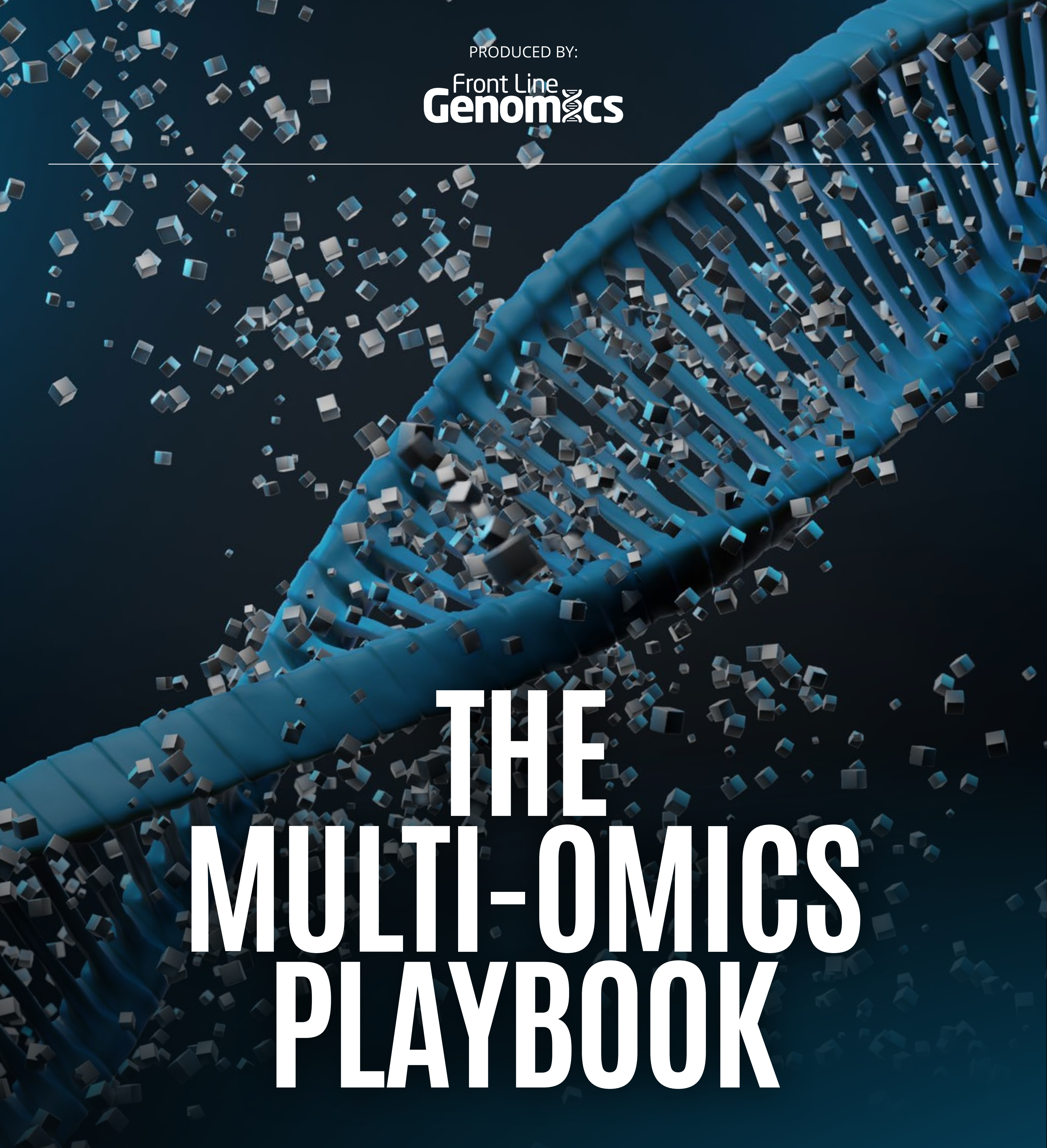
PRODUCED BY:

Front Line Genomics



THE MULTI-OMICS PLAYBOOK

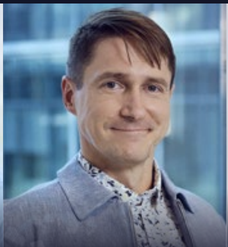
SPONSORED BY

10x GENOMICS

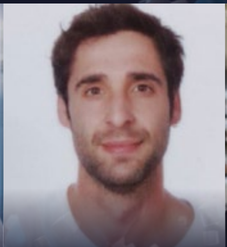
biomodal

CONTRIBUTORS

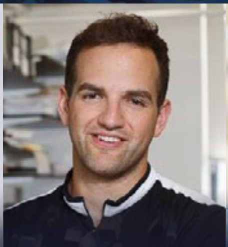
Nima Aghaeepour
Associate Professor, Anaesthesiology, Perioperative and Pain Medicine & Paediatrics - Neonatal and Developmental Medicine
Stanford University



Theodore Alexandrov
Team Leader, Structural and Computational Biology Unit
European Molecular Biology Laboratory (EMBL)

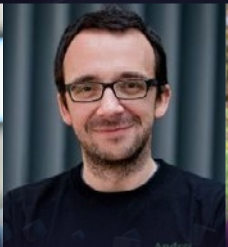


Ricard Argelaguet
Senior Research Scientist
Altos Labs



Mathew Chamberlain
Principal Scientist
Johnson & Johnson
Innovative Medicine

Rui Chen
Professor of Molecular and Human Genetics
Baylor College of Medicine



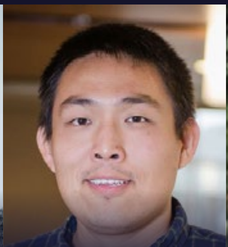
Andrea Corsinotti
Single-cell Multi-omics Facility Manager, Centre for Regenerative Medicine, Institute for Regeneration and Repair
University of Edinburgh

Mirjana Efremova
Group Leader
Barts Cancer Institute

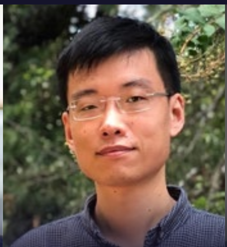
Shirley Greenbaum
Postdoctoral fellow, Department of Pathology
Stanford University
Resident, Department of Obstetrics and Gynaecology
Hadassah-Hebrew University Medical Center

Ingela Lanekoff
Professor, Department of Chemistry-BMC
Uppsala University

Holly-May Lewis
Senior Laboratory Technician (LC-MS)
University of Surrey



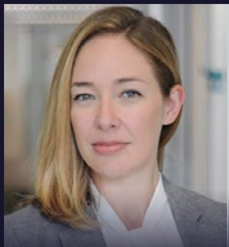
Qin Ma
Professor, Department of Biomedical Informatics
The Ohio State University



Zongming Ma
Professor, Department of Statistics and Data Science
Yale University

Iain Macaulay
Technical Development Group Leader
Earlham Institute

Pau Badia i Mompel
PhD Candidate, Saez-Rodriguez Group
Heidelberg University



Samantha Morris
Associate Professor of Development Biology and Genetics
Washington University School of Medicine in St. Louis

Sushmita Roy
Professor, Department of Biostatistics and Medical Informatics
University of Wisconsin-Madison
Faculty
Wisconsin Institute of Discovery

Xiaotao Shen
Postdoctoral Research Fellow, Snyder Lab
Stanford University

Suhas Vasaikar
Principal Scientist, Clinical Biomarker and Diagnostics
Seattle Genetics (Seagen)

Judith Zaugg
Group Leader
European Molecular Biology Laboratory (EMBL)

Bingjie Zhang
Postdoctoral Research Fellow, Satija Lab
New York Genome Center

FOREWORD

Matt Higgs

Science Writer
Front Line Genomics

Team Acknowledgements:
Lyndsey Fletcher and Rich Lumb from Front Line Genomics, for their support with editing and proofreading.

MULTI-OMICS IS A REALLY EXCITING METHODOLOGY. TRAPPED IN THE MONO-OMIC VIEWPOINT HAS PREVENTED US FROM TRULY EXPLORING THE INTRICATELY COMPLEX NATURE OF BIOLOGY. HOWEVER, WHILE MULTI-OMICS CAN PROVIDE MULTI-INSIGHTS IT ALSO PRESENTS MULTI-CHALLENGES.

Principal of which, it is uniquely challenging to keep track of because of the broad range of advancements that can be grouped under the multi-omics umbrella. This range extends from efforts to standardise classical multi-omics pairings such as proteomics and transcriptomics, to expanding multi-omics to include microbiomes, metabolomics, exposomes and so on.

This playbook is a unique resource to keep track of this breadth.

Within these pages, you will find an up-to-date overview of the integration landscape, in depth summaries of transcriptomic and proteomic methods and transcriptomic and epigenomic methods. Plus, an introduction to metabolomics and coverage of the various applications of multi-modal data in a clinical and pharmaceutical capacity.

While we cannot claim to have covered every nook and cranny, this playbook provides an up-to-date overview of multi-omics for 2024, specifically focusing on the advances in single-cell and spatial methodologies.

Furthermore, by interviewing a series of experts in the field, we have gained unique insights and guidance, which have shaped this playbook. Excerpts from our discussions with these experts are found throughout the chapters. Within them, you will find advice on how to get the most out of specific tools, hard fought wisdom gained from working with these technologies and designing new tools, as well as perspectives and views on current topics in multi-omics.

We would like to take this opportunity to thank all of our contributors for their time and insights when writing this playbook.

We would also like to thank the sponsors of this report, 10x Genomics and biomodal.

We hope you find this playbook a helpful resource.

Thank you for reading.

CONTENTS

CHAPTER 1

INTO THE MULTI-OMICS-VERSE

THIS CHAPTER PRESENTS A GENERAL INTRODUCTION TO MULTI-OMICS; WHAT IS IT, AND WHY PERFORM IT? WE BRIEFLY INTRODUCE SINGLE-CELL AND SPATIAL MULTI-OMICS, BEFORE ADDRESSING EACH 'OMIC' IN TURN, AND DISCUSSING THE INHERENT VALUE OF LOOKING AT OMICS TOGETHER IN THE SAME EXPERIMENT. WHAT CAN WE LEARN ABOUT THE SECRET LIFE OF CELLS FROM THE PURSUIT OF MULTIMODAL DATA?

What is multi-omics?

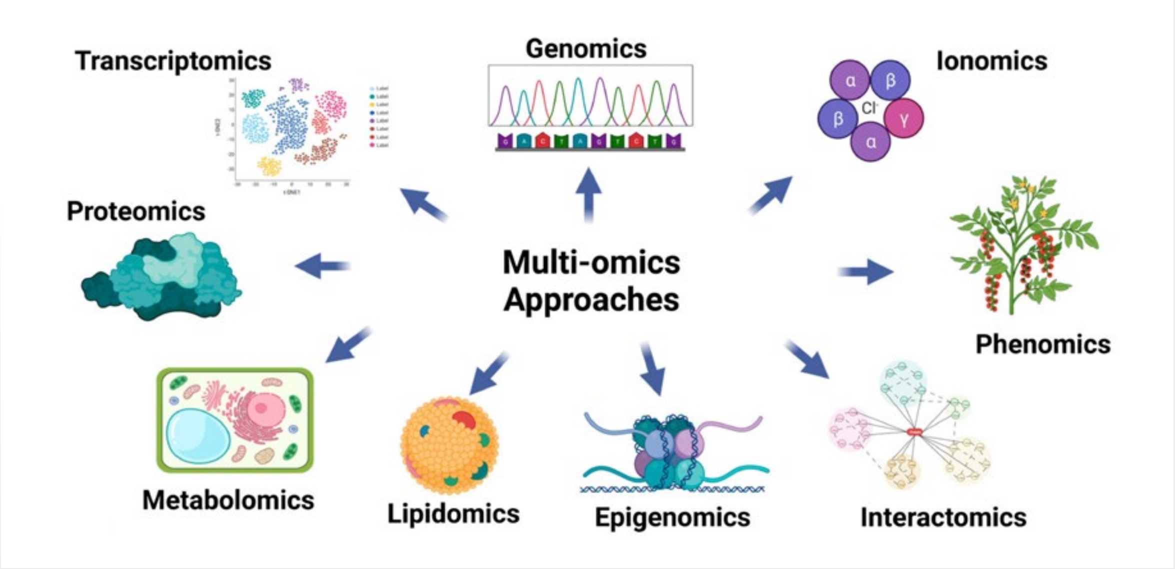
It is important to begin with a definition of multi-omics. Multi-omics is the acquisition, integration and analysis of 'omics' data from different molecular levels. Although, traditionally, this may have referred to taking information from different levels of the DNA-RNA-Protein genetic dogma, it can now also refer to additional layers such as the epigenome and metabolome. It can even include layers from outside the genomics paradigm, such as the microbiome, exposome and phenome (see Figure 1.1).

In the world of big omics data, capturing all the information at a particular level of the molecular processes of the organism is the goal.

Multi-omics experiments tend to take two forms. One form involves integrating data from different molecular layers that has been acquired from separate experiments. While the data was not gathered in the same experiment, sophisticated computational tools enable this data to be merged and insights about two molecular layers can be drawn on a broad level.[2,3]

FIGURE 1.1. A SELECTION OF THE MULTI-OMICS APPROACHES THAT ARE CURRENTLY AVAILABLE TO RESEARCHERS.

Image credit: Roychowdhury, et al.[1]



The second and ‘pure’ form of multi-omics involves integrating data from different molecular layers captured from the same sample/single cell. New tools and techniques are frequently being announced that can profile two or more ‘omics’ from one sample or one cell in the same experiment. These methods allow you to explore the interplay between two molecular layers within the confines of the same cell, a more refined approach for uncovering intracellular biology.

Why perform multi-omics?

Why go to the bother of integrating genomic, transcriptomic and proteomic information when one is transcribed/translated into the next? Furthermore, multi-omics methods are more costly and require a higher expertise to perform correctly. So, why do scientists continue to use them?

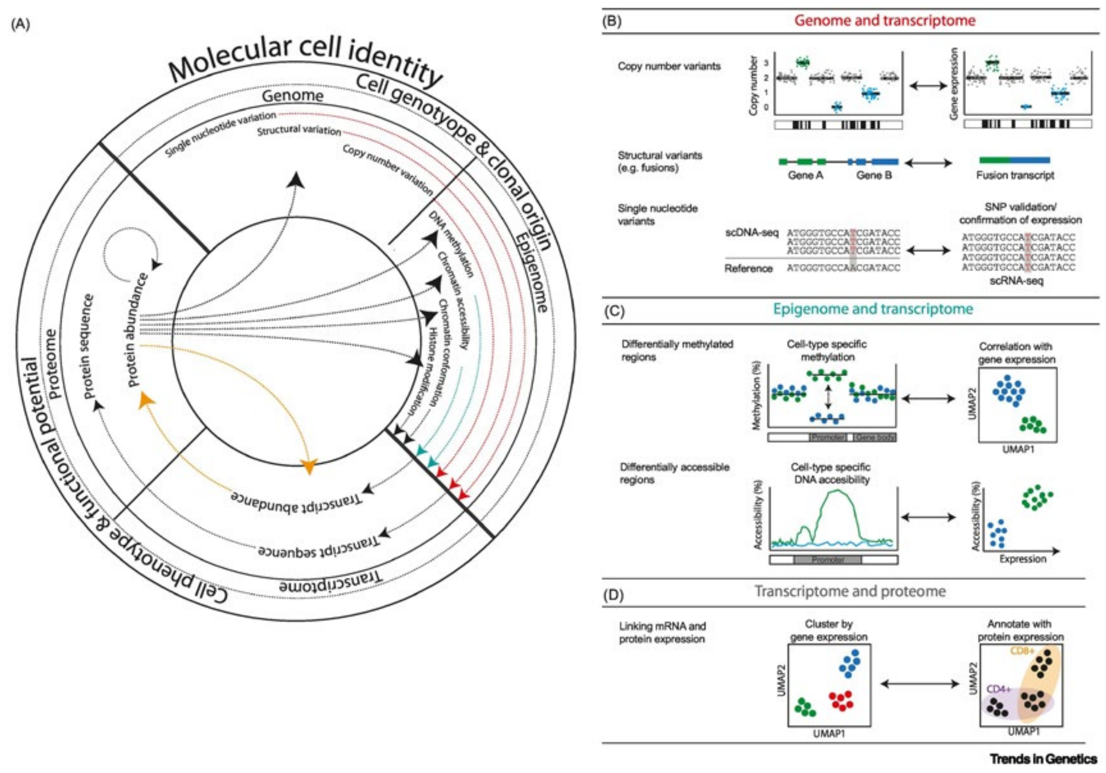
Ultimately, the biological processes that occur within each organism, each tissue and each cell are the result of complex networks. While each omic presents a valid and insightful partial picture of the inner workings of a cell or tissue, when used alone, one omic cannot capture the true intricacies. Not all open and accessible DNA regions will be transcribed into RNA, and not all RNA will be translated into protein. The puzzle can only be completely solved with all the pieces (see Figure 1.2).

From a pragmatic perspective, the adoption of multi-omics methods is part of the attempt to find more robust classifiers of biological samples (e.g., cancer subtypes) and improve patient stratification, but the biological implications are vast. For instance, how does genotype link to phenotype? Profile the genome alongside RNA and other omics to find out.

This enhanced view of biology is more than valuable enough for researchers to invest the time, money and expertise into designing multi-omics approaches.

FIGURE 1.2. MULTI-OMICS AND CELL IDENTITY.

*(**A**) The molecular cell identity can only be understood through the profiling an interaction between these many molecular layers. Multi-omics methods to achieve this are highlighted in (**B-D**). Image Credit: Ogbeide, et al.*[4]



Single-cell multi-omics

While initial methods to integrate multi-omics data began at the bulk level, it has been in the single-cell and spatial landscape that multi-omics has taken off in the last few years. Hence, single-cell and spatial multi-omics will be the focal point for this report.

As a reflection of this advancement, the Nature Method of the Year in 2019 was single-cell multimodal omics[5]. To map the convoluted landscape of DNA, RNA, protein and epigenomics, we need methods that can treat cells as individual entities rather than a homogenised cluster, and single-cell methodologies have delivered.

Single-cell RNA-sequencing has revealed much about cell types and cell states, but RNA by itself is not enough to fully resolve cell types[6]. The combination of transcriptomics and, for example, cell surface proteomics, allows better resolution of cell types with subtle transcriptomics profiles. Furthermore, scRNA-seq tells us little of a cell's lineage or trajectory, which can be crucial to track for cancer research. Incorporating epigenomics is necessary to broaden this picture.

Even though these technologies give us the power to visualise the genetic interplay within individual cells, new challenges originate from taking multi-omics measurements in single cells at scale. For instance, new integration strategies are needed, batch-effects must be carefully considered and data size has gotten incredibly large[7].

Spatial multi-omics

The latest developing area of multi-omics is spatial. Here, we have seen the development of a series of new tools that visualise omics data from two distinct modalities with spatial context[8-10]. This adds a whole new dimension to the data. RNA and cell surface proteins can now be visualised in cells. Spatial RNA and epigenomics based gene regulatory networks can be visualised for complex tissues within spatial context to each other. Furthermore, the multi-omics basis of cell-cell communication can now be exposed.

Chapter 3 will highlight some of the latest spatial multi-omics technologies to be released in the last few years including Spatial CITE-seq[11], DBiT-seq[12], SM-OMICS[13], DNA-MERFISH[14] and Spatial-CUT&TAG/ATAC-RNAseq[15].

Given the importance of spatial context for biology and disease, a pivotal example of which is the tumour microenvironment, it is hard to underestimate the importance that spatial multi-omics will have for personalised and precision medicine[16,17].

What is being profiled? - The Big Five Omics

Before covering the bulk, single-cell and spatial methodologies that attempt to profile multiple omics together, we will first cover each of the major 'omics' individually to highlight the nature of each omic and the value it provides when profiled individually.

First, we'll cover the big five – genomics, epigenomics, transcriptomics, proteomics and metabolomics (See Figure 1.3).

GENOME

By studying the genome, you are ultimately trying to make sense of the 3.2 billion base pairs of information encoded there (and that's just for humans!)[18]. Our DNA is the most immutable genomic layer to be interrogated with multi-omics approaches. It remains largely unaltered through life and is not going to adapt based on environmental exposure. When analysing the genome, we are looking for anomalous features such as single nucleotide variations (SNVs), indels, insertions, deletions, copy number variations (CNVs), duplications and inversions. Essentially, any feature on the genome that could be associated with disease or other outcome. Genome-wide association studies (GWAS), arrays and next generation whole genome sequencing all allow this information to be gathered[19].

EPIGENOME

Epigenomics is the study of the set of chemical changes to the DNA or histone proteins. These changes ultimately shape how available sections of DNA are to transcription machinery. Without accessible DNA, the associated gene cannot be expressed or influence the phenotype of a cell. Naturally, the epigenome is represented by several different data types:

- **We can measure the methylation status of the DNA**, a repressive regulator of expression caused by the addition of methyl groups on genomic regions called CpG islands, through methods such as bisulphite sequencing[20-22].
- **We can measure the modifications directly on histone proteins** that change how accessible sections of DNA are. This is mostly popularly performed with either ChIP-seq[23] or CUT&Tag[24]
- **We can directly measure how exposed a section of DNA is for transcription.** This is referred to as open-chromatin profiling and is most commonly performed using ATAC-seq[25].
- **Finally, we can measure the three-dimensional profile of the DNA within a cell**, ultimately exposing the sections of DNA that are in close contact and the sections that might be inaccessible in a 3D landscape. This is most popularly performed with Hi-C methodology[26].

The epigenomic profile of a cell or a tissue is distinct and useful for identification, just like transcriptomics (see next section). The epigenome is also malleable and can be changed upon exposure to the environment and as the result of disease, meaning it can be used to profile disease-related effects.

TRANSCRIPTOME

The omic workhorse of the past 20 years, micro-array and RNA-sequencing has seen RNA become a key marker for cell state, disease biomarkers and everything in between[27]. Ultimately, transcriptomics is the profiling of RNA transcripts that are produced from the genome. It is a measure of potential; an RNA could be on its way to translation, regulation or degradation. In any case, it's presence in your sample implies that the gene is needed in some way.

NGS and single-cell approaches have resulted in a state-of-play in which whole transcriptomes are being analysed from millions of cells at once. Because of the maturity of the field, and the depth of transcriptomics data, RNA tends to be found in most multi-omics approaches as the 'rock'. However, RNA is not without its limitations. It is a temporary construct in the cell and its link to phenotype is complex at best[28] (see this recent paper on the incongruence between RNA and cell perturbation in the heart[29]).

PROTEOME

The phenotype of an organism, tissue or cell is ultimately dependent on the proteins that are active within a cell. These proteins are transcribed from RNA. One would think that where you find RNA transcripts, you will find the resulting protein. However, this is not always the case, and RNA and protein levels do display discrepancies[30,31]. Proteomics is an important measure to truly assess phenotype. Recorded through mass spectrometry or targeted, antibody-based approaches, proteomics represents a different dimensionality datatype to transcriptomics and has its own challenges. Single-cell and spatial proteomics is a burgeoning field of study with exciting advances, see this review[32] and this review[33] for more.

METABOLOME

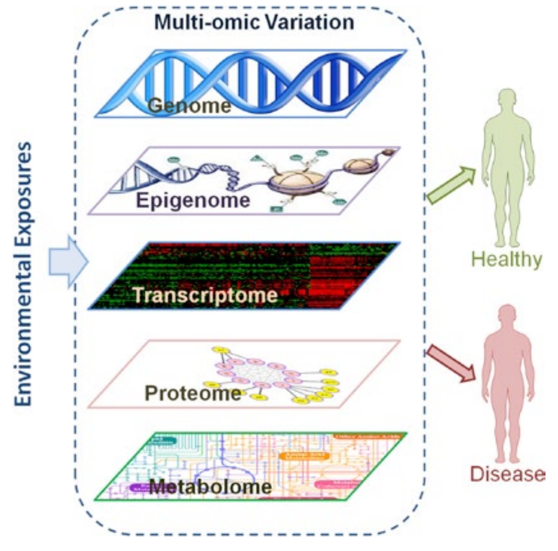
Then we get to the metabolome. The broadest of the major omics, it represents the profiling of every small molecule within tissues and cells, often referred to as metabolites. This includes sugars, fatty acids, lipids and amino acids - the building blocks of cell structures and the output of the metabolism. By profiling these molecules, we get the closest insight possible into what is actually happening within a cell, because we see the direct remnants of the processes.

Metabolites are short-lived, so we are truly seeing what has just happened in the cell. This makes metabolomics particularly valuable for healthcare and pharmaceutical uses, since it can inform the direct consequences of perturbations. Metabolites are profiled most commonly by mass spectrometry or NMR, and spatial and single-cell applications are becoming increasingly available[34-36]. Several excellent reviews have been released on the topic[34-36], and Chapter 6 of this report will cover this in detail.

Lipidomics is often considered its own 'omic' despite ultimately existing under the umbrella of metabolomics. Lipidomics is the specific study of the lipid profile of a cell or tissue. Measured by mass spectrometry, there is remarkable diversity in the lipidome, and it has also recently begun to be measured at the single-cell level[37]. **Glycomics** is in a similar boat. It is the specific study of the glycans, such as the cell surface glycoproteins and glycolipids, which are crucial for cell-cell recognition and are important disease markers. Glycomics is also assessed via metabolomics tools, namely mass spectrometry and high-performance liquid chromatography (HPLC)[38].

FIGURE 1.3. THE PRINCIPLE MULTI-OMICS LAYERS WHEN CONSIDERING HUMAN HEALTH AND BIOMEDICAL SCIENCE.

Image credit: Sun and Hu[39]



What is being profiled? - Extra-Omic Inclusions

Multi-omics, in its purest definition, could be restricted to the five modalities covered above. These are the modalities representing each component of the genomic processes within cells. DNA is the blueprint. The epigenome represents the long-term, mostly modifiable elements that co-ordinate genome expression. RNA is the first stage of gene expression and a modifying agent in its own right. Proteins are the biological machinery of the cell, the major output of the genome. Finally, metabolites are the output of all these processes,

However, there are other elements that we can measure in a multimodal approach that still add valuable information to the tale. Examples of these are below.

MICROBIOME

The human body is comprised of an impressive diversity of cell types. Even more impressive is the diversity of micro-organisms hosted within. It is now well known that these microbes are not hitchhikers catching a ride; they are integral parts of the human ecosystem, synthesising molecules that can enter our bloodstream and even our brain, ultimately modulating many processes within our body (check out parts 1 and 2 of the our coverage of the topic). Ultimately, this cluster of organisms represents a second genome to interrogate, meaning metagenomics, meta-transcriptomics, meta-proteomics and meta-metabolomics are all currently utilised to understand microbial diversity, function and activity in the same way that multi-omics is deployed for our own cells[40].

EXPOSOME

While perhaps not an 'omic', the exposome relates to all exposures an individual has had in their life that could influence their biology and lead them to this cellular phenotype[41,42]. This is particularly relevant for disease-based studies and ultimately is about incorporating patient data into multi-omics studies as an additional layer of information. The exposome could refer to exposure to air pollution, cigarette smoke, alcohol, stress, green spaces, exercise, phenols, phthalates, minerals, pesticides, for example.

PHENOME

Phenomics is the study of the phenotype of an individual. At first, this may seem a waste to define, since phenotype is at the core of every study. For example, does the individual have the disease or not? However, detailed phenotypic data is not necessarily abundantly available and given the wealth of omics data coming from the pipeline, there is now a need to produce the phenotypic data to match. This is the heart of phenomics[43].

A preview of what's to come

Before we dive deep into the multi-omics-verse, we would like to end this first chapter with a collection of responses from our contributors to the question – **What are some of the latest/most exciting things happening in multi-omics?**

While the responses below are varied, highlighting how wide-ranging and exciting this field currently is, there are some consistent trends, such as the power of machine learning, the emergence of single-cell and spatial metabolomics and the proliferation of new methods for multi-omic pairings. For all of these topics and more, you will find coverage within this playbook.

NIMA AGHAEEPOUR

Associate Professor, Anaesthesiology, Perioperative and Pain Medicine & Paediatrics - Neonatal and Developmental Medicine, **Stanford University**

FLG: What are some of the latest and exciting things happening in multi-omics?

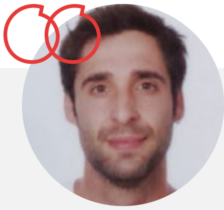
Nima: *I think foundational models are becoming more and more popular these days. The promise that they offer is that you don't have to be limited to the cohort that you paid for. You can start learning from public datasets and building large models that understand relationships between various omics datasets. So, when you have your own question, your machine learning algorithm doesn't have to start from scratch. It can bring all the knowledge that it has gained from public datasets to your multi-omics assay that increases the predictive power and reduces the number of patients that you're going to need to measure from.*

MIRJANA EFREMOVA

Group Leader
Barts Cancer Institute

FLG: What are some of the latest/exciting things happening in the multi-omics field?

Mirjana: *Single-cell multi-omics methods, by providing a holistic view of cells in health and disease, are revolutionising molecular cell biology research. I am excited about using transcriptomics in combination with chromatin accessibility, histone modifications and nucleosome organization to elucidate epigenetic processes involved in cancer progression. Spatial data, in addition, will be crucial for interrogating cell-cell communication networks and identifying the signals from the microenvironment that mediate or sustain specific cancer cell states.*



RICARD ARGELAGUET

Senior Research Scientist

Altos Labs

FLG: What are some of the latest and exciting things that you've seen happening in multi omics?

Ricard: *There are two or three things that I think are pushing quite hard now. One of them is mosaic data integration for gene regulatory networks, which is this idea that there are these gene regulatory network models that you can train and apply on one data set, and you get an answer. But they don't leverage any of the knowledge that is out there. So, models that are able to integrate all of this data across multiple resources to refine the prediction in a specific data set, these are going to be very powerful methods. And they're starting to come up now.*

Then along the same lines, there have been these foundational models all inspired by the GPT models etc. This is really pushing from the deep learning community, where they are trying to build these models on single-cell RNA-seq data, by training and leveraging millions of cells from many different data sources. And the same is being done right now for chromatin accessibility. When it comes to building models that bridge across different omics, I've not seen any, but I'm sure that people are actively working on these.

MATHEW CHAMBERLAIN

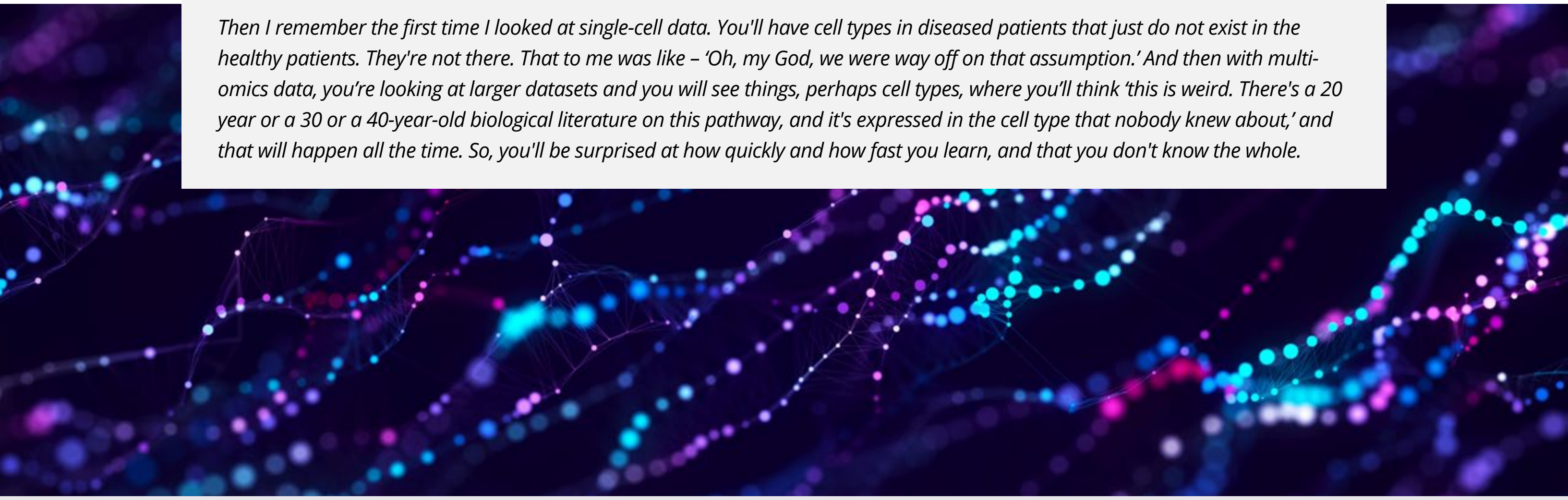
Principal Scientist

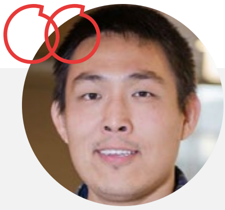
Johnson & Johnson Innovative Medicine

FLG: Are there any single-cell multi omics topics that really interest you? Maybe a technology or application that has caught your eye?

Mathew: *A lot of the data integration methods for CITE-seq catch my eye because, essentially, you can build machine learning models from these CITE-seq atlases that you're integrating that can predict protein levels from any single-cell RNA-seq dataset during integration. You can get CITE-seq data pretty much for free. I think a really clever way to utilise that is to run CITE-seq for a handful of proteins, and then just use the models to generate predictions. From that you get values for hundreds of proteins for free. That's something that I believe will be very useful.*

Then I remember the first time I looked at single-cell data. You'll have cell types in diseased patients that just do not exist in the healthy patients. They're not there. That to me was like – 'Oh, my God, we were way off on that assumption.' And then with multi-omics data, you're looking at larger datasets and you will see things, perhaps cell types, where you'll think 'this is weird. There's a 20 year or a 30 or a 40-year-old biological literature on this pathway, and it's expressed in the cell type that nobody knew about,' and that will happen all the time. So, you'll be surprised at how quickly and how fast you learn, and that you don't know the whole.





QIN MA

Professor, Department of Biomedical Informatics
The Ohio State University

FLG: When you're looking at the multi-omics field and the way things are progressing, what are some of the latest and exciting things that you're seeing in the field?

Qin: *I think the most exciting part is how close we are to real translational and clinical impact. In the early stage of my bioinformatics career, many people were focusing on gene finding, protein structure identification and gene regulatory network inference at the bulk level. Now, using single-cell data, we can mimic the real biological system, and that has to be done at the single-cell level. Whether that is the gene activity, protein activity or gene regulatory network activity, we have to consider them at the single-cell level, because that's the real case, that's not simulation. Having the real case and having the computational techniques derived to make sense from the big single-cell data... I think we are very close to the to the endpoint of clinical and translational applications.*

ZONGMING MA

Professor, Department of Statistics and Data Science
Yale University

FLG: What exciting things are happening in multi-omics?

Zongming: *What is very exciting to me is, if you just browse journals that cover multi-omics, you see new technologies measuring new modalities coming out on, say, a weekly basis. The landscape of this multi-omics world is changing rapidly and in a good way; you have simultaneous measurements of many different types of biomarkers available now. Using bimodal measurement technologies to understand the connection of new modality with older ones, and then, based on that understanding, creating new tools to eventually perform the aggregation and integration of many different modalities together, would be something remarkable. I personally think this is a very exciting direction to work on.*

JUDITH ZAUGG

Group Leader
European Molecular Biology Laboratory (EMBL)

FLG: Is there anything that's caught your eye in this field over the past year?

Judith: *What I find interesting is the increasing number of deep learning networks and machine learning models that are now producing very interpretable results and very strong predictions. Interpretable models, where you can actually interpret what drives the predictions. And I think this is going to be very powerful when applied to multi-omics datasets that are increasingly getting published. I think these models will become very powerful in understanding fundamental biological mechanisms, and potentially disease mechanisms.*

PAU BADIA I MOMPEL

PhD Candidate, Saez-Rodriguez Group
Heidelberg University

FLG: What are the exciting things happening in multi-omics right now?

Pau: *For starters, the whole explosion of methods that combined chromatin accessibility and transcriptomics for GRN inference, which we believe is going to provide a better representation of GRNs. It will trim down a lot of these false positives. The other thing would be the technologies that are starting to appear. I saw one from Satija et al.[44], where they combine phosphoproteomics with chromatin accessibility, although for now it's still antibody based targeted phosphoproteomics - it's not mass spectrometry. It would be really cool to combine transcriptomics, chromatin accessibility plus the functional state of your transcription factors based on phosphoproteomics for better GRN reconstruction.*

SUSHMITA ROY

Professor, Department of Biostatistics and Medical Informatics, **University of Wisconsin-Madison,** Faculty, **Wisconsin Institute of Discovery**

FLG: What have you seen in your field over the last six to nine months that excites you?

Sushmita: *There are several things. Going beyond single-cell to spatial is an interesting direction, and not only spatial transcriptomics, but spatial epigenomics, and also spatial proteomics and metabolomics. Those would be other types of modalities to incorporate, and I would really like to move in that direction. Single-cell proteomics is becoming more widely used, but it's not at the same production scale as single-cell RNA-seq. Hopefully, we will get there. I know a lot of people are excited about it. On the data side, getting better datasets from multimodal single-cell, and also spatial datasets, is very exciting.*

On the methodology side of things, really getting into models that tell us something about perturbations. How will a system actually behave when we perturb it in a particular way? How well can we predict that from just very minimalistic data? Can we build models that enable us to figure out what the minimum things we need to measure in order to make high level and accurate predictions? When we go to patient samples, we can't really measure so much data, so what can we do based on what can we predict, and so on? Those are some of the exciting directions where you get into causal models and causal representation learning, but it's very new.

SUHAS VASAIKAR

Principal Scientist, Clinical Biomarker and Diagnostics
Seattle Genetics (Seagen)

FLG: From your experience, what are some of the latest/exciting things happening in the multi-omics field?

Suhas: *There are several exciting developments happening in the field of multi-omics that have the potential to transform our understanding of complex biological systems. Here are some examples:*

Integration of single-cell data: Single-cell omics technologies are rapidly advancing, and researchers are now able to generate multi-omics data from individual cells. This has the potential to provide unprecedented insights into cellular heterogeneity, cell-to-cell communication and disease mechanisms.

Multi-omics data visualization: As the amount of multi-omics data being generated continues to increase, there is a growing need for effective data visualization tools. New visualization methods, such as interactive network-based visualization platforms, are now being developed to help researchers gain insights from complex multi-omics data sets.

Multi-omics biomarker discovery: Integrating data from multiple omics technologies can help identify biomarkers that are more accurate and reliable than those identified using a single technology. These biomarkers can be used for disease diagnosis, prognosis and treatment.

Deep learning approaches: Deep learning approaches, such as deep neural networks and convolutional neural networks, are now being applied to multi-omics data sets to identify complex patterns and relationships between different omics data types. These methods have the potential to reveal new insights into disease mechanisms and identify novel therapeutic targets.

Overall, the integration of multiple omics data types, along with advances in single-cell omics, data visualization, biomarker discovery and deep learning, are driving exciting developments in the field of multi-omics research.

BINGJIE ZHANG

Postdoctoral Research Fellow, Satija Lab
New York Genome Center

FLG: What's really exciting right now in the multi-omics space?

Bingjie: *I'm really excited about the metabolomics profiling methods. It's an essential missing part in our current field. Although still in its early stages, I have seen some work from Andrew Fraser's group, where they are using structure-switching aptamers to capture metabolites45. Currently, they only tested this at the bulk level with a few targets, but I am very excited about the potential for this to be used in single-cell experiments and to simultaneously profile hundreds of metabolites.*

Chapter 1 references

1. Roychowdhury, R. *et al.* **Multi-Omics Pipeline and Omics-Integration Approach to Decipher Plant's Abiotic Stress Tolerance Responses.** *Genes* **14**, 1281 (2023).
2. Efremova, M. & Teichmann, S.A. **Computational methods for single-cell omics across modalities.** *Nature Methods* **17**, 14-17 (2020).
3. Argelaguet, R., Cuomo, A.S.E., Stegle, O. & Marioni, J.C. **Computational principles and challenges in single-cell data integration.** *Nature Biotechnology* **39**, 1202-1215 (2021).
4. Ogbeide, S., Giannese, F., Mincarelli, L. & Macaulay, I.C. **Into the multiverse: advances in single-cell multiomic profiling.** *Trends in Genetics* **38**, 831-843 (2022).
5. **Method of the Year 2019: Single-cell multimodal omics.** *Nature Methods* **17**, 1-1 (2020).
6. Schier, A.F. **Single-cell biology: beyond the sum of its parts.** *Nature Methods* **17**, 17-20 (2020).
7. Zhu, C., Preissl, S. & Ren, B. **Single-cell multimodal omics: the power of many.** *Nature Methods* **17**, 11-14 (2020).
8. Vandereyken, K., Sifrim, A., Thienpont, B. & Voet, T. **Methods and applications for single-cell and spatial multi-omics.** *Nature Reviews Genetics,* 1-22 (2023).
9. Baysoy, A., Bai, Z., Satija, R. & Fan, R. **The technological landscape and applications of single-cell multi-omics.** Nature *Reviews Molecular Cell Biology,* 1-19 (2023).
10. Li, X. **Harnessing the potential of spatial multiomics: a timely opportunity.** *Signal Transduction and Targeted Therapy* **8**, 234 (2023).
11. Liu, Y. *et al.* **High-plex protein and whole transcriptome co-mapping at cellular resolution with spatial CITE-seq.** *Nature Biotechnology* (2023).
12. Liu, Y. *et al.* **High-Spatial-Resolution Multi-Omics Sequencing via Deterministic Barcoding in Tissue.** *Cell* **183**, 1665-1681.e18 (2020).
13. Vickovic, S. *et al.* **SM-Omics is an automated platform for high-throughput spatial multi-omics.** *Nature Communications* **13**, 795 (2022).
14. Su, J.-H., Zheng, P., Kinrot, S.S., Bintu, B. & Zhuang, X. **Genome-scale imaging of the 3D organization and transcriptional activity of chromatin.** *Cell* **182**, 1641-1659. e26 (2020).
15. Zhang, D. *et al.* **Spatial epigenome–transcriptome co-profiling of mammalian tissues.** *Nature* **616**, 113-122 (2023).
16. Hsieh, W.-C. *et al.* **Spatial multi-omics analyses of the tumor immune microenvironment.** *Journal of Biomedical Science* **29**, 96 (2022).
17. Lee, R.Y. *et al.* **The promise and challenge of spatial omics in dissecting tumour microenvironment and the role of AI.** *Frontiers in Oncology* **13**, 1172314 (2023).
18. Kruglyak, L. & Nickerson, D.A. **Variation is the spice of life.** *Nature genetics* **27**, 234-236 (2001).
19. Akiyama, M. **Multi-omics study for interpretation of genome-wide association study.** *Journal of Human Genetics* **66**, 3-10 (2021).
20. Moore, L.D., Le, T. & Fan, G. **DNA methylation and its basic function.** *Neuropsychopharmacology* **38**, 23-38 (2013).
21. Mulqueen, R.M. *et al.* **Highly scalable generation of DNA methylation profiles in single cells.** *Nat Biotechnol* **36**, 428-431 (2018).
22. Farlik, M. *et al.* **Single-cell DNA methylome sequencing and bioinformatic inference of epigenomic cell-state dynamics.** *Cell Rep* **10**, 1386-97 (2015).
23. Rotem, A. *et al.* **Single-cell ChIP-seq reveals cell subpopulations defined by chromatin state.** *Nature biotechnology* **33**, 1165-1172 (2015).
24. Bartosovic, M., Kabbe, M. & Castelo-Branco, G. **Single-cell CUT&Tag profiles histone modifications and transcription factors in complex tissues.** *Nature biotechnology* **39**, 825-835 (2021).
25. Buenrostro, J.D. *et al.* **Single-cell chromatin accessibility reveals principles of regulatory variation.** *Nature* **523**, 486-490 (2015).
26. Nagano, T. *et al.* **Single-cell Hi-C reveals cell-to-cell variability in chromosome structure.** *Nature* **502**, 59-64 (2013).
27. Jovic, D. et al. **Single-cell RNA sequencing technologies and applications: A brief overview.** *Clinical and Translational Medicine* **12**, e694 (2022).
28. Fletcher, M. **The complex relationship between cell transcriptomic state and phenotype.** *Nature Genetics* **55,** 1421-1421 (2023).
29. Fernández-Chacón, M. *et al.* **Incongruence between transcriptional and vascular pathophysiological cell states.** *Nature Cardiovascular Research* **2**, 530-549 (2023).
30. Brion, C., Lutz, S.M. & Albert, F.W. **Simultaneous quantification of mRNA and protein in single cells reveals post-transcriptional effects of genetic variation.** *Elife* **9**(2020).
31. Li, J., Zhang, Y., Yang, C. & Rong, R. **Discrepant mRNA and Protein Expression in Immune Cells.** *Curr Genomics* **21**, 560-563 (2020).
32. Bennett, H.M., Stephenson, W., Rose, C.M. & Darmanis, S. **Single-cell proteomics enabled by next-generation sequencing or mass spectrometry.** *Nature Methods* **20**, 363-374 (2023).
33. Mund, A., Brunner, A.-D. & Mann, M. **Unbiased spatial proteomics with single-cell resolution in tissues.** *Molecular Cell* **82**, 2335-2349 (2022).
34. Lanekoff, I., Sharma, V.V. & Marques, C. **Single-cell metabolomics: where are we and where are we going?** *Current opinion in biotechnology* **75**, 102693 (2022).
35. Saunders, K.D., Lewis, H.-M., Beste, D.J., Cexus, O. & Bailey, M.J. **Spatial single cell metabolomics: Current challenges and future developments.** *Current opinion in chemical biology* **75**, 102327 (2023).
36. Zhang, C., Le Dévédec, S.E., Ali, A. & Hankemeier, T. **Single-cell metabolomics by mass spectrometry: ready for primetime?** *Current Opinion in Biotechnology* **82**, 102963 (2023).
37. Li, Z. *et al.* **Single-cell lipidomics with high structural specificity by mass spectrometry.** *Nature Communications* **12**, 2869 (2021).
38. Ma, X. & Fernández, F.M. **Advances in mass spectrometry imaging for spatial cancer metabolomics.** *Mass Spectrometry Reviews,* e21804 (2022).
39. Sun, Y.V. & Hu, Y.-J. **Chapter Three - Integrative Analysis of Multi-omics Data for Discovery and Functional Studies of Complex Human Diseases.** *in Advances in Genetics,* Vol. 93 (eds. Friedmann, T., Dunlap, J.C. & Goodwin, S.F.) 147-190 (Academic Press, 2016).
40. Velten, B. *et al.* **Identifying temporal and spatial patterns of variation from multimodal data using MEFISTO.** *Nature Methods* **19**, 179-186 (2022).
41. Krausová, M. *et al.* **Understanding the chemical exposome during fetal development and early childhood: a review.** *Annual Review of Pharmacology and Toxicology* **63**, 517-540 (2023).
42. Maitre, L. *et al.* **Multi-omics signatures of the human early life exposome.** *Nature Communications* **13**, 7024 (2022).
43. Ying, W. **Phenomic studies on diseases: potential and challenges.** Phenomics, 1-15 (2023).
44. Blair, J.D. *et al.* **Phospho-seq: Integrated, multi-modal profiling of intracellular protein dynamics in single cells.** *bioRxiv* (2023).
45. Tan, J.H., Mercado, M.P. & Fraser, A.G. **A novel platform for metabolomics using barcoded structure-switching aptamers.** *bioRxiv,* 2023.06.09.544402 (2023).

