–1347, September 2021.

[155] Michael I Love, Wolfgang Huber, and Simon Anders. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology*, 15(12), December 2014.

[156] M. D. Robinson, D. J. McCarthy, and G. K. Smyth. edgeR: a bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, 26(1):139–140, November 2009.

[157] Saket Choudhary and Rahul Satija. Comparison and evaluation of statistical error models for scRNA-seq. *Genome Biology*, 23(1), January 2022.

[158] F. William Townes and Rafael A. Irizarry. Quantile normalization of single-cell RNA-seq read counts without unique molecular identifiers. *Genome Biology*, 21(1), July 2020.

[159] Douglas Curran-Everett. Explorations in statistics: the log transformation. *Advances in Physiology Education*, 42(2):343–347, June 2018.

[160] Robert B. O'Hara and D. Johan Kotze. Do not log-transform count data. *Methods in Ecology and Evolution*, 1(2):118–122, March 2010.

[161] Emma Pierson and Christopher Yau. ZIFA: Dimensionality reduction for zero-inflated single-cell gene expression analysis. *Genome Biology*, 16(1), November 2015.

[162] Davide Risso, Fanny Perraudeau, Svetlana Gribkova, Sandrine Dudoit, and Jean-Philippe Vert. A general and flexible method for signal extraction from single-cell RNA-seq data. *Nature Communications*, 9(1), January 2018.

[163] Valentine Svensson. Droplet scRNA-seq is not zero-inflated. *Nature Biotechnology*, 38(2):147–150, January 2020.

Chapter 4 :: Present Investigations

4.1 Summary

Article I describes a method to integrate single cell RNA-seq and *in situ* capture-based spatial transcriptomics data, effectively allowing the user to map cell types found in the former to the spatial domain characterized by the latter. The method is probabilistic in its character and models both data modalities as negative binomial distributed. In short, it leverages the annotated single cell data to learn cell type specific parameters, which are used to decompose the spatial gene expression profiles into contributions from said cell types. In this project, I formulated the theoretical model which the method relies upon, implemented the method in code (as a tool called *stereoscope*), analyzed the data as well as interpreted the results, and wrote the manuscript with input from all the contributing authors.

The method presented in **Article I** has been utilized in several projects both in-house and by other labs. One example being the study outlined in **Article II**, where we relied heavily on spatial mapping of single cell data to answer questions related to co-localization and regional enrichment of cell types. This study focused on HER2-positive breast cancer samples, surveyed with the first generation Spatial Transcriptomics (ST1K) technique. By combining the information obtained from the single cell mapping with an unsupervised gene expression-based analysis we managed to: (i) extract a set of core signatures representative of the samples in the study; (ii) define a TLS (tertiary lymphoid structure) specific gene signature with predictive power and clinical relevance; and (iii) identify a trifold interaction between two cell type subsets and a chemokine signal, which was also confirmed to be present in several external data sets. My role was to coordinate the project, but I also conducted the cell type related analysis, interpreted the complete set of results with help from an immunologist (Camilla Engblom), and wrote the majority of the manuscript. Entering the project at a later stage, I was not involved in the experimental design or sample collection.

The findings and work of **Article II** emphasized the importance of spatial patterns of cell types, sparking an interest in ways of finding genes with distinct spatial structures. While methods for finding spatially variable genes already existed, we wanted to approach the task from a slightly different angle. To us, a spatial pattern is defined by its lack of randomness, the more random a feature's spatial organization is, the less of a pattern it exhibits. From this definition, the fundamental concepts resulting in **Article III** emerged. In the paper, we present a method to find spatially variable features in an unsupervised manner by applying a numerical method (finite differences) to simulate diffusion of molecules in spatial transcriptomics data. To quantify initial randomness, the method measures the time until the system convergences (a homogeneous random state). Operating on the premise that expression profiles with a structured initial configuration will require more time to converge than those with a random spatial distribution, we are able to rank genes by their "patternedness". In the study we show competitive results with existing methods w.r.t. both accuracy and computational performance. In this project, I conceived the model, implemented it in code, analyzed the data as well as interpreted the

results, and wrote the manuscript with input from the other co-author.

In **Article IV**, we again used *stereoscope* to delineate the spatial distribution of cell types in mouse liver. However, we also sought to address more tissue specific questions and thus devised a method to investigate the features' dynamical signals (e.g., gene expression or cell type abundance) with respect to certain vein structures in the tissue. In brief, this strategy modeled the feature values as a function of the distance to the nearest vein structure, allowing us to further investigate so called zonation-patterns in the liver. In addition, a classifier (based on logistic regression) was constructed to predict the type of vein based on its (spatial) neighborhood expression profile, allowing us to computationally annotate ambiguous vein structures. I designed and implemented the aforementioned methods in this project, as well as being heavily involved in the writing process. Still, my role was secondary to the main author who interpreted the results, produced the data, and wrote most of the text.

In the final paper, **Article V**, we propose a new method to construct so-called common coordinate frameworks (CCFs) for spatial transcriptomics data. This is achieved by using a statistical approach relying on landmark annotation and Gaussian Process Regression. More specifically, a function relating gene expression to landmark distances is learnt and then used to transfer the observed data to any reference of choice. By this transfer, it is possible to compare local changes in gene expression between conditions or time points as well as performing more sophisticated forms of spatiotemporal modeling. A CCF also allows spatial gene expression from multiple samples to be represented jointly in a single reference, facilitating the identification of canonical structures or patterns. Similar to the other method development projects, I designed the underlying model, implemented it in code, conducted the analyses, and wrote the manuscript with input from all other authors. All the synthetic data was generated by me, but the new (unpublished) developmental heart data belongs to a larger project, where I was not involved in the experimental design or collection of data.