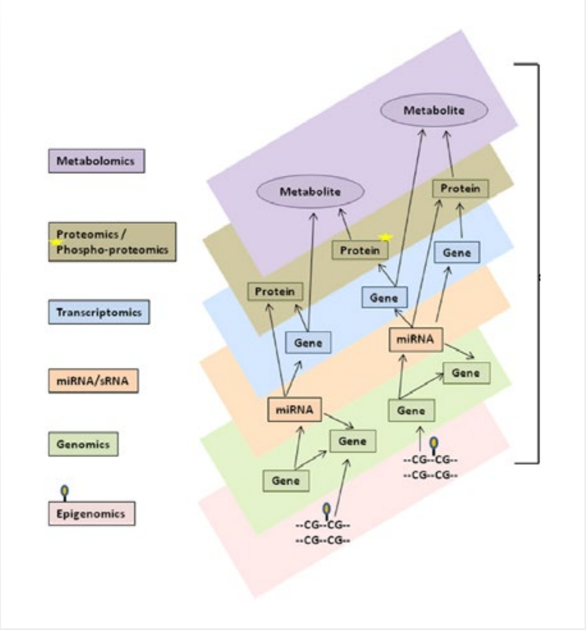
CHAPTER 2

BRINGING OMICS TOGETHER. INTEGRATING MULTI-OMICS DATA

THE PRINCIPAL CHALLENGE IN MULTI-OMICS IS INTEGRATING THE DATA FROM DIFFERENT OMES IN A BIOLOGICALLY MEANINGFUL WAY. IN THIS CHAPTER, WE WILL LOOK AT DIFFERENT COMPUTATIONAL TOOLS AND SERVICES TO INTEGRATE MATCHED AND UNMATCHED MULTI-OMICS DATA.

FIGURE 2.1. INTEGRATION OF MULTI-OMICS ACROSS THE MAJOR OMICS LAYERS.

Image Credit: Hossain, et al.[1]



Intro to integration

While each omic provides valuable data alone, in concert, new and valuable insights can be gained. Integrating multi-omics data can reveal new cell subtypes, cell interactions and interactions between the different omic layers leading to gene regulatory and phenotypic outcomes. Since each omic layer is causally tied to the next, multi-omics integration serves to disentangle this relationship to properly capture cell phenotype (Figure 2.1).

We are now in the era of acquiring data from millions of cells. The integration of this large, complex, multimodal data has the potential to reveal much about biological mechanisms and pathways, but this represents a considerable challenge to researchers[2]. Ultimately, sophisticated computational tools and methodologies are at hand to solve this issue.

But which ones to use? The principle focus of this chapter will be to outline the selection of methodologies that are available to the reader as of the writing of this report (November 2023). We will cover the different types of integration and the options available. We will also hear from several developers of these tools as they describe their innovations and, fundamentally, why multi-omics integration is a technique you have to get right for a successful trip into the Multi-omics-verse.

Why is integration still a challenge?

Ultimately, integration of multi-omics data is a moving target for which a one-size-fits-all approach will not work. Drawing insights from two specific omics requires unique strategies, since each omic has a unique data scale, noise ratio and, hence, its own preprocessing steps. Furthermore, these omics are not captured with the same breadth, meaning there is inevitably missing data. Specific tools are required for specific challenges, hence the variety of tools available. We asked some of experts in the field to explain why multi-omics integration is still such a challenge and whether there were any tools they recommend.

BINGJIE ZHANG

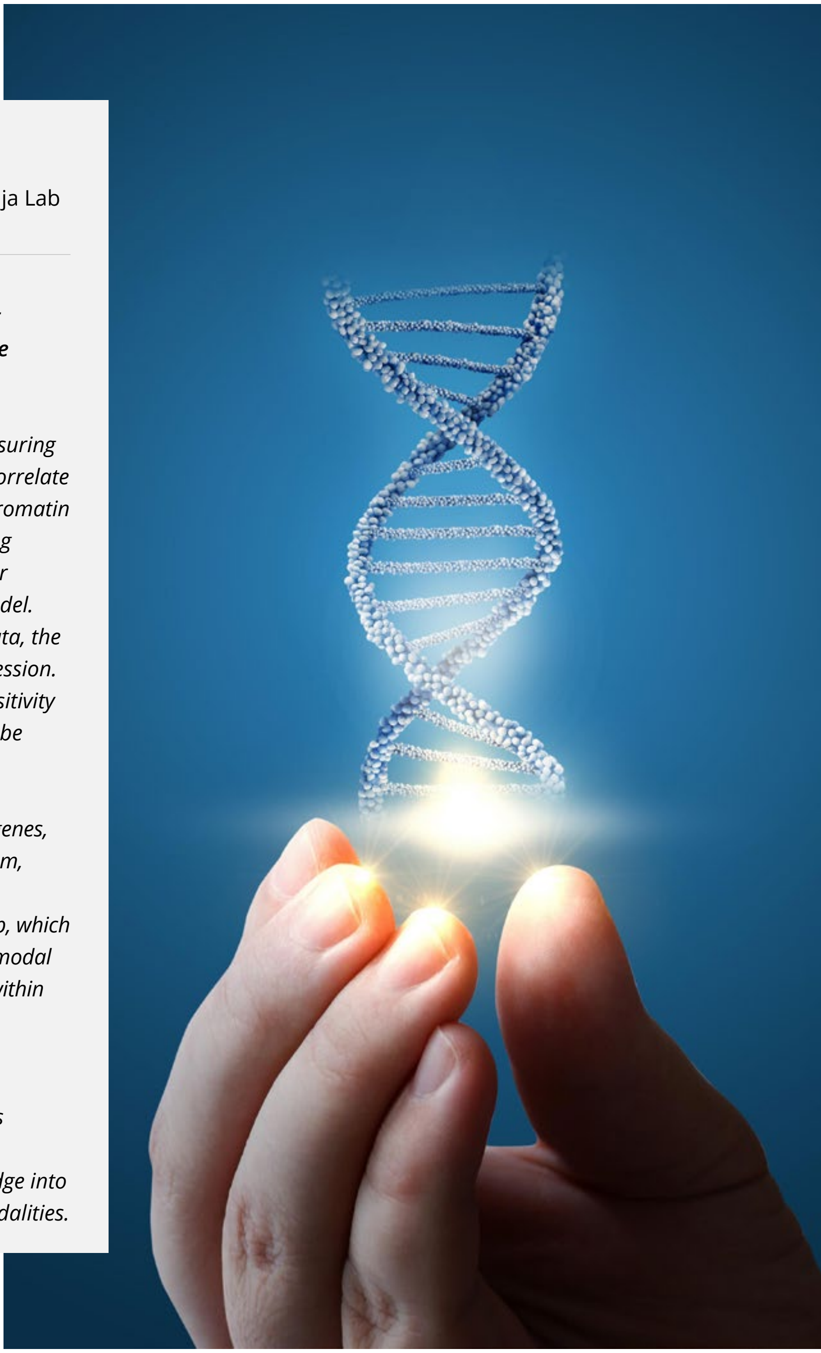
Postdoctoral Research Fellow, Satija Lab
New York Genome Center

FLG: Why do you think multi-omics integration is still such a challenge? And are there any approaches to integration that you particularly like (I know your lab recently released bridge integration)?

Bingjie: *Integration is challenging simply because we are measuring two different modalities, and our understanding of how they correlate with each other is not very clear. Mapping RNA-seq to open chromatin accessibility might be relatively easier because of the underlying assumption that actively transcribed genes should have greater open chromatin accessibility, which is a correlation we can model. However, for other modalities, such as RNA-seq and protein data, the most abundant protein may not correlate with high gene expression. This disconnect makes integration very difficult. Moreover, sensitivity remains an issue. A gene detected at the RNA level may simply be missing in the ATAC dataset.*

In terms of the scale, while scRNA-seq can profile thousands of genes, current proteomic methods may only measure a limited spectrum, often restricted to hundreds of proteins. Our solution is bridge integration[3]*, a method developed by Yuhan Hao in the Satija Lab, which utilizes a bridge dataset as a 'dictionary.' This dataset is a multimodal dataset that measures the two modalities we wish to integrate within the same cell, allowing us to use it as a biological "translator" to establish connections between those two modalities.*

As for other methods, I would recommend GLUE from Ge Gao's group[4]*. They employ an entirely different method from ours, incorporating graph-based methods to integrate prior knowledge into the model, thereby inferring a connection between the two modalities.*



ZONGMING MA

Professor, Department of Statistics and Data Science
Yale University

FLG: Can you briefly outline multi-omics integration and why it's still a problem to this day?

Zongming: *When we talk about multi-omics data integration, there are at least two different scenarios. One scenario is, when you can take multi-omics measurements directly at a single-cell level or close to single-cell resolution. Here the question is how you make sense of two different collections of features, which are, in many senses, not directly comparable. So, that is one type of multi-omics data integration, oftentimes called "vertical integration".*

The other scenario is when, for each cell, you can only measure one modality, say only RNA (gene expression levels) or only some proteins, or only epigenome information. The question then is, how can you combine different modalities from different cells? The aim here is to perform "diagonal integration" so that you can augment the set of measured features for each cell to the union of all modalities to be integrated. Personally, I think this is a more challenging multi-omics data integration scenario.

FLG: And is the integration challenge different for different omics? For instance, for RNA with protein or RNA with any combination of omics, is the challenge different?

Zongming: *There are some shared challenges and there are also unique ones, depending on the kinds of features one is measuring. If you want to integrate ATAC, or epigenome, information, with the whole transcriptome, it is relatively easy. In this setting, you can predict the expression levels of many genes simultaneously based on epigenome information. The prediction of each individual gene may not be super-accurate, but collectively, the cumulative information across many genes in such a prediction is usually sufficiently high for warranting a decent integration outcome.*

However, the challenge becomes quite different when you have a small, targeted panel of proteins together with the whole transcriptome. Even if you can predict the level of each protein relatively accurately, the limited number of features in the protein dataset means the cross-modality cell-cell similarity is more difficult to measure now, because the cumulative information across all features in at least one of the involved datasets is limited.

In both examples, it is of importance to improve the prediction accuracy of individual features. However, in the second one, the key is to improve the overall signal-to-noise ratio when measuring cross-modality cell-cell similarity.

FLG: What are some of the tools for multi-omics data integration that people should be paying attention to, perhaps looking at different types of integration?

Zongming: *I'd like to mention something for horizontal integration, that is the integration of multiple datasets with the same feature set. It could be multi-omics or just a single data modality. In this setting, if you have samples that are collected under similar scenarios, let's say it's all healthy samples and there's homogeneity among them, then methods like Seurat[5], Harmony[6], and LIGER[7] are very capable.*

However, it has become increasingly common in lab experiments and clinical trials that you have some healthy or control sample and you also have samples at different disease stages or under different treatments. If you need to perform horizontal integration, or sometimes also known as batch integration, on samples collected under such heterogeneous scenarios, these tools can be a bit too heavy-handed and can remove biologically meaningful signals. It is not the fault of these tools themselves, as they were not designed to leverage the distinction between control and treatment sample. So, Nancy Zhang at Penn and I have done some recent work trying to recover the biological signal while retaining a good integration result for batch correction, and we developed a new tool called CellANOVA[8].

SUHAS VASAIKAR

Principal Scientist, Clinical Biomarker and Diagnostics
Seattle Genetics (Seagen)

FLG: For our readers, can you briefly outline the challenge of multi-omics data integration, and, in your opinion, what are some of the tools/approaches for multi-omics data integration that people should be paying attention to?

Suhas: *Multi-omics data integration refers to the process of combining and analysing data from different types of omics technologies, such as genomics, transcriptomics, proteomics and metabolomics. The challenge of multi-omics data integration is that each omics data type provides a different perspective on the biological system being studied, and integrating these different types of data can be difficult due to differences in data structure, resolution and complexity.*

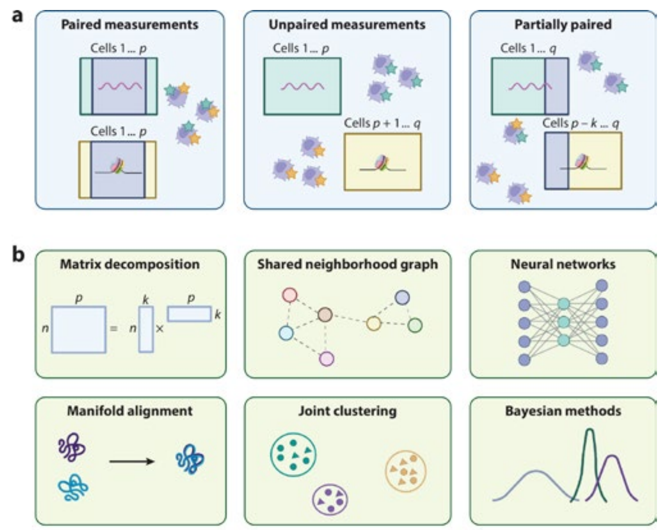
The challenge of multi-omics data integration lies in the complexity of the data, as well as the need to develop methods for integrating data across different omics platforms. There are several tools and approaches for multi-omics data integration that people should be paying attention to. Some popular tools for multi-omics data integration include:

- ***Data normalization and preprocessing tools**, such as limma, DESeq2 and edgeR, that can standardize different types of omics data and make them comparable.*
- ***Network-based analysis tools**, such as Cytoscape, that can visualize and analyse complex relationships between different types of omics data. It involves constructing molecular interaction networks from the different omics data sets and using these networks to identify key regulatory pathways and biological processes.*
- ***Machine learning and statistical modelling tools**, such as Random Forest, Gradient Boosting and Bayesian networks, which involve training algorithms to identify patterns in the multi-omics data and make predictions about biological outcomes that can identify patterns and predict outcomes based on multi-omics data.*
- ***Pathway and enrichment analysis tools**, such as GSEA and GOseq, that can identify enriched pathways and functional categories based on multi-omics data.*
- ***Some specific tools for multi-omics data integration** include Multi-Omics Factor Analysis (MOFA), Integrative Multi-Omics Analysis (IMOA) and Multi-Omics Correlation Analysis (MOCA).*

Overall, the field of multi-omics data integration is rapidly evolving, and researchers should pay attention to new computational tools and approaches that can help them extract meaningful insights from their data.

FIGURE 2.2. COMPUTATIONAL STRATEGIES FOR SINGLE-CELL MULTI-OMICS INTEGRATION

*(**A**) Levels of input data – paired, partially paired and unpaired. (**B**) Techniques deployed in integration. Image Credit: Flynn, et al.*[10]



Flynn E, et al. 2023
Annu. Rev. Biomed. Data Sci. 6:313–37

Types of Integration

When trying to divide up the computational tools for integration meaningfully, one principle distinction between strategies is whether the tool is designed for multi-omics data that is matched (recorded from the same cell) or unmatched (recorded from different cells). While modern methods are frequently able to create the more desirable, matched multi-omics data, much of the previous work in this area has worked on integrating unmatched data. Furthermore, not only can unmatched data refer to different cells from the same sample, but it can also involve integrating different cells from different samples of the same tissue taken at different times in different experiments (see Figure 2.2A).

Integration can also be seen as operating at the horizontal, vertical and diagonal level[9]. Horizontal integration is the merging of the same omic across multiple datasets. Several tools exist for this purpose and while it is technically a form of integration, it is not true multi-omics integration and so won't be considered further in this chapter.

Vertical integration merges data from different omics within the same set of samples, essentially equivalent to matched integration. The cell is leveraged as the anchor to bring these omics together.

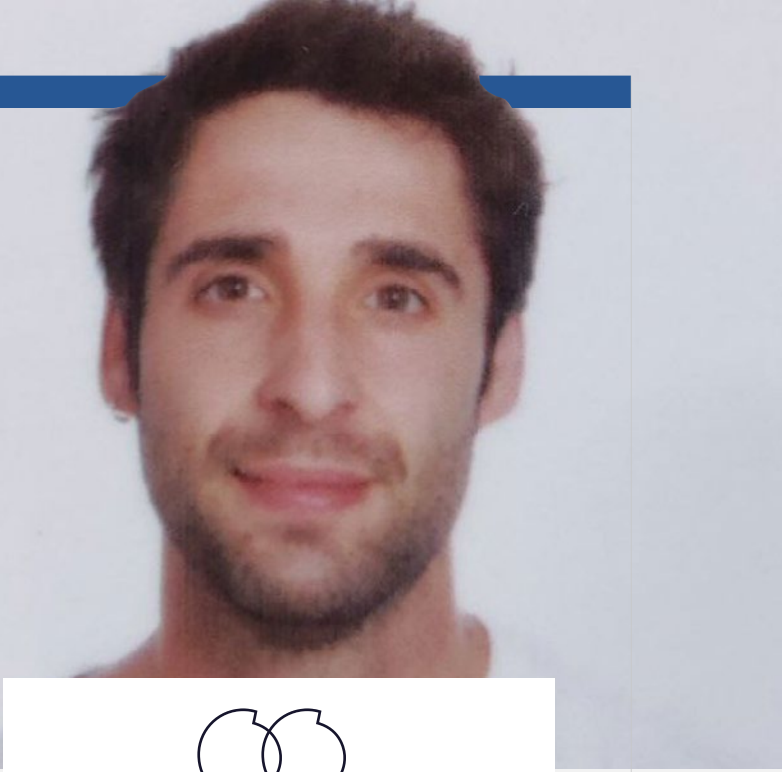
Diagonal integration is the final, and technically most challenging, form of integration. Here, different omics from different cells/different studies are brought together. The anchor can no longer be the cell and has to be some co-embedded space in which commonality between cells is found. This is essentially 'unmatched' integration.

We spoke to **Ricard Argelaguet**, lead author of the paper defining the terms horizontal, vertical and diagonal integration, about his perspective on multi-omics integration[9].

INTERVIEW: RICARD ARGELAGUET

SENIOR RESEARCH SCIENTIST

ALTOS LABS



FLG: ***Could you briefly introduce yourself, some of your research background, and then a summary of your current role.***

Ricard: I'm curarently working as a Senior Research Scientist in the machine learning group at Altos labs. I initially did my PhD in the European Bioinformatics Institute in Cambridge, UK, where I was working on the development and the application of multi-omics data integration methods. That's where I developed some of the models that are used currently, like MOFA, MEPHISTO and so on. I'm a biologist by training and I worked a lot on applying those methods into biological scenarios, more specifically, in the context of single-cell multi-omics datasets. For example, in the context of embryonic development, gastrulation. Now, interestingly, we're trying to shift all this knowledge that we've gathered working on this process and are applying it in the context of reprogramming, ageing and longevity, etc.

FLG: ***For our readers, can you just briefly outline the challenge of multi-omics integration and why it's still a challenge to this day?***

Ricard: I think that multi-omics data integration is a diverse set of challenges, and that's what makes it a unique problem. There are challenges first on the biological side; if you want to develop methods for multi-omics integration, what question do you want to answer? What useful biology can you learn from those methods? Doing methods for the sake of doing methods... I don't find this particularly useful. You need to have some biological question in mind.

> "WHAT USEFUL BIOLOGY CAN YOU LEARN FROM THOSE METHODS? DOING METHODS FOR THE SAKE OF DOING METHODS... I DON'T FIND THIS PARTICULARLY USEFUL. YOU NEED TO HAVE SOME BIOLOGICAL QUESTION IN MIND."

Second, there is the statistical complexity. Each one of those data modalities is fairly heterogeneous, and they have different statistical properties. Some datasets are binary, some datasets are continuous, and you have to model them under different data distributions, statistical assumptions, etc. Over the years we have developed methods to analyse each one of them separately, but when it comes to the integration, you need to make sure that your model is able to accurately model the different statistical principles for each of the datasets.

Finally, there is the computational challenge, because integrating this data is complex. First, you need all these different datasets stored in a meaningful way with the right metadata. You also want to make sure that your model is easy to use, runnable and useful. I think that's something we did quite well with MOFA. There were quite a few methods already back then, but they were not scalable, and they were hard to use. In the end, as a user, you want to have an API interface that is usable.

***FLG:* Can you also describe the terms that you set out in your 2021 paper for multi-omics integration, namely horizontal, vertical and diagonal integration and perhaps how your tools fit within that framework?**

Ricard: So, multi-omics integration doesn't refer to just one problem. It actually refers to different types of problems, and it depends on the experimental design that you have. I'll give you an example: If you have an experiment where you profile different omics from the same group of cells or the same group of samples, you know that this data is paired. You can match that dataset even though the features that you're measuring are completely different because the samples are exactly the same. This is really powerful because it gives you an anchor, it gives you something to anchor the different data modalities. It makes the modelling much easier and what you can learn from the data much more meaningful. We call this vertical integration, and this is the case for MOFA. The samples are assumed to be matching across the different assays.

Now we have other cases known as horizontal integration, which is almost the reverse problem. The features are exactly the same, but the samples are different. This could be an example where you're just measuring one data modality, like RNA-seq, in ten groups of samples in hospital one, and ten groups of samples in hospital two. Here, the genes are the same and you're going to anchor the different data matrices by using this gene space.

And finally, the more complex problem, which are the ones that people are more actively trying to address now, is what we call diagonal integration, or even mosaic integration. In this case, you don't have this pairedness anymore. In the case of diagonal, your samples are different, your features are different, and this makes it very, very challenging. It limits the amount of knowledge you can extract from this case. So, technically, you want to avoid this type of experimental design. Sometimes it's not avoidable, and that's why we have to develop methods for this. The last case is mosaic integration, where you really have a combination of matrices where some samples are paired, some features are paired, and you have to explore these anchors to try and learn as much as you can from the data.

***FLG:* Are there tools and approaches for diagonal or mosaic integration that you think are particularly good?**

Ricard: For diagonal integration, specifically for single-cell data, one of the ones I really like is called GLUE. In this approach, they have RNA-seq on one side and chromatin accessibility on the other side, which are provided from different cells. Then they develop a unifying matrix from which they infer a unifying model, where they assume relationships between the genes and the chromatin accessibility site. They use this extra matrix to anchor the original two matrices. It's a very powerful method with a very sensible application, which is easy to use. So, it's one of my favourite methods for diagonal integration.

In the case of mosaic integration, two methods come to my mind, specifically for single-cell data. One is called a StabMap that was developed by Shila Ghazanfar in the group of John Marioni. And then there's another one, which was called dictionary learning by Rahul Satija's lab in the context of the Seurat method that they have. Here again, the key is always to exploit all of these anchors that you have across the different datasets.

***FLG:* Could you describe your SingleCellMultiModal package the you published in the PLOS Computational paper?**

Ricard: One of the challenges for the multi-omics data field is getting the data in an easy to use format in a self-contained object where you can just query all the different modalities and the different samples. Some resources already do a good job at this. For example, TCGA has thousands of samples with different data modalities, and they have an API that makes it quite the easy to extract the different data that you need.

In the case of single-cell multi-omics data, it wasn't really the case. Instead, you have to download the data, you have to reprocess it from scratch, and there was no easy way of querying and fetching useful data to train your models. So, we tried to do a consistent reprocessing of the data, bringing everything into a self-contained object using the Bioconductor standards, such that any user could just quickly download this data. If they want to reprocess things by themselves, they can always go from the raw data, but at least this provides a processed data object with all the metadata that they can just run their models on.

Below you will find a table detailing many of the available multi-omics integration tools and whether they are for matched or unmatched integration. The following sections of this report will address each of these in turn, before concluding with a look at transfer-learning and spatial integration strategies.

TABLE 2.1. MULTI-OMICS INTEGRATION TOOLS SEPARATED BY MATCHED VS. UNMATCHED INTEGRATION CAPACITY

For each computational tool, the name, year of release, methodology of integration and the omics capacity are noted. This table was adapted from Baysoy, et al.[11]

Year	Name	Methodology	Integration capacity	Ref.
MATCHED INTEGRATION TOOLS (From same single cell)				
2019	SCHEMA	Metric learning-based method	Chromatin accessibility, mRNA, proteins, immunoprofiling, spatial coordinates	12
2020	Seurat v4	Weighted nearest-neighbour	mRNA, spatial coordinates, protein, accessible chromatin, microRNA	13
2021	DCCA	Variational autoencoders	mRNA, chromatin accessibility	14
2021	DeepMAPS	Autoencoder-like neural networks	mRNA, chromatin accessibility, protein	15
2019	citeFUSE	Network-based method	mRNA, protein	16
2020	MOFA+	Factor analysis	mRNA, DNA methylation, chromatin accessibility	17
2020	scMVAE	Variational autoencoder	mRNA, chromatin accessibility	18
2020	totalVI	Deep generative	mRNA, protein	19
2020	BREM-SC	Bayesian mixture model	mRNA, protein	20
2022	SCENIC+	Unsupervised identification model	mRNA, chromatin accessibility	21
2022	FigR	Constrained optimal cell mapping	mRNA, chromatin accessibility	22
2021	MIRA	Probabilistic topic modelling	mRNA, chromatin accessibility	23
2023	CellOracle	Modelling gene regulatory networks	mRNA, CRISPR screening, chromatin accessibility	24
2022	MultiVelo	Probabilistic latent variable model	mRNA, chromatin accessibility	25
UNMATCHED INTEGRATION TOOLS (From different single cells)				
2019	Spectrum	Weighted nearest-neighbour	microRNA, mRNA, protein	26
2020	BindSC	Canonical correlation	mRNA, chromatin accessibility	27
2019	MMD-MA	Manifold alignment	mRNA, chromatin accessibility, DNA methylation, imaging	28
2019	MuSiC	Unsupervised topic modelling	mRNA, CRISPR screening	29
2019	Seurat v3	Canonical correlation analysis	mRNA, chromatin accessibility, protein, spatial	5
2020	UnionCom	Manifold alignment	mRNA, DNA methylation, chromatin accessibility	30
2019	CloneAlign	Statistical method	mRNA, DNA	31
2021	Pamona	Manifold alignment	mRNA, chromatin accessibility	32
2022	GLUE	Variational autoencoders	Chromatin accessibility, DNA methylation, mRNA	4
2019	LIGER	Integrative non-negative matrix factorization	mRNA, DNA methylation	7
2022	StabMap	Mosaic data integration	mRNA, chromatin accessibility	33
2021	Cobolt	Multimodal variational autoencoder	mRNA, chromatin accessibility	34
2021	MultiVI	Probabilistic modelling	mRNA, chromatin accessibility	35
2022	Seurat v5	Bridge integration	mRNA, chromatin accessibility, DNA methylation, protein	3

Matched (Vertical) Integration

Vertical integration methods rely on technologies that profile omics data from two or more distinct modalities from within a single cell. From this position, the cell itself can be used as an anchor by which to integrate varying modalities.

Given that the majority of multi-omics tools either measure RNA and protein concurrently or RNA and epigenomic information (mainly via ATAC-seq) concurrently, the majority of tools for vertical integration focus on these pairs of modalities, as can be seen in Table 2.1.

Of the various approaches, there are matrix factorization methods (e.g., MOFA+[17]), neural network-based (e.g., scMVAE[18], DCCA[14], DeepMAPS[15]) and network-based methods (e.g. cite-Fuse, Seurat v4)[36]. See Figures 2.2B and 2.3 for an overview of the techniques. We will look at a few popular examples.

MOFA+[17] is a well-known approach for matched integration, using Bayesian Group Factor Analysis framework to jointly model variation across covariates.

Seurat has several integration methods including canonical correlation analysis, mutual nearest neighbours and weighted nearest neighbours (WNN)[13]. The latter of these was employed to integrate single-cell RNA and ATAC data and also to integrate RNA and proteomic data from CITE-seq. Due to the low number of proteins captured in multi-omics technology, specific tools such as totalVI[19] and sciPENN[37] are used to correct for these protein profiles.

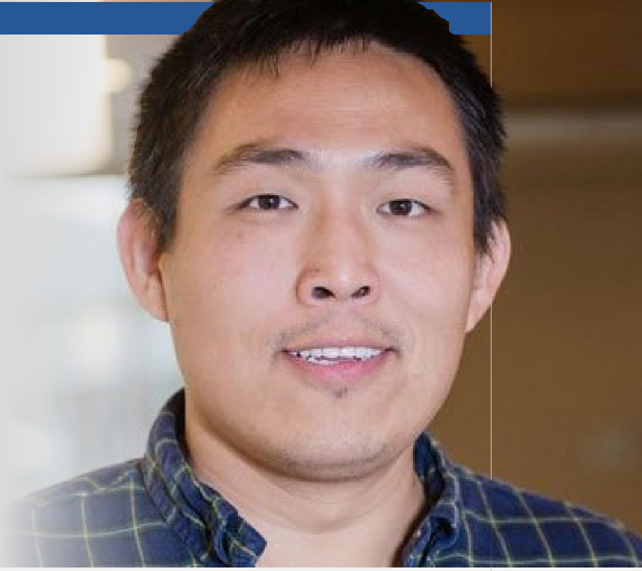
Several tools to integrate chromatin accessibility and mRNA are designed with gene regulatory network analysis in mind (e.g., SCENIC+[21], FigR[22], CellOracle[24]). These tools will be covered in Chapter 5.

Machine learning approaches are proving incredibly useful for looking at integration patterns in the complex multi-omics data. MarsGT[38] and DeepMAPS[15] are two recent examples, based on the successful scGNN[39] deep learning model. DeepMAPS uses a graph transformer neural network architecture. The attention mechanism in this architecture learns relevant omic-omic interaction networks and cell-cell similarities from integrating multi-omics data. We spoke to the senior author of the DeepMAPS publication, **Professor Qin Ma** about multi-omics integration and his group's latest computational tools.

INTERVIEW:

QIN MA

PROFESSOR, DEPARTMENT OF BIOMEDICAL INFORMATICS
THE OHIO STATE UNIVERSITY



***FLG:* Can you just briefly introduce yourself? Give some of your research background and some of your current research interests?**

Qin: I was trained in theoretical mathematics and during my PhD. I fell in love with applied mathematics, specifically in biology and the biomedical field. I did a bioinformatics postdoc and started a lab in 2015, focusing on bioinformatics and mathematical biosciences. Now, I'm a professor in the Department of Biomedical Informatics at The Ohio State University, in the College of Medicine. We are doing computational modelling and data analysis, and we are eager to have a strong clinical and translational impact. We have a single-cell emphasis in our research, using AI and machine learning.

***FLG:* Can you describe some of the computational challenges in multi-omics data analysis that your group is looking at?**

Qin: The challenges for single-cell are, first of all, the large scale. At bulk level, it's hard to accumulate hundreds or thousands of samples by an individual lab, that's too expensive. But at single-cell level, you easily get thousands of cells, even millions of cells, which means the sample size is pretty high and the raw data is huge.

Second is the data heterogeneity. At the bulk level, we take the average. For one cancer patient, you get the tumour tissue and with bulk RNA-seq you get one expression profile for one sample. In single-cell, we take a tissue, and we extract every single cell, then sequence them individually, and you see heterogeneity there. This matters for the mathematical and data analysis and interpretation of the results.

The last one is noise. At the single-cell level you get the 'dropout' issue. This means you will lose a lot of activities of genes just because of technology cannot capture the weak signals because we are only focusing on one cell at a time. When you gather data that is very sparse, e.g., in one cell, you can only capture hundreds of genes, and usually on the bulk level you can capture 10,000 - this is the noise issue.

These challenges together are what we encounter in one single modality. For single-cell multi-omics, if you are sequencing RNA in some cells and you are also sequencing epigenomics, like ATAC-seq, the modalities have different levels of noise. When you try to combine them to make a story by leveraging different modalities on a single-cell level, that's a challenge squared. I think that's why single-cell multi-omics attracts a lot of computational researchers. One very strong rationale for doing this is that all complex tissue, and all complex diseases, including cancer, need these single-cell approaches. They involve a system with multiple placeholders - immune cells with 10s of subtypes, tumour cells with 10s of subtypes, and they talk with each other. Just taking an average in a complex system would lose a lot of information. While single-cell multi-omics is expensive and has a lot of challenges, people still have no hesitation going there.

***FLG:* Let's cover the computation tools that you've published that are targeted towards multi-omics data. Can we start with DeepMAPS?**

Qin: We started with a paper scGNN to show how we apply AI and deep learning in biomedical informatics. That was the first single-cell level graph neural network model for modelling cellular and molecular level heterogeneity. Our second work developed a tool named scDEAL. This tool integrates bulk and single-cell RNA-seq data and uses a deep transfer learning model to predict cancer drug responses at the single-cell level, holding strong clinical impacts for drug selection and repurposing.

Based on scGNN, we developed DeepMAPS. DeepMAPS is a deep learning model for single-cell multi-omics, and compared to the entire computational field, it is one of the field's first models. Right after publishing DeepMAPS, it was elected by the Nature Portfolio as one of the 50 best papers in cancer research.

If we take a step back, you may ask, why scGNN and why DeepMAPS? If you remember, the challenges - huge data size, heterogeneity, dropout - these all result in a very high noise-to-signal ratio. Everything we mentioned here is a perfect fit for deep learning. Deep learning models need a lot of data to train themselves, and deep learning is well known to remove noise and increase the signal-noise ratio in a lot of fields, not only in biology. So, organically, this is a good solution for single-cell multi-omics. We first got started with a simple scenario with single-cell RNA-seq. There's a lot of data, let's try to see whether deep learning can analyse the data a get it into a better shape. That's the scGNN we published, and it demonstrates the value of leveraging AI/machine learning with single cell RNA-seq. This was the foundational evidence showing AI could potentially shape single-cell data analysis for the better.

In DeepMAPS, we are not targeting only one modality. We are targeting multiple RNA-seq datasets, RNA and ATAC, RNA and protein, various different combinations of the modalities as input. As an output, we address the cellular heterogeneity, it can find out how many cell clusters, cell types or subtypes can be identified. What's the underlying modality network, for example, you input RNA and ATAC, and then you can see the regulatory network for the specific cell population. If you input protein and RNA, you can get other networks. So, eventually, DeepMAPS can deliver cell-type specific gene networks.

So, how do you interpret the output? The interpretation is, what's the heterogeneity of your biological system, and how many clusters are there. For each of the clusters, what kind of marker and what kind of network is defining it and defining the cell type or cell state. This is DeepMAPS and this works as an end-to-end model, which means it is not a pipeline; we model everything together in one deep learning model. It's a heterogenous graph transformer, which is a robust solution for single-cell multi-omics.

***FLG:* Do you have any situations where it's been used by people since you released the paper to draw clinical insights or biological insights?**

Qin: In our DeepMAPS paper, we apply DeepMAPS to lymphoma, which is a very aggressive cancer type. We used it to identify a specific lymphoma cell subtype, which has not been identified by other computational frameworks. This is a potential mechanism of cancer progression, but if we want to show the clinical implementation and real translation impact, we should have a direct lab or physician partnership established so that we can apply those mechanisms derived at a single-cell level, to the population level. If at the patient-level, we had hundreds of lymphoma cancer patients, we could check whether the mechanism does indeed have differences or commonalities in the cohort, and then we are on target to design some treatments, or pretreatments. That's from the integrated efforts from both the computational and the wet lab groups.

***FLG:* I will jump straight to another one of your tools. Can you explain MarsGT?**

Qin: Well, scGNN, DeepMAPS and MarsGT have a similar research theme. In DeepMAPS, we focus on the major cell types - can we distinguish the major cell types, e.g., the cancer cells and the immune cells? We can then define their molecular mechanism and find out what is underlying these types. This is DeepMAPS, but one limitation of this and other similar tools is when we try to identify some rare population, which means the number of cells could be less than 3%, or only 1% of the entire population. For example, if you have 1,000 cells, a rare cell population may be less than 10. That means, if you want to target rare populations, cell numbers need to be high.

We know that there are times when a rare population will have a critical role in disease progression. For example, if you have cancer, take a drug, but then the cancer comes back, it's likely not related to the major cell types, but a very rare population who have drug resistance, which people called it the minimal residue disease. Rare populations are important, but from the computational point of view it's very hard, because there are very few data points in the whole population. If you want to identify your population, you need a specific strategy, otherwise you can easily find false positives. Sometimes, the rare population will have heterogeneity, which means that if you want to define a rare population, this will not usually be down to one gene or two. Instead, it should be a very complicated network to define the cell type.

After DeepMAPS, we quickly released the next version focusing on rare populations and we called it MarsGT. Once the biological network for that rare population can be identified we can drive exciting insights. There are a couple of case studies, specifically in immune oncology, but the manuscript is still under review so I'm not providing too many details.

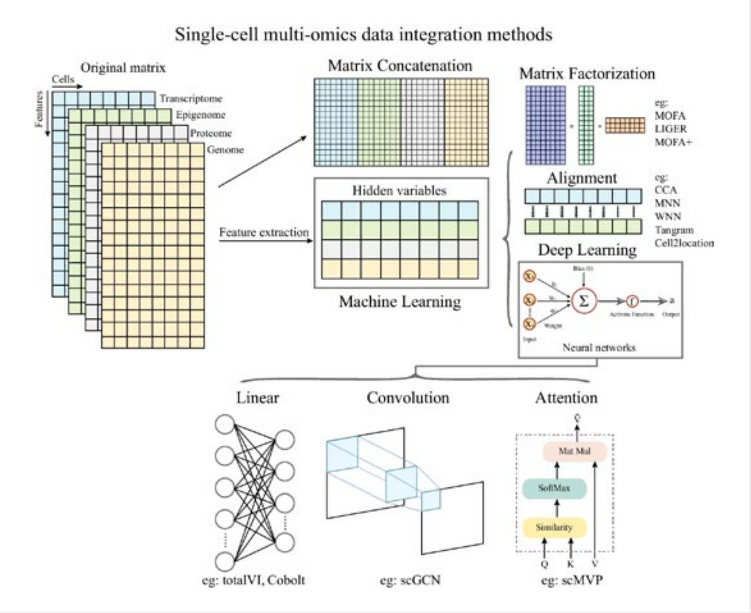
Unmatched (Diagonal) Integration

Unmatched integration highlights a different and more substantial challenge. Unmatched experiments are technically easier to perform because each cell can be treated optimally for the omic that is due to be analysed. Yet, because the omics data from different modalities are drawn from distinct populations/cells, the cell or tissue cannot be used as an anchor. An anchor has to be derived by some other means.

A general solution to this problem is to project cells into a co-embedded space/non-linear manifold to find commonality between cells in the omics space. Due to the learning-based nature of this task, this space is popularised by a number of machine learning and statistical methods designed to find the most appropriate anchor to align cells.

FIGURE 2.3. SINGLE-CELL MULTI-OMICS DATA INTEGRATION METHODS.

The easiest way to integrate multi-omics is to concatenate the original feature matrix of various omics data, but the noise and distinct meaning of values confuse the results of the integration. Machine learning methods extract features from the original matrix and then combine the features across multi-modalities. Deep learning algorithms have also been applied based on various types of networks; e.g., linear, convolution and self-attention. Image Credit: Wang, et al.[36]



A popular tool for unmatched integration recently introduced is GLUE[4], which stands for Graph-Linked Unified Embedding, and it can achieve triple-omic integration. Using graph variational autoencoder, GLUE can learn how to anchor features using prior biological knowledge, which it uses to link omic data.

scJoint[40] is another new method showing promise in this area. It uses transfer-learning to integrate atlas-scale heterogenous data, outperforming classical methods such as LIGER[7] and Seurat v3[5]. MultiDGD[41], a recent tool released in preprint in August 2023 from the Teichmann lab, looks to be the superior option here. Employing a Gaussian Mixture Model rather than an autoencoder, it has several advantages over previous methods such as more flexible and high quality representations of the data and shows high performance on the atlas-scale multi-omics data that is now widely available.

Mosaic Integration

Mosaic integration is an alternative to diagonal integration. This can be used when you have an experimental design in which each experiment has various combinations of omics that create sufficient overlap.

For example, if one sample was assessed for transcriptomics and proteomics, another for transcriptomics and epigenomics and a third for proteomics and epigenomics, there is enough in common between these samples to integrate the data.

Tools such as COBOLT[34] and MultiVI[35] present modern methods to integrate mRNA and chromatin accessibility in a mosaic fashion. They create a single representation of the cells across datasets to be used in downstream analysis.

A final tool here is MultiMAP[42]. It is a graph-based method that assumes a uniform distribution of cells across a latent manifold structure to integrate datasets with unique and shared features. We caught up with **Dr. Mirjana Efremova**, one of the senior authors of MultiMAP, to ask her about multi-omics integration.

INTERVIEW:

MIRJANA EFREMOVA

GROUP LEADER BARTS CANCER INSTITUTE

***FLG:* Can you introduce yourself, give some of your research background and your current research interests**

Mirjana: I am a computational biologist by training. During my PhD at the Medical University of Innsbruck, Austria, I investigated how the immune system shapes colorectal cancer evolution. For my postdoctoral research, I joined the Teichmann Lab (Wellcome Sanger Institute) where I co-led the development of the first human atlas of the maternal-fetal interface in early pregnancy using single-cell transcriptomics. I became interested in understanding how cells communicate and how signals from the environment mediate distinct cellular phenotypes, particularly in cancer, and I developed the cell-cell communication statistical framework CellPhoneDB for inference of cellular communication networks. I started my lab at the Barts Cancer Institute where we are focused on studying cancer cell plasticity and the intrinsic and extrinsic mechanisms that drive plasticity in metastasis and therapy resistance.

> “MULTIMAP INTEGRATES DATA BY CONSTRUCTING A NONLINEAR MANIFOLD ON WHICH DIVERSE HIGH DIMENSIONAL DATA RESIDE AND THEN PROJECTING THE MANIFOLD AND DATA INTO A SHARED LOW DIMENSIONAL EMBEDDING SPACE"

***FLG:* For our readers, can you briefly outline the challenge of multi-omics data integration and why it is still a challenge to this day?**

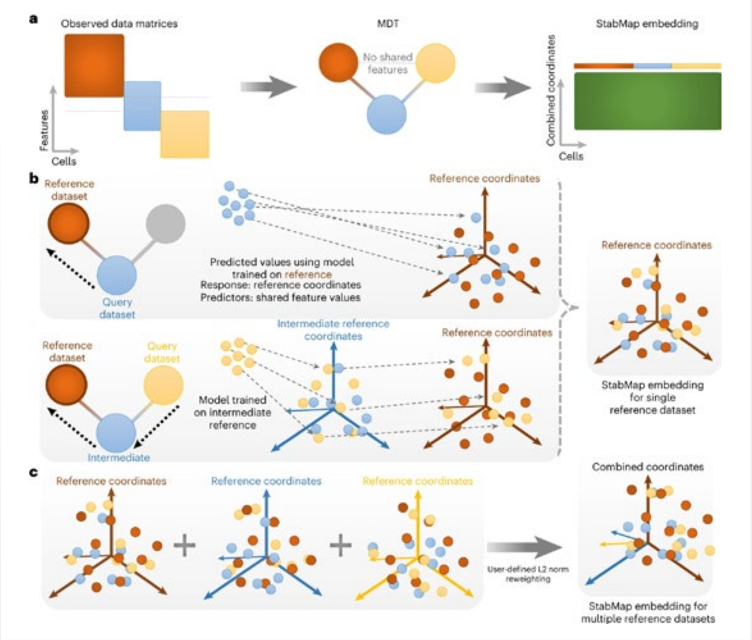
Mirjana: A major challenge in integrating data from different modalities is the distinct feature spaces of different modalities (for example, gene expression in scRNA-seq vs. accessible chromatin regions in scATAC-seq). Typically, different omics methods are measured independently and measure distinct unpaired features with different underlying distributions and properties. They can also have different noise and batch characteristics, which are challenging to identify and correct.

***FLG:* Can you describe the multi-omics tools you have been involved with, MultiMAP, the approach to multimodal data the tool adopts and how it fits into the landscape of multi-omics data integration?**

Mirjana: MultiMAP is an algorithm for dimensionality reduction and integration of multiple modalities. It basically generalises the UMAP algorithm to the setting of multiple datasets with different dimensions. Specifically, MultiMAP integrates data by constructing a nonlinear manifold on which diverse high dimensional data reside and then projecting the manifold and data into a shared low dimensional embedding space.

FIGURE 2.4. SINGLE-CELL MULTI-OMICS DATA INTEGRATION METHODS.

*(**A**) Data that StabMap is specialised for, multiple non-overlapping matrices, (**B**) Method for StabMap, cells are projected on the intermediate reference. (**C**) The process is repeated for the selected reference datasets to reach the final integration. Image Credit: Ghazanfar, et al.* [33]



Finally for mosaic integration, we would like to pay close attention to two very recent tools that were highlighted in Nature.[43] Both methods address the issue of having a low number of overlapping features between the datasets you want to integrate.

The first of these tools is StabMap[33] (Figure 2.4), which was demonstrated to integrate proteomics with an mRNA and chromatin accessibility dataset without the need for directly overlapping features. It implements a 'multi-hop' strategy by querying one dataset with intermediary datasets in a chain until the second dataset is queried. All query datasets are then projected onto the reference space.

The other tool is Bridge Integration, as part of Seurat v5[3]. This method is similar to StabMap but uses an intermediary dictionary dataset to unify the features of the query and reference datasets. This method achieves the computational efficiency necessary to run large-scale multi-omics integration on a personal computer.

Spatial Integration

With the increasing development of spatial multi-omics methods (see Chapter 3), new integration strategies are needed for this data. Principally, we are looking at vertical spatial integration as these spatial modalities naturally capture the omics within the confines of a cell or 'spot', which works as the anchor.

Existing spatial methods, such as ArchR[44], have been successfully deployed for spatial integration. The example here used the RNA modality to indirectly spatially map other modalities, specifically spatial transcriptome and epigenome integration[45]. Another example is Cell2location[46], which was successfully used to integrate spatial RNA and ATAC data in the human heart using a shared nearest neighbours (SNN) strategy[47].

Given the popularity of GLUE for diagonal integration of single-cell data, Dr. Jinmiao Chen has recently released SpatialGlue[48], a spatial version that allows the integration of omics on spatial sections.

Existing tools are also being modified to allow spatial analysis. For example, the developers of MOFA+ have recently released MEFISTO[49], which uses the same factor analysis approach with a new capability to handle both temporal and spatial components within the model

Ultimately, the development of paired and unpaired spatial integration methods is a space to watch for future developments, as more paired multi-omics methods are released.

Chapter 2 references

1. Hossain, M.S., Joshi, T. & Stacey, G. **System approaches to study root hairs as a single cell plant model: current status and future perspectives**. *Front Plant* Sci **6**, 363 (2015).
2. Miao, Z., Humphreys, B.D., McMahon, A.P. & Kim, J. **Multi-omics integration in the age of million single-cell data.** *Nature Reviews Nephrology* **17**, 710-724 (2021).
3. Hao, Y. *et al.* **Dictionary learning for integrative, multimodal and scalable single-cell analysis.** *Nature Biotechnology* (2023).
4. Cao, Z.-J. & Gao, G. **Multi-omics single-cell data integration and regulatory inference with graph-linked embedding.** *Nature Biotechnology* **40**, 1458-1466 (2022).
5. Stuart, T. *et al.* **Comprehensive integration of single-cell data.** *Cell* **177**, 1888-1902. e21 (2019).
6. Korsunsky, I. *et al.* **Fast, sensitive and accurate integration of single-cell data with Harmony.** *Nature Methods* **16**, 1289-1296 (2019).
7. Welch, J.D. *et al.* **Single-cell multi-omic integration compares and contrasts features of brain cell identity.** *Cell* **177**, 1873-1887. e17 (2019).
8. Zhang, Z. *et al.* **Signal recovery in single cell batch integration.** *bioRxiv,* 2023.05.05.539614 (2023).
9. Argelaguet, R., Cuomo, A.S.E., Stegle, O. & Marioni, J.C. **Computational principles and challenges in single-cell data integration.** *Nature Biotechnology* **39**, 1202-1215 (2021).
10. Flynn, E., Almonte-Loya, A. & Fragiadakis, G.K. **Single-Cell Multiomics.** *Annual Review of Biomedical Data Science* **6**, 313-337 (2023).
11. Baysoy, A., Bai, Z., Satija, R. & Fan, R. **The technological landscape and applications of single-cell multi-omics.** *Nature Reviews Molecular Cell Biology,* 1-19 (2023).
12. Singh, R., Hie, B.L., Narayan, A. & Berger, B. S**chema: metric learning enables interpretable synthesis of heterogeneous single-cell modalities.** *Genome biology* 22, 1-24 (2021).
13. Hao, Y. *et al.* **Integrated analysis of multimodal single-cell data.** *Cell* **184**, 3573-3587. e29 (2021).
14. Zuo, C., Dai, H. & Chen, L. **Deep cross-omics cycle attention model for joint analysis of single-cell multi-omics data.** *Bioinformatics* **37**, 4091-4099 (2021).
15. Ma, A. *et al.* **Single-cell biological network inference using a heterogeneous graph transformer.** *Nature Communications* **14**, 964 (2023).
16. Kim, H.J., Lin, Y., Geddes, T.A., Yang, J.Y.H. & Yang, P. **CiteFuse enables multi-modal analysis of CITE-seq data.** *Bioinformatics* **36**, 4137-4143 (2020).
17. Argelaguet, R. *et al.* **MOFA+: a statistical framework for comprehensive integration of multi-modal single-cell data.** *Genome biology* **21**, 1-17 (2020).
18. Zuo, C. & Chen, L. **Deep-joint-learning analysis model of single cell transcriptome and open chromatin accessibility data.** *Briefings in Bioinformatics* **22**, bbaa287 (2021).
19. Gayoso, A. *et al.* **Joint probabilistic modeling of single-cell multi-omic data with totalVI.** *Nature Methods* **18**, 272-282 (2021).
20. Wang, X. *et al.* **BREM-SC: a bayesian random effects mixture model for joint clustering single cell multi-omics data.** *Nucleic acids research* **48**, 5814-5824 (2020).
21. Bravo González-Blas, C. *et al.* **SCENIC+: single-cell multiomic inference of enhancers and gene regulatory networks.** *Nature Methods* **20**, 1355-1367 (2023).
22. Kartha, V.K. *et al.* **Functional inference of gene regulation using single-cell multi-omics.** *Cell genomics* **2**(2022).
23. Lynch, A.W. *et al.* **MIRA: joint regulatory modeling of multimodal expression and chromatin accessibility in single cells.** *Nature Methods* **19**, 1097-1108 (2022).
24. Kamimoto, K. *et al.* **Dissecting cell identity via network inference and in silico gene perturbation.** *Nature* **614**, 742-751 (2023).
25. Li, C., Virgilio, M.C., Collins, K.L. & Welch, J.D. **Multi-omic single-cell velocity models epigenome–transcriptome interactions and improves cell fate prediction.** *Nature Biotechnology* **41**, 387-398 (2023).
26. John, C.R., Watson, D., Barnes, M.R., Pitzalis, C. & Lewis, M.J. **Spectrum: fast density-aware spectral clustering for single and multi-omic data.** *Bioinformatics* **36**, 1159-1166 (2020).
27. Dou, J. *et al.* **Unbiased integration of single cell multi-omics data.** *biorxiv,* 2020.12. 11.422014 (2020).
28. Liu, J., Huang, Y., Singh, R., Vert, J.P. & Noble, W.S. **Jointly Embedding Multiple Single-Cell Omics Measurements.** *Algorithms Bioinform* **143**(2019).
29. Duan, B. *et al.* **Model-based understanding of single-cell CRISPR screening.** *Nature communications* **10**, 2233 (2019).
30. Cao, K., Bai, X., Hong, Y. & Wan, L. **Unsupervised topological alignment for single-cell multi-omics integration.** *Bioinformatics* **36**, i48-i56 (2020).
31. Campbell, K.R. *et al.* **clonealign: statistical integration of independent single-cell RNA and DNA sequencing data from human cancers.** *Genome biology* **20**, 1-12 (2019).
32. Cao, K., Hong, Y. & Wan, L. **Manifold alignment for heterogeneous single-cell multi-omics data integration using Pamona.** *Bioinformatics* **38**, 211-219 (2021).
33. Ghazanfar, S., Guibentif, C. & Marioni, J.C. **Stabilized mosaic single-cell data integration using unshared features.** *Nature Biotechnology* (2023).
34. Gong, B., Zhou, Y. & Purdom, E. **Cobolt: integrative analysis of multimodal single-cell sequencing data.** *Genome Biology* **22**, 351 (2021).
35. Ashuach, T. *et al.* **MultiVI: deep generative model for the integration of multimodal data.** *Nature Methods* **20**, 1222-1231 (2023).
36. Wang, X., Wu, X., Hong, N. & Jin, W. **Progress in single-cell multimodal sequencing and multi-omics data integration.** *Biophysical Reviews* (2023).
37. Lakkis, J. *et al.* **A multi-use deep learning method for CITE-seq and single-cell RNA-seq data integration with cell surface protein prediction and imputation.** *Nature Machine Intelligence* **4**, 940-952 (2022).
38. Wang, X. et al. MarsGT: **Multi-omics analysis for rare population inference using single-cell graph transformer.** *bioRxiv,* 2023.08.15.553454 (2023).
39. Wang, J. *et al.* **scGNN is a novel graph neural network framework for single-cell RNA-Seq analyses.** *Nature Communications* **12**, 1882 (2021).
40. Lin, Y. *et al.* **scJoint integrates atlas-scale single-cell RNA-seq and ATAC-seq data with transfer learning.** *Nature Biotechnology* **40**, 703-710 (2022).
41. Schuster, V., Dann, E., Krogh, A. & Teichmann, S.A. **multiDGD: A versatile deep generative model for multi-omics data.** *bioRxiv,* 2023.08.23.554420 (2023).
42. Jain, M.S. *et al.* **MultiMAP: dimensionality reduction and integration of multimodal data.** *Genome biology* **22**, 1-26 (2021).
43. Lee, M.Y.Y. & Li, M. **Integration of multi-modal single-cell data.** *Nature Biotechnology* (2023).
44. Granja, J.M. *et al.* **ArchR is a scalable software package for integrative single-cell chromatin accessibility analysis.** *Nature genetics* **53**, 403-411 (2021).
45. Foster, D.S. *et al.* **Integrated spatial multiomics reveals fibroblast fate during tissue repair.** *Proceedings of the National Academy of Sciences* **118**, e2110025118 (2021).
46. Kleshchevnikov, V. *et al.* **Cell2location maps fine-grained cell types in spatial transcriptomics.** *Nature biotechnology* **40,** 661-671 (2022).
47. Kuppe, C. *et al.* **Spatial multi-omic map of human myocardial infarction.** *Nature* **608**, 766-777 (2022).
48. Long, Y. *et al.* I**ntegrated analysis of spatial multi-omics with SpatialGlue.** *bioRxiv,* 2023.04.26.538404 (2023).
49. Velten, B. *et al.* **Identifying temporal and spatial patterns of variation from multimodal data using MEFISTO.** *Nature Methods* **19**, 179-186 (2022).

MEET THE MACHINERY. THE LATEST MULTI-OMICS TECHNOLOGIES

THE MULTI-OMICS REVOLUTION HAS BEEN LARGELY ENABLED BY TECHNOLOGICAL DEVELOPMENT. THIS HAS PROGRESSED ON TWO FRONTS; (1) IN THE LABS AND RESEARCH INSTITUTIONS, PROVIDING NOVEL WAYS OF PROFILING MULTIPLE OMICS, AND (2) ON THE COMMERCIAL FRONT, PROVIDING KITS AND INSTRUMENTS CAPABLE OF PRODUCING MULTI-OMICS DATA IN A SCALABLE, REPLICABLE AND ACCESSIBLE WAY. THIS CHAPTER WILL GIVE AN UPDATE ON THE LATEST AVAILABLE TECHNOLOGIES, SPECIFICALLY THE COMMERCIAL OFFERINGS, AND ALSO FOCUS ON THE LATEST SPATIAL MULTI-OMICS INSTRUMENTS.

In the previous chapter, we looked at the computational challenge of integrating two types of omics data together. We saw that for many applications of multi-omics, profiling data from the same cell in the same experiment is extremely valuable. The cell can be used as an anchor, and the dynamics between two omics can be compared directly.

However, omics data from the same cell is not easily acquired, and calls for specialised methodologies and commercialised technologies to achieve data of sufficient quality. Here, we will first review single-cell multi-omics methods, both in-house and commercial, before concentrating on the new wave of spatial multi-omics methodologies, principally for transcriptomics and proteomics. Finally, we will hear from two multi-omics group/core leaders about their work and experiences with these methodologies.

Single-cell Multi-omics Methods

We begin with a brief overview of the single-cell multi-omics methods that have been released, as well as a summary of the commercially available multi-omics solutions. For a full in-depth review of mutli-omics methods, please refer to the excellent reviews referenced here[1-5].

GENOMICS-BASED

Methods to profile the single-cell genome alongside the transcriptome were some of the earliest multi-omics techniques. Examples include G&T-seq[6], DR-seq[7], SIDR[8], Target-seq[9] & DNTR-seq[10]. G&T-seq is the only method that offers CNV, SNV and fusion genomic data with full length transcriptomics, while DR-seq does the same with 3′ or 5′ transcriptomics. These methods are all plate-based but vary in throughput, with DNTR-seq allowing the highest cell throughput. Very recently, we have seen the release of scONE-seq[11], following a seq-spilt approach specialised for long-term stored frozen samples.

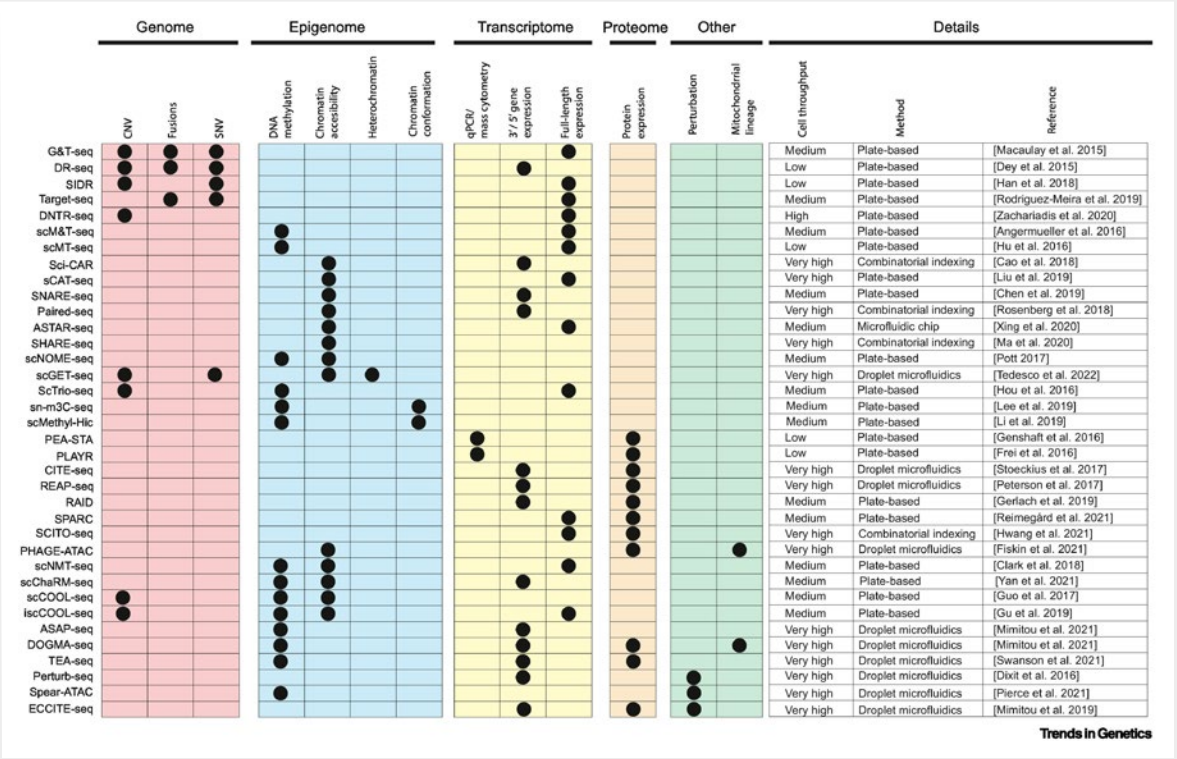
TRANSCRIPTOMICS & PROTEOMICS

Multi-omics approaches using proteomics currently rely on targeted, antibody-based profiling alongside single-cell RNA sequencing. Typically, these approaches profile extracellular proteins on a cell using antibodies, although some approaches exist for intra-cellular and intranuclear tagging. Early approaches used qPCR alongside proteomics such as PEA-STA[12] and PLAYR.

The field was revolutionised with the introduction of CITE-seq[13] and REAP-seq[14], which profiled surface proteomics alongside whole transcriptome (3′/5′). These methods work with high throughput using microfluidics and have been widely used since their inception. Cite-seq allows over 100 surface proteins to be profiled alongside the transcriptome.

FIGURES 3.1. SINGLE-CELL MULTI-OMICS METHODS LISTED BY OMICS THEY PROFILE, CELL THROUGHPUT AND SINGLE-CELL METHODOLOGY.

Image credit: Ogbeide, et al. [3]



	Genome			Epigenome				Transcriptome			Proteome	Other		Details		
	CNV	Fusions	SNV	DNA methylation	Chromatin accesibility	Heterochromatin	Chromatin conformation	qPCR/ mass cytometry	3'/5' gene expression	Full-length expression	Protein expression	Perturbation	Mitochondrial lineage	Cell throughput	Method	Reference
G&T-seq	●	●	●							●				Medium	Plate-based	[Macaulay et al. 2015]
DR-seq	●	●	●						●					Low	Plate-based	[Dey et al. 2015]
SIDR	●		●							●				Low	Plate-based	[Han et al. 2018]
Target-seq		●	●							●				Medium	Plate-based	[Rodriguez-Meira et al. 2019]
DNTR-seq	●									●				High	Plate-based	[Zachariadis et al. 2020]
scM&T-seq				●						●				Medium	Plate-based	[Angermueller et al. 2016]
scMT-seq				●						●				Low	Plate-based	[Hu et al. 2016]
Sci-CAR					●				●					Very high	Combinatorial indexing	[Cao et al. 2018]
sCAT-seq					●					●				Very high	Plate-based	[Liu et al. 2019]
SNARE-seq					●				●					Medium	Plate-based	[Chen et al. 2019]
Paired-seq					●				●					Very high	Combinatorial indexing	[Rosenberg et al. 2018]
ASTAR-seq					●					●				Medium	Microfluidic chip	[Xing et al. 2020]
SHARE-seq					●									Very high	Combinatorial indexing	[Ma et al. 2020]
scNOME-seq				●	●									Medium	Plate-based	[Pott 2017]
scGET-seq	●		●		●	●								Very high	Droplet microfluidics	[Tedesco et al. 2022]
ScTrio-seq	●			●						●				Medium	Plate-based	[Hou et al. 2016]
sn-m3C-seq				●			●							Medium	Plate-based	[Lee et al. 2019]
scMethyl-Hic				●			●							Medium	Plate-based	[Li et al. 2019]
PEA-STA								●			●			Low	Plate-based	[Genshaft et al. 2016]
PLAYR								●			●			Low	Plate-based	[Frei et al. 2016]
CITE-seq									●		●			Very high	Droplet microfluidics	[Stoeckius et al. 2017]
REAP-seq									●		●			Very high	Droplet microfluidics	[Peterson et al. 2017]
RAID									●		●			Medium	Plate-based	[Gerlach et al. 2019]
SPARC										●	●			Medium	Plate-based	[Reimegård et al. 2021]
SCITO-seq										●	●			Very high	Combinatorial indexing	[Hwang et al. 2021]
PHAGE-ATAC					●						●		●	Very high	Droplet microfluidics	[Fiskin et al. 2021]
scNMT-seq				●	●					●				Medium	Plate-based	[Clark et al. 2018]
scChaRM-seq				●	●				●					Medium	Plate-based	[Yan et al. 2021]
scCOOL-seq	●			●	●									Medium	Plate-based	[Guo et al. 2017]
iscCOOL-seq	●			●	●					●				Medium	Plate-based	[Gu et al. 2019]
ASAP-seq				●					●					Very high	Droplet microfluidics	[Mimitou et al. 2021]
DOGMA-seq				●					●		●		●	Very high	Droplet microfluidics	[Mimitou et al. 2021]
TEA-seq				●					●		●			Very high	Droplet microfluidics	[Swanson et al. 2021]
Perturb-seq									●			●		Very high	Droplet microfluidics	[Dixit et al. 2016]
Spear-ATAC				●								●		Very high	Droplet microfluidics	[Pierce et al. 2021]
ECCITE-seq									●		●	●		Very high	Droplet microfluidics	[Mimitou et al. 2019]

Trends in Genetics

RAID-seq[15] is a similar plate-based methodology and allows intracellular protein tagging of six targets. QuRIE-seq[16] represents an advanced version of this protocol, using microfluidics based transcriptomics and expanding the protein repertoire to 80 intra- or extracellular proteins.

Recent tools SPARC[17] and SCITO-seq[18] couple protein expression with full-length RNA expression; the former is also targeted for intracellular proteins, while the latter works at very high throughput via combinatorial indexing. Perturbation-based methods, e.g., ECCITE-seq[19] and Perturb-CITE-seq[20], and nuclear proteomics based methods, e.g., inCITE-seq[21] using single nuclei, have also recently been introduced recently, increasing the available options for multi-omics using proteomics.

TRANSCRIPTOMICS & EPIGENOMICS

Epigenomics has most often been coupled with transcriptomics. For example, DNA methylation can be co-profiled with RNA using scMT-seq[22] and scM&T-seq[23]. Profiling histone modifications with scCUT&Tag[24] has been paired with several other omics measurements such as transcriptome (CoTECH[25] & Paired-tag[26]) and surface proteins (scCUT&Tag-Pro).[27] We spoke to **Dr. Bingjie Zhang**, one of the lead developers of scCUT&Tag-pro, to get more insights into multi-omics tool development.

> "RECENT TOOLS SPARC17 AND SCITO-SEQ18 COUPLE PROTEIN EXPRESSION WITH FULL-LENGTH RNA EXPRESSION; THE FORMER IS ALSO TARGETED FOR INTRACELLULAR PROTEINS, WHILE THE LATTER WORKS AT VERY HIGH THROUGHPUT VIA COMBINATORIAL INDEXING."

INTERVIEW:

BINGJIE ZHANG

POSTDOCTORAL RESEARCH FELLOW, SATIJA LAB

NEW YORK GENOME CENTER

> "I DID NOT EVEN INTEND TO DEVELOP A MULTI-OMICS METHOD.THE CUT&TAG METHOD FROM HENIKOFF'S LAB IS VERY MATURE; I ONLY WANTED TO ADAPT HIS METHOD TO THE 10X PLATFORM. BUT THINGS DID NOT GO AS WELL AS I EXPECTED. THE DATA I GOT WERE SIMPLY TOO SPARSE TO YIELD MEANINGFUL BIOLOGICAL CONCLUSIONS."

***FLG:* Can you just briefly introduce yourself, give some of your research background, and some of your current research interests and projects?**

Bingjie: I have always been very interested in epigenetics. I completed my PhD at Tsinghua University in China, where I started working on epigenome reprogramming during mouse embryo development. I then moved to the US for my postdoctoral research, where I have been focusing on developing single-cell profiling methods for histone modifications, and also single-cell multi-omics methods. In the future, I would like to apply these methods to explore the epigenetic gene regulation during lineage differentiation in the immune system, and the related disease.

***FLG:* I wanted to ask you about your recently released multi-omics method - scCUT&Tag-pro - a method for histone modifications and cell surface protein measurement. Could you just describe why you created this method and a little bit about how it works?**

Bingjie: In my Ph.D. study, I realized how dynamic the epigenetic landscape can be, so it was very natural for me to anticipate future explorations of the epigenome at a single-cell resolution. In my postdoctoral research, I joined Rahul Satija's lab, focusing on investigating the dynamics of histone modification within a cell. Initially, I did not even intend to develop a multi-omics method. The CUT&Tag method from Henikoff's lab is very mature; I only wanted to adapt his method to the 10x platform. But things did not go as well as I expected. The data I got were simply too sparse to yield meaningful biological conclusions.

I wasn't satisfied with just developing novel techniques that didn't further biological insights. Besides, with more data, we began to realize that, due to the nature of histone modification itself, there is no guarantee that we could achieve consistent cell type annotation, which is essential for the downstream integrated analysis. So, as an alternative solution, inspired by the ASAP-seq from NYGC Innovation Lab, we decided to introduce cell surface proteins to assist with data integration and eventually developed the single-cell CUT&Tag-pro. We performed the cell surface protein staining first and then conducted the CUT&Tag. Because we were using the same panel of antibodies, we could easily annotate the cell types by reference mapping and integrate all the different experiments together.

***FLG:* Could you describe the integration strategy between those two modalities, and a little bit maybe about the downstream analysis strategy – the scChromHMM platform?**

Bingjie: In this project, we profiled six different histone modifications. As we aimed to explore all these modifications within the same cell type, integration was crucial. We utilized a computational workflow called reference mapping. The reference is a well-annotated human PBMC CITE-seq dataset containing more than 100,000 cells and also about 200 antibodies. Because the antibody panel used in our CUT&Tag-pro largely overlaps with that in the reference dataset, we can use the protein information from CUT&Tag-pro to project them onto the reference and transfer a consistent set of cell type annotations. With this method, we can project any query dataset into a space defined by the reference. Eventually, we have a single harmonized atlas that contains all the different modalities: RNA, protein, chromatin accessibility and six histone modifications.

To assign the chromatin states at single-cell resolution, we first need to generate single-cell profiles with measurements of six histone marks. Previously, Satija Lab described an anchoring workflow to 'transfer' modalities across experiments. Since we can integrate all different modalities together into a common space, we should be able to impute all modalities into the same set of cells. Thus, we applied a similar procedure to interpolate 20,000 single-cell profiles, each consisting of six histone modifications. My collaborator, Avi Srivastava, developed scChromHMM to annotate the chromatin state within a cell, which can be seen as an extension of ChromHMM. To run this, there are basically two steps: The first step is to run ChromHMM using the pseudobulk profiles from the CUT&Tag-pro to learn parameter estimates and obtain a list of possible chromatin states. Secondly, we run the forward-backward algorithm on the interpolated single-cell profiles. So eventually, for each cell, we would get a probability of each chromatin state for each 200bp window.

***FLG:* What applications do you have in mind for this technology?**

Bingjie: I think it's particularly useful for immune cells, as we have very good knowledge about their cell surface protein markers. Although in our paper we used 173 antibodies, if you aren't concerned with very rare cell populations and are aware of the protein markers for your cell type of interest, then just a dozen or so protein markers can give good clustering results. So, if you're working on the immune system and would like to investigate epigenetic gene regulation, I do believe our method would be very useful.

Chromatin accessibility (ATAC-seq) can also be profiled with RNA, with methods such as SNARE-seq[28], Paired-seq[29], SHARE-seq[30] and, most recently and most sensitively, ISSAAC-seq[31]. New methods such as ICICLE-seq[32] and Phospho-seq[33] pair ATAC-seq with surface proteomics and intracellular proteomics, respectively.

Epigenomic methods have also been paired together in multi-omics form to create a more holistic epigenomic profile of individual cells. For example, scCool-Seq[34] is a method that profiles DNA methylation and chromatin accessibility with a medium throughput and high coverage. The updated version, iscCOOL-seq[35], improves that throughput further.

The transcriptome has been paired with two forms of epigenetic measurements in one tool in ScNMT-seq[36] and, more recently, scNOMeRe-seq[37], snmCAT-seq and scChaRM-seq[38]. These methods profile the transcriptome with methylation levels and chromatic accessibility.

Nuclear organisation (HiC) has been paired with methylome profiling using methods such as scMethyl-Hic[39] and snm3C-seq[40].

TRIO-OMICS AND MORE

Three or more omics in one assay has been the limit of current methodologies, but a variety of methods have been created with this purpose in mind.

The original amongst the list is ScTrio-seq[41], a method for genomics (CNV, SNV, somatic mutations), transcriptomics and epigenomics (methylation levels). Genomics and epigenomics have been paired with the proteome in recent tools, PHAGE-ATAC[42] and ASAP-seq[43]. Both methods profile mitochondrial DNA alongside chromatin accessibility and intra- or extracellular proteins in a high throughput manner. TEA-seq[32] and NEAT-seq[44] represent recent tools linking the transcriptome, proteome and chromatin accessibility. TEA-seq works with 46 surface proteins while NEAT-seq works with intracellular proteins with a focus on transcription factor binding motif accessibility. DOGMA-seq[43] represents the merge of these methods, profiling mtDNA, RNA and >200 intra- and extracellular proteins alongside chromatin accessibility.

COMMERCIALISED SINGLE-CELL & BULK MULTI-OMICS

There is a selection of commercialised kits available for single-cell multi-omics that provide all the support to profile two or more omics in one experiment. Examples of these include:

- The **10x Genomics Multiome kit** allows the simultaneous profiling of gene expression and ATAC-based chromatin accessibility using the 10x Chromium controller. This kit means that, for each cell, you get two readouts.
- **Mission Bio** recently released the third version of their **Tapestri®** platform. With DNA as its primary analyte, it allows the analysis of the genotype of a cell such as CNVs, SNVs, plus proteins. The new version allows up to four times more cells captured per sample, which increases the ability to detect rare cells.
- **BioSkryb Genomics' ResolveOME™ kit** allows the near-complete survey of the genome and mRNA transcriptome at single-cell resolution. With their associated BaseJumper™ data analysis software, this setup creates a unified workflow for DNA and RNA interrogation at single-cell level.
- **Singleron's PromoScope™** kit allows the simultaneous quantification of the whole transcriptome, as well as protein glycosylation at the single-cell level. Relying on their SCOPE-chip® technology, this is the first kit to quantify protein modifications alongside transcriptomics.
- **BD Biosciences** offers a different solution. Their single-cell **AbSeq kit** allows whole transcriptome and protein detection in single cells used with their BD Rhapsody™ Single-Cell Analysis Systems.
- **Isoplexis** also allow transcriptomics and functional proteomics through their **Duomic kit.** Duomic is a single chip and allows these measures to be connected from single cells including their in vivo proteomic methods.
- **Biolegend** produce a **TotalSeq™ kit** that also allows proteomics alongside transcriptomics (performed by another kit such as 10x Genomics Chromium), allowing 100s of proteins to be detected
- **Qiagen** offer a DNA and RNA - QIAseq multimodal kit necessary to analyse both analytes in 12 samples: SNVs, CNVs and inDels from the genome and gene expression transcriptomics from each sample.
- A genomic and epigenomic kit is offered by **biomodal** called the **duet multi-omics solution +modC.** This kit allows full genome alongside methylation screening.

We spoke to **Dr. Iain Macaulay,** the first of our multi-omics group leaders about his experience with these tools.

INTERVIEW: IAIN MACAULAY

TECHNICAL DEVELOPMENT GROUP LEADER EARLHAM INSTITUTE

***FLG:* Can you briefly introduce yourself and give some of your research background in multi-omics.**

Iain: I'm Iain Macaulay, I'm a Group Leader at the Earlham Institute in Norwich. Our group works almost entirely on single-cell technology methods developments and applications. We've done lots of work with multi-omics - starting nearly 10 years ago, when we started developing methods for parallel DNA and RNA sequencing from the same cell which became the G&T-seq protocol. And we've adapted and developed that in lots of different ways to enable some epigenetic measurements as well as transcriptional measurements. As a group, we work with almost every biological system you can imagine. Our main focus, and my main research interest, is in blood cell development, but our group works within our institute to apply multi-omics technology to anything from plants, microbes to human, mouse and chicken developmental biology.

***FLG:* Do you have to adapt your multi-omics protocols for those different biological systems?**

Iain: Not so much actually. It's quite interesting. For single-cell DNA and RNA sequencing, in general, we haven't had to adapt anything particularly significant. If you can get the DNA and RNA out of the cell, you can read it out and sequence it. So, in any vertebrate system, things usually work very well between organisms. When we move to plants, there's a little bit of a challenge in getting the cell wall off. But once you've done that, then it also works fairly easily. Bacteria are a little bit different, and we haven't done so much multi-omics work on bacteria, just single omics stuff, but that can work really well for bacteria In general, it's quite nice that once we set the methods up, we can take on projects in really different areas with the same method.

***FLG:* You were part of the team that made some of the earliest methods to sequence two omics in parallel. Could I get your perspective on how this field has developed from those early methods to now. What's been your experience of it?**

Iain: So, the main thing shifting the whole single-cell field, since the early days, is scale. Scale has become the thing that people want. Early on, there was a lot of emphasis on getting as much information as possible from every single cell, and the nature of the microfluidics platforms etc. drove the field towards scale, over completeness of information per cell. The objective is to get as many cells as possible and then you'll understand a very big system. The same is true for almost all of the multi-omics approaches that are coming in. So, 10x's Multiome, and others, are still based on many 1,000s of cells, and that's really cool.

Where it falls down a little bit is that you're not getting all of that information from every cell. When you do an ATAC-seq library from a single cell by 10x, you're just doing 50,000 reads per cell. So, the genome coverage is quite low. It still enables you to do a lot, but if you really think about why we are doing multi-omics, there's something missing in terms of really understanding how one cell or a small number of cells are using their genomes to produce a transcriptome. In this way, high throughput single-cell can enable you to see patterns, but maybe not completely understand the biology.

Another thing people have struggled with a bit is analysis. Making sense of multi-omics data is a challenge, because you have a lot of information from each cell, even if it's not a high throughput experiment. People might want to link epigenetic measurements with transcriptional measurements, which is a really complex process in the cell. Hence, there's no simple formula like 'you've got open chromatin, you must have an expressed gene.' That's sometimes the case, but it's not exactly the case for every gene - there'll be some variation. People look for high level patterns in the data, but understanding how individual genes are regulated is maybe less of a concern for people, but it probably will be in the future.

Once you've generated all that data, visualising it and sharing with people is still quite a problem. There's standard processes for clustering, and that's great, but sharing the connections between the different omes, or, if in a G&T-seq experiment, we see a single nucleotide variant or an extra copy of a chromosome, how do we explain to people what that does to that cell's transcriptome? If you gain a copy of the chromosome, you'll have extra expression of the genes on that chromosome. But there are also off-chromosome effects on what those genes are regulating. You end up with quite complex data that you immediately have to start waving your hands around to explain. I think some of the visualisation tools are not there. A lot of cool tools have emerged, but I think a lot of the biological interpretation and visualisation tools still have some work to do.

FLG: There are a lot of multi-omics methods, how would you go about choosing a method? What questions are you asking to help researchers sift through the different methods?

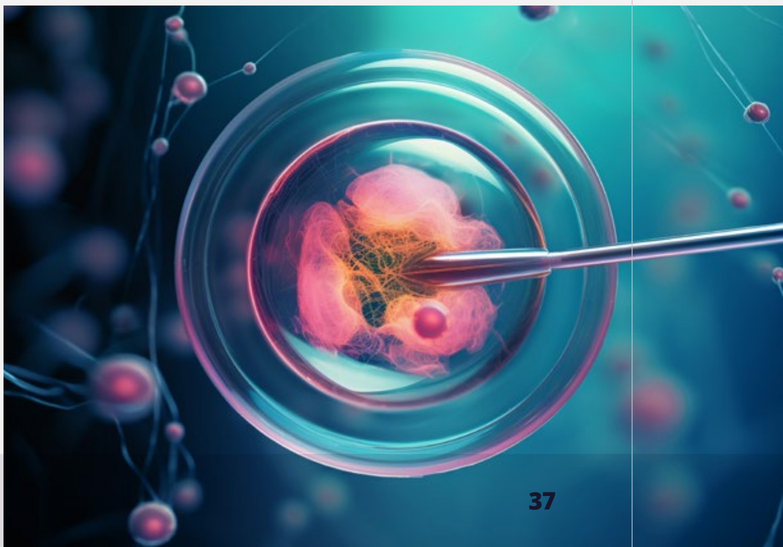
Iain: Some criteria are defined by really basic things such as accessibility of resources, how much money they have and whether we have a 10x instrument on site. At the moment we have a project where they have used 10x Multiome and they want to focus in on fewer cells and get better coverage. So, we're actually looking at developing an approach - which has probably already been done - where you can try and figure out a bespoke method for the biological problem.

In the middle ground, if they want RNA and open chromatin, typically we talk people through the coverage issue. Some people have assumed they'll get their promoter region, that their 20 promoter regions will be detectable in every cell. So, we go through that expectation management. And then we get people with a specific type of question, usually a developmental biology question, where there are large changes happening, and differentiation programmes being measured. Doing 10x Multiome on developing chicken or mouse embryos, you start to see patterns emerging. Those are the projects that have gone forward, where there's a real strong motivation to look at organisation of gene expression.

For picking the best method, I think it really comes down to accessibility of platforms and resources. I run a research group, but we also run a facility or a platform that is accessible to people. Within the research group, we can do whatever work we get funding to support. But for services for other people, you don't want to a) set up new methods for people or b) invest a lot of time in something that you have to validate before you can offer it to someone. A lot of what drives our decision making is – 'we've got the 10x instrument, it will do this, does that work for your experimental plans?' Which is maybe less exciting than – 'wow, this new method has come out, we have to try it.' Because often we look at the list of methods, and we think – 'well, do we really need to try all of them?'

FLG: What would you like to see in multi-omics over the coming years?

Iain: I would like to see spatial methylation, but I don't think we will for a long time. If you think about the end users of spatial technologies, where it will really benefit humanity will be in histopathology and clinical activity, and understanding somatic mutations and their impact on tissue architecture and tumour architecture. The other thing that would be really interesting is the lineage tracing cell barcoding studies that are able to read out which cell made which cell, cell lineage on top of spatial information. I think that's going to be really cool. I think that will be something that will transform a lot of developmental biology. If you think about the end users, I think a lot of people will be happy with spatial transcriptomics and proteomics, proteomics integrating with spatial would be quite useful too.



Spatial Multi-omics Methods

Spatial multi-omics methods were first developed and released in 2018 with seqFISH[45], profiling the transcriptome and proteome in a probe-based manner using in-situ hybridisation, achieving subcellular resolution. Transcriptomics and proteomics have been spatially profiled using several more technologies since then, including with another probe-based method, MERFISH[46].

DBiT-seq[48] was a minor revolution in this area, using barcode capture based transcriptomics with NGS antibody spatial proteomics (for 22 surface proteins) and boasting a 10-20 micrometer resolution. Spatial RNA and proteomics is also available through commercial offerings from 10x in the Visium and Xenium and Nanostring in the GeoMx and CosMx, which will be explored in the next section.

The latest offerings for spatial proteomics and transcriptomics include SPOTS[49], SM-OMICS[50] and Spatial-CITE-seq[51], the latter of which allows the profiling of almost 200 surface proteins at 20 micrometer resolution. Spatial multi-omics with proteomics has historically been limited to a handful of surface protein markers, but these recent technologies have allowed the inclusion of substantially more proteins.

Epigenomics has also been incorporated spatially, initially through probe-based methods such as DNA-MERFISH[52] and spatial DNAseqFISH+[53], which achieve subcellular resolution, and through OligoFISSEQ[54], which links epigenomics to proteomics through fluorescence in situ sequencing.

Most recently, the emergence of spatial ATAC-RNA-seq[55] and spatial CUT&Tag-RNA-seq[55] using the DBiT-seq methodology have allowed barcode based profiling of the two epigenomic measurements alongside RNA in a spatially resolved manner.

FIGURE 3.2. DEPICTION OF THE USES OF SPATIAL MULTI-OMICS.

Visualising multiple omics types in one spatial image allows advanced profiling of spatial architecture, TME structure, spatial cell-cell communication and, ultimately, improved clinical outcomes. Image Credit: Wu, et al.[47]

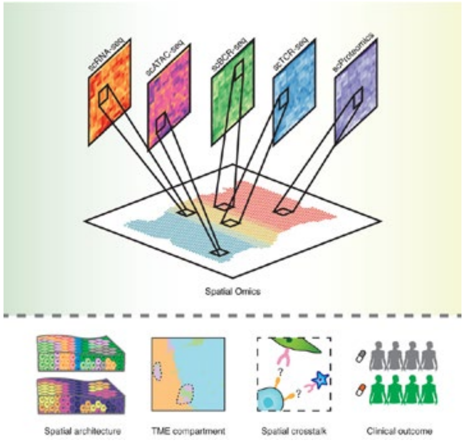
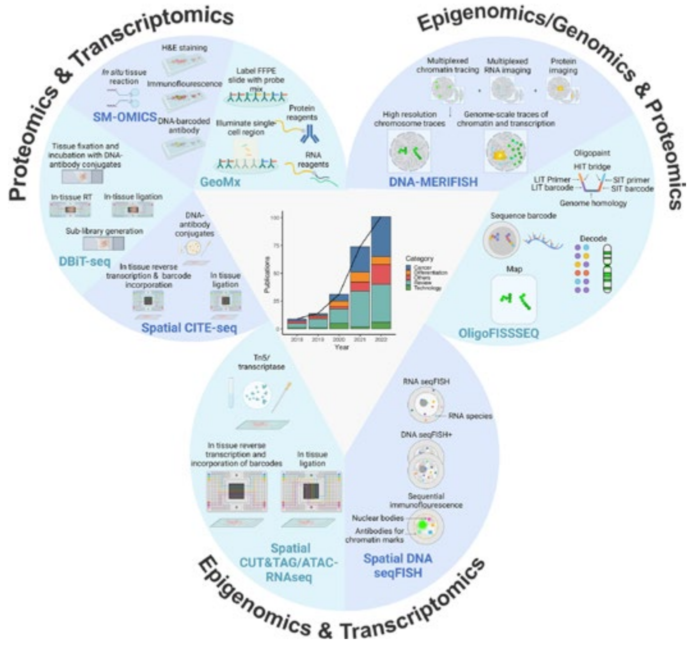


FIGURE 3.3. THE EXPLOSION OF MULTI-OMICS SPATIAL TECHNOLOGIES OVER THE LAST 5 YEARS.

Representative spatial technologies for the different multi-omics assessments. In the centre is a bar chart of the number of publications reporting spatial multi-omics methods in different categories. Image Credit Li[56]

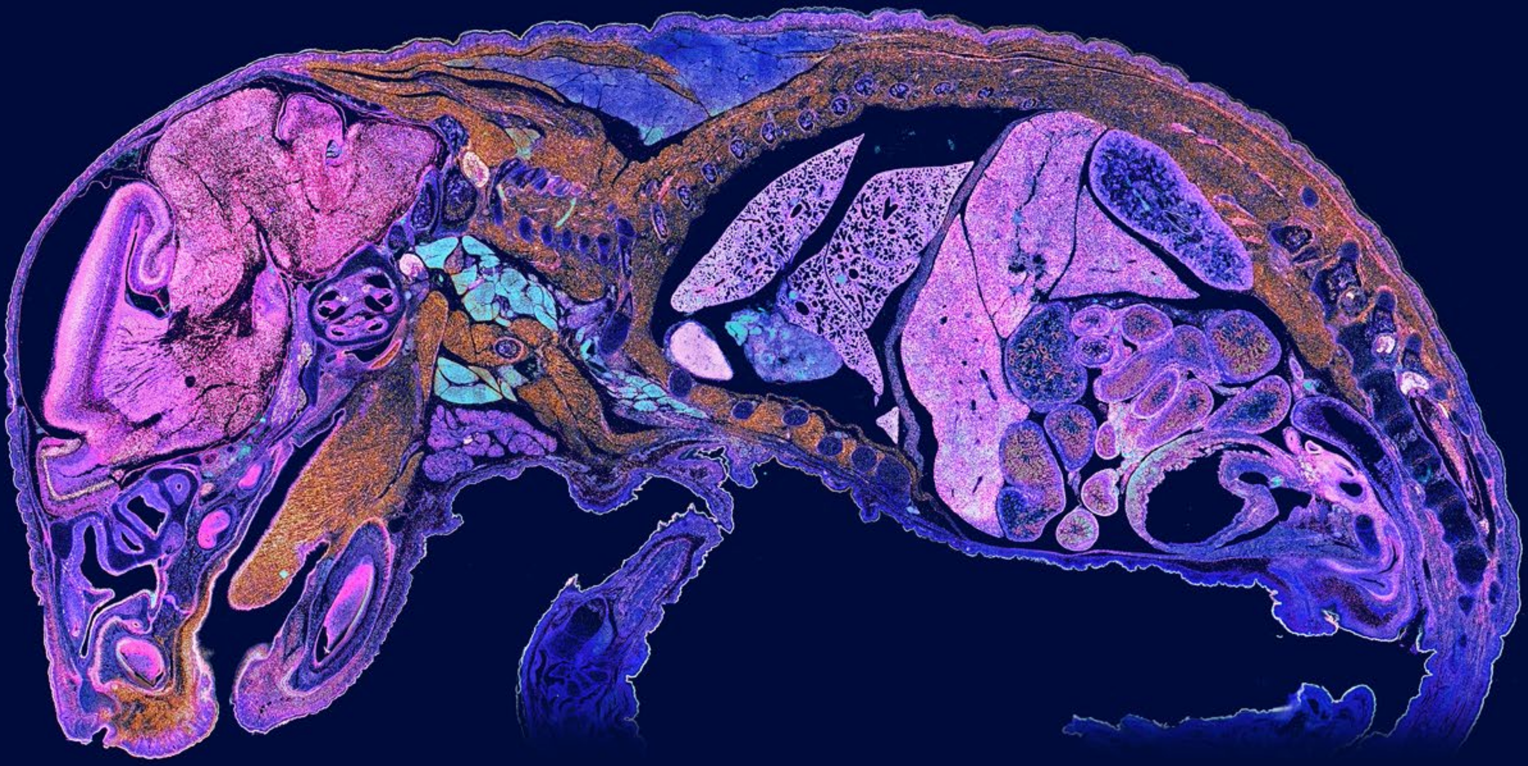


COMMERCIALISED SPATIAL MULTI-OMICS

There are a range of commercialised kits available for spatial multi-omics, allowing you to profile two or more omics in one experiment with spatial resolution. Examples of these include:

- **10x Genomics** provide both the **Visium** + Cyt Assist™ for whole transcriptome and 31-plex protein assays at a ~50 micron resolution and the **Xenium™**, which enables the profiling of 1000s of RNAs using a probe-based strategy alongside proteins at subcellular resolution.
- **Nanostring** provide both the **CosMx™ SMI** - the highest plex in situ imager with 1000-plex RNAs and 64-plex proteins analysed in the same tissue at subcellular resolution - and the **GeoMx™**, which has broader protein and RNA capability with lower resolution.
- **Vizgen MERSCOPE™**, the commercial realisation of MERFISH, launched a protein co-detection kit that allows users to take full advantage of the subcellular hi-plex nature of the instrument while detecting up to five proteins.
- **Akoya Bioscience's Phenocycler-Fusion** was released in January 2022 and offers a high throughput workflow at sub-cellular resolution for 100+ markers, either RNA or protein biomarkers. It is the fastest single-cell spatial biology solution, able to map a million cells in 10 minutes.
- Companies are also beginning to work in combination to create multi-omics possibilities. An example of this is **ACD Biotechne** and **Standard Biotools**, who have created a workflow to combine the 12-plex RNAscope™ assay with the 40-plus protein Imaging Mass Cytometry™ assay, to create RNA and protein multi-omics results.
- Outside of transcriptomics and proteomics, **AtlasXOmics** is the commercialised spin-off from DBiT-seq and currently allows users to incorporate epigenomic information in a spatial context alongside proteomics and transcriptomics, like the original technology.

We spoke to **Dr. Andrea Corsinotti**, our second Multi-omics core leader about his experience with these tools.


Mouse pup

INTERVIEW: ANDREA CORSINOTTI

SINGLE-CELL MULTI-OMICS FACILITY MANAGER, CENTRE FOR REGENERATIVE MEDICINE, INSTITUTE FOR REGENERATION AND REPAIR
UNIVERSITY OF EDINBURGH

***FLG:** Let's jump straight in with a question about your current role. What multi-omics capacity does your facility have?*

Andrea: Most of the work that we do at the moment is with reagents and workflows from 10x Genomics, which is the main player in the market for single-cell analysis. They have a plethora of reagents and kits and workflows for measuring different things, which ranges from just gene expression, ATAC-seq, CITE-seq for surface proteins, CRISPR screens, PCR profiling and many other applications. We also use reagents and workflows from Parse Biosciences, a different company offering similar solutions. For spatial biology, we use 10x Visium and BGI stereo-seq. We just bought a NanoString CosMx, and we will soon get the 10x Xenium.

***FLG:** As a core facility, what kind of questions do you get a lot of?*

Andrea: On our website we have a question and answer section. You'll find questions there such as – 'How many samples? How many genes/how many cells am I going to get?' Another very typical question we get is – 'How do I prepare my samples?' This is a tricky question, because if you work on barley flowers, or if you work on human brain, you are the experts in preparing cells from barley flowers or from human brain. I have a background in stem cell biology, so I can only help you prepare stem cells, in vitro cultures. The second very common question is - How much is this going to cost?' I wish there was a way to simply answer both questions, but there isn't.

***FLG:** Is sample prep something that people consistently have problems with even now?*

Andrea: It depends on the sample. For new experiments, most of the work that we do goes into the optimization of the sample prep. For all these technologies, what you put in is the main determinant of the quality of the data that you produce. There are very few things that we can change during the workflow. So, most of our expertise goes into helping people find the best ways to prepare their cells. The easiest way to do it is to have a goal, e.g., we need a single-cell suspension with these characteristics or a single-nuclei suspension with these characteristics. Then we can guide them to choose from the protocols or tools that are available through the various alternatives, until we get to a sample that is of good enough quality for us to process in our experiments.

***FLG:** Would you say that most of the multi-omics kits do what they say on the tin, or have you had any troubles with them?*

Andrea: When we work with commercially available products, they tend to deliver what they promise. There tends to be two obstacles. One: if you want to do an experiment in which, on top of gene expression, you want to screen a lot of surface proteins, there are very robust workflows for this type of experiment. The success depends on whether you have good antibodies to detect all the proteins that you want to detect. The kit will deliver, but there are other limitations that are not strictly related to the kit that can interfere with the results. Two: is it's not always that easy to make sense of the data that you're getting. The classical example in terms of multi-omics is a combination of single-cell RNA-seq with single-cell ATAC-seq from the same sample. Although the workflow works fine and we get what we are promised, people that I know that have been using this workflow often struggle making sense of the data because of the complexity that is associated with it.

FLG: Are there new challenges that emerge from specific multi-omics protocols as opposed to the mono-omics alternatives?

Andrea: One of the things that we still cannot do very well, at the single-cell level, relates to proteins. If you want an unbiased approach that doesn't require antibodies, the tools that are available are still in an early phase of development compared to RNA-seq. This is mostly because of the sequencing versus other technologies that are needed to reveal RNA vs proteins. One thing that people have asked me in terms of multi-omics is whether it is possible to perform gene expression analysis from a cell while also looking at intracellular proteins. This is very difficult or almost impossible to do at the moment. My background is gene regulation, so looking at how transcription factors interfere with a gene expression programme of a cell is a very interesting question. The fact that there are no well-established tools to do both things at once, at single-cell level, is a little bit frustrating. You can look for papers in which they use a home-made method to do something similar, but as a facility, we need to provide a service that is scalable and reliable. We need to work with something that is benchmarked, rather than something that has worked in one lab in one place.

FLG: What multi-omics combinations are people approaching you asking to do?

Andrea: It really depends. Sometimes people have a specific question, and they want to do an experiment in which you are combining two measurements. In that case, the answer can be yes or no, depending on the tools that are available. In other cases, people don't know about the existence of multi-omics tools. In that case, we try to suggest – 'Why don't you do this? Or why don't you also do that?' Sometimes people think that they will gain some information by adding measurements, and then we do the reality check – 'If you do these things in the easy way, you're going to get some data and probably some answers to both your questions. If you try to over complicate it too much, you may have more problems with the execution of the experiment and with the analysis of the data'. So, we try to deter them.

Unfortunately, it has to do with what is doable and what is not doable. With the current technologies, we know that single-cell RNA-seq is something that we can do very easily and generate such a large amount of data that people struggle to analyse. To add in additional measurements and combinations with other omic types really needs to be carefully thought out to avoid these kinds of situations in which the experiment is technically challenging, very expensive, and then you get data that is difficult to analyse and to understand.

> "WHAT HAPPENS TO THE OTHER COMPONENT IN EITHER METHOD DOESN'T MATTER, BECAUSE YOU'RE NOT GOING TO USE IT. YOU CAN SACRIFICE THE RNA TO PRESERVE THE CHROMATIN AND YOU CAN SACRIFICE THE CHROMATIN TO PRESERVE THE RNA."

FLG: There are tools now that allow you to profile two omics from the same cell or same nucleus. Another option is to split your cells and profile each omic from separate populations. Is it cheaper to do both on the same cell and is there a disadvantage to the split method?

Andrea: The main advantage of doing both in the same cell is that you can properly match the information in exactly the same style. One thing that we have noticed, for example, when we do single-cell RNA-seq and single-cell ATAC-seq, is that the starting material for the two assays is different. RNA-seq requires good quality RNA and ATAC-seq requires good quality chromatin. It is often difficult to have a sample preparation method that is gentle enough to maintain both types of materials.

Here, you can do the experiment together or you can do the experiment separately. If your interest is knowing exactly what happens in one cell, looking at the two bits of information together is the way to do it. If your aim is to have datasets that are meaningful individually, and that can be integrated, doing it separately is more successful. You will use a method to prepare cells for single-cell RNA-seq, which looks at RNA, and the method for single-cell ATAC-seq that looks at the chromatin. What happens to the other component in either method doesn't matter, because you're not going to use it. You can sacrifice the RNA to preserve the chromatin and you can sacrifice the chromatin to preserve the RNA. I think the costs are quite similar if you do them together or if you do them separately. It always goes back to the type of question that one has, rather than the cost.

Chapter 3 references

1. Baysoy, A., Bai, Z., Satija, R. & Fan, R. **The technological landscape and applications of single-cell multi-omics.** *Nature Reviews Molecular Cell Biology,* 1-19 (2023).
2. Vandereyken, K., Sifrim, A., Thienpont, B. & Voet, T. **Methods and applications for single-cell and spatial multi-omics.** *Nature Reviews Genetics,* 1-22 (2023).
3. Ogbeide, S., Giannese, F., Mincarelli, L. & Macaulay, I.C. **Into the multiverse: advances in single-cell multiomic profiling.** *Trends in Genetics* **38**, 831-843 (2022).
4. Dimitriu, M.A., Lazar-Contes, I., Roszkowski, M. & Mansuy, I.M. **Single-cell multiomics techniques: from conception to applications.** *Frontiers in Cell and Developmental Biology* **10**, 854317 (2022).
5. Wang, X., Wu, X., Hong, N. & Jin, W. P**rogress in single-cell multimodal sequencing and multi-omics data integration.** *Biophysical Reviews* (2023).
6. Macaulay, I.C. et al. **G&T-seq: parallel sequencing of single-cell genomes and transcriptomes.** N*ature methods* **12**, 519-522 (2015).
7. Dey, S.S., Kester, L., Spanjaard, B., Bienko, M. & Van Oudenaarden, A. **Integrated genome and transcriptome sequencing of the same cell.** *Nature biotechnology* **33**, 285-289 (2015).
8. Han, K.Y. *et al.* **SIDR: simultaneous isolation and parallel sequencing of genomic DNA and total RNA from single cells.** *Genome research* **28**, 75-87 (2018).
9. Rodriguez-Meira, A., O'Sullivan, J., Rahman, H. & Mead, A.J. **TARGET-Seq: a protocol for high-sensitivity single-cell mutational analysis and parallel RNA sequencing.** *STAR protocols* **1**, 100125 (2020).
10. Zachariadis, V., Cheng, H., Andrews, N. & Enge, M. **A highly scalable method for joint whole-genome sequencing and gene-expression profiling of single cells.** *Molecular Cell* **80**, 541-553. e5 (2020).
11. Yu, L. *et al.* **scONE-seq: A single-cell multi-omics method enables simultaneous dissection of phenotype and genotype heterogeneity from frozen tumors.** *Science Advances* **9**, eabp8901 (2023).
12. Genshaft, A.S. *et al.* **Multiplexed, targeted profiling of single-cell proteomes and transcriptomes in a single reaction.** *Genome biology* **17**, 1-15 (2016).
13. Stoeckius, M. *et al.* **Simultaneous epitope and transcriptome measurement in single cells.** *Nature methods* **14**, 865-868 (2017).
14. Peterson, V.M. *et al.* **Multiplexed quantification of proteins and transcripts in single cells.** *Nature biotechnology* **35**, 936-939 (2017).
15. Gerlach, J.P. *et al.* **Combined quantification of intracellular (phospho-) proteins and transcriptomics from fixed single cells.** *Scientific reports* **9**, 1469 (2019).
16. Rivello, F. *et al.* **Single-cell intracellular epitope and transcript detection reveals signal transduction dynamics.** *Cell Reports Methods* **1**(2021).
17. Reimegård, J. *et al.* **A combined approach for single-cell mRNA and intracellular protein expression analysis.** *Communications Biology* **4**, 624 (2021).
18. Hwang, B. *et al.* **SCITO-seq: single-cell combinatorial indexed cytometry sequencing.** *Nature Methods* **18**, 903-911 (2021).
19. Mimitou, E.P. *et al.* **Multiplexed detection of proteins, transcriptomes, clonotypes and CRISPR perturbations in single cells.** *Nature methods* **16,** 409-412 (2019).
20. Frangieh, C.J. *et al.* **Multimodal pooled Perturb-CITE-seq screens in patient models define mechanisms of cancer immune evasion.** *Nature genetics* **53**, 332-341 (2021).
21. Chung, H. *et al.* **Joint single-cell measurements of nuclear proteins and RNA in vivo.** *Nature methods* **18,** 1204-1212 (2021).
22. Hu, Y. *et al.* **Simultaneous profiling of mRNA transcriptome and DNA methylome from a single cell.** *Single Cell Methods: Sequencing and Proteomics,* 363-377 (2019).
23. Angermueller, C. *et al.* P**arallel single-cell sequencing links transcriptional and epigenetic heterogeneity.** *Nature methods* **13**, 229-232 (2016).
24. Bartosovic, M., Kabbe, M. & Castelo-Branco, G. **Single-cell CUT&Tag profiles histone modifications and transcription factors in complex tissues.** *Nature biotechnology* **39**, 825-835 (2021).
25. Xiong, H., Luo, Y., Wang, Q., Yu, X. & He, A. **Single-cell joint detection of chromatin occupancy and transcriptome enables higher-dimensional epigenomic reconstructions.** *Nature Methods* **18**, 652-660 (2021).
26. Zhu, C. *et al.* **Joint profiling of histone modifications and transcriptome in single cells from mouse brain.** *Nature methods* **18**, 283-292 (2021).
27. Zhang, B. *et al.* **Characterizing cellular heterogeneity in chromatin state with scCUT&Tag-pro.** *Nature biotechnology* **40**, 1220-1230 (2022).
28. Chen, S., Lake, B.B. & Zhang, K. **High-throughput sequencing of the transcriptome and chromatin accessibility in the same cell.** *Nature biotechnology* **37**, 1452-1457 (2019).
29. Zhu, C. *et al.* **An ultra high-throughput method for single-cell joint analysis of open chromatin and transcriptome.** *Nature structural & molecular biology* **26**, 1063-1070 (2019).
30. Ma, S. *et al.* **Chromatin potential identified by shared single-cell profiling of RNA and chromatin.** *Cell* **183**, 1103-1116. e20 (2020).
31. Xu, W. *et al.* **ISSAAC-seq enables sensitive and flexible multimodal profiling of chromatin accessibility and gene expression in single cells.** *Nature Methods* 19, 1243-1249 (2022).
32. Swanson, E. *et al.* **Simultaneous trimodal single-cell measurement of transcripts, epitopes, and chromatin accessibility using TEA-seq.** *Elife* **10**, e63632 (2021).
33. Blair, J.D. *et al.* **Phospho-seq: Integrated, multi-modal profiling of intracellular protein dynamics in single cells.** *bioRxiv* (2023).
34. Guo, F. *et al.* **Single-cell multi-omics sequencing of mouse early embryos and embryonic stem cells.** *Cell research* **27**, 967-988 (2017).
35. Gu, C., Liu, S., Wu, Q., Zhang, L. & Guo, F. **Integrative single-cell analysis of transcriptome, DNA methylome and chromatin accessibility in mouse oocytes.** *Cell research* **29**, 110-123 (2019).
36. Clark, S.J. *et al.* **scNMT-seq enables joint profiling of chromatin accessibility DNA methylation and transcription in single cells.** *Nature communications* **9**, 781 (2018).
37. Wang, Y. *et al.* **Single-cell multiomics sequencing reveals the functional regulatory landscape of early embryos.** *Nature communications* **12**, 1247 (2021).
38. Yan, R. *et al.* **Decoding dynamic epigenetic landscapes in human oocytes using single-cell multi-omics sequencing.** *Cell Stem Cell* **28**, 1641-1656. e7 (2021).
39. Li, G. *et al.* **Joint profiling of DNA methylation and chromatin architecture in single cells.** *Nature methods* **16**, 991-993 (2019).
40. Lee, D.-S. *et al.* **Simultaneous profiling of 3D genome structure and DNA methylation in single human cells.** *Nature methods* **16**, 999-1006 (2019).
41. Hou, Y. *et al.* **Single-cell triple omics sequencing reveals genetic, epigenetic, and transcriptomic heterogeneity in hepatocellular carcinomas.** *Cell research* **26**, 304-319 (2016).
42. Fiskin, E. *et al.* **Single-cell profiling of proteins and chromatin accessibility using PHAGE-ATAC.** *Nature Biotechnology* **40**, 374-381 (2022).
43. Mimitou, E.P. *et al.* **Scalable, multimodal profiling of chromatin accessibility, gene expression and protein levels in single cells.** *Nature biotechnology* **39**, 1246-1258 (2021).
44. Chen, A.F. *et al.* **NEAT-seq: simultaneous profiling of intra-nuclear proteins, chromatin accessibility and gene expression in single cells.** *Nature Methods* **19**, 547-553 (2022).
45. Shah, S., Lubeck, E., Zhou, W. & Cai, L. **seqFISH accurately detects transcripts in single cells and reveals robust spatial organization in the hippocampus.** *Neuron* **94,** 752-758. e1 (2017).
46. Xia, C., Fan, J., Emanuel, G., Hao, J. & Zhuang, X. **Spatial transcriptome profiling by MERFISH reveals subcellular RNA compartmentalization and cell cycle-dependent gene expression.** *Proceedings of the National Academy of Sciences* **116**, 19490-19499 (2019).
47. Wu, Y., Cheng, Y., Wang, X., Fan, J. & Gao, Q. **Spatial omics: Navigating to the golden era of cancer research.** *Clinical and Translational Medicine* **12**, e696 (2022).
48. Liu, Y. *et al.* **High-Spatial-Resolution Multi-Omics Sequencing via Deterministic Barcoding in Tissue.** *Cell* **183**, 1665-1681.e18 (2020).
49. Ben-Chetrit, N. *et al.* **Integration of whole transcriptome spatial profiling with protein markers.** *Nature Biotechnology* **41**, 788-793 (2023).
50. Vickovic, S. *et al.* **SM-Omics is an automated platform for high-throughput spatial multi-omics.** *Nature Communications* **13,** 795 (2022).
51. Liu, Y. *et al.* **Spatial-CITE-seq: spatially resolved high-plex protein and whole transcriptome co-mapping.** *Research Square* (2022).
52. Xia, C., Fan, J., Emanuel, G., Hao, J. & Zhuang, X. **Spatial transcriptome profiling by MERFISH reveals subcellular RNA compartmentalization and cell cycle-dependent gene expression.** *Proc Natl Acad Sci U S A* **116**, 19490-19499 (2019).
53. Takei, Y. *et al.* **Integrated spatial genomics reveals global architecture of single nuclei.** *Nature* **590**, 344-350 (2021).
54. Nguyen, H.Q. *et al.* **3D mapping and accelerated super-resolution imaging of the human genome using in situ sequencing.** *Nature Methods* **17**, 822-832 (2020).
55. Zhang, D. *et al.* S**patial epigenome–transcriptome co-profiling of mammalian tissues.** *Nature* **616**, 113-122 (2023).
56. Li, X. **Harnessing the potential of spatial multiomics: a timely opportunity.** *Signal Transduction and Targeted Therapy* **8**, 234 (2023).

CHAPTER 4

DECONVOLUTING DOGMA. DNA, RNA AND PROTEIN MULTI-OMICS

DNA IS TRANSCRIBED TO RNA, WHICH IS TRANSLATED TO PROTEIN. THIS IS THE GENOMIC DOGMA. IN THIS CHAPTER, WE WILL OUTLINE THE APPLICATIONS OF, AND METHODS TO, SEQUENCE THE GENOME, TRANSCRIPTOME AND/OR PROTEOME CONCURRENTLY. WE WILL SPECIFICALLY FOCUS ON THE EMERGING TECHNOLOGIES AND COMPUTATIONAL TOOLS FOR SINGLE-CELL AND SPATIAL TRANSCRIPTOMICS, AND PROTEOMICS.

Some of the first multi-omics tools were methods to co-profile DNA and RNA. Consequently, this has become a widely relied upon multi-omics method, with distinct insights.

As briefly covered in Chapter 3, transcriptomics and proteomics is a newer area of development, and new tools are consistently being released. Uniting protein data with RNA is also not straightforward, with several specific tools for RNA/protein integration already mentioned in Chapter 2. On a base level, incorporating proteomics with transcriptomics allows for more robust cell typing than can be gained from transcriptomics alone, because the proteome reflects ongoing cellular processes rather than 'primed' processes (see Figure 4.1). However, as this chapter will show, there is more utility to this multi-omic method than first meets the eye.

Linking Genotype to Phenotype

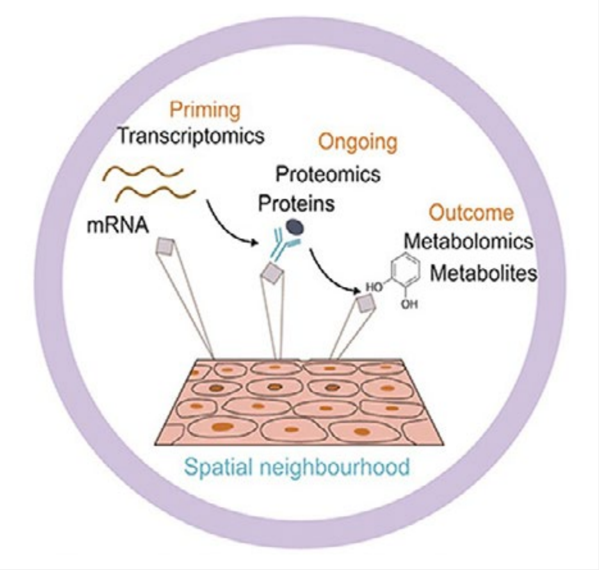
We'll begin with a brief overview of the advantages of looking at DNA information alongside RNA signals. This can be performed with methods described in Chapter 3, such as G&T-seq[2], scONE-seq[3] and Mission Bio's Tapestri v3 platform for single-cell.

The direct benefit of combining DNA and RNA measurements in a single cell is the ability to link DNA-based functional genetic variants to specific cell-type variation in gene expression[4]. This means that the downstream effects of disease-associated genetic loci can be linked to cellular consequences, to ultimately provide deeper insights into the molecular and cellular mechanisms involved with disease risk[5].

For diseases in which somatic genetic variation plays a role, such as cancer[6], Alzheimer's[7] and Parkinson's disease[8], these methods are important for understanding disease pathogenesis. Findings from these studies have shown phenomena such as distinct transcriptional consequences to acquired DNA copy number aberrations, showing that genotype does not carry directly into transcriptional phenotype[9].

FIGURE 4.1. DNA IS THE CODE, TRANSCRIPTOMICS SHOWS US WHAT IS PRIMED, PROTEOMICS TELLS US WHAT IS ONGOING, AND METABOLOMICS TELLS US WHAT HAS HAPPENED.

Image Credit: Fangma, et al. [1]



These tools are also valuable for studying the efficacy and safety of genome editing in the germline. They can assess the on-target and off-target genome edits from CRISPR-Cas[9] alongside the phenotypic consequences of both[10].

Linking Transcriptomics to Proteomics

The focus for the rest of this chapter will be on the technologies and applications that link transcriptomics to proteomic information. All cellular process and functions revolve around proteins. For example, they make up the structure of cells, they perform biochemical process via their role as enzymes, and they are the receptors and ligands of cellular communications[5,11].

As previously stated, proteomics ensures more robust cell typing than is produced from transcriptomics. Capturing transcriptomics and proteomics can also provide functional information that cannot be captured by genomics alone[12]. Proteomics has revealed the effects of genetic variants in conditions otherwise undetectable from RNA analysis[13]. We direct readers to this specific review from 2023 on the latest transcriptomic and proteomic methods[14].

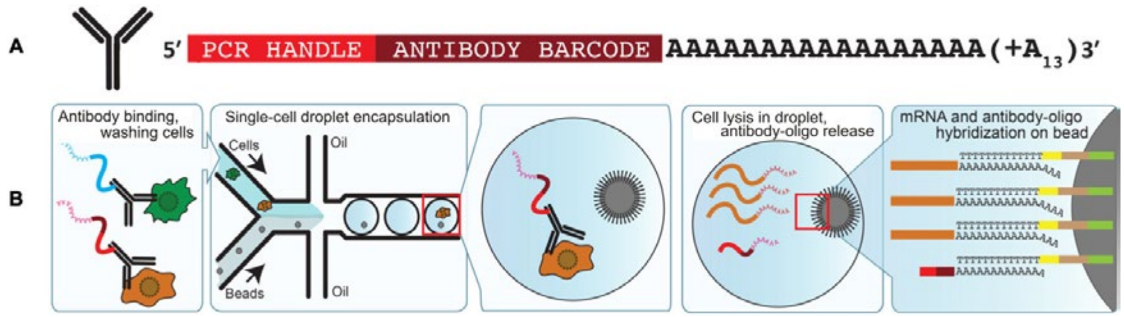
Transcriptomics and proteomics has proved a popular multi-omics combination at the single-cell level, with tools such as CITE-seq[15] (see Figure 4.2) and REAP-seq[16] proving very popular. This is also an exciting combination at the spatial level, with commercial offerings such as 10x's Visium and Xenium and NanoString's GeoMx and CosMx, alongside the latest in-house tools such as SPOTS[17], SM-Omics[18] and Spatial-CITE-seq[19].

One of the issues with proteomic data is the difficulty in profiling the whole 'proteome'. Most methods can only profile a number of pre-selected protein targets. These methods also tend to only target extracellular proteins, since these are easily isolated alongside the transcriptome. By profiling these proteins, these techniques provide additional information about cell identity and cell state depending on which proteins are present on the surface of the cell. These methods are a long shot from an un-biased profiling of all intra and extracellular proteins.

The first transcriptomic/proteomics methods used fluorescent antibodies to label proteins. CITE-seq and REAP-seq changed the game by introducing oligonucleotide-conjugated antibodies (see Figure 4.2A), which contained a PCR handle making the protein measurements compatible with sequencing technologies. This is now the most widely used method. A very recent method, PHAGE-ATAC[21], uses a unique protein tagging method. The protein recognition is based on nanobody-displaying phages instead, which allows the reliable detection of cell-surface proteins across thousands of cells.

FIGURE 4.2. OVERVIEW OF THE CITE-SEQ WORKFLOW.

*(**A**) oligonucleotide-conjugated antibody. (**B**) RNA-seq and antibody tagging workflow. Image Credit: Timp and Timp [20]*


A
5′
PCR HANDLE
ANTIBODY BARCODE
AAAAAAAAAAAAAAAAA(+A13)3′
B
Antibody binding, washing cells
Single-cell droplet encapsulation
Cells
Oil
Beads
Oil
Cell lysis in droplet, antibody-oligo release
mRNA and antibody-oligo hybridization on bead

Another distinction between the current single-cell methods for RNA/protein measurements is the type of protein targeted. Most methods target extracellular proteins but some target intranuclear proteins such as inCITE-seq[22] and NEAT-seq[23], and intracellular proteins SPARC[24] and RAID-seq[25]. InCITE-seq and NEAT-seq are performed on nuclei in which nuclear membranes are made permeable with a light formaldehyde treatment to access the proteins. Methods such as SPARC profile intracellular proteins by removing a protein-containing supernatant from the cell to profile.

In April 2023, a new method was released called TRAPS-seq[26]. This method capture proteins in a novel manner. Proteins that are secreted by a cell (i.e., the proteins a cell releases to influence other cells) can be captured on the cell surface as they are released. These are then probed by oligonucleotide-barcoded antibodies and are sequenced alongside the transcriptome of the cell in a time-resolved manner. This method was used to measure cytokine secretion and presents a new tool for seamless integration of secretomics and transcriptomics[27].

Imaging the spatial transcriptome and proteome

Recent advances in spatial transcriptomics and proteomics means that imaging either or both at subcellular resolution and/or at hi-plex is now possible[28].

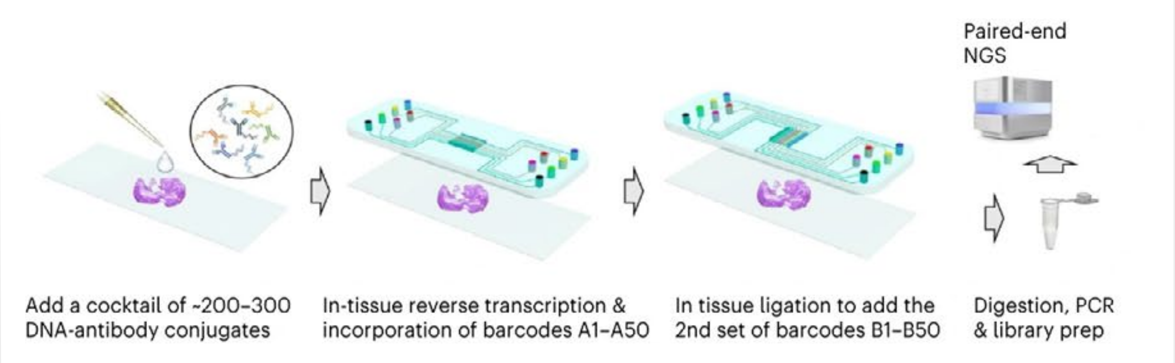
Subcellular profiling has been achieved with non-proteomic spatial multi-omics technologies such as seqFISH[29] and MERFISH[30,31] and DNAseqFISH+[32]. Recently, Stereo-CITE-seq[33] was published, which is a brand new method based on the Stereo-seq nanoball technology[34] that allows very high spatial resolution profiling of 10s of proteins alongside whole transcriptomics. This brings proteotranscriptomics to subcellular resolution.

When looking at hi-plex, SPOTS[17] is a new tool which supplements the 10x Visium process by recording a panel of >30 intra- and extra-cellular proteins for simultaneous proteomics. SM-Omics[18] is a similar new tool based on the same DNA-barcoded antibodies but allows up to 64 extracellular protein targets.

The marriage of number of markers and high resolution is currently found with Spatial-CITE-seq[35], published in early 2023. This is a very high-plex protein and whole transcriptome co-mapping method (see Figure 4.3) since this method allows the profiling of 200-300 extracellular proteins at 20 micron resolution. The technology behind NanoString's CosMx Spatial Molecular Imager (SMI)[36] also allows ~ 100 of extracellular proteins to be profiled and at subcellular resolution. There a comparatively fewer targets but it achieves a better spatial resolution.

FIGURE 4.3. OVERVIEW OF SPATIAL-CITE-SEQ METHODOLOGY.

Image Credit: Liu, et al. [35]


Paired-end
NGS
Add a cocktail of ~200–300
DNA-antibody conjugates
In-tissue reverse transcription &
incorporation of barcodes A1–A50
In tissue ligation to add the
2nd set of barcodes B1–B50
Digestion, PCR
& library prep

It is still the case that current spatial transcriptomic and proteomic methods are limited in several regards, either in resolution, in number of protein targets, or the fact they are based on serial characterisation of the two modalities rather than parallel profiling. However, 2023 has seen improvements in both resolution and target number, suggesting these limitations will still be improved come 2024.

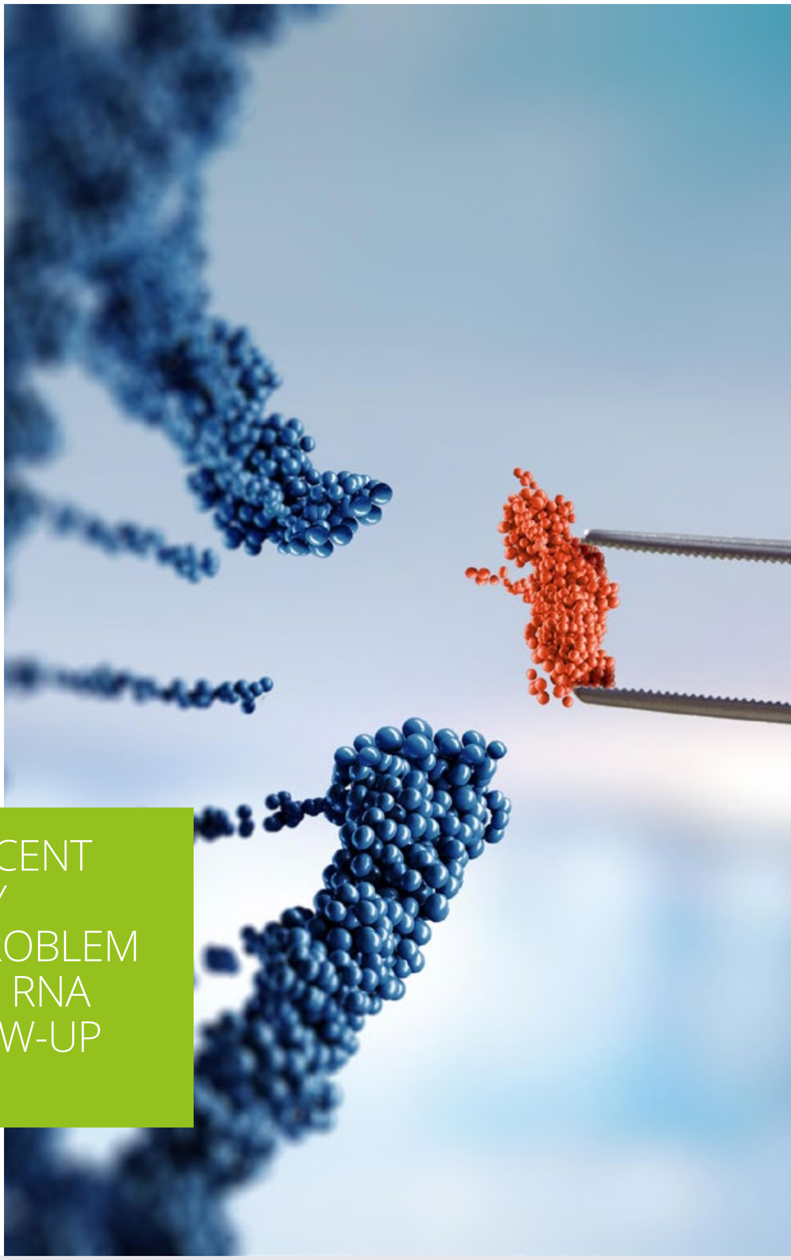
Computational Tools for integrated transcriptomic and proteomic insights

Some methods that we have already covered for integration, such as DeepMAPS[37], work across different modalities, but there are tools specifically addressing the challenge of integrating RNA and protein. Early tools such as citeFUSE[38] and BREM-SC[39] work via network-based and Bayesian mixture models respectively.

Slightly more recent tools, such as totalVI[40] and sciPENN[41], rely on AI models. totalVI is a deep generative model and addresses specific problems with each data type, such as the differences in noise levels and the elevated background of protein antibody-based methods. It is fast becoming a relied upon methods for CITE-seq data. sciPENN is a more recent tool, also using deep learning, but is specialised for handling the batch effects and issues with minimal overlap that come from integrating multiple CITE-seq datasets.

A promising and recent method specifically designed for the problem of integrating protein and RNA data is MARIO, and its follow-up MaxFuse. These tools are designed for the situation that current transcriptomic and proteomic methods end up with weak linkage. This refers to the situation in which there are very few matching points between two datasets, i.e., the situation in which a handful of RNAs have had proteins profiled too, and hence operate as the weak linkage.

We spoke to **Professor Zongming Ma**, one of the senior authors of both MARIO and MaxFuse, to learn more about how his computational tools work.



“A PROMISING AND RECENT METHOD SPECIFICALLY DESIGNED FOR THE PROBLEM OF INTEGRATING PROTEIN AND RNA DATA IS MARIO, AND ITS FOLLOW-UP MAXFUSE. "

INTERVIEW:

ZONGMING MA

PROFESSOR, DEPARTMENT OF STATISTICS AND DATA SCIENCE YALE UNIVERSITY

FLG: ***Could you describe your research journey and current research interests?***

Zongming: Sure. I come from a statistics and machine learning background. Earlier in my career, my research focused on the theory side of data analysis. About five years ago I started to also work on real data motivated method development. In the last three years, I have spent a lot of time and energy working on data integration, especially data integration related to multi-omics, from both algorithm and theory perspectives. This is my path from statistics to multi-omics data analysis.

FLG: ***Next I wanted to talk about some of the tools you've been involved with, MARIO[42] and MaxFuse[43]. From my understanding, Mario comes before MaxFuse, so could you maybe start with Mario and the approach it takes for multi-omics data integration, and why you chose to produce this tool?***

Zongming: In a sense, MARIO is the predecessor of MaxFuse, but each tool is addressing a different challenge. For Mario, in collaboration with Garry Nolan's lab at Stanford and Sizun Jiang at Harvard, we considered the setting where one has measurements of proteins in targeted panels on different cells or from different datasets, and the number of proteins measured by both datasets is small.

> "THE MAJOR DIFFICULTY IS THIS – IF YOU HAVE THE SAME BIOMARKER BUT MEASURED DIFFERENTLY, AND THE NUMBER OF BIOMARKERS THAT YOU MEASURE AT THE SAME TIME IS LIMITED... HOW CAN YOU PERFORM INTEGRATION IN SUCH A SETTING?"

Suppose you only have 10 or 15 proteins that are measured in both datasets, potentially by different technologies. For example, if you have CITE-seq data from sequencing and if you have a spatial proteomics dataset from immune-fluorescence imaging, then you're measuring the same features, but the measurement technologies are different. So, the challenge is, how do you align the different measurements of the same features and then try to match cells based on this alignment? The major difficulty is this – if you have the same biomarker but measured differently, and the number of biomarkers that you measure at the same time is limited... how can you perform integration in such a setting? This is the challenge addressed by MARIO.

FLG: ***And then where does MaxFuse go from there?***

Zongming: Once you have MARIO, then as long as you have single-cell CITE-seq data on a certain tissue sample, you can try to link cells in the CITE-seq dataset with those in a related spatial protein dataset collected from a comparable tissue sample. After you use the protein part of the CITE-seq data and completed this integration, you could then map the RNA information of each cell to its match in the spatial dataset, and thus create a spatial transcriptomic dataset in silico.

However, single-cell CITE-seq is not as widely adopted as single-cell RNA-seq. So, in order to make this kind of tool really suitable for the most common type of single-cell dataset, the natural question to ask is, can you do this without having protein measurements in the sequencing domain? Can you directly map RNA-seq information onto a spatial protein dataset? So, this is the motivating question we tried to answer when developing MaxFuse with Garry and Nancy Zhang at Penn.

FLG: ***Can you briefly explain how MaxFuse works?***

Zongming: Sure. MaxFuse has three major steps.

The first step is to obtain an informative prediction of protein abundances based on the expression levels of their coding genes. Direct linear prediction is not ideal. It's a relatively weak prediction, because there are other things that are regulating the translation process, but at least this gives us something that we can hinge on to start with. Then because this prediction is not going to pinpoint the cells directly, we adopt the idea of 'shrinkage', also known as 'smoothing', to improve it. That is, you try to bring in a bit of extra information by looking around at those "who are close to you", and see how they behave, and then try to average within neighbourhoods so that you can effectively increase the signal-to-noise ratio in prediction. For that purpose, having a lot of RNA features is helpful, because, with the entire transcriptome, you can define cell state at a very fine scale, and better distinguish "who are close to you". This provides a sufficiently good prediction that allows a crude initial matching between two modalities.

In the second step, we iteratively improve this matching by cycling through canonical correlation co-embedding, smoothing, and matching.

In the final step, we take the output of the iterative refinement procedure, and produce the final matching and the final integration.

FLG: ***And MaxFuse was used in the HuBMAP human intestine study[44]. Can you describe how it was used in that?***

Zongming: HuBMAP is a consortium level data collection effort. Teams at Stanford (Snyder, Nolan, and Greenleaf labs), working with human intestines, have CODEX images for certain sections and then they have separate, single-nucleus ATAC-seq and single-nucleus RNA-seq, and some 10x Multiome measurements - where both ATACs and RNAs are collected at the same time at the single-cell level. However, there's no simultaneous collection of protein and RNA, or protein and ATAC in this data collection process. The natural question to ask is, can we create certain spatial transcriptomic maps given the amount of data that people have spent a lot of time, money, and energy collecting?

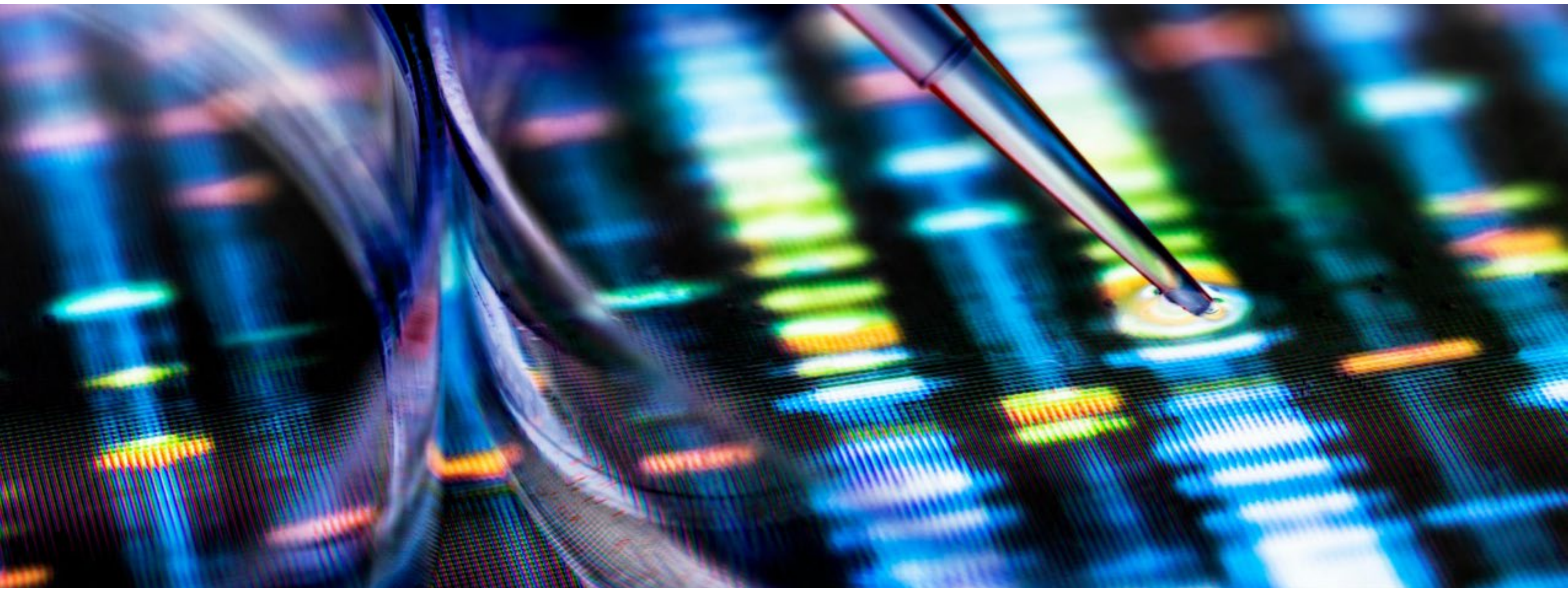
What we showed was, yes, you can do it by mapping the RNA information onto these CODEX images. And so, in the HubMAP human intestine paper, we mapped RNA onto the spatial CODEX data and then in our MaxFuse paper, we also mapped epigenome information onto the CODEX data.

Other computational tools have been specifically designed for the unique challenge of analysing proteomic and transcriptomic data. One example is CITEMO[45], a flexible framework to comprehensively explore single-cell multi-omics data. This tool helps identify cell subtypes and cell states from this combined data.

Other tools to help with clustering and cell identification have been released in the last 12-18 months. DEMOC[46] is a deep embedded learning model that jointly clusters CITE-seq data. It outperforms existing methods, achieving a more stable performance. MMoCHi[47] is a cell classifier reconciling gene and protein expression without reliance on atlases. CellCharter[48] is a scalable algorithmic framework for identifying, characterising and comparing cellular niches between heterogenous spatial transcriptomic and proteomic data samples.

Another interesting set of tools can use existing CITE-seq data to predict cell surface protein expression from the scRNA-seq data. A recent example of this tool is called CrossmodalNet[49], which is an interpretable deep learning model that can accurately predict cell surface protein abundance based on transcriptomic data. A similar example is TransPro[50], which can predict cell-specific chemical proteomic profiles after chemical perturbation even though the model is only trained on scRNA-seq data.

Finally, the need for stable reference data is becoming necessary with the influx of high-throughput multi-omics data. Reference data allows researchers to calibrate the accuracy and reproducibility of their omics workflows. A recent example of reference data has been provided for the human transcriptome and proteome[51].



Examples of applications

Given the newness of technologies to profile the transcriptome and proteome, and the unique challenge with integrating the data, use cases are not that abundant. We are seeing increasing numbers of studies each year. However, most still integrate transcriptomic and proteomic data from separate experiments.

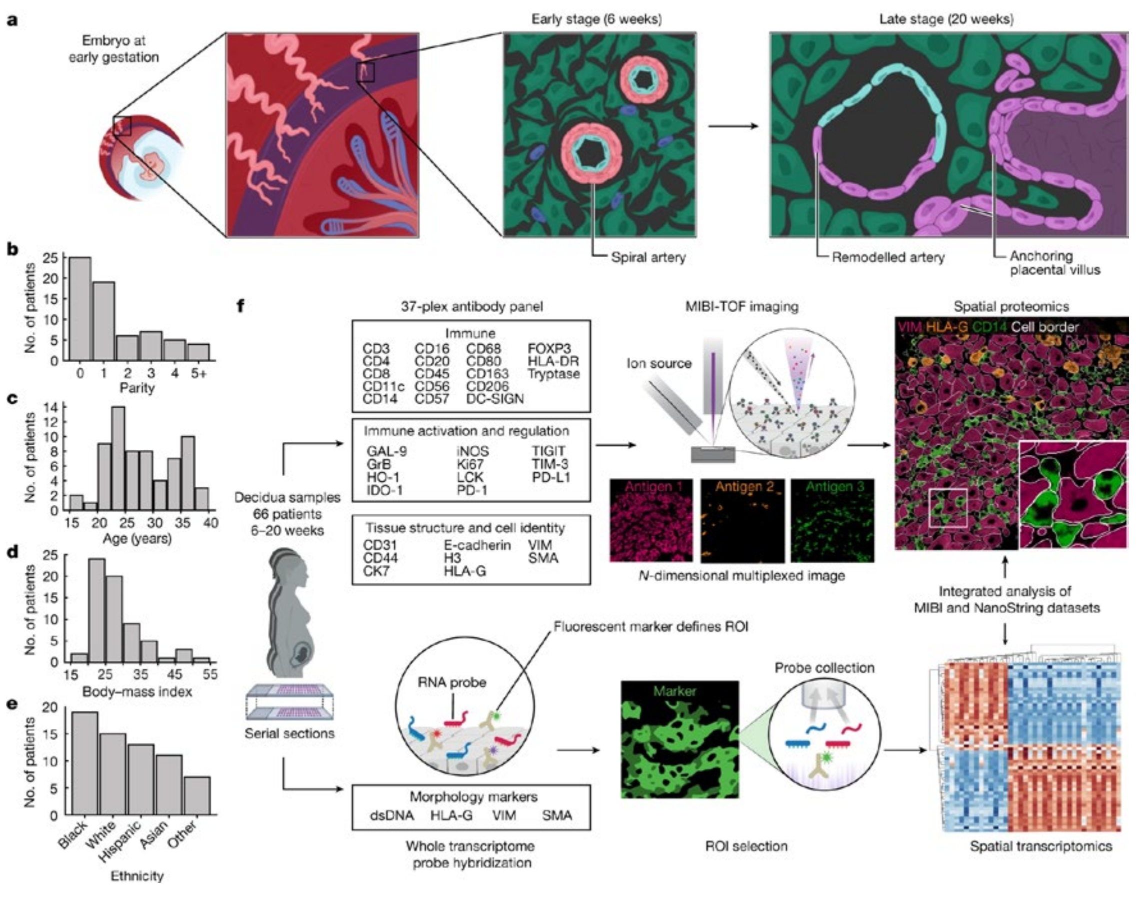
High-profile example cases of proteotranscriptomics have included several projects from the HuBMAP consortium, who released two studies in July 2023 showcasing spatial transcriptomics and proteomics. Furthermore, the HuBMAP consortium have recently released a uniform CITE-seq processing pipeline[52] in a bid for uniformly processing and indexing multi-omics single-cell data.

The first of these two studies[44] was the study on the human intestine mentioned by Professor Zongming Ma. Here, 10x Multiome data and proteomics data from CODEX were profiled (54 markers) and the proteomics data and 10x Multiome data were used in conjunction to provide highly accurate cell type distinctions. Since this was an atlasing project, this multi-omics approach is there to provide others with rich cell profiles to work from in the future.

In the other study[53], researchers combined the strengths of spatial transcriptomics and proteomics to construct a spatiotemporal atlas of the human maternal-fetal interface in the first half of pregnancy. 500,000 cells from 66 individuals were profiled with a MIBI-TOF 37-plex panel providing subcellular proteomics and Nanostring DSP spatial transcriptomics (see Figure 4.4). We recently spoke to **Dr. Shirley Greenbaum**, first author of this study about what she gained from a multi-omics approach here.

FIGURE 4.4. OVERVIEW OF MATERNAL-FETAL INTERFACE MULTI-OMICS STUDY[53]

*(**A**), Diagram of a human embryo in utero at 6 weeks of gestation. Left, the maternal–fetal interface consisting of decidua basalis (purple) with maternal spiral arteries (light pink) and fetal chorionic villi in the intervillous space (bottom right corner). Middle and right, early-stage (6 weeks) unremodelled spiral artery and progression to late-stage (20 weeks) remodelled artery and anchoring fetal villi. (**B**), Cohort parity distribution. (**C**), Cohort age distribution. (**D**), Cohort distribution of body–mass index. (**E**), Cohort ethnicity distribution. (**f**), TMA construction and serial sections for multi-omics workflow. Top, antibody panel, MIBI acquisition and spatial proteomics data extraction. Bottom, morphology marker panel and probe diagram, NanoString DSP ROI selection and spatial transcriptomics data extraction. Image & Caption Credit: Greenbaum, et al.* [53]



INTERVIEW:
SHIRLEY GREENBAUM

POSTDOCTORAL FELLOW, DEPARTMENT OF PATHOLOGY, **STANFORD UNIVERSITY**, RESIDENT, DEPARTMENT OF OBSTETRICS AND GYNAECOLOGY, **HADASSAH-HEBREW UNIVERSITY MEDICAL CENTER**

***FLG:* Can you describe the multi-omics approach that you deployed in your Nature maternal-fetal interface study?**

Shirley: The core approach of our study was measuring the expression of proteins at the maternal-fetal interface at the single cell level. Processes in any tissue are driven by its cells. By better understanding these cells – their size, shape and the proteins they express – we can gain deeper insights. Cells use these proteins to communicate with their neighbouring cells, and by studying single cells in the placenta we can begin to understand what drives placentation and immune tolerance during pregnancy.

Until very recently we were very limited in the number of protein targets we were able to measure for each cell. The novelty of using Multiplexed Ion Beam Imaging (MIBI) is that we can detect almost 40 targets simultaneously for each and every cell, and also capture the spatial arrangement of these cells in the tissue in relation to tissue features (e.g., arteries, glands). So, for example, if before we were only able to say that around half of all cells in the tissue are immune cells, now we can determine whether each individual cell is a T cell, what T cell sub-population it belongs to, whether that specific cell is showing signs of “exhaustion”, whether it is inducing immune tolerance, and what proteins it is expressing to do so.

In the second part of our study, we used Nanostring GeoMx® DSP to measure expression of genes by interstitial and intravascular EVTs. These two complementary methods enabled us to depict, with a very high resolution, the processes that take place in the tissue during this critical time of placentation.

***FLG:* What were some of the major findings using this spatial multi-omics approach?**

Shirley: One of our main findings was deciphering the enigmatic relationship between the remodelling of maternal vessels and the invasion of fetal cells into the uterus. We know that in normal pregnancy, these arteries, which are normally coiled and constricted, dilate extensively to become wide flaccid vessels that can transfer low velocity, low pressure blood flow to the placenta (this process, for example, does not occur smoothly in preeclampsia). Using the single-cell data generated by MIBI, we were able to establish the relationship between the invasion of fetal cells to these arteries, and to the remodelling of these arteries.

It was surprising to discover that the maternal artery remodelling process was not driven by adjacent maternal immune cells, but rather by the invasion of fetal cells. This implies that perhaps it is the fetus that is driving the remodelling of its mother's arteries, and not the mother; because abnormal remodelling of arteries is a pathological characteristic of preeclampsia. These findings may lead to a better understanding of this disease.

***FLG:* How do you hope people will use this spatio-temporal atlas?**

Shirley: I really hope that researchers will make use of the huge amount of data that is included in the maternal fetal atlas to answer other clinical and basic science questions. That is why I am so happy that the data is available to the scientific community through the HuBMAP project. The HuBMAP Consortium is a joint effort of several groups from Stanford University and other leading research institutions, set out to establish a global and open portal of single-cell datasets. This portal contains more than 1900 datasets from more than 30 organs that were collected using various single-cell technologies of the highest quality. It’s an amazing platform that enables researchers to query and visualize the data in a very intuitive and accessible way.

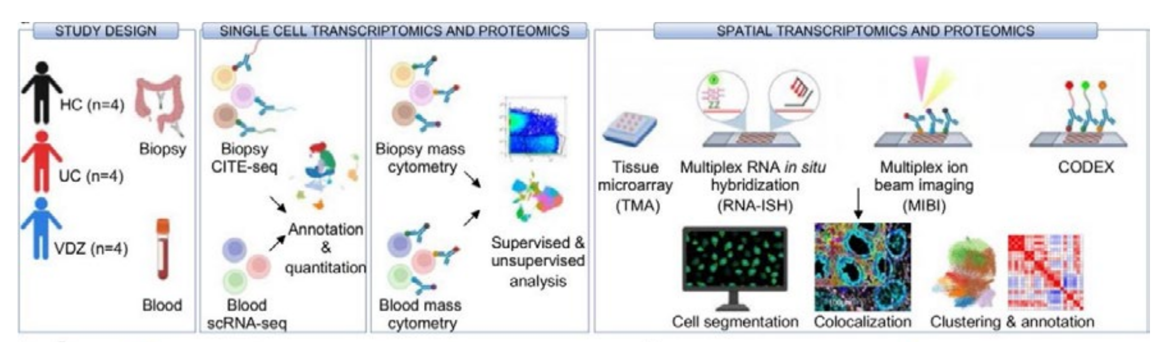
Beyond atlasing, we have seen multi-omics used more functionally, revealing clinical insights. In multiple sclerosis, CITE-seq revealed inhibition of specific pathway (NF-κB) in B cells when patients are treated with vitamin D[54]. In another example, in the original publication of SPOTS there was a spatial bias for immune dysfunction in spleen tumours[17].

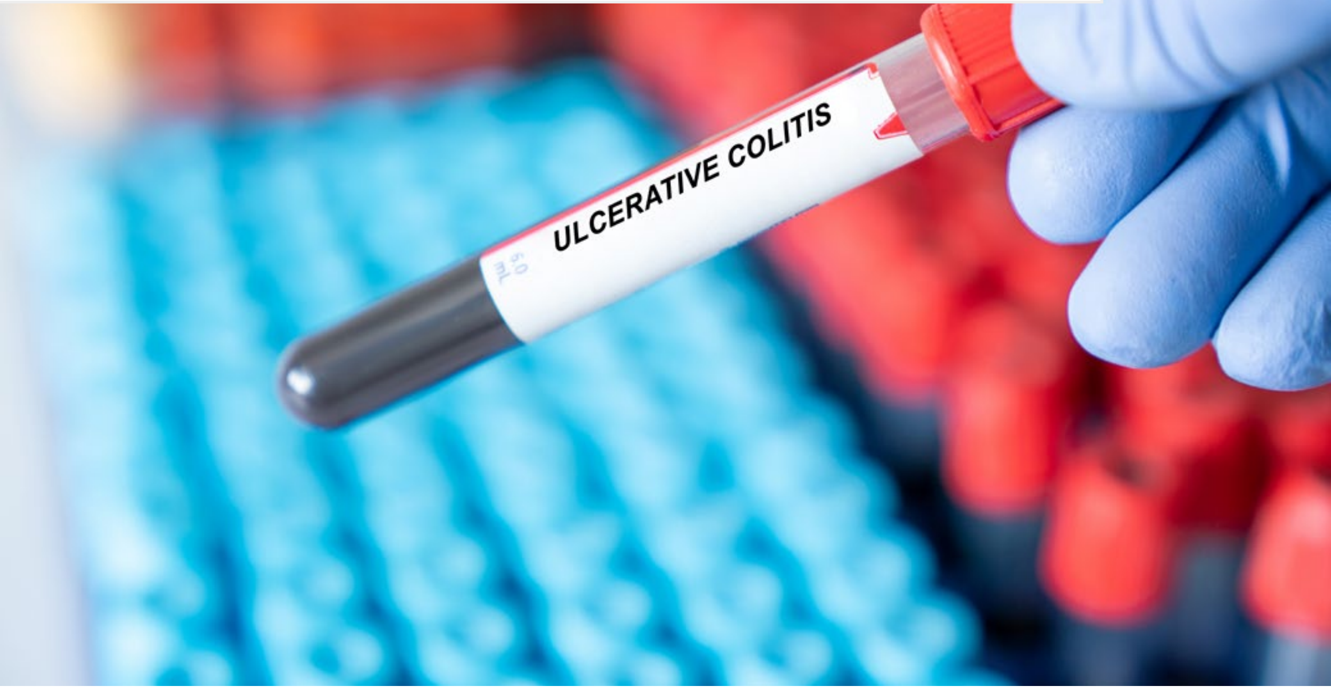
An example from 2023 is the proteo-transcriptomic map of non-alcoholic fatty liver disease (NAFLD) signatures[55]. Here, patient plasmas were analysed with the SomaScan, and a matched liver biopsy was sequenced for transcriptomics. The resulting multi-omic dataset was used to find a set of 31 proteo-transcriptomic features to define NAFLD, which can be used as biomarkers. Further, the single-cell RNA-seq revealed the cell types likely to contribute to the proteomic changes that occur with disease progression.

Another example used CITE-seq, RNA-ISH, MIBI and CODEX on Ulcerative Colitis (UC) samples[56]. This single-cell and spatial, transcriptomic and proteomics study revealed important therapeutic implications such as identify mononuclear phagocytes as a key target for anti-integrin therapy in UC (see Figure 4.5).

FIGURE 4.5. OVERVIEW OF ULCERATIVE COLITIS MULTI-OMICS STUDY.

2 groups of patients and healthy controls underwent a collection of single-cell and spatial transcriptomics and proteomics assays as depicted.
Image Credit: ***Mennillo, et al.*** [56]


STUDY DESIGN
HC (n=4)
UC (n=4)
VDZ (n=4)
Biopsy
Blood
SINGLE CELL TRANSCRIPTOMICS AND PROTEOMICS
Biopsy CITE-seq
Annotation & quantitation
Blood scRNA-seq
Biopsy mass cytometry
Supervised & unsupervised analysis
Blood mass cytometry
SPATIAL TRANSCRIPTOMICS AND PROTEOMICS
Tissue microarray (TMA)
Multiplex RNA in situ hybridization (RNA-ISH)
Multiplex ion beam imaging (MIBI)
CODEX
Cell segmentation
Colocalization
Clustering & annotation



Chapter 4 references

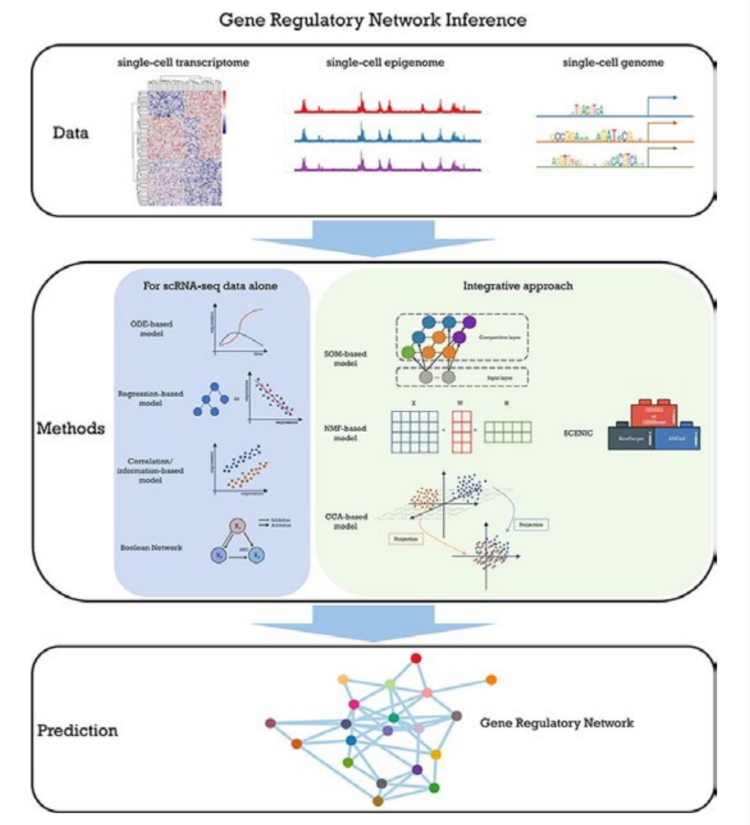
1. Fangma, Y., Liu, M., Liao, J., Chen, Z. & Zheng, Y. **Dissecting the brain with spatially resolved multi-omics.** *Journal of Pharmaceutical Analysis* **13**, 694-710 (2023).
2. Macaulay, I.C. *et al.* **G&T-seq: parallel sequencing of single-cell genomes and transcriptomes.** *Nature methods* **12**, 519-522 (2015).
3. Yu, L. et al. scONE-seq: **A single-cell multi-omics method enables simultaneous dissection of phenotype and genotype heterogeneity from frozen tumors.** *Science Advances* **9**, eabp8901 (2023).
4. Ogbeide, S., Giannese, F., Mincarelli, L. & Macaulay, I.C. **Into the multiverse: advances in single-cell multiomic profiling.** *Trends in Genetics* **38**, 831-843 (2022).
5. Vandereyken, K., Sifrim, A., Thienpont, B. & Voet, T. **Methods and applications for single-cell and spatial multi-omics.** *Nature Reviews Genetics,* 1-22 (2023).
6. Rambow, F. *et al.* **Toward minimal residual disease-directed therapy in melanoma.** *Cell* **174**, 843-855. e19 (2018).
7. Miller, M.B. *et al.* **Somatic genomic changes in single Alzheimer's disease neurons.** *Nature* **604**, 714-722 (2022).
8. Bizzotto, S. & Walsh, C.A. **Genetic mosaicism in the human brain: from lineage tracing to neuropsychiatric disorders.** *Nature Reviews Neuroscience* **23**, 275-286 (2022).
9. Zachariadis, V., Cheng, H., Andrews, N. & Enge, M. **A highly scalable method for joint whole-genome sequencing and gene-expression profiling of single cells.** *Molecular Cell* **80**, 541-553. e5 (2020).
10. Bekaert, B. *et al.* **CRISPR/Cas gene editing in the human germline.** *in Seminars in Cell & Developmental Biology Vol.* 131 93-107 (Elsevier, 2022).
11. Efremova, M. & Teichmann, S.A. **Computational methods for single-cell omics across modalities.** *Nature Methods* **17**, 14-17 (2020).
12. Babu, M. & Snyder, M. **Multi-Omics Profiling for Health.** *Molecular & Cellular Proteomics* **22**, 100561 (2023).
13. Battle, A. *et al.* **Impact of regulatory variation from RNA to protein.** *Science* **347**, 664-667 (2015).
14. He, L., Wang, W., Dang, K., Ge, Q. & Zhao, X. **Integration of single-cell transcriptome and proteome technologies: Toward spatial resolution levels.** *VIEW* **4**, 20230040 (2023).
15. Stoeckius, M. *et al.* **Simultaneous epitope and transcriptome measurement in single cells.** *Nature methods* **14**, 865-868 (2017).
16. Peterson, V.M. *et al.* **Multiplexed quantification of proteins and transcripts in single cells.** *Nature biotechnology* **35**, 936-939 (2017).
17. Ben-Chetrit, N. *et al.* **Integration of whole transcriptome spatial profiling with protein markers.** *Nature Biotechnology* **41**, 788-793 (2023).
18. Vickovic, S. *et al.* **SM-Omics is an automated platform for high-throughput spatial multi-omics.** *Nature Communications* **13**, 795 (2022).
19. Liu, Y. et al. **High-plex protein and whole transcriptome co-mapping at cellular resolution with spatial CITE-seq.** *Nature Biotechnology* (2023).
20. Timp, W. & Timp, G. **Beyond mass spectrometry, the next step in proteomics.** *Science Advances* **6**, eaax8978 (2020).
21. Fiskin, E. *et al.* **Single-cell profiling of proteins and chromatin accessibility using PHAGE-ATAC.** *Nature Biotechnology* **40**, 374-381 (2022).
22. Chung, H. *et al.* **Joint single-cell measurements of nuclear proteins and RNA in vivo.** *Nature methods* **18**, 1204-1212 (2021).
23. Chen, A.F. *et al.* **NEAT-seq: simultaneous profiling of intra-nuclear proteins, chromatin accessibility and gene expression in single cells.** *Nature Methods* **19**, 547-553 (2022).
24. Reimegård, J. *et al.* **A combined approach for single-cell mRNA and intracellular protein expression analysis.** *Communications Biology* **4**, 624 (2021).
25. Gerlach, J.P. *et al.* **Combined quantification of intracellular (phospho-) proteins and transcriptomics from fixed single cells.** *Scientific reports* **9**, 1469 (2019).
26. Wu, T., Womersley, H.J., Wang, J.R., Scolnick, J. & Cheow, L.F. **Time-resolved assessment of single-cell protein secretion by sequencing.** *Nature Methods* **20**, 723-734 (2023).
27. Cheow, L.F. & Wu, T. **Single-cell measurement of dynamic protein secretion and transcriptome.** *Nature Methods* (2023)
28. Christopher, J.A., Geladaki, A., Dawson, C.S., Vennard, O.L. & Lilley, K.S. **Subcellular Transcriptomics and Proteomics: A Comparative Methods Review.** *Molecular & Cellular Proteomics* **21**(2022).
29. Shah, S., Lubeck, E., Zhou, W. & Cai, L. **seqFISH accurately detects transcripts in single cells and reveals robust spatial organization in the hippocampus.** *Neuron* **94**, 752-758. e1 (2017).
30. Su, J.H., Zheng, P., Kinrot, S.S., Bintu, B. & Zhuang, X. **Genome-Scale Imaging of the 3D Organization and Transcriptional Activity of Chromatin.** *Cell* **182**, 1641-1659.e26 (2020).
31. Chen, K.H., Boettiger, A.N., Moffitt, J.R., Wang, S. & Zhuang, X. **RNA imaging. Spatially resolved, highly multiplexed RNA profiling in single cells.** *Science* **348**, aaa6090 (2015).
32. Takei, Y. *et al.* **Integrated spatial genomics reveals global architecture of single nuclei.** *Nature* **590**, 344-350 (2021).
33. Liao, S. *et al.* **Integrated Spatial Transcriptomic and Proteomic Analysis of Fresh Frozen Tissue Based on Stereo-seq.** *bioRxiv,* 2023.04.28.538364 (2023).
34. Chen, A. *et al.* **Spatiotemporal transcriptomic atlas of mouse organogenesis using DNA nanoball-patterned arrays.** *Cell* **185**, 1777-1792. e21 (2022).
35. Liu, Y. *et al.* **High-plex protein and whole transcriptome co-mapping at cellular resolution with spatial CITE-seq.** *Nature Biotechnology* **41**, 1405-1409 (2023).
36. He, S. *et al.* **High-plex imaging of RNA and proteins at subcellular resolution in fixed tissue by spatial molecular imaging.** *Nature Biotechnology* **40**, 1794-1806 (2022).
37. Ma, A. *et al.* **Single-cell biological network inference using a heterogeneous graph transformer.** *Nature Communications* **14**, 964 (2023).
38. Kim, H.J., Lin, Y., Geddes, T.A., Yang, J.Y.H. & Yang, P. **CiteFuse enables multi-modal analysis of CITE-seq data.** *Bioinformatics* **36**, 4137-4143 (2020).
39. Wang, X. *et al.* **BREM-SC: a bayesian random effects mixture model for joint clustering single cell multi-omics data.** *Nucleic acids research* **48**, 5814-5824 (2020).
40. Gayoso, A. *et al.* **Joint probabilistic modeling of single-cell multi-omic data with totalVI.** *Nature Methods* **18**, 272-282 (2021).
41. Lakkis, J. *et al.* **A multi-use deep learning method for CITE-seq and single-cell RNA-seq data integration with cell surface protein prediction and imputation.** *Nature Machine Intelligence* **4**, 940-952 (2022).
42. Zhu, B. et al. **Robust single-cell matching and multimodal analysis using shared and distinct features.** Nature Methods **20**, 304-315 (2023).
43. Chen, S. *et al.* **Integration of spatial and single-cell data across modalities with weak linkage.** *bioRxiv,* 2023.01.12.523851 (2023).
44. Hickey, J.W. *et al.* **Organization of the human intestine at single-cell resolution.** *Nature* **619,** 572-584 (2023).
45. Hu, H. *et al.* CITEMO(XMBD): **A flexible single-cell multimodal omics analysis framework to reveal the heterogeneity of immune cells.** *RNA Biol* **19**, 290-304 (2022).
46. Zou, G., Lin, Y., Han, T. & Ou-Yang, L. **DEMOC: a deep embedded multi-omics learning approach for clustering single-cell CITE-seq data.** *Briefings in Bioinformatics* **23**(2022).
47. Caron, D.P. *et al.* **Multimodal hierarchical classification of CITE-seq data delineates immune cell states across lineages and tissues.** *bioRxiv,* 2023.07.06.547944 (2023).
48. Varrone, M., Tavernari, D., Santamaria-Martínez, A. & Ciriello, G. **CellCharter: a scalable framework to chart and compare cell niches across multiple samples and spatial -omics technologies.** *bioRxiv,* 2023.01.10.523386 (2023).
49. Yang, Y. *et al.* I**nterpretable modeling of time-resolved single-cell gene–protein expression with CrossmodalNet.** *Briefings in Bioinformatics* **24**(2023).
50. Wu, Y., Liu, Q. & Xie, L. **Hierarchical multi-omics data integration and modeling predict cell-specific chemical proteomics and drug responses.** *Cell Reports Methods* **3**(2023).
51. Lu, S. *et al.* **A multi-omics dataset of human transcriptome and proteome stable reference.** *Scientific Data* **10**, 455 (2023).
52. Lu, X. & Ruffalo, M. **HBM-CITEseq: a uniform CITE-seq processing pipeline for the HuBMAP Consortium.** *bioRxiv,* 2022.12.19.521058 (2022).
53. Greenbaum, S. *et al*. **A spatially resolved timeline of the human maternal–fetal interface.** *Nature* **619**, 595-605 (2023).
54. Galoppin, M. *et al.* **CITE-seq reveals inhibition of NF-κB pathway in B cells from vitamin D-treated multiple sclerosis patients.** *bioRxiv,* 2023.09.25.559400 (2023).
55. Govaere, O. *et al.* **A proteo-transcriptomic map of non-alcoholic fatty liver disease signatures.** *Nature Metabolism* **5**, 572-578 (2023).
56. Mennillo, E. *et al.* **Single-cell and spatial multi-omics identify innate and stromal modules targeted by anti-integrin therapy in ulcerative colitis.** *bioRxiv,* 2023.01.21.525036 (2023).

GETTING TO GRIPS WITH GENE REGULATION. TRANSCRIPTOMICS AND EPIGENOMICS

BY LOOKING AT GENOMIC, TRANSCRIPTOMIC AND, IMPORTANTLY, EPIGENOMIC DATA, WE CAN BEGIN TO SEE HOW GENE REGULATION IS CO-ORDINATED. THIS CHAPTER WILL INVESTIGATE MULTI-OMICS METHODS FOR DNA, RNA AND THE EPIGENOME, AS WELL AS THE VARIETY OF TOOLS AVAILABLE FOR MAPPING GENE REGULATION AND GENE REGULATORY NETWORKS (GRNS).

FIGURE 5.1. GENE REGULATORY NETWORK INFERENCE METHODOLOGY.

Through single-cell RNA, DNA and epigenomic information, gene regulatory networks can be predicted through integrating the data and deploying various models to create the network. Image Credit: Hu, et al. [2]



Gene expression is tightly regulated by a complex interplay of regulatory interactions with other genes and signalling molecules. Specific proteins known as transcription factors (TFs) can regulate the expression of genes in these networks by binding to DNA regions and having repressive or negative effects on transcription rates.

The field of gene regulatory inference is around two decades old, and the technique has been performed in the micro-array, NGS, bulk and single-cell eras and, most recently, in the multi-omics era. Bulk and single-cell RNA sequencing data alone does allow for the inference of gene regulation in principle, since the RNA expression of TFs can inform you of their functionality[1]. However, regulatory processes are too complex to reliably model with transcriptomic data alone.

Epigenomic data, specifically chromatin accessibility measurements through ATAC-seq[3], ChIP-seq[4] and CUT&Tag[5], can provide information about the accessibility of TF binding sites and adds important information to the networks drawn from transcriptomics data. While ChIP-seq and CUT&Tag would be the preferred methods, profiling TF binding in this way is costly and limited to TFs with available antibodies. Instead, it is ATAC-seq that allows one to infer TF binding site availability and is most commonly used in GRN inference.

To introduce this topic, we first spoke to **Professor Sushmita Roy** from the Wisconsin Institute of Discovery about gene regulatory network analysis and her experience from the field over its 20-year transition to single-cell multi-omics. Furthermore, Professor Roy introduces us to our first multimodal GRN inference tool of this chapter – scMTNI[6].

INTERVIEW: SUSHMITA ROY

PROFESSOR, DEPARTMENT OF BIOSTATISTICS AND MEDICAL INFORMATICS, **UNIVERSITY OF WISCONSIN-MADISON**, FACULTY, **WISCONSIN INSTITUTE OF DISCOVERY**

***FLG:* Can you first introduce yourself and some of your research background and current research projects?**

Sushmita: I am Sushmita Roy. I am a Professor in the Biostatistics and Medical Informatics Department at the University of Wisconsin-Madison. And I am faculty at the Wisconsin Institute for Discovery, which is an interdisciplinary institute comprised of computational and experimental researchers looking at a range of problems across different systems. My research interests are, and have always been, at the interface of computation and biology. So, we have been developing and applying methods coming from machine learning, to look at problems in gene regulation and gene regulatory networks. Previously, before the dawn of single-cell genomics, we were using a lot of bulk data, and we actually still do. We've also developed tools to look at multi-omics data, as well as just gene expression-based approaches to combine data to get insights into regulation and regulatory networks. More recently, we've been developing methods to look at different problems in integrating single-cell multi-omics data, to get a better understanding of cell type and cell fate specific gene regulatory networks. We engage in collaborations across different organisms e.g., plants and mammals and fish

***FLG:* Can you just explain to our readers what gene regulatory networks are? Why so many people are interested in them and why they are challenging to define.**

Sushmita: I and many others in the field think about gene regulation and gene regulatory networks as the control machinery in cells. Our DNA is like a book, it has all of the information, and it works as the instruction manual on how to make an organism. But it is really the regulation of gene expression that makes something work, by interpreting what is in the genome of an individual.

In many organisms, like human or mammalian genomes, a lot of what is contributing to regulation is coming from the noncoding DNA. Oftentimes, when we think about organisms, we might ask questions such as – 'What are the proteins?' 'What are the genes in that organism?' Maybe that makes them something more or less complex, but it's really how they are controlled. This combinatorial control is what defines cell-context-specific expression patterns. It determines what, when and where sets of genes get expressed or activated, and that helps cells to do different things. It allows cells to function differently to respond to different types of stresses, or to differentiate in different ways or change their fate in different ways to enter a disease state or not. That's why we are very interested in regulation, because its influence is everywhere. It's also important in evolutionary processes, and changes in gene regulation have been associated with morphological diversity.

There are many reasons why we think it is important, but it's a really hard problem. In particular for gene network inference, which is the problem of trying to figure out who controls a particular gene – that is, find the regulators for that gene. They are many possibilities for how that particular gene can be regulated and there are many levels of control. So, we need the combination of technology to be able to measure the different levels or layers of regulation. We also need computational methods to really try to make sense of these different types of measurements that come from these technologies.

"THE OTHER WAY IS WHAT WE LIKE TO DO, AND IT INVOLVES USING ACCESSIBILITY TO INFORM OUR GRAPH NETWORK. SO, WE NOT ONLY USE CHROMATIN FOR AN INITIAL SKELETON NETWORK, BUT WE ALLOW THE ADDITION OF REGULATORS THAT MAY NOT HAVE ACCESSIBILITY SUPPORT."

FLG: Although a lot of gene regulatory network work is now being done in single-cell, there's a history of it being done in bulk. Could you talk about the computational tools that were made in the bulk era and the directions the field has taken? Finally, how has single-cell has changed the game?

Sushmita: The field of gene network inference is, I would say, almost 20 years old, perhaps more than that. It basically started with microarrays and researchers realised – 'Oh, now we have this high dimensional measurement of the level of gene expression, which is a readout of what is happening inside the cell.' From this data, we can we reverse engineer what the network might be. So, methods were developed and many of them are still being used. They may be called different things, and maybe adapted differently, but many of them are still being used.

I often use a slide in my talks that has this timeline of how the tools developed. Probabilistic graphical models, specifically, Bayesian networks were some of the early methods that were used. There were also other methods like Boolean networks, and Information theoretic methods, some modelling only discrete expression and some modelling only pairs of genes. Bayesian networks are a type of Probabilistic graphical model and are quite powerful in the sense that they work with noisy data, and they deal well with uncertainty. They also give you an interpretable model of a regulatory network, which can be learned from data. That's basically what we want. We want models that are interpretable and learnable from data.

As the field evolved, people developed methods to model perturbations and to model temporal dynamics and context specificity. There are several tools of this nature out there. But it was really with single-cell genomics that we could really get into cell type specificity, and really get into fine grained dynamics. For example, we are now able to look at new cell populations that we didn't even know existed.

The other big thing that has happened with single-cell genomics is the sample size - every cell contributes to a measurement that we can use for inferring a GRN. Before, and we've done this, to get sufficient sample size you had to combine data from different experiments, for example from gene expression omnibus, a public database of gene expression datasets. That was a pain. Now, a single-cell dataset gives you 1000s or tens of 1000s of measurements, and that has really helped. However, the data are sparse, and we now have to worry about these issues. But many methods have been adapted, for example, GENIE3, which was developed for bulk, is applied for single-cell and performs well. We have also applied methods we developed for bulk e.g., MERLIN, and it has given us a lot of mileage for single-cell as well.

FLG: What popular tools would you recommend for doing gene regulatory network analysis in both bulk and single-cell?

Sushmita: Gene regulatory networking is a hard problem and there are lots of tools out there. We like to use our own tools because we know how they work and how they fail. If something is wrong, we go back and try to figure out if there is something wrong with the data or modelling assumptions. For our tools, we use MERLIN and scMTNI, these are the tools that we've been developing. Some other tools that we think are really powerful are GENIE3, or SCENIC, which is a random forest based method. We've also benchmarked some of the popular methods, such as Inferelator, which comes from Richard Bonneau's group. This is also a pretty good method that tries to incorporate auxiliary data. That's another way the field has been progressing, asking how to go beyond gene expression and how to take other auxiliary data to inform the gene regulatory network model. So, there was work from their group and also from our group to try to incorporate these priors to get better networks and so on.

FLG: That leads to my next question, I know that integrating data beyond transcriptomics has helped this field a lot. Can you talk about how chromatin accessibility and other auxiliary information been brought into models of gene regulation?

Sushmita: There are two ways in which people have done this. One is using accessibility data to give you a structure of the network. But this is dependent upon knowing where the transcription factors bind, the sequence specificity. It is then used to build a skeleton network that people then try to remove edges from.

So that's one way and it's quite informative, but it is limited in the sense that you can only use those transcription factors that have known sequence-specific motifs.

The other way is what we like to do, and it involves using accessibility to inform our graph network. So, we not only use chromatin for an initial skeleton network, but we allow the addition of regulators that may not have accessibility support. This leverages the best of both worlds in a sense. People have also used ATAC and RNA to better define cell types and for better cell clustering and data integration, which can also influence the ultimate end goal of trying to infer gene regulatory networks.

***FLG:* I want to talk to you specifically about your tool - scMTNI. Could you describe how that one works and why people would opt to use the tool for single-cell multi-omics GRN inference?**

Sushmita: We think scMTNI is quite flexible, if you have single-cell RNA sequencing, and single-cell ATAC-seq data. One of the things that we incorporated in scMTNI is the relationship across cell types. In single-cell data there are multiple cell populations, and we have to infer cell-type specific networks. Others have also done this, but one thing that we bring in is the ability to incorporate how the cell types are related.

Let's consider an embryonic stem cell that is becoming a neuronal progenitor and then becoming a neuronal cell. For that progression, how do you incorporate it into the GRN inference task? Well, that's something that scMTNI gives you. It explicitly incorporates the cell lineage structure.

The other thing is that, even if you don't have accessibility data, you can still use our tool, you can just use gene expression and use sequence-specific motifs to inform the GRN structure. That's where this idea of using accessibility or sequence motifs as priors in the model is useful. But if you have accessibility, you can also incorporate that into the model.

***FLG:* With your work on integration, are you working primarily on RNA and ATAC? And can you talk about the challenges of trying to integrate RNA and ATAC data?**

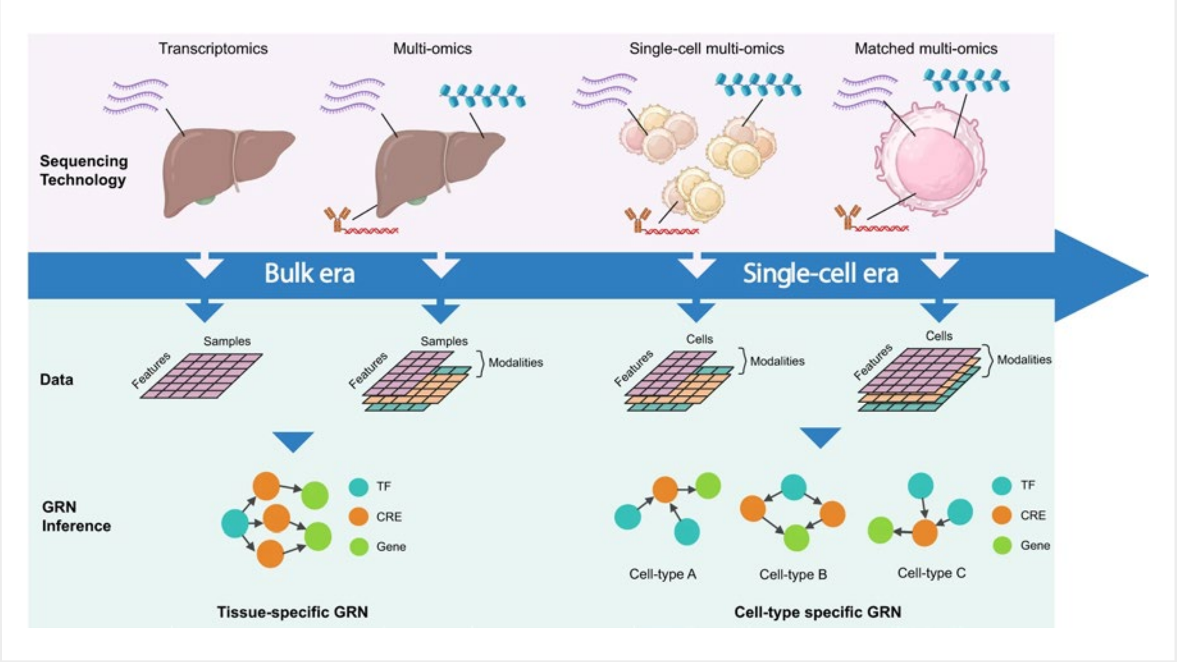
Sushmita: Yes, we are primarily looking at RNA and ATAC, but also RNA datasets across different samples. Integration can occur over two levels - across multi-sample data sets, but also multiome samples - and we are interested in incorporating relatedness across different samples. For example, when people have a time course or you have related population cohorts across related experiments, can we incorporate that to better do dimensionality reduction to define the cell clusters?

In terms of RNA and ATAC, I would say one of the problems is that many approaches are using a gene centric approach, there are some approaches trying to relax that, but accessibility gives you so much more. By gene-centric, what I mean is that you need some way to map cells measured for RNA to cells with ATAC. That mapping comes by looking at the accessibility of the gene compared to its expression. Some methods actually just stop there. Some methods try to incorporate more of the accessibility profile, but the vast majority of the methods are really just using the accessibility at the gene-level.

For me, trying to use the entire spectrum of accessibility measures is very important, as well as bringing in things such as long-range gene regulation. We are actually interested in integrating other data beyond RNA and ATAC, but primarily RNA and ATAC, because that's the largest type of data that is available.

FIGURE 5.2. THE TRANSITION OF GRN INFERENCE FROM BULK TRANSCRIPTOMICS TO SINGLE-CELL MATCHED MULTI-OMICS MEASUREMENT.

Image Credit: Kim, et al.[7]



As covered by Professor Roy and visualized in Figure 5.2, gene regulatory inference is a broad field expanding beyond the multi-omics approaches. Single-cell multi-omics has enabled GRN inference in new and exciting ways.

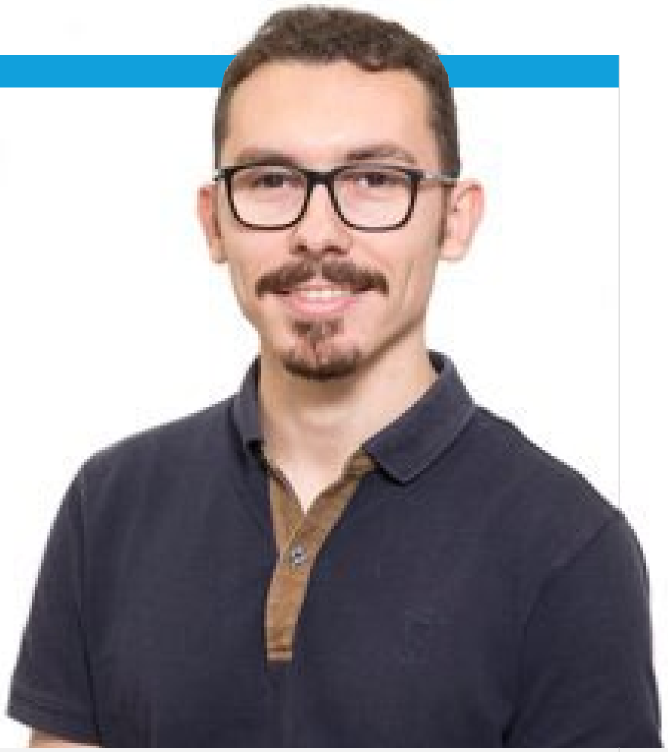
Reflecting this, there are some excellent reviews on the topic of multi-omics and GRNs[2,7,8]. We recently spoke with **Pau Badia i Mompel**, first author of a recent Nature Reviews Genetics review, which explores how single-cell multi-omics has shaped modern gene regulatory network inference.

INTERVIEW:

PAU BADIA I MOMPEL

PHD CANDIDATE, SAEZ-RODRIGUEZ GROUP

HEIDELBERG UNIVERSITY



***FLG:* Could you describe what is meant by gene regulatory networks and why people are interested in them?**

Pau: Gene Regulatory Networks (GRNs) are mathematical representations of what we think gene regulations should be. These representations are done in a graph format in which the nodes, that are genes, are connected through edges to other genes. Some genes have regulatory capabilities, meaning that they can affect the downstream transcription of other genes, and they are called transcription factors. The rest of the genes in the network are other protein-coding genes. People are interested in these models because gene regulation is a very complex process, and it is nice to have a systems biology definition that we can then model and analyse the structure of.

***FLG:* Gene regulation could be mapped with bulk technologies. What do single-cell and multi-omics technologies actually bring to the study of gene regulatory networks?**

Pau: This is something that I discussed in our recent review. What we're doing right now with multimodal GRNs could have been done during the bulk era. The problem with doing this in bulk is it would be a more costly process, because you need to secure more samples. Also, if you want cell type specificity, you need to FACS sort, which also requires time, expertise and prior knowledge. The nice thing about single-cell multi-omics is that we can now profile two technologies in a very unbiased way. We just blend the tissue, we sequence it, and then we have an unbiased profile of what's going on in the cell. Additionally, you don't need that many tissue samples, because in one single-cell assay, you get multiple observations that you can use for GRN inference modelling to generate networks. Before, you would have needed large patient cohorts to do that.

***FLG:* What specific omics are really useful to bring in when you're trying to map gene regulatory networks in the multi-omics era?**

Pau: On top of RNA, the obvious is chromatin state. So, if we try to infer GRNs from only transcriptomics, we have many false positives in those networks. By starting with just classic gene co-expression networks, the problem is that you basically connect everything with everything. You can trim down this based on prior knowledge, for example by identifying genes that should not be regulatory. By distinguishing between TFS and non-TFS, you can already prune a lot of false interactions. Another step, and this is what classic SCENIC did, is motif enrichment without chromatin accessibility.

This reinforces the point, if you don't have other omics, you can use prior knowledge. But the cool thing now is doing multi-omics with chromatin accessibility, where at least we now know if genes are open or closed.

Another layer that I would add here, but I think we're still quite far away, would be phosphoproteomics. This would profile the actual active state of all these transcription factors, and this is the actual link that is missing. The next step would be to also include cell receptor presence. This would be CITE-seq – basically, proteomics of signalling receptors.

> "ONE OF THE MAIN CHALLENGES IN GRN INFERENCE IS THAT, ALTHOUGH MANY OF THESE METHODS CLAIM THEY ARE BUILT FOR SINGLE-CELL, NONE OF THEM ACTUALLY MODEL THE SPARSITY OF THIS DATA."

If we were able to profile all these different aspects in one single, multi-omics technology without being too sparse, that would be amazing, but I think it's still far away.

***FLG:* For people interested in mapping GRNs, what downstream applications are possible when they've done this multimodal analysis and mapped the GRNs?**

Pau: It really depends on the biological question at hand. I would say the most tangible one would be deciding cell fate. For example, if you're studying trajectory analysis and you want to know what's driving some cells to go into a specific cell state or another cell type, it's very interesting to infer GRNs. Then you can try to identify which regulatory programmes are triggering these cells to shift to another cell type/cell state. This is relatively easy to validate with knockout experiments.

This has many different applications. For example, imagine you want to transform fibroblasts into healthy cells so that they start being functional in the specific tissue where they are. Or you're working in ageing, and you want to stop skin cells from transitioning to an old phenotype. This kind of analysis.

***FLG:* In your review paper, there is a table of computational tools to infer GRNs. Do you have any guidance for deciding which tool to use? Some tools were for matched, others for unmatched multimodal data, which is one distinction, but is there a gold standard?**

Pau: Currently, there are no real standards, and this is what I'm working on right now. With other computational groups we're building a benchmark to try to identify which methods perform better than others. But right now, I cannot say which method is better. I would say, pick your champion, because, in the end, it comes down to user accessibility. If you see a method that you can run, just use it. Personally, I would use SCENIC+, in my case, because they provide the biggest TF motif database for now. But we still don't know if the actual model of SCENIC+ outperforms all other methods.

***FLG:* Your review also covered some of the remaining challenges for GRN inference analysis. Can you briefly summarise what you think is perhaps the main challenge?**

Pau: One of the main challenges in GRN inference is that, although many of these methods claim they are built for single-cell, none of them actually model the sparsity of this data. Single-cell data is really cool, but it's very sparse. It's empty matrices mostly. One thing that we're exploring is how preprocessing of the sparsity affects GRN inference. The idea is to use pseudo-bulks and meta cells and other aggregation strategies to try to see if we can improve GRN inference. Which is funny, because we started with bulk, then we go to single-cell, and now we we're back to bulk. But in the end, it makes richer profiles.

As covered by our contributors, there are challenges in working with ATAC-seq and RNA-seq data simultaneously, and modelling this sparsity in the data is key to successful computational inferences. We will now take a look at some of the current popular computational tools to leverage transcriptomic and epigenomic data.

GRN multi-omics tools

There are a large variety of GRN inference tools that utilise single-cell and bulk multi-omics data. Below, you will find a table of computational methods for GRN analysis using multi-omics data.

As you will see from Table 5.1, there are many tools for the job of gene network inference. We have already covered some of these tools such as DeepMAPS[12], GLUE[16] and scMTNI[6]. Here, we will cover three interesting and widely-used computational tools from this list, starting with GRaNIE[17].

GRaNIE & GRaNPA & ENHANCER BASED NETWORKS

Enhancers are genomic locations the play an important role in cell-type-specific gene regulation. These enhancers are regulated by TFs and epigenetic mechanisms. GRaNIE is a tool to specifically build enhancer-based GRNs using multi-omics data. Accompanied by GRaNPA, which can assess the biological relevance of the generated GRNs, this is a unique tool suite for GRN inference.

We recently spoke to **Dr. Judith Zaugg,** Group Leader at EMBL and lead developer of GRaNIE and GRaNPA, about how multi-omics is enhancing her work on enhancer-based gene regulatory networks and how her computational tools can help researchers in this space.

TABLE 5.1. LIST OF MULTI-OMICS GRN INFERENCE COMPUTATIONAL METHODS.

Methods are organised by the various differences in the methods process. Table is adapted from: Badia-i-Mompel, et al. [8]

Tools	Possible inputs	Type of multimodal data	Type of modelling	Type of interactions	Statistical framework	Ref
ANANSE	Groups, contrasts	Unpaired	Linear	Weighted	Frequentist	9
CellOracle	Groups, trajectories	Unpaired	Linear	Signed, weighted	Frequentist or Bayesian	10
DC3	Groups	Unpaired	Linear	Binary	Frequentist	11
DeepMAPS	Groups	Paired or integrated	Linear	Weighted	Frequentist	12
Dictys	Groups, trajectories	Unpaired/paired or integrated	Linear	Signed, weighted	Frequentist	13
DIRECT-NET	Groups	Paired or integrated	Non-linear	Binary	Frequentist	14
FigR	Groups	Paired or integrated	Linear	Signed, weighted	Frequentist	15
GLUE	Groups	Paired or integrated	Non-linear	Weighted	Frequentist	16
GRaNIE	Groups	Paired or integrated	Linear	Weighted	Frequentist	17
Inferelator 3.0	Groups	Unpaired	Linear or non-linear	Weighted	Frequentist or Bayesian	18
IReNA	Trajectories	Unpaired	Linear	Signed, weighted	Frequentist	19
MAGICAL	Groups, contrasts	Unpaired	Non-linear	Weighted	Bayesian	20
MICA	Groups	Unpaired	Non-linear	Signed, weighted	Frequentist	21
Pando	Groups	Paired or integrated	Linear or non-linear	Signed, weighted	Frequentist or Bayesian	22
PECA	Groups	Paired or integrated	Linear	Weighted	Bayesian	23
Regulatory Motifs	Groups	Paired or integrated	Linear	Signed	Frequentist	24
RENIN	Groups	Paired or integrated	Linear	Signed, weighted	Frequentist	25
scAI	Groups	Paired or integrated	Linear	Weighted	Frequentist	26
sc-compReg	Groups, contrasts	Unpaired	Linear	Binary	Frequentist	27
SCENIC+	Groups, contrasts, trajectories	Paired or integrated	Linear	Signed, weighted	Frequentist	28
scMEGA	Trajectories	Paired or integrated	Linear	Weighted	Frequentist	29
scMTNI	Groups, trajectories	Unpaired	Linear or non-linear	Weighted	Bayesian	6
SOMatic	Groups	Unpaired	Linear	Binary	Frequentist	30
Symphony	Groups	Unpaired	Linear	Signed, weighted	Bayesian	31
TimeReg	Groups, trajectories	Paired or integrated	Linear	Binary	Frequentist	32
TRIPOD	Groups	Paired or integrated	Non-linear	Signed, weighted	Frequentist or Bayesian	33

INTERVIEW: JUDITH ZAUGG

GROUP LEADER EUROPEAN MOLECULAR BIOLOGY LABORATORY (EMBL)

***FLG:* Could you just begin by introducing yourself and introducing your lab and what the research aims of your lab are?**

Judith: I can start with the broad vision of our group, which is, essentially, to understand how genetic and epigenetic states of a cell may determine its response to signals or interactions with other cells, and through that, give rise to complex phenotypes. So, we are interested in understanding complex phenotypes, but from a very molecular point of view. When we talk about complex traits in humans, when we talk about diseases, we want to consider the genetic variation, which is one part of what may predispose somebody to certain disease, but we also want to consider the epigenetic variation. And to understand disease mechanisms, we also need to consider the cell type in which a certain mis regulation or aberrant response is actually happening. The cell type, and also the developmental or differentiation trajectory.

The system that we are particularly interested in, is the immune system and, specifically, in the home of the immune cells, which is the bone marrow. The hematopoietic stem cells within this tissue produce 500 billion hematopoietic cells per day, giving rise to all blood and immune cells. There's this entire niche surrounding the stem cells, that is influencing the cellular decisions in differentiation trajectories. Importantly, differentiation is driven by gene regulation, by transcription factors, by specific enhancers and so on. That's where the enhancers, gene regulation and multi-omics comes in into our work.

***FLG:* Could you describe how your lab has transitioned to starting to use multi-omics and how multi-omics has helped answer the research questions of your lab?**

Judith: The reason I am very keen on multi-omics is because we are very interested in transcription factors. Transcription factors are proteins that are interacting with each other to gain their function, they are post-translationally modified (phosphorylated) and they go in and out of the nucleus. So, looking at the RNA molecule of a transcription factor is often not very helpful for understanding their function because they really have to get modified. By using, for example, ATAC-seq (chromatin accessibility) or some kind of active histone mark (CHIP-seq) we are able to measure these epigenetic modifications genome-wide. This means we can actually map transcription factor binding sites based on the motif across the genome. The first tool from my group, diffTF (differential transcription factor activity), essentially takes accessibility as a readout of transcription factor activity.

You may argue, since you could just take accessibility for that, you wouldn't need RNA, so you wouldn’t need multi-omics. But, one caveat with transcription factors is that the binding sites are very similar. If you just look at accessibility across binding sites, you cannot distinguish many transcription factors. So, it is key to integrate the expression level of a transcription factor. That's where we have been using multi-omics a lot, to actually understand the activity of transcription factors and try to gauge how cell type specific it is.

***FLG:* What other multi-omics would you like to see in this field, is there an option to bring proteomics into this as well?**

Judith: I think proteomics will be very useful. At the moment, specifically for transcription factors, it's very challenging because transcription factors tend to be very lowly expressed. And it's hard to actually capture them in proteomics.

> “IT IS WELL DOCUMENTED IN THE FIELD THAT WHEN YOU TRY TO LOOK AT TRANSCRIPTION FACTORS, FROM A PROTEOMICS POINT OF VIEW, YOU TEND TO FIND VERY SMALL FRACTIONS OF THEM."

It is well documented in the field that when you try to look at transcription factors, from a proteomics point of view, you tend to find very small fractions of them. A lot of the genomics assays have moved to the single-cell basis whereas proteomics is still lagging a little bit behind. What is actually quite exciting are the assays for surface proteomics with barcode antibodies, and these assays also becoming available for intracellular proteins. So, that is very promising. Another field that that we're usually ignoring in genomics communities is metabolomics. Metabolites and metabolism are on a very different time scale. So, that's even more challenging to properly integrate.

***FLG:* And I would like to jump straight into your computational tools, GRaNIE, and GRaNPA. Could you give a brief overview of how they work, and how they build these enhancer-based gene regulation networks?**

Judith: GRaNIE stands for Gene Regulatory Network Inference including Enhancers. Gene regulatory networks is a very long-standing field, but when we started this work most of the methods and most of the databases were connecting transcription factors directly to their target genes. Whereas, we were really interested in integrating the regulatory elements like enhancers and promoters, because that's where most of the genetic variants are lying that are associated with disease. If you're interested in disease variants, you need to include those enhancers.

What I like about the GRaNIE is that it is very interpretable because it uses a very simple model. We use co-variation across individuals, using transcription factor expression, motifs and accessibility to make the transcription factor to enhancer links, and then we use the proximity and variation across individuals, again, to link enhancers to genes. So, it's a very simple network framework, and we've extensively validated that based on molecular evidence.

I think when you build gene regulatory networks, it's really important to know, what is your network based on? Is it based on variation across individuals, across cell types, across cell states? We have shown, in other work, that variation across individuals is good at predicting response to any type of environmental stimulus. Hence, I think variation across individuals is a great way of integrating overall variation that is relevant for disease.

Another thing that bothers me in the gene regulatory network field. It's very easy to build a network, but it's very hard to actually functionally validate your network, because there's essentially no gold standard. It's really hard to benchmark your network, except for picking a couple of known interactions. You can do a couple of knockouts, you can look at a ChIP-seq peaks for a couple of transcription factors and so on, but it's never global. So, that’s what motivated us to develop GRaNPA which is the Gene Regulatory Network Performance Analysis or Prediction Analysis. We initially developed this as a performance analysis, so that, whenever we have a gene regulatory network connecting transcription factors to genes, we want this network to capture cell-type specific differential expression. We want our networks to be able to predict differential expression in a specific cell type and not in another cell type. And this should be better than a random network based on the same gene and transcription factor pairings. Essentially, GRaNPA is a predictive tool to predict differential expression based on a cross validation and random forest approach, where we can then be more confident that our network is really capturing some of the underlying biology that we are interested in.

Now, having developed this, we realised that with the random forest network framework, we can also very easily look at the features that are important for the prediction. And that then gives us a tool to identify the transcription factors that are important for certain differential expression response. that's what we're using GRaNPA for most of the time. First of all, we want to check whether the network is good. And then we want to understand what are the transcription factors that are driving, for example, an infection response in an autoimmune disease.

***FLG:* So GRaNIE and GRaNPA are built for bulk data? Is there an option to use them for single-cell data?**

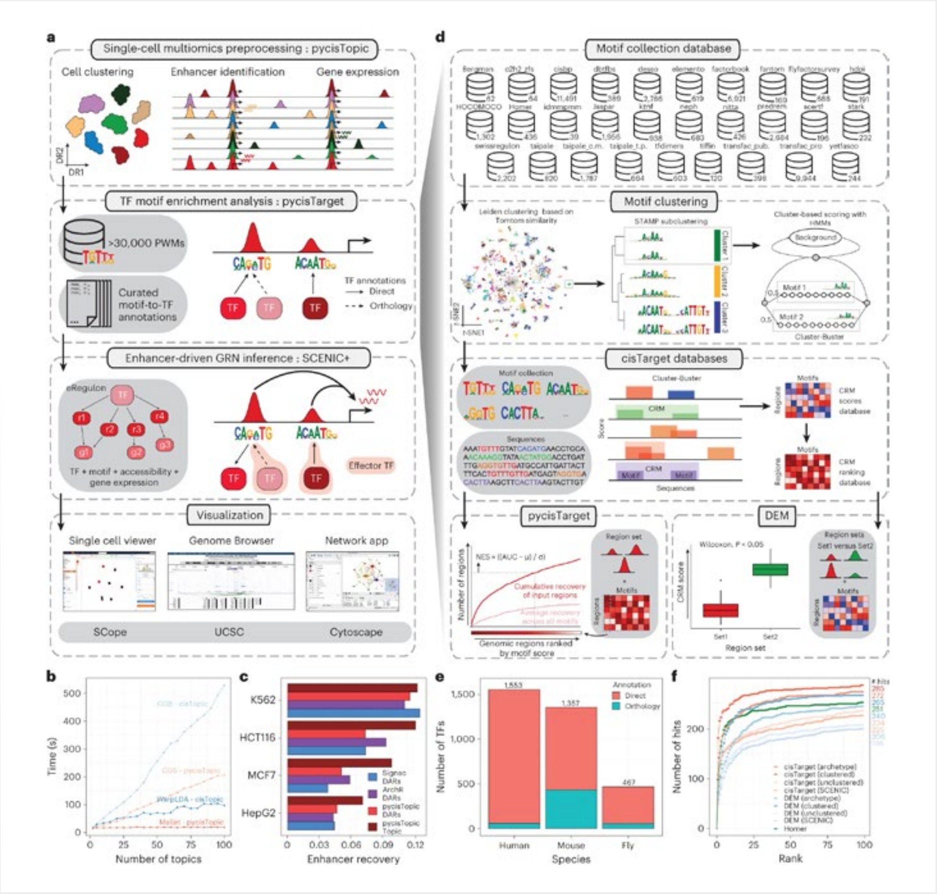
Judith: Actually, we have a vignette on our website showing how we can use it for single-cell data, and we’re currently writing a follow up manuscript. To apply to single-cell, we still do pseudo-bulk on single cells. Essentially, we split cells into different groups in silico and it does perform well.

SCENIC+

SCENIC[1] is a very popular and well-liked tool produced by Professor Stein Aerts group. This tool leverages single-cell RNA-seq data to model GRNs. Their new tool, SCENIC+[28], expands the original SCENIC methodology into the multi-omics space. It works with paired and unpaired single-cell multi-omics data and has the most comprehensive meta-database of TF binding DNA motifs for humans, mice and flies. Benchmarking within the SCENIC+ study showed its superior performance compared to other tools for predicting TF binding sites and target genes. Figure 5.3 highlights the SCENIC+ workflow.

FIGURE 5.3. SCENIC+ WORKFLOW

SCENIC+ infers GRNs using pycisTopic preprocessing followed by using the Motif collection database comprised of ~35,000 unique motifs. Image Credit: Bravo González-Blas, et al.[28]



CELLORACLE & CELL LINEAGE TRACKING

SCENIC+ introduces a computational perturbation algorithm to predict cellular changes or cell fate upon TF KO. However, another well-known tool with in-silico perturbation methodology is CellOracle[10] from the lab of Professor Sam Morris. CellOracle is less concerned with the gene networks than it is about predicting the shifts in GRNs and cell identity that occur when specific developmental regulators are altered. It is all about predicting changes using an in-silico perturbation approach.

Furthermore, the Morris lab has recently released CellTag-multi[34], an extension of their 'cell tagging' method that tracks cell origin following differentiation[35]. This represents another method merging epigenomics and transcriptomics to learn more about cell identity and gene regulation.

We recently caught up with **Professor Sam Morris** to ask her about both her multi-omics lineage tracking tool and her popular GRN in-silico perturbation tool, CellOracle.

INTERVIEW: SAMANTHA MORRIS

ASSOCIATE PROFESSOR OF DEVELOPMENT BIOLOGY AND GENETICS

WASHINGTON UNIVERSITY SCHOOL OF MEDICINE IN ST. LOUIS

***FLG:* Can you describe a little bit about what your lab has been up to in the last few years in terms of the genomics of cell lineage, cell identity and cell reprogramming?**

Sam: We got into genomics broadly about 12 years ago and we came into the reprogramming field. And with the Yamanaka factors, and reprogramming to pluripotency, there was all this excitement – 'We can make any cell type on demand.' We built a gene regulatory network-based platform that used bulk expression data, called CellNet, to measure the identity of the cells that people were engineering in vitro against their actual in vivo counterparts. We were able to see that people were not producing the cell types that they were claiming to have made. And there were two problems with that; the approach isn't scalable, but also, the resulting cells are developmentally immature.

For almost 10 years now, we've been taking these technologies into the single-cell era, using single-cell multi-omics to try and measure cell identity in a completely unbiased way. We apply these methods reprogramming protocols and ask – 'What have we actually made here?', 'Have any cell types reached their target destination?', 'When do we see off target trajectories?', 'How do we block those?' Complementing this approach is our CellOracle software in which we use simple machine learning approaches to nominate factors that can push cells closer to their target destinations.

On the experimental genomics technology side, we have developed lineage tracing approaches to explore – 'What steps did a cell go through to get to its target destination?' So, finding out what set a cell on a specific path very early on. If we get to those early stages, understand the mechanisms, then we can push more cells on the right trajectory. The end goal here is to improve reprogramming to eventually get cells into the clinic, to use them for disease modelling, toxicology testing. That's a very broad overview of how the genomics is intermingled with the lineage reprogramming.

***FLG:* I first wanted to ask you about your cell tagging technology and your recently released CellTag-multi, which is expanding the capacity to multi-omics. Can you explain to the readers how CellTag works, and how you've broadened it out into CellTag-multi?**

Sam: CellTagging is a very simple lineage tracing method. We use lentivirus and deliver random barcodes into cells.

With the original CellTagging method, each cell in a population receives a combination of barcodes. The barcodes are inherited, expressed, and recovered in parallel with the single-cell transcriptome. Using this labelling method, we map which cells are clonally related to each other, and with sequential rounds of labelling, we can build lineage trees. Using this lineage information, we can track which cells successfully reprogrammed or went into a dead end. We can go back and look at the ancestors to ask – 'well, what were these cells doing early on in this process?'. What were the early changes that contributed to them successfully reprogramming or to them doing something undesirable? This was the original 2018 technology and that allowed us to define the reprogramming landscape.

While some cells successfully reprogram, others enter an 'off-target' trajectory. These cells are reprogrammed, but they're just not the right identity that we're looking for. We were looking earlier and earlier in the reprogramming process to find the origins of these off-target cells, and we weren't seeing many transcriptional differences in early stages, but we knew that they had very different reprogramming outcomes. We thought, first of all, it could be technical, just because there's dropout in single-cell RNA sequencing, or it could be that we don't see transcriptional differences yet at those stages. This is when we started looking toward chromatin accessibility through single-cell ATAC-seq. Changes in chromatin accessibility precede changes in gene expression. So, just having that view, and identifying which areas of the genome are becoming accessible, we've been able to see, as early as day three, what sets cells on a specific trajectory.

FLG: ***Have you seen CellTag-multi adopted for other systems?***

Sam: We've seen the adoption of CellTagging across different systems, particularly in the context of cancer. For example, groups are trying to understand why some cells become drug resistant. That's where I think the chromatin accessibility information on top of the RNA could be really helpful. Within the CellTag-multi paper, we also apply it to haematopoiesis, just because there's so much unknown about that system that it's ideal to validate a new method.

FLG: ***Within the CellTag-multi, did you merge the RNA and ATAC information?***

Sam: Yes, we merge at the level of clones. We tried the 10x Multiome kit, but this approach relies on capturing the lineage barcodes from the RNA side of the pipeline to link the lineage and chromatin accessibility. However, with the Multiome kit, it's very difficult to pick up the lineage barcodes and we don't yet know why. However, early on, we made a conscious decision to capture lineage across independent modalities. With CellTag-multi, you don't have to rely on capture of RNA, to readout lineage with chromatin accessibility. It's a complete standalone technology. Thus, it gives the user more flexibility in experimental design.

> “RIGHT NOW, ONE OF THE DISADVANTAGES OF CELLTAGGING IS THAT WE HAVE SERIALLY TRANSDUCE CELLS TO BUILD LINEAGES. SOME CELL TYPES DON'T TRANSDUCE WELL, SOME CELL TYPES AREN'T ACCESSIBLE IN VIVO."

We're now expanding CellTag-multi to histone state, and DNA methylation capture. We do think there's only so much high-quality information that a single cell can give up, it will just deteriorate the more you try and get out of a cell. Obviously, the limitation there is that you need big enough clones to be able to power these analyses. But we think that higher quality information will be more desirable in this instance.

***FLG:* Alright, let's get straight over to CellOracle and gene regulatory networks. Could you first describe why GRN’s are important to map and what they mean for cell identity and cell lineages?**

Sam: We can think of GRNs as master regulators of cell identity. They tell us how transcription factors control cell identity, which transcription factors connect to which genes. With CellOracle, we take this gene regulatory network perspective, because we're interested in how transcription factors control cell identity. Traditionally, with GRN inference, you produce these hairballs of ‘a transcription factor that's connected to many, many genes’. So, how do you start interpreting what that means for cell identity? This is why we built CellOracle.

I think that people see the CellOracle paper and they see GRNs, and they think, ‘Oh, CellOracle is built to infer GRNs.’ Well, we use that approach, but really, the reason we're doing that is so that we can then perturb transcription factor expression in silico in these networks, to ask – ‘If we lose this transcription factor from a network or we gain this transcription factor, then how will cell identity shift?’ Hence, we've been able to make predictions of how cell identity changes once transcription factors are perturbed, and it's the GRNs that underlie this approach.

***FLG:* So, just bringing it back to multi omics. CellOracle can make use of RNA and ATAC data, what value does a multi-omics approach bring here?**

Sam: We included single-cell ATAC data, and you can use bulk data as well. The first step in CellOracle creates a ‘base gene regulatory network’ for each species. Really, we're using the chromatin accessibility data there to create a map of all biologically feasible connections in the network. And because you could start from the perspective of ‘transcription factor X can connect to all genes,’ and since every single transcription factor can connect to every gene, it's very chaotic and noisy.

So, why start from that principle, when you can use the ATAC-seq data to know which of these connections are biologically feasible? It helps clean up and remove the noise from the GRNs. And then in the second step, we take that base GRN and then use the actual single-cell RNA-seq data for each cluster to define the connections that are actually active in the network for each defined cell type and state.

***FLG:* Wha are your hopes for different kinds of multi-omics technology for your research questions?**

Sam: Right now, one of the disadvantages of CellTagging is that we have serially transduce cells to build lineages. Some cell types don't transduce well, some cell types aren't accessible in vivo. We're working on methods to increase lineage tracing resolution, looking at methods to mutate barcodes gradually over time. High-resolution lineages are going to give us a lot more insight into the dynamics of reprogramming.

There have been several notable new tools released in 2023 that we haven't covered here. Examples include Dictys[13], a dynamic method taking context-specificity into account, KiMONo[36] for inferring GRNs in the presence of missing data, MICA[21], a non-linear approach with superior performance to classic methods and Normi[37] for inferring GRNs with non-redundant mutual information.

Given this ever-increasing list of methods, there are attempts to benchmark these multi-omics methods using approaches such as ground truth, synthetic data or regulatory databases. Follow these references for an in-depth coverage of the topic[8,38-40], including a recently released web-based benchmarking platform to compare your data to real data with various noise levels.

Applications of Epigenomics, Transcriptomics and GRNs

These multi-omics approaches summarised above are incredibly valuable for understanding cell identity and tracking cell fate during periods of transition, either during development or during a cell's transition into a disease state.

Understanding cell identity is a key challenge for the single-cell field and combining epigenomic and transcriptomics is a useful strategy to accomplish this. Organs with complex cellular makeups such as the brain[41,42] and eye[43] have benefited from this approach to resolve cell identities.

We recently spoke to **Professor Rui Chen** at Baylor College of Medicine about his work atlasing the human and mouse retina using transcriptomics and chromatin accessibility measures[43].

> "GIVEN THIS EVER-INCREASING LIST OF METHODS, THERE ARE ATTEMPTS TO BENCHMARK THESE MULTI-OMICS METHODS USING APPROACHES SUCH AS GROUND TRUTH, SYNTHETIC DATA OR REGULATORY DATABASES."

INTERVIEW: RUI CHEN

PROFESSOR OF MOLECULAR AND HUMAN GENETICS BAYLOR COLLEGE OF MEDICINE

***FLG:* How have multi-omics methods enhanced the work that you're trying to do?**

Rui: Of course, single-cell RNA-seq is really powerful, you can use the data to identify cell types and look at the transcriptome. But, from a genetic point of view, we're also interested in not only the cell type and gene transcription, but also in how the genes are regulated, and the potential impact that variants have on gene regulation. A lot of genetic burden is not in the coding region of the genome. I think the Multiome is a way for us to identify not only the gene transcriptome, but also identify potential elements in the same cell for that gene.

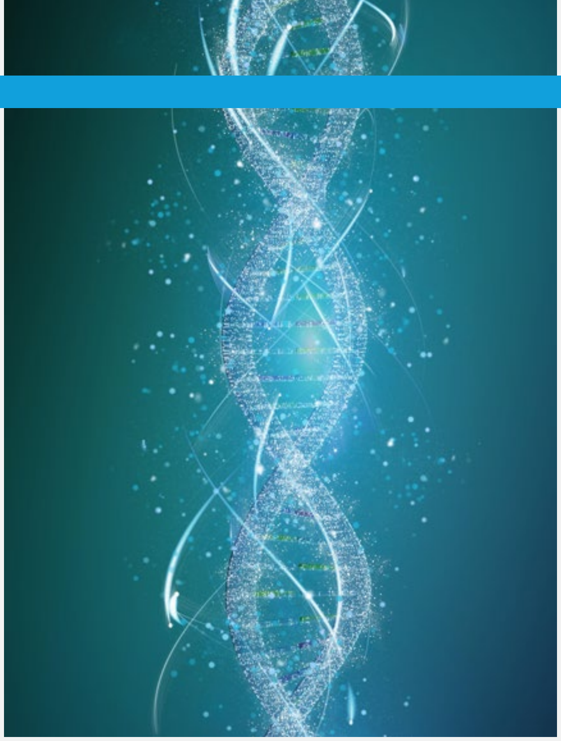
***FLG:* You've recently published a paper performing RNA and epigenome profiling of the retina. Can you describe this multi-omics approach for atlasing the retina?**

Rui: So, that paper is actually an intermediate product - we have a new one to be submitted in the next couple of weeks. The major goal for the cell atlas is to identify all cell types in the tissue. The retina is a part of the central nervous system. It's simpler compared to the brain, of course. There are only three layers of neurons that are well organised. The neuron types in the retina include photoreceptors cells, bipolar cells, retinal ganglion cells (RGCS), horizontal cells and the Muller cells.

From previous studies based on the morphology, the physiology and some limited molecular markers, we know the types for each cell class. For example, for the RGCs, in the mouse, there may be 40 different types and they might play different roles. From a neuroscience point of view, it's important to know which cell types are out there, how they're connected and how they function together to build a circuit board. To know the circuits, you need to know the components first, and then you can understand how they're connected and then how they function.

> "THAT'S WHY WE NEED A LARGE NUMBER OF CELL PROFILES IN ORDER TO IDENTIFY THE RARE ONES, OTHERWISE YOU WOULD JUST MISS THEM."

So, I think the first goal is to identify all the parts. There are a large number of cell types that exist in the retina, although it is relatively simple. Coupled with the fact that not all the neuron types and cell types are equally represented in the tissue, for example, the retinal ganglion cells together only account for 1% of all neurons in the retina, but they play a critical role. They are the cell that relays the signal from eye to the brain and without them, you're not going to see anything. So, if the mouse has 40 different types, we need to think about 40 times 1%, so each one is represented on average at 0.02%. This is only going to be worse in humans since the RGCs are not equally represented and they could differ by a hundredfold in terms of abundancy.



That's why we need a large number of cell profiles in order to identify the rare ones, otherwise you would just miss them. So, the atlas we have published, there are a quarter of a million cells. This is not a straight quarter million either, since we used markers to enrich for the rare cell types to improve their potential percentage. In the new one we're just about to submit, there are about 2 million cells in there.

***FLG:* Why did you also include a population that you had ATAC-seq for? Why make this a multi-omics study? And have you done that for the new atlas that's going to be published?**

Rui: In previous studies, people have used ATAC-seq, although this is mostly bulk ATAC-seq, and a very small amount of ATAC-seq, for example, 10-20 thousand cells of the atlas. I would say that 20,000 ATAC-seq cells will not give you high enough resolution. It will allow you to get the major class of a cell, but among them, there's so many different types and they have distinct chromatin profiles. To better understand how genes are regulated and to find the gene networks and the motifs, you need the chromatin profile. So, that's why in the paper we did 100,00 nuclei with ATAC-seq. We can get the resolution of class and sub-class, but this number will not be enough for rare cell types. In this new study, we increase that 100,000 nuclei to 400,000. This is still not enough, but we have a much better resolution. To see extremely rare cells, we would need to increase that fivefold or tenfold.

***FLG:* What computational tools do you tend to use in your lab for analysing ATAC and RNA and for integrating them together?**

Rui: We used BindSC, single-cell bind, for integration. There are many tools that do integration. And we also use scGLUE, but at the time of the paper, we uses BindSC since there were fewer tools available at the time. And then we actually developed this new software, BindSC, and we found when we compare the performance it looks pretty good. The modality of RNA-seq and ATAC is quite different, ATAC is much sparser.

***FLG:* Can you just briefly explain to our readers why you included a cross-species comparison for the retina?**

Rui: There's multiple reasons, but for the atlas point of view, the primary reason is to help the annotation of cell clusters. In the human retina there are roughly 110 cell types. Most of them, we don't have a name for it, we don't know the function and we don't know how it connects. In contrast, the mouse retina has been well studied, and there has also been some work in the primate - physiology, staining, morphology, etc. So, we try to borrow the names of these cell types for the human types that have no name. This means that the people who study it will be able to cross reference relatively easily.

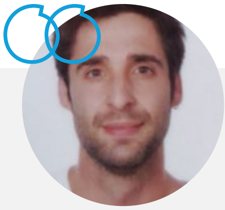
We also know there is similarity, but there is also divergence between species, you cannot always find an ortholog cluster between human and mouse. In fact, a lot of the time, we cannot. This gives you a chance to study the differences. The third thing is, by mapping this out, not based on fragmented information but with everything on a plate for two species or three species, then this allows us to make the best comparison between that, because it's more comprehensive and it's not based on one or two markers. Also, people use mice as a disease model and there is always the question of whether what you learn from mice can be translated to humans. If you study cell types in the mouse, does it even existing humans, how exactly can you translate it?

***FLG:* Do you have any future plans for different types of multi-omics profiling of the retina? How about proteomics and metabolomics?**

Rui: All of the above. The transcriptome is great, and chromatin is great, but it's only one angle or facet of the cell. Cells have protein, metabolites, also the spatial aspect; I think of all of them as just different angles. We're dealing with very high dimensional data, the more you have, the better holistic understanding of the cell function state you have. We are working on the metabolome, and also really watching out for the proteome. I know it's coming.

There have been recent examples of disease insights gained from GRN multi-omics in type 2 diabetes[44], placental-mediated low birth weight[45], heart failure[46] and pancreatic disease[47]. These approaches will eventually allow the engineering of cell fate[48] and ultimately disease prevention[49] through this understanding of how cellular identity is established and maintained.[8]

When it comes to development, these tools have the power to map developmental processes in unprecedented ways. For example, Argelaguet, et al. [50] recently profiled mouse embryos during organogenesis and used the RNA and ATAC-seq readouts to map the TFs and gene regulatory networks that underpin the lineage commitment events for different organs. We spoke to Ricard Argelaguet again about this study that he was first author for.



RICARD ARGELAGUET

Senior Research Scientist
Altos Labs

FLG: I wanted to ask you about the single-cell multi-omics work you did in mouse embryos and gene regulation. What has single-cell multi-omics done for gene regulatory network inference?

Ricard: *Since the early days of bioinformatics, there has been a substantial interest in learning gene regulatory networks, essentially trying to link transcription factors to their target genes. Initially, most of this was constrained to just RNA-seq data and people would do co-expression networks to find correlations between the transcription factors and potential target genes. But it was really hard to know, in a systematic way, which genes are targeted by which transcription factors. Now, by the inclusion of another data modality, it really improved this process, and this other data modality is something that tells you about where those transcription factors are binding where they are active.*

The ideal modality would be something like ChIP-seq, where for each transcription factors, you know exactly where it's binding, so you know which genes it's potentially targeting. The issue is we cannot do ChIP-seq for all transcription factors because this will be way too expensive and there's no way of doing this experiment. But, with the addition of ATAC-seq (chromatin accessibility), you can get an estimate of where transcription factors are binding. By looking at the accessibility of the chromatin, and then looking at motifs in the DNA, you can get a rough idea of where transcription factors are binding. People have shown that by combining these two data modalities, it gives you a reasonably good estimate for gene regulatory networks. This was good in the bulk RNA sequencing era but for this type of estimation, you really need lots of measurements. So, when the single-cell multi-omics field kicked in, specifically RNA-seq and ATAC-seq from the same cell, it gives you the power for doing these gene regulatory inference models.

FLG: And what did you find in mouse embryos?

Ricard: *So, as I said, what really made this new era of gene regulatory network inference possible is the chromatin accessibility and RNA expression measurements from the same cell. And we were quite lucky to have access very early on to the new 10x Multiome Kit, which essentially allows you to profile these at high throughput. Our lab had experience on working with mouse embryos for a few years and mapping RNA expression, methylation and chromatin dynamics. And one of the main analyses that we performed was this gene regulatory network inference. We were able to map the networks that essentially determine the cell fate transitions, going from one cell type to another cell type in the mouse embryo. And one of the beautiful things about these models is that they allow you to make predictions, such as, if you knock out or manipulate a specific transcription factor, what will actually happen to the cell state, will the state move in this direction, or will it move in that direction?*

Chapter 5 references

1. Aibar, S. *et al.* **SCENIC: single-cell regulatory network inference and clustering.** *Nature Methods* **14**, 1083-1086 (2017).
2. Hu, X., Hu, Y., Wu, F., Leung, R.W.T. & Qin, J. **Integration of single-cell multi-omics for gene regulatory network inference.** *Computational and Structural Biotechnology Journal* **18**, 1925-1938 (2020).
3. Buenrostro, J.D. *et al.* **Single-cell chromatin accessibility reveals principles of regulatory variation.** *Nature* **523**, 486-490 (2015).
4. Rotem, A. *et al.* **Single-cell ChIP-seq reveals cell subpopulations defined by chromatin state.** *Nature biotechnology* **33,** 1165-1172 (2015).
5. Bartosovic, M., Kabbe, M. & Castelo-Branco, G. **Single-cell CUT&Tag profiles histone modifications and transcription factors in complex tissues.** *Nature biotechnology* **39**, 825-835 (2021).
6. Zhang, S. *et al.* **Inference of cell type-specific gene regulatory networks on cell lineages from single cell omic datasets.** *Nature Communications* **14**, 3064 (2023).
7. Kim, D. *et al.* **Gene regulatory network reconstruction: harnessing the power of single-cell multi-omic data.** *Systems Biology and Applications* **9**, 51 (2023).
8. Badia-i-Mompel, P. *et al.* **Gene regulatory network inference in the era of single-cell multi-omics.** *Nature Reviews Genetics* **24**, 739-754 (2023).
9. Xu, Q. *et al.* **ANANSE: an enhancer network-based computational approach for predicting key transcription factors in cell fate determination.** *Nucleic acids research* **49**, 7966-7985 (2021).
10. Kamimoto, K. *et al.* **Dissecting cell identity via network inference and in silico gene perturbation.** *Nature* **614**, 742-751 (2023).
11. Zeng, W. *et al.* **DC3 is a method for deconvolution and coupled clustering from bulk and single-cell genomics data.** *Nature communications* **10**, 4613 (2019).
12. Ma, A. *et al.* **Single-cell biological network inference using a heterogeneous graph transformer.** *Nature Communications* **14**, 964 (2023).
13. Wang, L. *et al.* **Dictys: dynamic gene regulatory network dissects developmental continuum with single-cell multiomics.** *Nature Methods* **20**, 1368-1378 (2023).
14. Zhang, L., Zhang, J. & Nie, Q. **DIRECT-NET: An efficient method to discover cis-regulatory elements and construct regulatory networks from single-cell multiomics data.** *Science Advances* **8**, eabl7393 (2022).
15. Kartha, V.K. *et al.* **Functional inference of gene regulation using single-cell multi-omics.** *Cell Genom* **2**(2022).
16. Cao, Z.-J. & Gao, G. **Multi-omics single-cell data integration and regulatory inference with graph-linked embedding.** *Nature Biotechnology* **40**, 1458-1466 (2022).
17. Kamal, A. *et al.* **GRaNIE and GRaNPA: inference and evaluation of enhancer-mediated gene regulatory networks.** *Molecular systems biology* **19**, e11627 (2023).
18. Skok Gibbs, C. *et al.* **High-performance single-cell gene regulatory network inference at scale: the Inferelator 3.0.** *Bioinformatics* **38**, 2519-2528 (2022).
19. Jiang, J. *et al.* **IReNA: integrated regulatory network analysis of single-cell transcriptomes and chromatin accessibility profiles.** *Iscience* **25**(2022).
20. Chen, X. *et al.* **Mapping disease regulatory circuits at cell-type resolution from single-cell multiomics data.** *Nature Computational Science* **3**, 644-657 (2023).
21. Alanis-Lobato, G. *et al.* **MICA: a multi-omics method to predict gene regulatory networks in early human embryos.** *Life Science Alliance* **7**(2024).
22. Fleck, J.S. *et al.* **Inferring and perturbing cell fate regulomes in human brain organoids.** *Nature*, 1-8 (2022).
23. Duren, Z., Chen, X., Jiang, R., Wang, Y. & Wong, W.H. **Modeling gene regulation from paired expression and chromatin accessibility data.** *Proceedings of the National Academy of Sciences* **114**, E4914-E4923 (2017).
24. Zenere, A., Rundquist, O., Gustafsson, M. & Altafini, C. **Using high-throughput multi-omics data to investigate structural balance in elementary gene regulatory network motifs.** *Bioinformatics* **38**, 173-178 (2022).
25. Ledru, N. *et al.* **Predicting regulators of epithelial cell state through regularized regression analysis of single cell multiomic sequencing.** *bioRxiv,* 2022.12.29.522232 (2022).
26. Jin, S., Zhang, L. & Nie, Q. **scAI: an unsupervised approach for the integrative analysis of parallel single-cell transcriptomic and epigenomic profiles.** *Genome biology* **21**, 1-19 (2020).
27. Duren, Z. *et al.* **Sc-compReg enables the comparison of gene regulatory networks between conditions using single-cell data.** *Nature Communications* **12**, 4763 (2021).
28. Bravo González-Blas, C. *et al.* **SCENIC+: single-cell multiomic inference of enhancers and gene regulatory networks.** *Nature Methods* **20**, 1355-1367 (2023).
29. Li, Z., Nagai, J.S., Kuppe, C., Kramann, R. & Costa, I.G. **scMEGA: single-cell multi-omic enhancer-based gene regulatory network inference.** *Bioinformatics Advances* **3**, vbad003 (2023).
30. Jansen, C. *et al.* **Building gene regulatory networks from scATAC-seq and scRNA-seq using linked self organizing maps.** *PLoS computational biology* **15**, e1006555 (2019).
31. Bachireddy, P. *et al.* **Mapping the evolution of T cell states during response and resistance to adoptive cellular therapy.** *Cell reports* **37**(2021).
32. Duren, Z., Chen, X., Xin, J., Wang, Y. & Wong, W.H. **Time course regulatory analysis based on paired expression and chromatin accessibility data.** *Genome research* **30**, 622-634 (2020).
33. Jiang, Y. *et al.* **Nonparametric single-cell multiomic characterization of trio relationships between transcription factors, target genes, and cis-regulatory regions.** *Cell Systems* **13**, 737-751. e4 (2022).
34. Jindal, K. *et al.* **Single-cell lineage capture across genomic modalities with CellTag-multi reveals fate-specific gene regulatory changes.** *Nature Biotechnology* (2023).
35. Biddy, B.A. *et al.* **Single-cell mapping of lineage and identity in direct reprogramming.** *Nature* **564,** 219-224 (2018).
36. Henao, J.D. *et al.* **Multi-omics regulatory network inference in the presence of missing data.** *Briefings in Bioinformatics* **24**(2023).
37. Zeng, Y., He, Y., Zheng, R. & Li, M. **Inferring single-cell gene regulatory network by non-redundant mutual information.** *Briefings in Bioinformatics* **24**(2023).
38. Pratapa, A., Jalihal, A.P., Law, J.N., Bharadwaj, A. & Murali, T. **Benchmarking algorithms for gene regulatory network inference from single-cell transcriptomic data.** *Nature methods* **17**, 147-154 (2020).
39. Seçilmiş, D., Hillerton, T. & Sonnhammer, E.L. **GRNbenchmark-a web server for benchmarking directed gene regulatory network inference methods.** *Nucleic Acids Research* **50**, W398-W404 (2022).
40. Uzun, Y. **Approaches for benchmarking single-cell gene regulatory network inference methods.** *arXiv preprint arXiv:2307.08463* (2023).
41. Zhu, C. *et al.* **Joint profiling of histone modifications and transcriptome in single cells from mouse brain.** *Nature methods* **18**, 283-292 (2021).
42. Luo, C. *et al.* S**ingle nucleus multi-omics identifies human cortical cell regulatory genome diversity.** *Cell genomics* **2**(2022).
43. Liang, Q. *et al.* **A multi-omics atlas of the human retina at single-cell resolution.** *Cell genomics* **3**(2023).
44. Liu, J. *et al.* **Uncovering the gene regulatory network of type 2 diabetes through multi-omic data integration.** *Journal of Translational Medicine* **20**, 604 (2022).
45. Tekola-Ayele, F. *et al.* **Placental multi-omics integration identifies candidate functional genes for birthweight.** *Nature Communications* **13**, 2384 (2022).
46. Zhou, X., Zhang, S., Zhao, Y., Wang, W. & Zhang, H. **A multi-omics approach to identify molecular alterations in a mouse model of heart failure.** *Theranostics* **12**, 1607-1620 (2022).
47. Augsornworawat, P. *et al.* **Single-nucleus multi-omics of human stem cell-derived islets identifies deficiencies in lineage specification.** *Nature Cell Biology* **25**, 904-916 (2023).
48. Su, E.Y., Spangler, A., Bian, Q., Kasamoto, J.Y. & Cahan, P. **Reconstruction of dynamic regulatory networks reveals signaling-induced topology changes associated with germ layer specification.** *Stem Cell Reports* **17**, 427-442 (2022).
49. Claringbould, A. & Zaugg, J.B. **Enhancers in disease: molecular basis and emerging treatment strategies.** *Trends in Molecular Medicine* **27**, 1060-1073 (2021).
50. Argelaguet, R. *et al.* **Decoding gene regulation in the mouse embryo using single-cell multi-omics.** *bioRxiv*, 2022.06.15.496239 (2022).

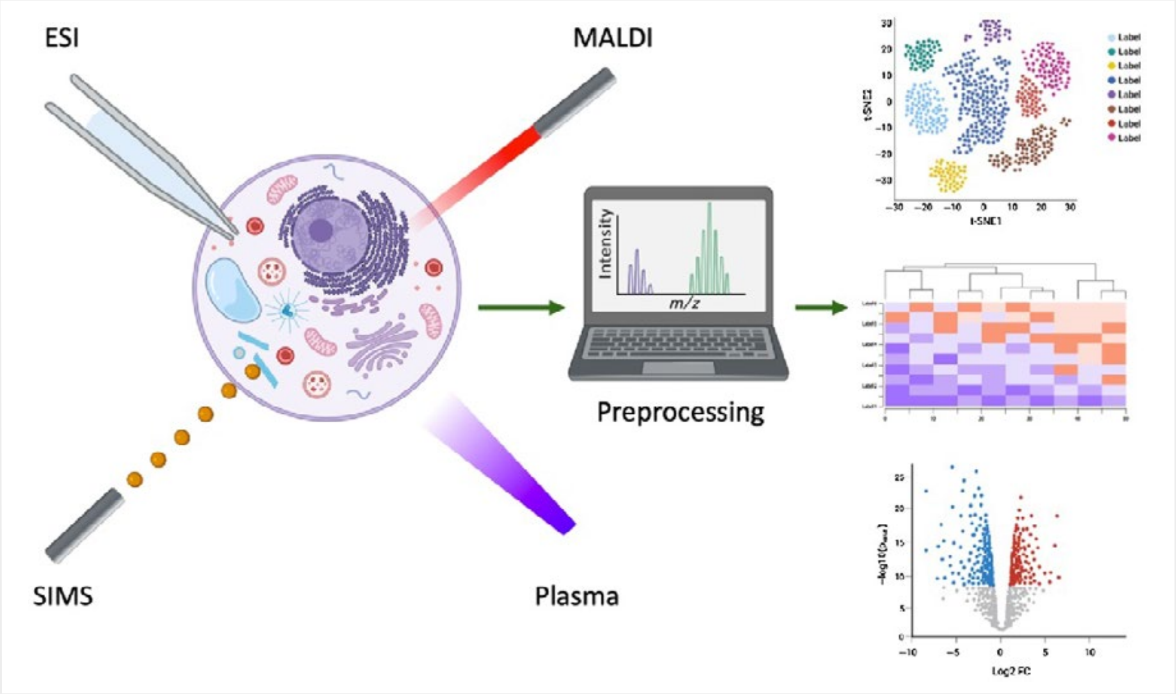
CHAPTER 6

BACK TO BUILDING BLOCKS: THE RISE OF METABOLOMICS

PROFILING DNA, RNA OR PROTEINS CAN PROVIDE VALUABLE INSIGHTS INTO THE ORIGIN AND NATURE OF CELLS, BUT THEY CAN'T TELL US WHAT IS HAPPENING IN THE CELL AT THAT MOMENT. METABOLITES SUCH AS SUGARS, LIPIDS, NUCLEOTIDES AND AMINO ACIDS CAN ALL BE MEASURED AND CAN SHOW US WHAT METABOLIC PROCESSES ARE CURRENTLY GOING ON INSIDE A CELL. THIS CHAPTER INTRODUCES AND COVERS THE MAJOR METHODS (SEE FIGURE 6.1) USED TO VISUALISE THE SMALL MOLECULES OFTEN LEFT OUT IN GENOMICS STUDIES – COLLECTIVELY KNOWN AS THE METABOLOME

FIGURE 6.1. THE FOUR MAJOR IMAGING MASS SPECTROMETRY METHODS AND A GRAPHIC OVERVIEW OF THE PREPROCESSING AND DATA ANALYSIS PROCESS.

Image Credit: Liu, et al. [1]



What is metabolomics?

Metabolites are considered any small biomolecule within a biological system under 2,000 daltons in weight, typically sugars, fatty acids, steroids, drugs and amino acids. These metabolites are the precursors, the intermediates and the end products of cellular processes[2]. They are the closest omic to the cell's phenotype. This makes them instrumental in signaling and modulating a cell's phenotype, but it also means that the cell's metabolome is dynamic, shifting rapidly in response to the environment.

INGELA LANEKOFF

Professor, Department of Chemistry-BMC
Uppsala University

FLG: For human cells, what information does single-cell metabolomics give you that other omics do not?

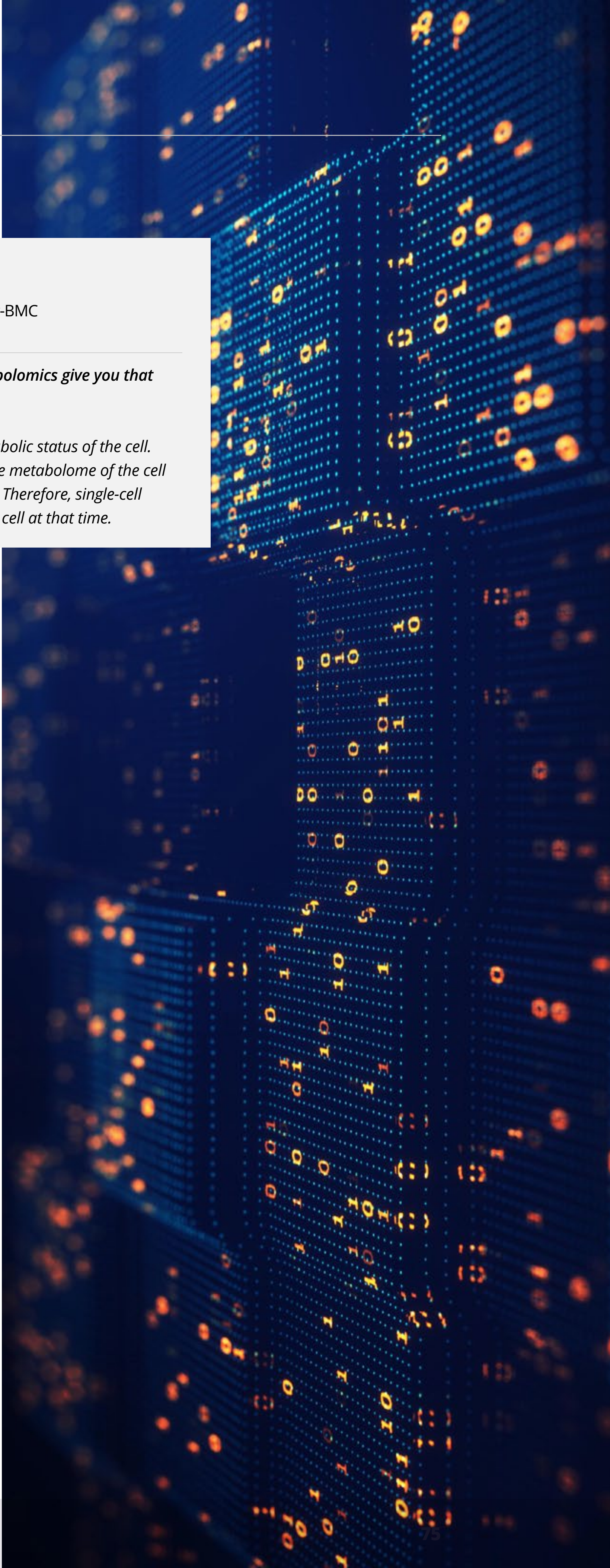
Ingela: *Single-cell metabolomics provides insights into the metabolic status of the cell. Compared to the individual cell's proteome or transcriptome, the metabolome of the cell is highly dynamic and can be altered in milliseconds to seconds. Therefore, single-cell metabolomics can provide an instant view of what occurs in the cell at that time.*

The goal of metabolomics is to identify and quantify these metabolites.

For transcriptomics and proteomics, the technology progressed from bulk to single-cell to spatial applications. For metabolomics, it is single-cell that currently represents the cutting edge, while spatial metabolomics has a slightly longer history of development.

We recently caught up with **Dr. Theo Alexandrov,** Team Leader at **EMBL** to introduce us to the field of metabolomics, and to discuss the latest advances in spatial and, most recently, single-cell metabolomics.

> "FOR METABOLOMICS, IT IS SINGLE-CELL THAT CURRENTLY REPRESENTS THE CUTTING EDGE, WHILE SPATIAL METABOLOMICS HAS A SLIGHTLY LONGER HISTORY OF DEVELOPMENT."



INTERVIEW: THEODORE ALEXANDROV

TEAM LEADER, STRUCTURAL AND COMPUTATIONAL BIOLOGY UNIT EUROPEAN MOLECULAR BIOLOGY LABORATORY

***FLG:* Can you briefly introduce yourself to the readers, your current position and explain a little bit about how you got into metabolomics?**

Theo: Absolutely. My name is Theodore, I go by Theo, and I work at EMBL, the European Molecular Biology Laboratory, which is an amazing place to work. It is an inter-governmental organisation for life sciences in Europe. You can think of it as like CERN, but for molecular biology. So, we have a very vibrant place to work, and I'm leading a research team that mainly develops methods and technologies for spatial and single-cell metabolomics. I also work in a metabolomics core facility, which is much more routine, but it was very useful for me to learn the basics and the inside-outs of metabolomics. This is because it's actually not my background. On top of this, I'm also a scientific lead at the Bio-Innovation Institute in Copenhagen, Denmark, where we are incubating a startup in single-cell metabolomics.

My background is computational. My PhD was in mathematics and statistics. I then went to Germany for a postdoc in a city called Bremen. You might not have heard of it, and not many people have, other than for the Town Musicians of Bremen fairy tale. Surprisingly, this is actually the world capital for mass spectrometry. Two out of five major mass-spec vendors have R&D departments there, which I didn't know then. My first postdoc was on econometrics on predicting transactions for MasterCard. Then someone approached me from the Bruker mass spec company – 'Can you do some machine learning predictions for something called mass spectra?' That's how I got into mass spectrometry.

It was computational first, but then I was asking myself – 'Why do we need these mass spectra?' And the answer was metabolomics, because metabolomics uses mass spectrometry to get a read out about small molecules. Back then it was an up-and-coming technology, and I went to UCSD to learn it. I got hosted by Pieter Dorrestein and learned a bunch from him and his team. This is how I transitioned more and more towards metabolomics, and molecular questions and molecular biology.

***FLG:* Could you provide an overview of how the field of metabolomics has developed to where we are now with single-cell and spatial metabolomics methods?**

Theo: I probably can't speak for the whole field of metabolomics. I see myself as a newcomer and there was so much amazing work done by many people, which I didn't see. I can talk only about my experiences, roughly starting from 2010. I came into this field through imaging mass spectrometry, which is a spatially resolved way of doing mass spectrometry. You can think of it as pixelated mass spectrometry, where for every pixel you have a mass spectrum. This is almost like a quantified barcode - which molecules are there and their relative intensities.

Imaging mass spectrometry was initially developed for proteomics, but it was not really delivering for proteomics, to these molecular machines. Then there was the introduction of high-resolution mass spectrometry, and this changed things completely. People realised that with high mass resolution one can resolve small molecules, metabolites, and the whole field went towards spatial metabolomics. This is how spatial metabolomics emerged and got reinforced by this amazing technology.

If we start talking about omics - so, genomics, transcriptomics, proteomics, metabolomics, epigenetics and so on - we can think about genomics as 'what can happen'. Transcriptomics is 'what might happen'; i.e., the part of the blueprint that will get activated and be turned into reality. Proteomics is about 'what makes it happen', because these molecular machines are like cogs in our cells that turn small molecules. Metabolomics is very interesting because it is about 'what is happening right now.' Metabolites have extremely rapid turnover, probably the most rapid out of all molecular entities. At the same time, there are usually a lot of them, but sometimes there is not a lot. It's all about stability, balance and very rapid reprogramming. It's about putting things into perspective.

Metabolomics has existed for decades. Spatial metabolomics has existed for about 15 years. Single-cell metabolomics is a completely different story. Single-cell metabolomics takes a molecular profile for an individual cell, and does it for many different cells. For every cell you will get a profile. Single-cell technologies, in particular genomics and transcriptomics, have paved the way over the past decade. But for metabolomics, it was very tricky. Only a couple of groups were able to do it, say 10 years ago, because it was almost impossible for metabolomics; conventional bulk metabolomics requires 1 million cells. Then mass spectrometry got much more sensitive, and now it's possible to do single-cell metabolomics. Now it's a rapidly moving technology with advancing experimental and computational applications. Everything's bubbling and booming and there is so much new stuff going on.

***FLG:* Is single-cell spatial metabolomics the gold standard that people are aiming for?**

Theo: Ultimately, I'm envisioning a future where metabolomics will be done routinely, spatially and in single-cell. We just had a review in Molecular Systems Biology[3] where we put forward a dream that saying, 'spatial metabolomics' would be as awkward as saying 'spatial microscopy'. You're not saying, 'spatial microscopy', because it's natively spatial. I think the same will be true for other omics as well once the technology becomes accessible. We'll be saying – 'I did metabolomics of tissue sections', and we'll mean single-cell/spatial.

***FLG:* Could you describe some of the developments that your lab has had in this area?**

Theo: We first came into spatial metabolomics from a computational angle. We first started developing statistical methods to help others mine through this data. The data that's generated is pretty big, it can reach up to several hundred gigabytes per tissue section. We developed a number of methods. Then, I was very fascinated to find the bottleneck – what is stopping the whole field from developing and growing. Critical back then, which was about five to ten years ago, was the problem of metabolite identification. How to find which molecules are encoded in these gigabytes of data. An image that we generate with spatial metabolomics has up to a million different channels. Some of these channels represent molecules, some do not.

To help with this, we developed software called METASPACE[4], where you can put data in and get molecules out, you can get images up and you can get additional imputation out. This is one of our key contributions to this field. It's now a cloud-based software and it is free and open source. There are now more than 2,000 users that upload data and we help them annotate that. Also, lots of people actually started sharing their data through this online platform, and this created more than 10,000 public datasets. Now anyone can go there and ask – 'Is this molecule adaptable,' 'In which context,' and so on. It's creating a knowledge base with not only molecular information but also associated with metadata. Just recently, we developed the new version for this metabolite identification, which is still one of the key bottlenecks of the field. It's machine learning-based. It allows us to find more molecules with a high confidence, and it's trained on the public data that people put into METASPACE. It's like a reinforcing cycle; when you share your datasets publicly, they can be used for training better methods, and thus you contribute to advancing the field.

> "IT'S MACHINE LEARNING-BASED. IT ALLOWS US TO FIND MORE MOLECULES WITH A HIGH CONFIDENCE, AND IT'S TRAINED ON THE PUBLIC DATA THAT PEOPLE PUT INTO METASPACE. IT'S LIKE A REINFORCING CYCLE; WHEN YOU SHARE YOUR DATASETS PUBLICLY, THEY CAN BE USED FOR TRAINING BETTER METHODS, AND THUS YOU CONTRIBUTE TO ADVANCING THE FIELD."

Then we got into the field of single-cell metabolomics, and this goes back to my first PhD student, Luca Rappez. Back then, we were not doing single-cell metabolomics. When Luca was searching for his PhD project, he asked – 'Can we do single-cell metabolomics with mass spec?' And I told him – 'You can try.' And he tried, and this completely changed how we do things, and it completely changed the focus of our team. Most of our developments now are in single-cell metabolomics.

The major foundation for us is a method called SpaceM[5]. It integrates imaging mass spectrometry, that we also use in spatial metabolomics, and integrates it with microscopy and with the computational methods that Luca has developed. Of course, since then, we have refined the method and improved it further. Now it's applicable to pretty much any cell type on a glass slide to then produce single-cell metabolomics profiles. This opened a lot of very interesting opportunities for tech development, for applications, and even for the commercialization opportunity that I mentioned. A lot of people in the world, and also companies, are interested in getting single-cell metabolomics profiles to make their decisions, either for developing better drugs, for developing better therapies or maybe even for diagnostics.

FLG: ***What is the progress in integrating metabolomics with other omics?***

Theo: Metabolomics is very challenging to combine with other omics because it works on a completely different layer. You can't use sequencing and you don't have amplification. So, you need to figure out how to do it. In this respect, spatial and single-cell provide opportunities that were not available for bulk. If you do homogenization, then you pretty much have a smoothie, and out of this smoothie you can isolate and analyse only a fraction.

For spatial metabolomics, there were recently some very interesting developments and technologies and there are now even commercial products for combining metabolomics with antibody-based protein detection. It opens the possibilities for any antibody-based technology, e.g., multiplex immuno-fluorescence, cyclic or non-cyclic, any metal tagged antibodies, as well as commercially-suggested peptide tagged antibodies. As long as you first do spatial metabolomics, then there is enough material left intact enough to do antibody analysis. So, you can combine beautiful spatial metabolomic images with protein detection using antibodies on the same tissue section. This is amazing because you can delineate cell types and tissue compartments.

Secondly, people are asking - 'What about transcriptomics?' Because single-cell transcriptomics has shown us so many new cell types. How awesome would it be to overlay this information on top of spatial metabolomics? However, spatial metabolomics is often done with a laser, and what happens with our transcripts? RNA can be degraded very rapidly, and one would think that if we shoot it with the laser, we will just break it into parts and fragments. But it was very convincingly shown, in the context of microbiology, that you can detect transcripts after you've done metabolomics. Metabolomics needs to be done first, but afterwards, it was shown that one can do RNA-FISH or spatial transcriptomics analyses. A notable example is a recent paper in Nature Biotech from the Andren and Lundeberg groups, combining spatial metabolomics and Visium spatial transcriptomics[6]. This opens the door to multimodal spatial omics combining spatial metabolomics with detecting transcripts or proteins.

Single-cell and Spatial Metabolomics methods

SPATIAL METABOLOMICS

Spatial metabolomics provides regional information on metabolites in cells and tissues. While spatial metabolomics is possible with several approaches, such as Raman spectroscopy and fluorescence lifetime imaging, the method with the most momentum is Mass Spectrometry Imaging (MSI)[3].

With MSI, to quantify metabolites, the sample is first ablated with a desorption/ionization source. Frequently adopted examples include, desorption electrospray droplets (DESI[7]), Secondary Ion beam Mass Spectrometry (SIMS[8]) or a matrix-assisted laser desorption/ionization (MALDI[9]). The sample is divided into regional pixels, and within each pixel, the molecules that have been desorbed by the ionization source are used to create a mass spectrum for each pixel.

SIMS based MSI approaches have the highest spatial resolution at nanometer scale, but the high energy of the primary ion beam can cause molecular fractionation, complicating analysis[10] (see Figure 6.2) . By contrast, MALDI liberates a greater proportion of intact molecules but typically achieves a resolution of > 10 microns. Recent advances have shown that better resolution is possible[11]. One downfall of these methods is that they are typically operated in a vacuum, meaning cells are not analysed in their native state. DESI methods offer analysis under ambient conditions, but tend to not be able to get resolution much below 50 microns (see Figure 6.3).

FIGURE 6.2. PRINCIPLES OF MAIN MSI TECHNOLOGIES AND TYPICAL WORKFLOW OF MALDI.

*(**A**) Illustration of the ionization methods of three major ion sources used in MSI technologies: SIMS, MALDI, and DESI. (**B**) The typical workflow of MALDI-based MSI metabolomics. Take brain tissue as an example Image Credit: Pang and Hu [10]*

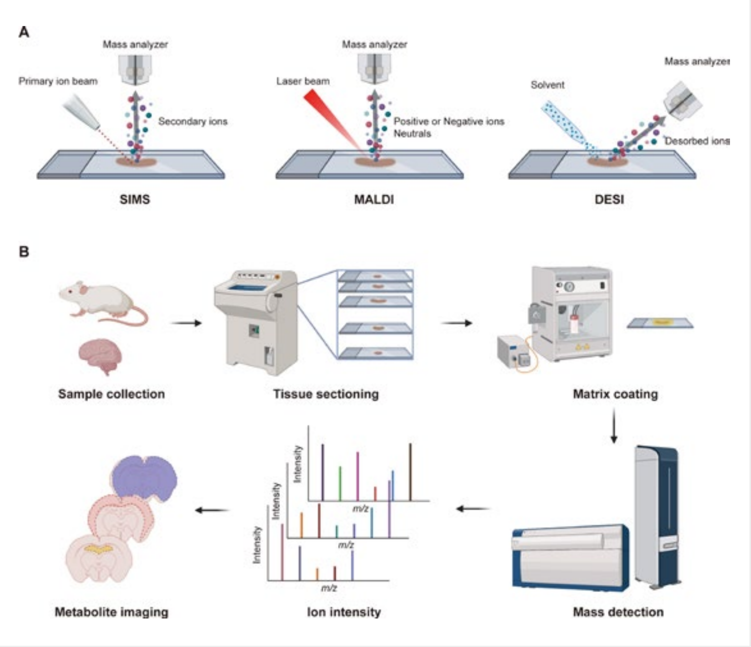
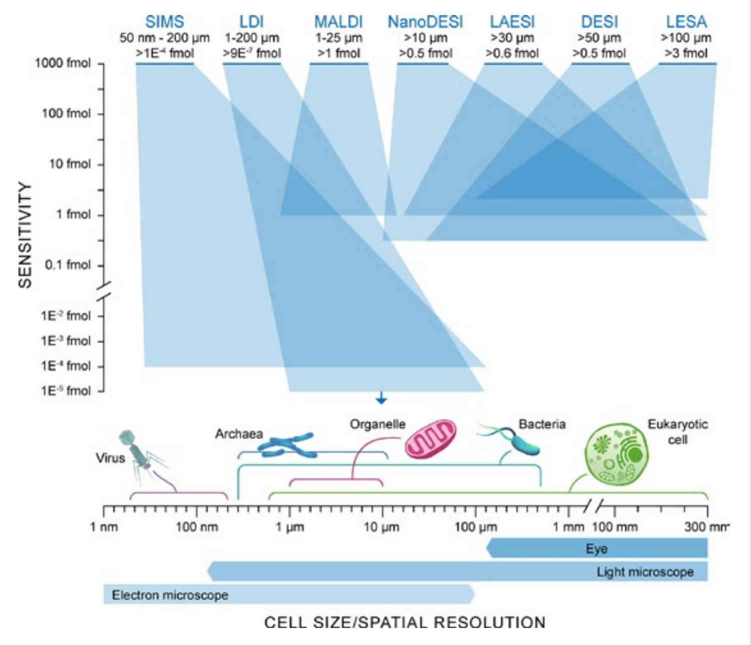


FIGURE 6.3. REPRESENTATION OF THE TECHNIQUES AVAILABLE FOR MS-BASED SPATIAL METABOLOMICS.

The range of sensitivities in femtomoles (y-axis) is compared against the spatial resolution range (x-axis) for these spatial-MS approaches. The spatial dynamic range is illustrated by the transparent blue boxes. Cell size dimensions and the lateral resolution of other structural imaging techniques are displayed along the x-axis for comparison. Image & Caption Credit: Taylor, et al. [12]



Spatial metabolomics is not without its issues. The broad diversity of molecules being measured, and the sensitivity of the ionization sources, mean that capturing and identifying even a small fraction of the available information is hugely challenging.

Hence, a prominent issue for all metabolomics experiments (whether spatial or single-cell) is the issue of annotating metabolites[13]. Most metabolites detected in these untargeted mass spectrometry methods end up unannotated due to the lack of reference libraries or databases for them. Despite the advances in technology, knowing the metabolites you have captured is still a challenge.

There are many recent advances in computation to address this challenge including deep learning models such as: CANOPUS[14], DarkNPS[15] and MSNovelist[16] and network approaches such as: GNPS[17], NetID[18] and KGMN[19]. But this is still a big challenge for the field. See this 2023 review for recent advances in computational metabolomics including annotation, visualization and integration[20].

We spoke to **Dr. Xiaotao Shen** about his pseudo mass spec imaging method[21], which can help address some of these mass spec challenges listed above.

XIAOTAO SHEN

Postdoctoral Research Fellow,, Snyder Lab
Stanford University

FLG: I would like to ask you about metabolomics. You did some work with deep learning-based pseudo mass spec imaging, could you describe how it works and how it helps metabolomic analysis?

Xiaotao: *Mass spectrometry is a high sensitivity instrument for small compounds, proteins, etc. The raw mass spec data provides chemical information, it's not biological information. We only know some information about the features; for example, their accurate mass. But we don't know what they are. To get this data, we need to use a component annotation programme. This is a challenge for the metabolomics field. Typically, we can get 100,000 features from the mass spectrometry raw data. But we can only identify 100 or so of the features. This is why I want to work on this method for diagnosis, because one of the most promising applications of metabolomics is for diagnosis.*

So, I wanted to know whether we could convert the mass spec data into an image and then use the image for diagnosis. If you see the raw mass spec data, it has three dimensions, an axis of time and an axis of accurate mass. So, I converted this into an image, and we used deep learning to process it. Deep learning is very powerful for imaging processing. This image contains all the information from that raw data. Using a machine learning model, we can then predict whether this mass spec is associated with disease.

Another challenge for metabolomics mass spectrum data is batch effect. For example, we have 100s of 1,000s of samples and we cannot measure all the samples in one day. Sometimes, it can take months or even a year to measure all the samples. And mass spectrometry is not so robust. Between today and tomorrow, maybe the intensity or the sensitivity will shift. So, we need to correct the batch of data, but it's very difficult. However, if we convert the raw data to an image, the shift will just cause changes in the levels of dark or light for the image. But the whole profile is still there. So, this is another advantage of this method - we can overcome the batch effects in the metabolomics data, so it can increase the performance and accuracy of the predictive model.

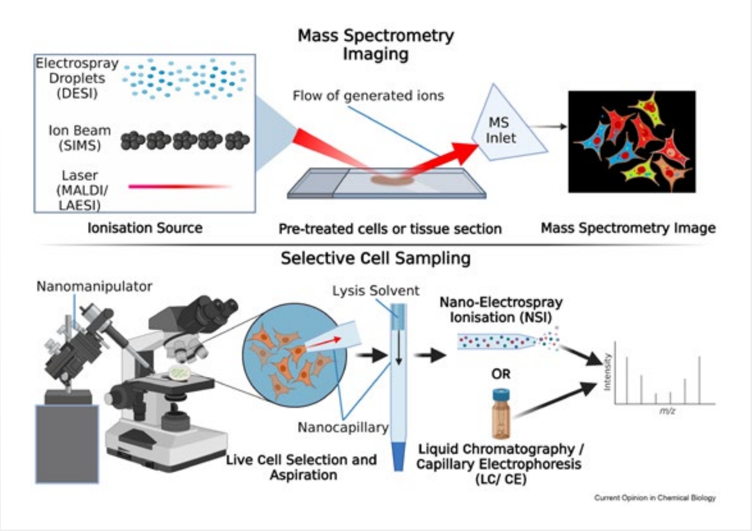
SELECTIVE CELL SAMPLING METABOLOMICS

Ways to produce single-cell level metabolomics have been explored with many reviews in the last year, which readers should refer to.[1,2,12,22-25]

Single-cell metabolomics is achieved using mass spec with a collection of sampling strategies and ionization techniques. Whereas spatial methodologies have to view the sample as 'pixel', single-cell methodologies capture metabolomics insights from the individual biological unit, the single cell and its internal biochemical processes[2] (see Figure 6.4)

FIGURE 6.4. REPRESENTATION OF MASS SPECTROMETRY IMAGING (MSI) AND SELECTIVE CELL SAMPLING (SCS) AS STRATEGIES FOR SPATIAL AND SINGLE-CELL METABOLOMICS RESPECTIVELY.

Image Credit: Saunders, et al.[25]

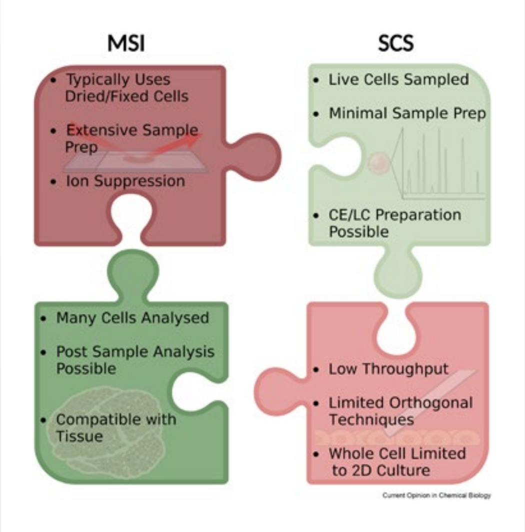

Mass Spectrometry Imaging
Electrospray Droplets (DESI)
Ion Beam (SIMS)
Laser (MALDI/ LAESI)
Flow of generated ions
MS Inlet
Ionisation Source
Pre-treated cells or tissue section
Mass Spectrometry Image
Selective Cell Sampling
Nanomanipulator
Lysis Solvent
Nano-Electrospray Ionisation (NSI)
OR
Intensity
m/z
Nanocapillary
Live Cell Selection and Aspiration
Liquid Chromatography / Capillary Electrophoresis (LC/ CE)
Current Opinion in Chemical Biology

Methods to draw single-cell insights from MALDI-MSI and other spatial metabolomics data also exist. SpaceM5 is an example, as already discussed in the interview above. This method can precisely estimate the cell parts that were ablated by the laser with subcellular precision. This method can then detect > 100 metabolites from > 1,000 individual cells per hour, giving a spatio-molecular matrix for each cell with a normalized metabolomic profile.

However, by using live cell selection followed by mass spectrometry, it is possible to truly capture a single cell's metabolome but only for cells in culture. These methods involve using a micropipette to sample a cell directly (or even a subcellular compartment of a cell) before transferring them to the mass spectrometer. This is termed live single-cell MS or Direct Analyte Probe Nanoextraction (DAPNe).

FIGURE 6.5. KEY SYNCHRONICITIES BETWEEN MSI AND SCS.

Image Credit: Saunders, et al.[25]


MSI
SCS
Typically Uses Dried/Fixed Cells
Extensive Sample Prep
Ion Suppression
Live Cells Sampled
Minimal Sample Prep
CE/LC Preparation Possible
Many Cells Analysed
Post Sample Analysis Possible
Compatible with Tissue
Low Throughput
Limited Orthogonal Techniques
Whole Cell Limited to 2D Culture
Current Opinion in Chemical Biology

Since the cell sampling and ionization are now separate processes, this gives the opportunity of manipulating the extracted cellular material prior to MS analysis. Hence, liquid chromatography LC-MS[26] or capillary electrophoresis and MS[27] can be used to separate analytes prior to analysis.

Single-cell metabolomics presents its own challenges[23]. Principle among these is the low picolitre volume of the material available in one cell for analysis. Furthermore, given the metabolome's nature to radically shift and the huge diversity of metabolites, accurately measuring the metabolome of an individual cell is challenging[22]. Finally, cell selection methods are low throughput. Finding the bridge between the advantages of the spatial metabolomics and single-cell metabolomics methods is crucial for the field to advance (Figure 6.5)[25].

We asked two of our experts some questions about their work in single-cell mass spectrometry.

HOLLY-MAY LEWIS

Senior Laboratory Technician (LC-MS)
University of Surrey

FLG: What practical advice might you have for someone getting into mass spec?

Holly: *Working with mass spectrometers regularly comes with challenges when it comes to optimization and troubleshooting. I would recommend taking time in the optimization stage and incorporating quality controls to always confirm your instrumentation is working optimally.*

FLG: What are some of the most popular methods for profiling the metabolome of human cells and is there a current 'gold-standard'?

Holly: *Single cell analysis is a new and exciting area of mass spectrometry. There have recently been a growing number of sampling approaches including mass spectrometry imaging, capillary sampling and microfluidics. I don't think there is a 'gold-standard', just lots of different exciting approaches for different studies.*

FLG: Can you briefly describe the DAPNe-LC-MS method for spatial Lipidomics and how it compares to other LC-MS methods?

Holly: *Many spatial sampling techniques are direct-MS methods, where there is no chromatographic separation of analytes prior to ionisation. This means that analytes are ionised simultaneously, which can lead to ion suppression and can limit sensitivity. DAPNe-LC-MS was a way of incorporating the capillary sampling approach of single cell sampling with the chromatographic separation of LC-MS.*

INGELA LANEKOFF

Professor, Department of Chemistry-BMC
Uppsala University

FLG: How are metabolites and lipids analysed, and how can that information be used clinically?

Ingela: *Similar to proteins, metabolites and lipids are most often analyzed with mass spectrometry that is usually coupled to a liquid or gas chromatography separation system for separation of the molecules prior to mass spectrometry. However, metabolites and lipids can also be detected with mass spectrometry alone in a direct infusion mode, where there is no prior separation.*

FLG: Can you briefly overview the most popular methods for single-cell profiling the metabolome of human cells?

Ingela: *Lipids are the most commonly analyzed metabolites of individual cells and the main techniques for this use direct infusion mass spectrometry with either sampling and ionization using MALDI (matrix assisted laser desorption ionization) or in-house built tools coupled to electrospray ionization. Profiling the metabolome of individual cells is a young field that is still exploring techniques to ensure a high coverage and high detection of metabolites and lipids. I expect to see a significant growth in the field, with creative ways to analyze the metabolome of individual cells being available in the near future!*

FLG: What are some of the developments that your lab has been working on in this area?

Ingela: *In my lab, we are focusing on using miniature liquid extraction of metabolites and lipids from individual cells that reside on the surface of glass slide. For this, we are developing and building probes that first extract the metabolites from the cell, and then transfer the liquid extract to the inlet on the mass spectrometer for electrospray ionization. This provides us with an on-line extraction for direct infusion mass spectrometry where we see all kinds of metabolites, including amino acids, energy metabolites, and a range of lipid species from each cell.*

FLG: What are the efforts like to produce multi-omics technology with metabolomics alongside other omics?

Ingela: *This is an important part of moving forward to understand the chemical mechanisms that occur in a cell. Due to the emerging field of single-cell metabolomics, where the focus still to some degree is technique development, there are not so many studies that employ multi-omics. However, I am confident that this will soon become more widely used.*

INTEGRATION OF METABOLOMICS WITH OTHER OMICS

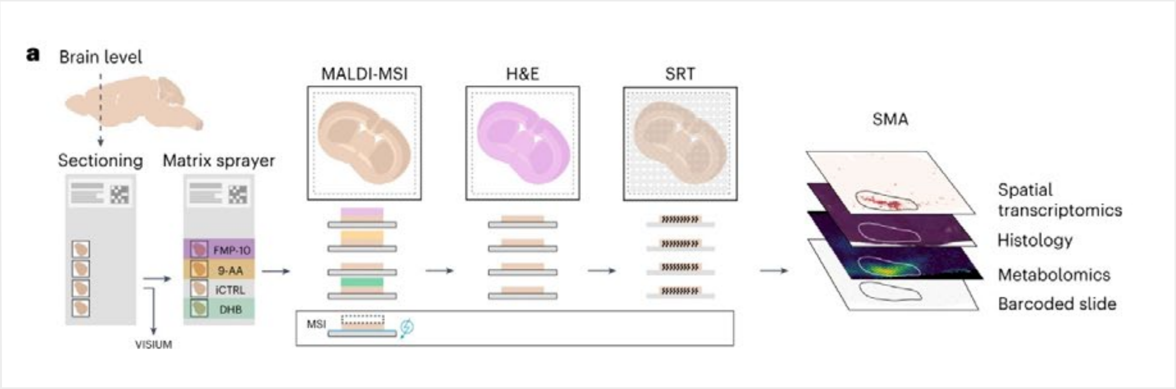
Spatial metabolomics offers a number of advantages over other omics analysed spatially. The data is much closer to cell state than gene or protein levels. It allows the incorporation of the exposome, microbiome and drug levels since exogenous compounds can be analysed too. There are practical advantages too, it is faster and substantially cheaper to operate[3]. This makes metabolomics a must-include omic in multi-omics studies for perturbation, clinical applications and drug-discovery.

As highlighted by the contributors above, vertical integration of metabolomics methodologies with other omics is quite challenging. There are numerous examples of diagonal integration, in which metabolomics profiling is performed concurrently from an adjacent section with other omics and integrated post-hoc[28]. What is more novel, is the genuine integration of spatial metabolomics methods with antibody-based proteomics workflows and commercial transcriptomics workflows.

An example of this integration was published in September 2023[6], in which mass spectrometry-based imaging is performed first using the MALDI-MSI system, followed by H&E staining, and then spatial transcriptomics using the 10x Genomics Visium platform. This was performed in mouse brain tissue. This method presents a way to simultaneously profile small molecules and gene expression within a tissue section, making spatial metabolomics a true multi-omic (see Figure 6.6).

FIGURE 6.6. THE SMA WORKFLOW AND QUALITY CONTROL DESIGN.

Nonembedded, snap-frozen samples were sectioned and thaw-mounted onto noncharged, barcoded Visium Gene Expression arrays. Tissue sections were then sprayed with MALDI matrices and MSI is performed. This was followed by H&E staining and imaging with bright field microscopy. Finally, sections were processed for SRT. Image Credit: Vicari, et al. [6]



Applications of metabolomics

Single-cell and spatial metabolomics has a similar portfolio of utility as the other major omics (transcriptomics, proteomics etc.), a topic we will cover in depth in the next chapter. However, before we get there, we would like to take some time to cover some specific uses that metabolomics has, to further illustrate the value of this emerging omic[29].

Cancer metabolism is a key area of work in metabolomics. Tumour cells have incredible flexibility in reprogramming their metabolism to support the various states of cancer evolution and in becoming resistant to therapies[30]. Understanding this process and identifying ways to interfere and prevent cancer metabolism is the goal within this field. A recent example of metabolomics in colorectal cancer[31,32] identified new metabolic vulnerabilities in patients and hence drug targets to improve patient outcomes.

In drug research, metabolomics presents a way to identify drug targets and elucidate the mode of action of drugs as well as the pharmacokinetics, pharmacodynamics[33,34]. This improves drug repurposing and efforts to identify drug-drug interactions[10]. Lipidomics has shown to be particularly valuable for drug discovery and development[35]. Furthermore, new methods such as Nanocapillary sampling LC-MS, which can be used to detect the uptake of a particular drug molecules in a single cell, provide precise data on drug uptake in individual cells[26].

HOLLY-MAY LEWIS

Senior Laboratory Technician (LC-MS)
University of Surrey

FLG: Can you describe the Nano-capillary sampling[26] method and the value of measuring drug-uptake in single cells?

Holly: *The benefit of measuring drugs in single cells being able to investigate inter-cell drug uptake/penetration heterogeneity, which could lead to the effective treatment of both cancer and infectious diseases. This could also be applied to radiation of cells and investigating bystander effects.*

FLG: Can you describe some of your work with metabolomics in COVID19[36], what kind of diagnostic power does metabolomics have for conditions such as COVID-19?

Holly: *I am currently more involved in metabolomic analysis of clinical samples using mass spectrometry. A person's metabolomic profile can reflect clinical disturbance when a person is infected with a virus such as COVID-19. Metabolic profiling therefore can identify biomarkers and could be used as both a diagnostic and prognostic tool.*

Clinical lipidomics is particularly useful since lipids are the most commonly analysed metabolite. Lipid are major components of the cell membrane and lipoproteins are distributed throughout the bodies tissues, meaning that the status of these lipoproteins can serve as a good indicator of an individual's metabolic state[37]. A recent report established >800 unique lipid species, of which many were shown to have associations to conditions such as ageing, diabetes and inflammation[38], and even a lipidomic state for heart failure[39].

Another major use of metabolomics is to investigate the role of the microbiome in human functioning and disease[40]. The microbes that live within us secrete small molecules into our systems to change our biology. Metabolomics allows us to track these molecular interlopers. The crosstalk between the microbiome and cancer is also being explored, a biological process that needs metabolomic methods to truly elucidate[41,42].

Hopes for Metabolomics

Finally, we asked two of our experts what they thought were some of the exciting things happening in metabolomics, and what their hopes were for the future of metabolomics.

INGELA LANEKOFF

Professor, Department of Chemistry-BMC
Uppsala University

FLG: What are your hopes for single-cell and spatial metabolomics?

Ingela: *I hope that the sensitivity can be increased further to enable a higher coverage of the metabolome and to allow for detection of low abundant metabolites or metabolites with poor ionization efficiency. With the constant improvement of sensitivity offered by modern mass spectrometers I hope we will get there soon - as long as we can afford to pay for the instruments.*

FLG: What are some of the latest/exciting things happening in the single-cell and spatial metabolomics?

Ingela: *It is exciting times in the field of single-cell metabolomics, with the community exploring ways to detect metabolites from a single cell. This can be done by sucking out a small portion of the cell, analyzing the cell intact through liquid extraction, or even putting an individual cell into a very fine pipette. I think it is exciting that more groups are seeing more small metabolites in addition to lipids!*

For spatial metabolomics I think the most exciting thing right now is the drive to differentiate between isomeric molecules without having to homogenize your sample. There are several groups, including mine, that are working with new creative strategies to fragment molecules in the mass spectrometer to identify the exact position of double bonds or functional groups.

HOLLY-MAY LEWIS

Senior Laboratory Technician (LC-MS)
University of Surrey

FLG: What are your hopes for single-cell and spatial metabolomics?

Holly: *I think the technology available will continue to improve and will continue gain significant attention in the mass spectrometry world. There are more and more single-cell mass spectrometry meetings happening, meaning that the community is communicating and collaborating, which will certainly further the field.*

Chapter 6 references

1. BLiu, Q., Martínez-Jarquín, S. & Zenobi, R. **Recent Advances in Single-Cell Metabolomics Based on Mass Spectrometry.** *CCS Chemistry* **5**, 310-324 (2023).

2. Duncan, K.D., Fyrestam, J. & Lanekoff, I. **Advances in mass spectrometry based single-cell metabolomics.** *Analyst* **144,** 782-793 (2019).

3. Alexandrov, T. **Spatial metabolomics: from a niche field towards a driver of innovation.** *Nature Metabolism* **5**, 1443-1445 (2023).

4. Nguyen, D.D. *et al.* **Facilitating Imaging Mass Spectrometry of Microbial Specialized Metabolites with METASPACE.** *Metabolites* **11**(2021).

5. Rappez, L. *et al.* S**paceM reveals metabolic states of single cells.** *Nature Methods* **18**, 799-805 (2021).

6. Vicari, M. *et al.* **Spatial multimodal analysis of transcriptomes and metabolomes in tissues.** *Nature Biotechnology* (2023).

7. Wang, Z. *et al.* **In situ metabolomics in nephrotoxicity of aristolochic acids based on air flow-assisted desorption electrospray ionization mass spectrometry imaging.** *Acta pharmaceutica sinica* B **10**, 1083-1093 (2020).

8. Pareek, V., Tian, H., Winograd, N. & Benkovic, S.J. **Metabolomics and mass spectrometry imaging reveal channeled de novo purine synthesis in cells.** *Science* **368**, 283-290 (2020).

9. Good, C.J. *et al.* **High spatial resolution MALDI imaging mass spectrometry of fresh-frozen bone.** *Analytical chemistry* **94**, 3165-3172 (2022).

10. Pang, H. & Hu, Z. **Metabolomics in drug research and development: The recent advances in technologies and applications.** *Acta Pharmaceutica Sinica B* **13**, 3238-3251 (2023).

11. Cuypers, E. *et al.* **'On the spot'digital pathology of breast cancer based on single-cell mass spectrometry imaging.** *Analytical Chemistry* **94**, 6180-6190 (2022).

12. Taylor, M.J., Lukowski, J.K. & Anderton, C.R. **Spatially Resolved Mass Spectrometry at the Single Cell: Recent Innovations in Proteomics and Metabolomics.** *Journal of the American Society for Mass Spectrometry* **32**, 872-894 (2021).

13. Singh, A. **Annotating unknown metabolites.** *Nature Methods* **20**, 33-33 (2023).

14. Dührkop, K. *et al.* **Systematic classification of unknown metabolites using high-resolution fragmentation mass spectra.** *Nature Biotechnology* **39**, 462-471 (2021).

15. Skinnider, M.A. *et al.* **A deep generative model enables automated structure elucidation of novel psychoactive substances.** *Nature Machine Intelligence* **3**, 973-984 (2021).

16. Stravs, M.A., Dührkop, K., Böcker, S. & Zamboni, N. **MSNovelist: de novo structure generation from mass spectra.** *Nature Methods* **19**, 865-870 (2022).

17. Nothias, L.-F. *et al.* **Feature-based molecular networking in the GNPS analysis environment.** *Nature Methods* **17**, 905-908 (2020).

18. Chen, L. *et al.* **Metabolite discovery through global annotation of untargeted metabolomics data.** *Nature Methods* **18**, 1377-1385 (2021).

19. Zhou, Z. *et al.* **Metabolite annotation from knowns to unknowns through knowledge-guided multi-layer metabolic networking.** *Nature Communications* **13**, 6656 (2022).

20. Ebbels, T.M.D. *et al.* **Recent advances in mass spectrometry-based computational metabolomics.** *Current Opinion in Chemical Biology* **74**, 102288 (2023).

21. Shen, X. *et al.* **Deep learning-based pseudo-mass spectrometry imaging analysis for precision medicine.** *Briefings in Bioinformatics* **23**, bbac331 (2022).

22. Lanekoff, I., Sharma, V.V. & Marques, C. **Single-cell metabolomics: where are we and where are we going?** *Current opinion in biotechnology* **75**, 102693 (2022).

23. Zhang, C., Le Dévédec, S.E., Ali, A. & Hankemeier, T. **Single-cell metabolomics by mass spectrometry: ready for primetime?** *Current Opinion in Biotechnology* **82**, 102963 (2023).

24. Ali, A. *et al.* **Single cell metabolism: current and future trends.** *Metabolomics* **18**, 77 (2022).

25. Saunders, K.D., Lewis, H.-M., Beste, D.J., Cexus, O. & Bailey, M.J. **Spatial single cell metabolomics: Current challenges and future developments.** *Current opinion in chemical biology* **75**, 102327 (2023).

26. Lewis, H.-M. *et al.* **Nanocapillary sampling coupled to liquid chromatography mass spectrometry delivers single cell drug measurement and lipid fingerprints.** *Analyst* **148**, 1041-1049 (2023).

27. Liao, H.-W., Rubakhin, S.S., Philip, M.C. & Sweedler, J.V. **Enhanced single-cell metabolomics by capillary electrophoresis electrospray ionization-mass spectrometry with field amplified sample injection.** *Analytica chimica acta* **1118**, 36-43 (2020).

28. Sun, C. *et al.* **Spatially resolved multi-omics highlights cell-specific metabolic remodeling and interactions in gastric cancer.** *Nature Communications* **14**, 2692 (2023).

29. Guo, S., Zhang, C. & Le, A. **The limitless applications of single-cell metabolomics.** *Current Opinion in Biotechnology* **71**, 115-122 (2021).

30. Danzi, F. *et al.* **To metabolomics and beyond: a technological portfolio to investigate cancer metabolism.** *Signal Transduction and Targeted Therapy* **8**, 137 (2023).

31. Lee, M.Y. & Tam, W.L. **Multimodal metabolomics pinpoint new metabolic vulnerability in colorectal cancer.** *Nature Metabolism,* **1-3** (2023).

32. Vande Voorde, J. *et al.* **Metabolic profiling stratifies colorectal cancer and reveals adenosylhomocysteinase as a therapeutic target.** *Nature Metabolism* **5**, 1303-1318 (2023).

33. Alarcon-Barrera, J.C., Kostidis, S., Ondo-Mendez, A. & Giera, M. **Recent advances in metabolomics analysis for early drug development.** *Drug discovery today* **27**, 1763-1773 (2022).

34. Garana, B.B. & Graham, N.A. **Metabolomics paves the way for improved drug target identification.** *Molecular Systems Biology* **18**, e10914 (2022).

35. Kostidis, S., Sánchez-López, E. & Giera, M. **Lipidomics analysis in drug discovery and development.** *Current Opinion in Chemical Biology* **72**, 102256 (2023).

36. Spick, M. et al. **An integrated analysis and comparison of serum, saliva and sebum for COVID-19 metabolomics.** *Scientific Reports* **12**, 11867 (2022).

37. Salihovic, S., Lamichane, S., Hyötyläinen, T. & Orešič, M. **Recent advances towards mass spectrometry-based clinical lipidomics.** *Current opinion in chemical biology* **76**, 102370 (2023).

38. Hornburg, D. *et al.* **Dynamic lipidome alterations associated with human health, disease and ageing.** *Nature Metabolism,* 1-17 (2023).

39. Ren, J. *et al.* **Mass Spectrometry Imaging-Based Single-Cell Lipidomics Profiles Metabolic Signatures of Heart Failure.** *Research* **6**, 0019 (2023).

40. Bauermeister, A., Mannochio-Russo, H., Costa-Lotufo, L.V., Jarmusch, A.K. & Dorrestein, P.C. **Mass spectrometry-based metabolomics in microbiome investigations.** *Nature Reviews Microbiology* **20**, 143-160 (2022).

41. Ganesan, R., Yoon, S.J. & Suk, K.T. **Microbiome and metabolomics in liver cancer: scientific technology.** *International Journal of Molecular Sciences* **24**, 537 (2022).

42. Mukherjee, A.G. *et al.* **The crosstalk of the human microbiome in breast and colon cancer: a metabolomics analysis.** *Critical Reviews in Oncology/Hematology* **176**, 103757 (2022).

MALLEABLE MULTI-OMICS. THE VARIOUS APPLICATIONS OF MULTI-MODAL DATA

MULTI-MODAL DATA HELPS US GET A CLEARER UNDERSTANDING OF THE MOLECULAR PROCESSES IN CELLS. THIS IN TURN MAKES IT EXTREMELY USEFUL FOR SCIENTISTS. IN THIS CHAPTER WE WILL REVIEW THE APPLICATIONS OF MULTI-OMICS IN BOTH THE CLINIC AND IN DRUG DISCOVERY, AND HOW AI AND DATA ANALYTICS ARE USED TO IMPROVE OUTCOMES.

Multipurpose multi-omics: applications in a broad range of diseases

Multi-omics has the potential to change the role of the clinician[1]. The profiling of genomics, transcriptomics, epigenomics, proteomics and metabolomics offers the opportunity for holistic investigation and contextual molecular understanding of disease.

Many if not all disease research and treatment plans could benefit from a multi-omics approach, but we’ve seen recent advances in some key areas. Neuropsychiatry, early-life and pregnancy, cancer and COVID-19 will be considered in separate sections below.

NEUROPSYCHIATRY

One area in which multi-omics technologies have been consistently used is in neurodegeneration and neuropsychiatry. As we’ve already covered, the brain is inherently complex and to understand one of the most pressing diseases of our age, dementia, that complexity needs to be managed.

Studies using various combinations of genomics, transcriptomics, lipidomics, proteomics, epigenomics and metabolomics have been published for Alzheimer's disease over the last few years in a bid to make more sense of the complexity[2].

Studies from this year have unraveled things such as unique molecular signatures of disease progression and therapeutic targets using transcriptomics, proteomics and metabolomics[3] and brain regional GRNs for Alzheimer’s and COVID-19 phenotypes[4].

Furthermore, there was a recent review on the application of multi-omics methods for psychiatry[5]. It highlighted the limitations of mono-omic work for looking into such a complex system as the brain, and the ways to generate meaningful rich biological markers for diagnosis and treatment.

EARLY-LIFE AND PREGNANCY

Early life and pregnancy related disorders are also benefiting from the use of multi-omics[6]. This takes the form of understanding more about pregnancy-associated conditions and untangling the complex web of effects that occur during pregnancy that can lead to offspring outcomes in later life.

On the former, we spoke to **Dr Shirley Greenbaum** to shed some further light on this from her perspective of working in preeclampsia.

SHIRLEY GREENBAUM

Postdoctoral fellow, Department of Pathology, **Stanford University,** Resident, Department of Obstetrics and Gynaecology, **Hadassah-Hebrew University Medical Center**

FLG: What are your hopes for multi-omics methods in a clinical setting?

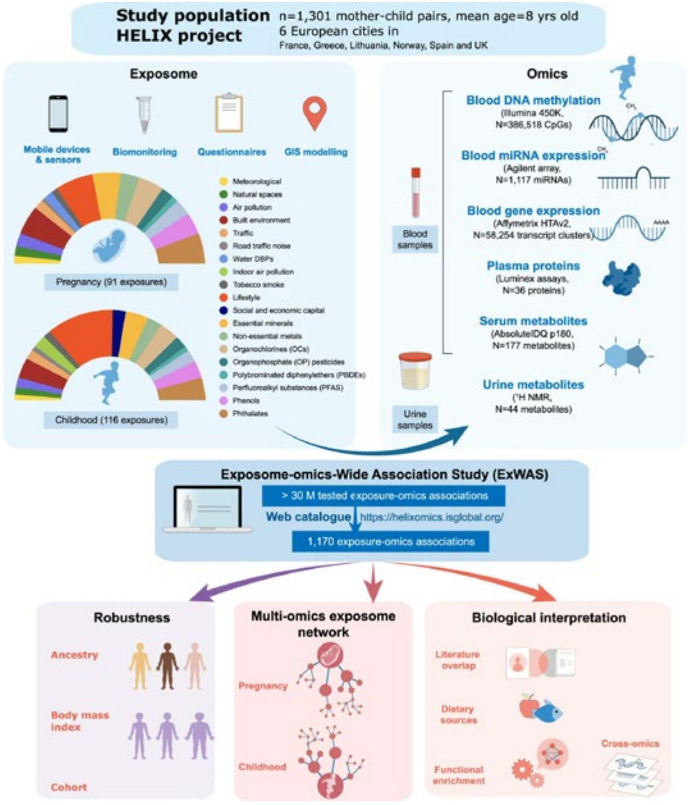
Shirley: *In the clinical world, preeclampsia remains one of our most significant challenges. This obstetric complication, affecting 5-8% of pregnancies, can escalate into a catastrophic condition with severe consequences for both the mother and the fetus. During my work as an obstetrician, I've encountered many patients affected by preeclampsia. Diagnosed during pregnancy, this life-threatening disease currently lacks a definitive treatment. Often, we're left with the difficult decision to induce labour prematurely.*

Despite many years of research in the field, preeclampsia is not completely understood. This might be because it is not a "one cell-one protein disease". In other words, we haven't found the one cell type to blame because there just isn't one. It is probably a multitude of things that go wrong, and that's where the multiomics approach becomes so essential. Now that we've been able to thoroughly enumerate and describe the various populations at the maternal-fetal interface, I'm optimistic about advancing our understanding of preeclampsia. I am really excited to see how the use of multiomics can assist us in doing that.

On the latter, a very large-scale multi-omics study called the HELIX project recently profiled an impressive array of the exposome as well as DNA, RNA, proteins and metabolites[7]. They used this resource to tie pregnancy and childhood exposures to genomic features, such as changes in the methylation status of genes in childhood, which impact later-life phenotypes. These kinds of resources will be pivotal for deconstructing the interaction between early-life exposures and later-life disease risk.

FIGURE 7.1. OVERVIEW OF THE HELIX PROJECT.

Source: Maitre, et al. [7]

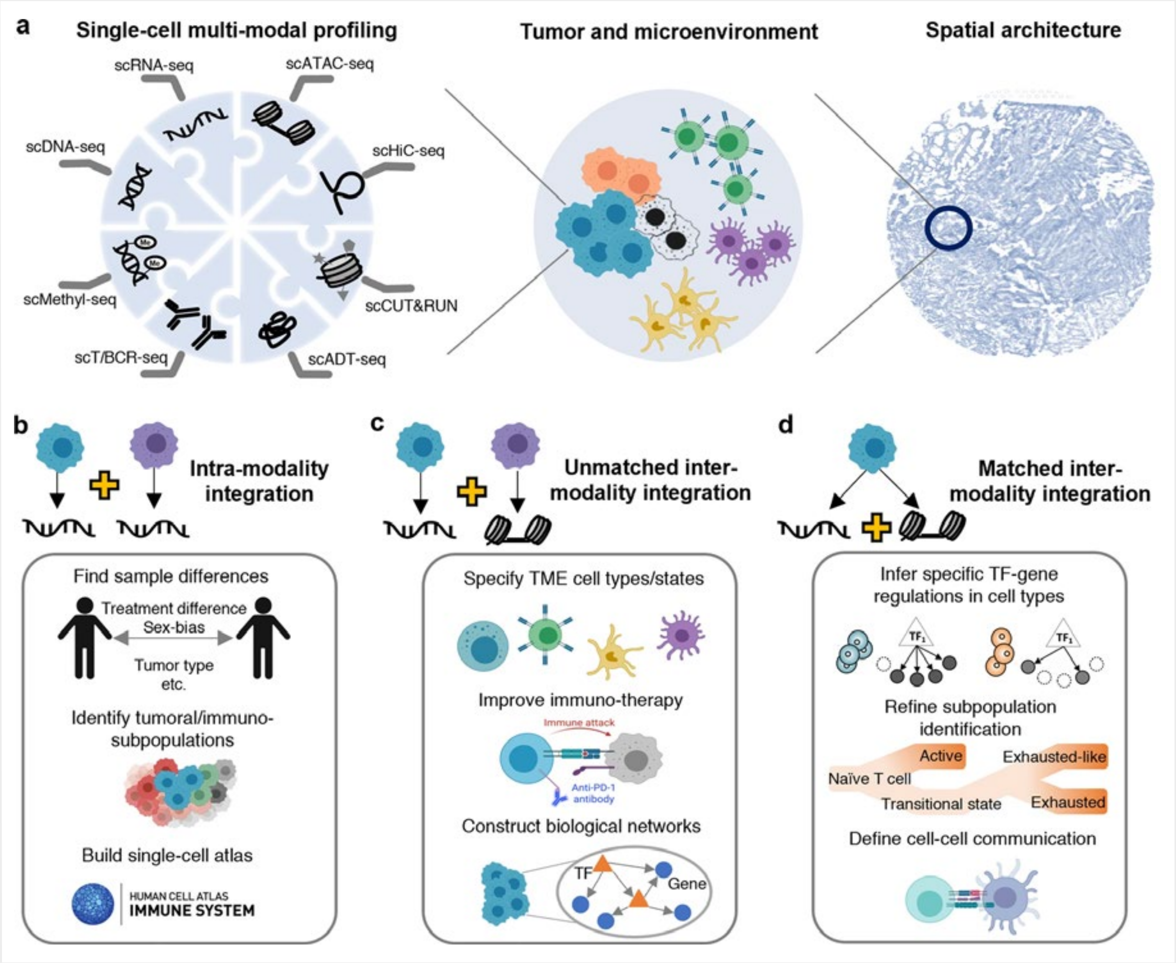

Study population
HELIX project
n=1,301 mother-child pairs, mean age=8 yrs old
6 European cities in
France, Greece, Lithuania, Norway, Spain and UK
Exposome
Mobile devices & sensors
Biomonitoring
Questionnaires
GIS modelling
Meteorological
Natural spaces
Air pollution
Built environment
Traffic
Road traffic noise
Water DBPs
Indoor air pollution
Tobacco smoke
Lifestyle
Social and economic capital
Essential minerals
Non-essential metals
Organochlorines (OCs)
Organophosphate (OP) pesticides
Polybrominated diphenylethers (PBDEs)
Perfluoroalkyl substances (PFAS)
Phenols
Phthalates
Pregnancy (91 exposures)
Childhood (116 exposures)
Omics
Blood samples
Urine samples
Blood DNA methylation
(Illumina 450K, N=386,518 CpGs)
Blood miRNA expression
(Agilent array, N=1,117 miRNAs)
Blood gene expression
(Affymetrix HTAv2, N=58,254 transcript clusters)
Plasma proteins
(Luminex assays, N=36 proteins)
Serum metabolites
(AbsoluteIDQ p180, N=177 metabolites)
Urine metabolites
(¹H NMR, N=44 metabolites)
Exposome-omics-Wide Association Study (ExWAS)
> 30 M tested exposure-omics associations
Web catalogue https://helixomics.isglobal.org/
1,170 exposure-omics associations
Robustness
Ancestry
Body mass index
Cohort
Multi-omics exposome network
Pregnancy
Childhood
Biological interpretation
Literature overlap
Dietary sources
Functional enrichment
Cross-omics

CANCER

We'll begin by looking at some of the latest work using multi-omics to aid our understanding of the tumour microenvironment (TME) (See Figure 7.2)[8].

FIGURE 7.2. OVERVIEW OF THE USES OF SINGLE-CELL MULTI-OMICS DATA.

*(**A**) shows the different data types that could be integrated to investigate the tumour and the microenvironment (**B-D**) highlight the different uses of multi-omics data integration for immune-oncology research. Image Credit: Ma, et al.* [8]


a
Single-cell multi-modal profiling
scRNA-seq
scATAC-seq
scDNA-seq
scHiC-seq
Me
Me
scMethyl-seq
scCUT&RUN
scT/BCR-seq
scADT-seq
Tumor and microenvironment
Spatial architecture
b
Intra-modality integration
Find sample differences
Treatment difference
Sex-bias
Tumor type
etc.
Identify tumoral/immuno-subpopulations
Build single-cell atlas
HUMAN CELL ATLAS
IMMUNE SYSTEM
c
Unmatched inter-modality integration
Specify TME cell types/states
Improve immuno-therapy
Immune attack
Anti-PD-1 antibody
Construct biological networks
TF
Gene
d
Matched inter-modality integration
Infer specific TF-gene regulations in cell types
TF₁
TF₁
Refine subpopulation identification
Active
Exhausted-like
Naïve T cell
Transitional state
Exhausted
Define cell-cell communication

By profiling tumours with multi-omics, researchers open doors to effectively stratify patients by risk, to identify molecularly targeted personalized treatments and to effectively monitor treatment response and tumour evolution[9]. See this review[9] for an in-depth look at the role of multi-omics in precision oncology.

Research progress is fast and this year we have seen clear multi-omics advances in cancer subtyping[10], identifying metastasis[11,12], liquid biopsy[13] and in the identification of potential therapeutic targets[14]. Furthermore, spatial multi-omics is becoming instrumental in truly deconvoluting the tumour microenvironment[15], a topic reviewed in-depth here[16]. This ultimately is set to improve cancer prognosis, diagnosis and treatment[17].

We spoke to some of our experts about their use of multi-omics methods for cancer research and the exploration of immune/cancer cell-cell communication.

MIRJANA EFREMOVA

Group Leader

Barts Cancer Institute

FLG: How are you using multi-omics methods in your work on cancer metastasis and therapy resistance?

Mirjana: *In our research, we employ multi-omics methods to profile colorectal cancer primary and metastatic samples, obtaining both gene expression and chromatin accessibility from the same cell. This enables us to characterize the phenotypically heterogeneous cancer cell states, as well as infer gene regulatory networks and identify specific cis-regulatory interactions that drive these states. We are hoping that this will lead to potential therapeutic targets that could be used to impair metastasis.*

FLG: Cell-cell communication is something you've worked on a lot, how will multi-omics measurements (as opposed to just transcriptomics) help with studying this and what are your hopes for multi-omics in this area?

Mirjana: *Integration of spatial transcriptomics and scRNA-seq data helps us identify cellular neighbourhoods or niches of cells that colocalise together, so that we can focus our cell-cell communication analysis on those cells that are in close proximity in the tissue. In addition, joint gene expression and chromatin accessibility enables us to expand our prediction of enriched ligand-receptor expression across cell types to also predicting which ligands would activate downstream signalling in the responsive cells, providing us with a filtered list of putative "active" ligands on which to focus our further in vitro validation analysis.*

BINGJIE ZHANG

Postdoctoral Research Fellow, Satija Lab

New York Genome Center

FLG: Your methods have helped find biomarkers for recovery from severe diseases. Could you perhaps speak a bit more broadly about the promise of multi-omics for precision medicine?

Bingjie: *I do believe that single-cell multi-omics technologies represent a key advancement for personalized medicine. I can't think of a better way to gain a more comprehensive understanding of what is happening in vivo. Imagine a scenario where, if I were to get sick, clinicians could analyse a blood sample or surgical specimen to decipher the complexities of my condition. This would involve identifying any genetic mutations, pinpointing disrupted epigenetic regulation, and even predicting the most effective therapeutic interventions for my case. It is this level of tailored healthcare that I believe we are advancing towards.*

FLG: What are your hopes for single-cell multi-omics technology for exploring cell interaction networks in the immune system?

Bingjie: *For the immune system, cell-cell communication is vitally important. Several computational methods, such as CellPhoneDB, CellChat, and iTALK, are already very popular. More recently, experimental methods like SPEAC-seq and LIPSTIC have been developed to directly measure cell-cell interactions. However, both require genetic engineering, which poses challenges for studying human primary cells. In the future, I would really like to see computational researchers utilize multi-omics data, particularly spatial data, to make use of physical distances for inferring cell-cell communication. Also, it would be great if we could develop more experimental methods that can be targeted for in vivo cells that work without the necessity for genetic engineering.*

biomodal

CASE STUDY: MULTIOMIC DATA ALLOWS READING OF MODIFIED CYTOSINE BASES AND SIMULTANEOUS MEASUREMENT OF GENOMIC MUTATIONS IN CANCER CELLS

RESEARCHERS IN DR SAM APARICIO'S GROUP AT THE BRITISH COLUMBIA CANCER RESEARCH CENTRE (BCCRC) AND THE UNIVERSITY OF BRITISH COLUMBIA, BC, CANADA, UTILISED 5-LETTER SEQUENCING TECHNOLOGY, DUET MULTIOMICS SOLUTION +MODC, TO INVESTIGATE 'EPIGENETIC REWIRING' IN BREAST CANCER CELLS.

In this study, duet multiomics solution +modC helped to reveal:

- the epigenetic landscape of untransformed diploid breast epithelial cells with wild-type, p53-/-BRCA1-/- and p53-/-BRCA2-/- genetic backgrounds
- significant activation of stem cell enhancers through reduced DNA methylation in p53-/-BRCA1-/- cells only
- similar activation of stem cell enhancers in a triple negative breast cancer (TNBC) patient xenograft sample
- the epigenetic rewiring caused by BRCA1-/-, identifying it as a crucial gene for this type of cancer pathogenesis

Challenge

In this case study, we highlight the Aparicio group's research on decoding the relationship between genomic mutational background and epigenomic, or nongenomic, transcriptomic contributions to the fitness of cancer cells. This research study measures both the state of the genome in cancer cells, as well as decodes the state of the epigenome and the transcriptome. Making it vital to capture the epigenetic information encoded in modified cytosine bases in DNA, as a component of their investigations.

Solution

OVERCOMING CHALLENGES TO PERFORM SIMULTANEOUS GENOMIC AND EPIGENETIC SEQUENCING OF BREAST CANCER CELLS

To overcome previous challenges associated with the utilisation of low-sample volumes and multiple workflows, Dr Gurdeep Singh, postdoctoral fellow, at the Aparicio lab employed **duet multiomics solution +modC** to simultaneously investigate the genetic and epigenetic landscape of breast cancer cells known to exhibit homologous recombination deficiency (HRD) which drives genomic instability and cancer pathogenesis. This cancer cell trait or driving mechanism is involved in both triple-negative breast cancer (TNBC) and high-grade serous ovarian cancer.

duet multiomics solution +modC enabled Dr Singh to investigate epigenetic rewiring - known to play a critical role in cancer pathogenesis, cancer advancement, and cancer drug resistance - in breast tumour cells.

HRD mutational instability is known to be dependent on BRCA and p53 mutations. In this study, Dr Singh used untransformed diploid 184hTERT breast epithelial cell lines deficient in genes often mutated in HRD cancer (p53−/−BRCA2−/− and p53−/−BRCA1−/−), with wild-type genomic background (WT184hTERT) cells as a control.

> "Having a single-workflow method that allows reading of modified cytosine bases and simultaneous measurement of genomic mutations is a game-changer for us."
>
> **Dr Sam Aparicio**

About BCCRC

BC CANCER RESEARCH
Provincial Health Services Authority

The BC Cancer Research Centre's mission is to pursue world-class research that aims to transform the lives of patients by exploring basic mechanisms and technology developments in all areas of cancer research including cancer control, clinical studies and trials, cancer surveillance, and population health and services. The research portfolio also supports facilities and platforms in genomics, bioinformatics, imaging, drug development, and tissue banking.

The Aparicio group studies the genomic and phenotypic behaviour of breast and other cancers. They integrate leading technologies to support their efforts to better understand how cancer clones evolve and to identify novel strategies for cancer treatment and predictors of response.

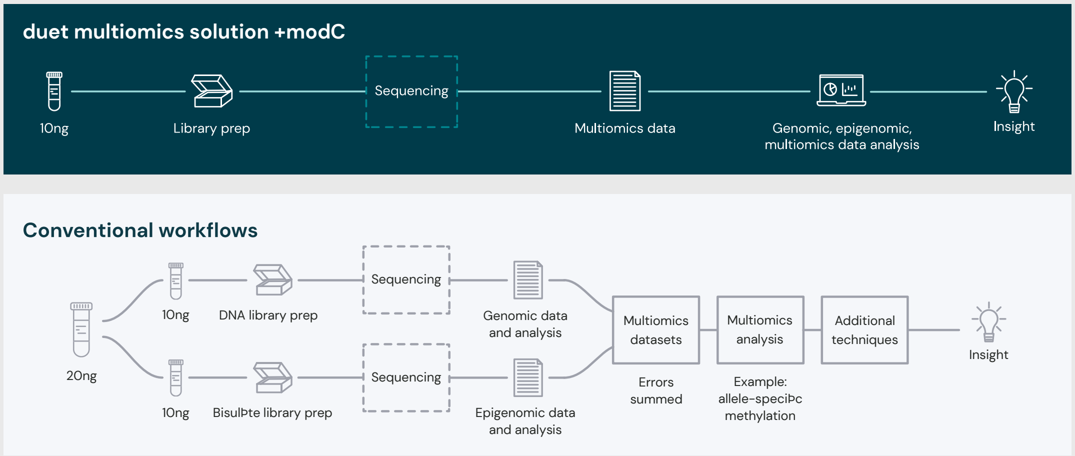


Figure 1. The duet multiomics solution +modC single workflow vs conventional genomic and epigenomic sequencing workflows. (Top panel) Following the pre-sequencing workflow, and device-agnostic sequencing, the post-sequencing bioinformatics pipeline aligns epigenetic and genomic sequencing data for analysis, interrogation, and insight.

(Bottom panel) Conventional epigenomic and genomic sequencing methods require multiple workflows, are more prone to errors, require more DNA sample (20ng), and multiple datasets to gain insights.

Method

DECODING EPIGENETIC REWIRING USING DUET MULTIOMICS SOLUTION +MODC

Firstly, DNA methylation using long-read sequencing was used to compare and confirm the DNA methylation landscape seen with **duet multiomics solution +modC** for WT184hTERT. The resulting data revealed strong Pearson correlation.

The next step was to interrogate the genomic and epigenetic landscape of all the cell types using **duet multiomics solution +modC.** The single workflow approach enabled researchers to glean greater insights from small amounts of sample DNA (Fig1).

Interestingly, using **duet multiomics solution +modC**, revealed that only the p53-/-BRCA1-/- 184hTERT, and not the untransformed diploid 184hTERT breast epithelial cell lines (WT184hTERT) or the p53-/-BRCA2-/- 184hTERT, showed significant activation of stem cell enhancers through reduced DNA methylation, and hence cancer-associated epigenetic reprogramming.

In a second step, Dr Singh used **duet multiomics solution +modC** on a reference TNBC patient-derived xenograft (PDX) sample, which also showed significant activation of stem cell enhancers through DNA methylation changes.

Results

EMPOWERING GAME-CHANGING RESEARCH IN A SINGLE WORKFLOW

In this study, the Aparacio lab used duet multiomics solution +modC to analyse in vitro breast cancer cell lines, then compared these findings to cells from a patient biopsy. They found strong correlation and alignment from the resultant comparative genomic and epigenetic data and were able to inform their research on cancer progression in triple-negative breast cancer. The findings illustrate that BRCA1-/- is crucial for HRD-specific cancer pathogenesis, where it also drives genomic instability signatures, and while BRCA2-/- drives genomic instability, it alone may not be able to drive the necessary epigenetic rewiring for cancer progression.

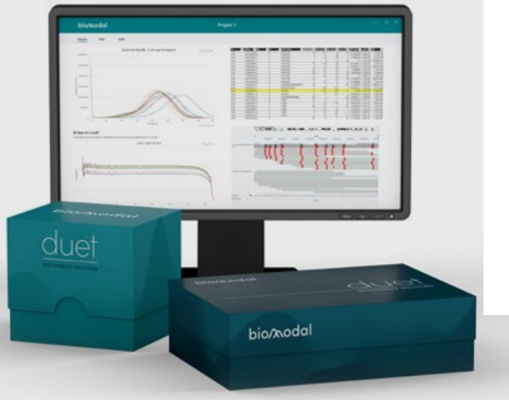
Researcher Spotlight

Dr Sam Aparicio, BM, BCh, PhD, FRCPath, FRSC

Dr Samuel Aparicio is the Nan & Lorraine Robertson Chair in Breast Cancer Research, holds the Canada Research Chair (Tier 1) in Molecular Oncology, and is the recipient of the 2014 Aubrey J Tingle Prize. He is also Head of the Department of Breast and Molecular Oncology at BC Cancer Research, part of the Provincial Health Services Authority, and a Professor in the Department of Pathology and Laboratory Medicine at UBC.

Dr Gurdeep Singh, PhD

Dr Gurdeep Singh is Post-Doc in Dr Samuel Aparicio's lab at BC Cancer, decoding the epigenetic basis of cancer pathogenesis and drug-resistance using CpG methylation & epigenomic landscape, and defining/testing the responsible transcriptional regulators. Dr Singh received his PhD in 2021 from The University of Toronto where he identified the genome sequence code that confers enhancer activity in embryonic stem cells, and other tissues, using functional genomics experiments and computational approaches.



JUDITH ZAUGG

Group Leader

European Molecular Biology Laboratory (EMBL)

FLG: Could you just talk a little bit about how your approaches can be applied to precision medicine?

Judith: *Sure, I can give you some examples. We have been using these multi-omics approaches to investigate relapse versus remission, for example, upon allogeneic stem cell transplantation. And we have identified specific surface markers based on analysing loads of different angles and using gene regulatory networks, etc. There turns out to be an important marker gene on one of the T cell populations that is quite predictive of patient relapse. Now, this is not mechanistic, this is a biomarker, but of course, it's maybe useful to stratify population to assess risk, and so on.*

I see my contribution to precision medicine as identifying these types of mechanisms that we can try to translate into precision medicine approaches. What we like to do is integrate genetic variants that are giving you a causal link, because these are genetic associations where there is no doubt about the causality direction, what remains to be understood is what's the mechanism of the causality. We then use the more descriptive gene regulatory networks to get the disease angle; i.e., what actually happens in disease in terms of differential expression, or differential accessibility. So, combining these two layers [RNA and ATAC], is a very powerful approach in our lab to understand mechanisms.

COVID-19

Multi-omics experiments have recently revealed new insights into another complex disease: COVID-19. Understanding the immune system's response to infection and the biological mechanisms for the different patient outcomes is paramount.

Examples of such approaches include a study using single-cell transcriptomics and proteomics, which revealed important biomarkers such as immune subpopulations and responses liked to COVID-19 pathogenesis[18]. Cytokines, metabolomics and proteomics revealed biomarkers of long COVID[19] and anti-inflammatory patterns that could be used for diagnosis.

One very recent example[20] used a combination of proteomics, transcriptomics and chromatin accessibility to identify vaccine-induced T cell populations and defined them multi-omically as potential treatment targets. We spoke to first author, **Dr. Bingjie Zhang** about her multi-omics study of COVID-19 vaccine response on PBMC.

BINGJIE ZHANG

Postdoctoral Research Fellow, Satija Lab
New York Genome Center

FLG: I wanted to ask you about your recent Nature Immunology paper, looking at COVID-19 Vaccine response with multi-omics. Could you give a brief overview of the multi-omics work you performed on PBMCs in that paper, focus on the contributions that CITE-Seq and ASAP-Seq had and the benefits of using proteomics, RNA, and chromatin all in one study?

Bingjie: *It's a very good example that demonstrates the application of single-cell multi-omics. The CITE-seq experiment in the project was used to identify antigen-specific T cells. In the vaccination study, we are particularly interested in the T cell response and aim to characterize the antigen-specific T cells responsive to vaccination. It's actually quite difficult to capture these rare cell populations in peripheral blood. Only with the assistance of cell surface protein data were we able to identify this small group of T cells that were induced by vaccination, and we further proved that they are specific to the SARS-CoV-2 antigen using DNA-oligo-tagged peptide-MHC class I multimers. For the ASAP-seq, we can use bridge integration to identify the same antigen-specific T cell population as identified in the CITE-seq study.*

The beauty of a multimodal dataset is that we can use one modality for computational predictions and another to validate whether those predictions are correct. In our case, we utilized ATAC data in the bridge integration and confirmed that the predicted antigen-specific T cells indeed had the expected protein markers. Then, we can use the chromatin accessibility data to identify enhancers specific to the antigen-specific T cells, and also the potential regulators for the induction and maintenance of these T cells.

Finally, given the long-term nature of COVID-19 infection, longitudinal approaches are necessary to truly catch molecular patterns for diagnosis. A recent example of a multimodal longitudinal study[21] took biological samples at timepoints across patients of different severities. Through GWAS, proteomics, transcriptomics and metabolomics, the molecular states of different patient groups were identified, along with the distinct temporal changes linked to different disease outcomes.

It is this type of large-scale multi-omics effort that will really get to the heart of complex disease. We recently spoke with **Suhas Vasaikar**, PhD MBA, Principal Scientist at Seagen, about PALMO[22], a platform that enables longitudinal multi-omics insights for just such investigations.

INTERVIEW: SUHAS VASAIKAR

PRINCIPAL SCIENTIST, CLINICAL BIOMARKER AND DIAGNOSTICS

SEATTLE GENETICS (SEAGEN)

FLG: ***Briefly introduce yourself, with some of your research background and a summary of your current role.***

Suhas: I am a researcher in the field of computational biology and bioinformatics. I completed my PhD in neuronal disorders at the Indian Institute of Technology Delhi. During my PhD, I developed a novel algorithm (SSG, JCN 2014) to analyse large-scale genomic and proteomic network Boolean data.

Over the past few years, I have focused on cancer, especially in the area of multi-omics where multiple forms of omics data were collectively used to infer the mechanism of cancer growth and progression. One such example from our proteogenomic approach provided crucial insights into the pathophysiology of colon cancer (Cell 2019), which were not attainable solely through earlier genomic and RNA expression data. This approach has also helped to identify potential novel therapeutic strategies for the treatment of colon cancer.

Recently I, along with the team from the Allen Institute for Immunology, have developed a tool that provides a way to explore longitudinal multi-omics data at single-cell level in different disease indications. The major goal was to provide a comprehensive yet simple-to-use software tool to extract insightful information from longitudinal omics data, which was not sufficiently addressed before.

Currently, I am a Principal Scientist at Seattle Genetics, with a focus on clinical biomarkers and diagnostics. At Seagen I am using my expertise in the multi-omics field to understand the cancer drug treatment data and provide insightful details to support ADCs, which can improve patient lives.

> “ONE SUCH EXAMPLE FROM OUR PROTEOGENOMIC APPROACH PROVIDED CRUCIAL INSIGHTS INTO THE PATHOPHYSIOLOGY OF COLON CANCER."

FLG: ***Can you describe your PALMO platform? What omics can it measure longitudinally and what kind of insights can be found using PALMO on multi-omics longitudinal data?***

Suhas: PALMO stands for Platform for Analysing Longitudinal Multi-omics data. Longitudinal multi-omics data refers to the collection of multiple types of biological data over time, such as genomics, transcriptomics, proteomics and metabolomics. This data can provide a more comprehensive understanding of biological processes and disease progression than single-omics data.

PALMO is a platform (https://github.com/aifimmunology/PALMO) that contains five analytical modules to examine longitudinal bulk and single-cell multi-omics data from multiple perspectives. These include decomposition of sources of variations within the data, collection of stable or variable features across timepoints and participants, identification of up- or down-regulated markers across timepoints of individual participants, and investigation of samples from same participants for possible outlier events (Nat Comm 2023).

Analysing longitudinal multi-omics data can provide insights into how various biological processes change over time, how they are affected by environmental factors, and how they relate to disease development and progression. For example, using PALMO we were able to show that longitudinal multi-omics analysis can help us to identify stable/variable features across PBMC immune cell types, changes in blood plasma over time, stable across time in cell-types (STATIC) genes from mouse brain tissue and heterogenous immune responses among COVID-19 patients.

The ultimate goal is to identify early warning signs of disease, track disease progression and predict treatment response. By integrating multiple types of data, researchers can gain a more holistic understanding of biological processes and identify new targets for therapeutic intervention. The insights gained from longitudinal multi-omics data can have far-reaching implications for personalized medicine and precision health.

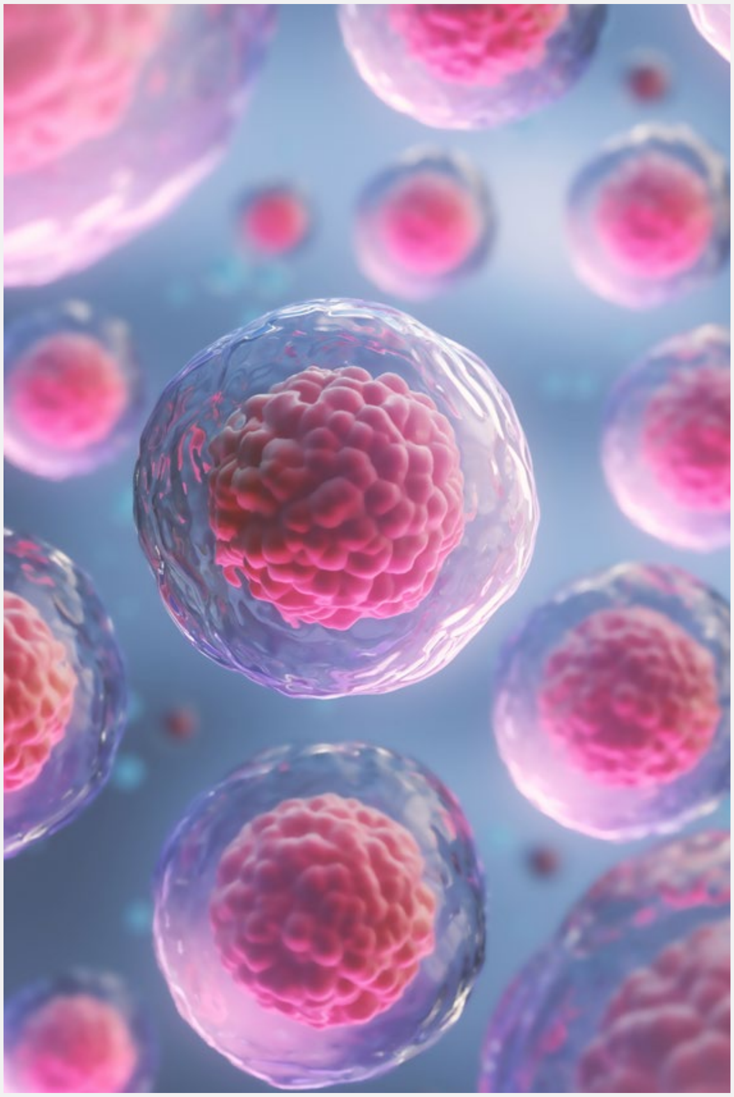
***FLG:* Can you discuss the value of applying multi-omics methods to clinical problems in your career, such as COVID-19 and cancer?**

Suhas: Multi-omics methods have significant potential for advancing our understanding of complex diseases such as COVID-19 and cancer. In the context of COVID-19, multi-omics approaches can help to identify key molecular pathways and biological processes that are involved in the disease progression and severity.

Our study used single-cell analyses of peripheral blood cells, serum proteomics, virus-specific cellular and humoral immune responses and clinical metadata to characterize signals associated with recovery and convalescence to define and validate a new signature of inflammatory cytokines, gene expression and chromatin accessibility that persists in individuals with post-acute sequelae of SARS-CoV-2 infection (PASC). This is great example to depict how multi-omics data can help us to identify the genetic and molecular factors that contribute to the severity of COVID-19 symptoms, which can in turn inform the development of targeted therapies (Nat Comm, 2023).

Similarly, multi-omics approaches can also be applied to cancer research to identify key genetic and molecular factors that contribute to the development and progression of cancer. By integrating data from multiple omics platforms, we were able to identify novel pan-cancer survival associated signatures and potential drug targets in 12 cancer types (http://www.linkedomics.org, NAR 2018) that can be used for personalized cancer treatment.

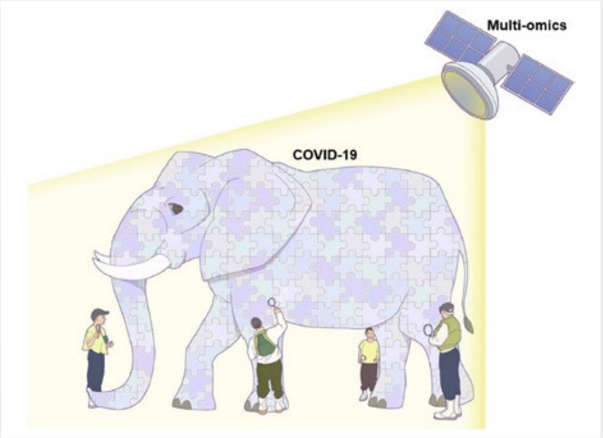
In short, in both COVID-19 and cancer research, multi-omics methods can provide a more comprehensive understanding of the disease mechanisms and enable the development of more effective diagnostic and therapeutic strategies. Additionally, these methods can also help to identify key molecular signatures and pathways that can be used for disease prognosis and monitoring of treatment response.



Ultimately, multi-omics presents an approach that can reveal the underlying molecular structure of the pathogen and molecular host response to the virus and vaccines, revealing the whole elephant rather than a trunk or a leg (see Figure 7.3)[23]. For deep reviews on this topic, please refer to the following references[24-26].

FIGURE 7.3. OBSERVING ONLY PART OF THE ELEPHANT ANALOGY FOR MULTI-OMICS.

Image Credit: Lu, et al. [23]



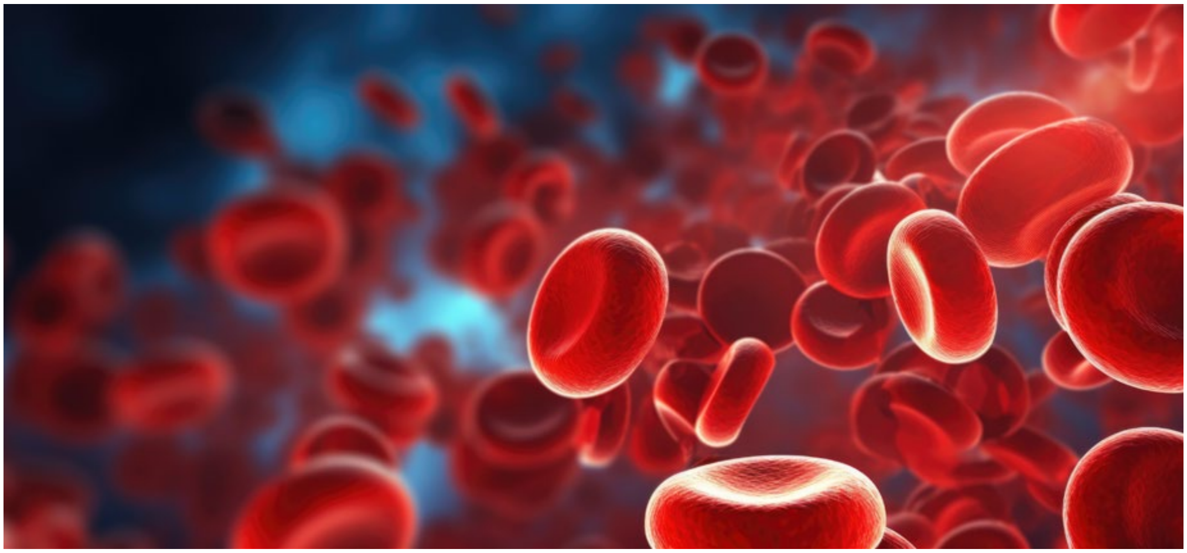
Using multi-omics and advanced computation to improve diagnosis

Multi-omics presents a unique strategy for early diagnosis and biomarker detection.

Bringing multi-omics to the clinic in real diagnostic situations is a challenge due to costs and feasibility. However, a very recent nation-wide implementation of multi-omics in Australia demonstrated the benefit of investing in such approaches[27]. Published in June 2023[27], the Acute Care Genomics program has integrated ultra-rapid whole-genome sequencing with other omics to improve diagnosis of critically ill children with rare disease and delivered these benefits on a national level. While genomics alone improved diagnosis, the incorporation of functional other omics, such as proteomics, improved diagnostic rates and outcomes.

One avenue to improve diagnostics is to increase the spread of markers. Another avenue is to find cost-effective ways to scale down a multi-omics assay into the most valuable mono-omic diagnostic test. Recent studies have shown that conditions such as preeclampsia[28] and preterm birth[29], when profiled with a multi-omics approach, can be reduced to a diagnosis involving a simple urine metabolite assessment. This has major implications for diagnosing these conditions in low and middle-income countries[30].

We recently spoke to **Professor Nima Aghaeepour** at Stanford University about his research developing computational approaches for multi-omics and personalized medicine, and his work using multi-omics to find scalable and cost-effective diagnostic markers for the conditions highlighted above.



INTERVIEW:

NIMA AGHAEEPOUR

ASSOCIATE PROFESSOR, ANESTHESIOLOGY, PERIOPERATIVE AND PAIN MEDICINE & PEDIATRICS - NEONATAL AND DEVELOPMENTAL MEDICINE STANFORD UNIVERSITY

***FLG:* Can you briefly introduce yourself, your research background and a summary of what your lab has been up to?**

Nima: I'm Nima Aghaeepour, I'm an Associate Professor at Stanford University School of Medicine. My background is in machine learning and artificial intelligence. Our laboratory works on a broad range of medical problems using various technologies, anything from genomics, proteomics, metabolomics, single-cell assays, to wearable devices to electronic health records. Currently, the topic that I'm particularly excited about is the intersection of electronic health records and biological modalities.

***FLG:* Could you give a very brief overview of how multi-omics technology has developed and been integrated into health monitoring and for precision medicine?**

Nima: Initially, 20 years ago, we were at a place where we were doing epidemiology in one universe. And, in a different universe, we were doing very limited biological studies, with really limited assays that we had. Gradually, these assays became capable of measuring more, and at the same time, our computing power increased, and we were able to handle more. These two advances need to improve with each other. At the time, when we started sequencing the human genome, the assays started outgrowing the computational power, and for the first time we had more data than we knew what to do with.

Then, after all those genomics projects, we quickly learned that biology doesn't stop at the genome, you need to go to higher levels. So, we started developing assays for proteomics, metabolomics, the microbiome and so on. Each of them is their own field and requires their own specialised computation. Then we realised that biology doesn't happen at one layer at a time, you need to study all of these layers simultaneously. So, to study all of these layers simultaneously, not only do you need a lot of computational power, but you also need people who are simultaneously experts in all of them and can use specialised computational pipelines for each. Then you need to build a layer on top of that, that can put everything together.

Now we are getting to a place where we are understanding that even that is not enough, because humans don't live in a petri dish. Humans are interacting with their environment on a regular basis. You can't capture that with self-reported questionnaires, you need real-time wearable devices for that. You can't put people into categories of cases and controls, everything is nuanced. Everything depends on what happened to you 10 years ago, what medications you're taking, how you interact with the healthcare system, how many surgeries you have had and how many times you've been pregnant. We need to go beyond that and connect all this sophisticated biology to real time environmental factors, wearable devices and health records and so on.

***FLG:* Multi-omics tends to refer to work on two omics, maybe RNA and epigenomics or RNA and protein. For some of your work it's been an integration of RNA, metabolism, lipidome, microbiome plus all this patient data. Integration is a challenge for the field but is it a new problem entirely when you're trying to integrate all this data together to produce clinical insights?**

Nima: If you are looking at RNA and proteins, it's not that sophisticated, because you still have your genes, and you can focus on the gene level and go up and down.

That's fairly straightforward. But once you start adding the microbiome, once you start adding metabolites to that, things get a lot more complicated. It's no longer measuring the same genes up and down at various layers of biology, you're in a completely different universe. Once you start adding an imaging modality, or a wearable device or an electronic health record, then really all bets are off, because you're not just looking at different universes of measurement, you're also looking at different times. You're looking at databases that have information on your subjects from the last 10 years. Staying in the genomics-adjacent modalities is fairly straightforward. After that, it gets pretty complex.

***FLG:* And is this why you're using machine learning based methodologies? Could you discuss how you have been leveraging machine learning to handle that issue?**

Nima: Yes, so we think multi-omics integration needs to happen in an interdisciplinary and collaborative way, you can't come up with one-size fits all algorithms that can do multi-omics integration for all problems at all times.

If you're trying to understand biological mechanisms, there is a way to do that across various modalities by looking at shared latent spaces, correlations across different modalities, directional correlations, some people call them causal relationships. And they will help you find biological mechanisms.

If you're trying to predict outcomes, you need specialised pipelines that can leverage each omics dataset correctly, according to the state of the art in that field. And then you need higher level models that can pull from those lower-level models to make the final predictions.

If you're trying to use multi-omics data to build a low-cost assay, companion diagnostics are something that need to be deployed at scale. Then you have a cost factor, because all of these omics assays have a different cost, and you need to teach your machine learning algorithm that so that it doesn't just reduce it to a small model that still needs your most expensive omics. You need it to find a surrogate model, using your cheaper assay so that you can scale it up and commercialise it.

So, there is not going to be a magic silver bullet that does everything for everybody. It needs to be interdisciplinary; the machine learning scientists need to talk to the biologist and to the clinicians to understand the nuances of the problem and deploy a pipeline that makes sense for that.

***FLG:* Do you have examples from your work of bringing together multiple different omics to help produce new clinical diagnostics?**

Nima: For example, in our article that was published a couple of years ago, in JAMA Network Open, we showed that we can build an assay for prediction of preterm birth in the first trimester of pregnancy. Preterm birth is the single largest cause of death of children under five years of age. In that paper, we first started with a multi-omics assay but then we reduced that to a simpler assay. We can use urine metabolites to build a surrogate for proteins in plasma, that enables us to make significantly cheaper assays that can be scaled up in low- and middle-income countries.

Similarly, in a paper in Patterns last year, we showed that we can use a multi-omics approach to build a model for preeclampsia. But then, again, we were able to reduce it to a urine-based metabolomics assay that can predict preeclampsia, almost as accurate as a full multi-omics assay. If you first start at the multi-omics level, and you figure out what it is that you need to project a clinical outcome, then you can reduce it to a cheap urine-based assay that can be deployed at scale.

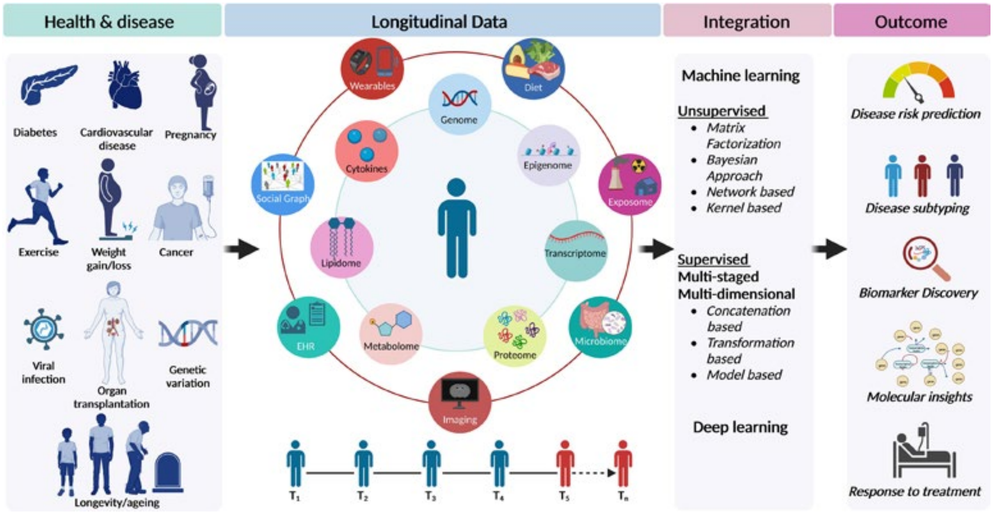
***FLG:* I also wanted to ask you about your CORALS tool for multi-omics data, would you mind introducing that for me?**

Nima: As we measure more and more features through different biological assays and combine them with each other, these datasets become pretty large. And it becomes really difficult to build large correlation networks that can measure correlations within and across omics assays. CORALS reduces the amount of memory and time that is required to build these correlation networks. It also provides an approach for measuring the difference between two different correlation networks. If you have a large correlation network of multi-omics data in healthy individuals, and you have a similar one in sick individuals, you can subtract one from the other, to see what new correlations are showing up and what correlations are disappearing when somebody is sick.

Building large correlation networks is a foundational task that is used in many different types of analysis, including now with graph neural networks and other modern deep learning approaches. So, the applications are broad. It depends on the individual users to decide why they need a correlation network.

FIGURE 7.4. LONGITUDINAL MULTI-OMICS DATA AND WEARABLE DATA ENABLE DEEP PHENOTYPING FOR PRECISION MEDICINE

Image Credit: Babu and Snyder [31]



Wearables, micro sampling & health monitoring

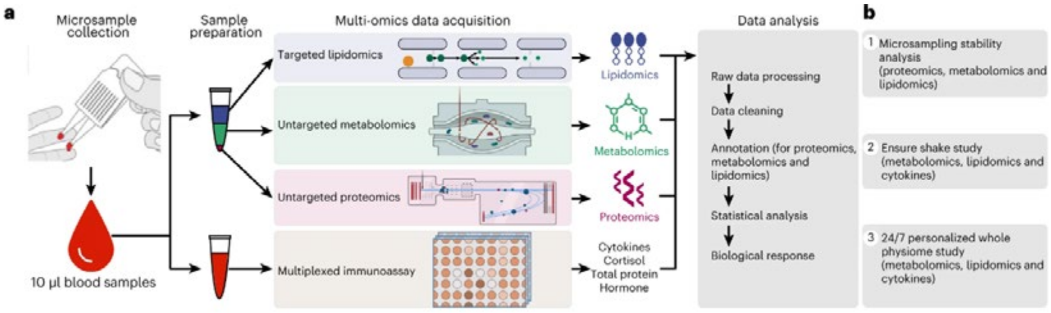
An exciting area of multi-omics for personalized medicine is the current approach to profiling thousands of different analytes from individuals. Even more impressive are the efforts to profile these metabolites from micro-samples, combined with wearables to provide near constant health monitoring[31].

By incorporating thousands of different omics measurements with detailed health data from electronic records and the longitudinal physiological monitoring from wearable devices, we can create a holistic phenotypic profile of an individual for faster and more effective diagnosis[32,33] (See Figure 7.4).

We recently spoke to Dr. Xiaotao Shen, a postdoctoral fellow in the Snyder Lab at Stanford University. Xiaotao was recently first author on the 2023 study[34] profiling thousands of multi-omic analytes from micro-samples of blood, collected by the participants in the comfort of their own home (see Figure 7.5). This approach, combined with wearables, provides a future of dynamic health profiling which would ultimately allow diagnosis at the earliest possible instance.

FIGURE 7.5. MULTI-OMICS DATA ACQUIRED FROM THE MICROSAMPLE FOLLOWED BY THE OUTLINE OF THE DATA ANALYSIS PIPELINE.

Image Credit: Shen, et al. [34]

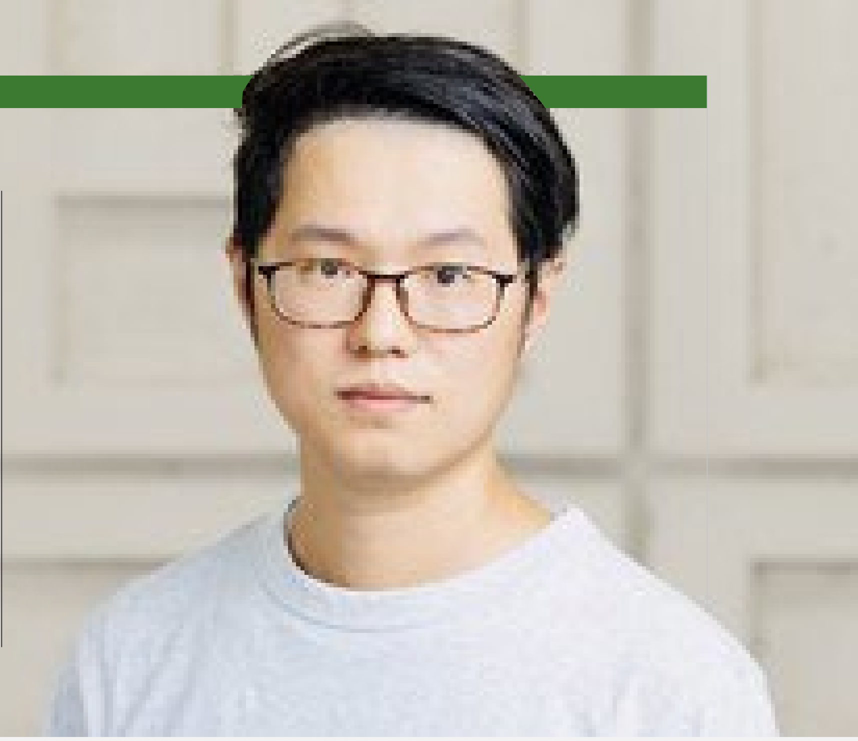

a
Microsample collection
Sample preparation
Multi-omics data acquisition
Data analysis
b
10 µl blood samples
Targeted lipidomics
Untargeted metabolomics
Untargeted proteomics
Multiplexed immunoassay
Lipidomics
Metabolomics
Proteomics
Cytokines
Cortisol
Total protein
Hormone
Raw data processing
Data cleaning
Annotation (for proteomics, metabolomics and lipidomics)
Statistical analysis
Biological response
1 Microsampling stability analysis (proteomics, metabolomics and lipidomics)
2 Ensure shake study (metabolomics, lipidomics and cytokines)
3 24/7 personalized whole physiome study (metabolomics, lipidomics and cytokines)

INTERVIEW:

XIAOTAO SHEN

POSTDOCTORAL RESEARCH FELLOW, SNYDER LAB

STANFORD UNIVERSITY



***FLG:** Can you just start by briefly introducing yourself and give some of your research background and your current research interests and projects?*

Xiaotao: Sure. Thank you so much for having me. My name is Xiaotao Shen and I'm a postdoc in Michael Snyder's lab at the Stanford School of Medicine. My background is in biology and mass spectrometry, mostly in data processing and analysis in mass spectrometry. A mass spectrometer is an instrument that can be used to measure small compounds, peptides and proteins from different biological samples. It's a very powerful instrument.

My background has been to use this instrument to profile biological samples and use this data for biomarker discovery and diagnosis of diseases. I got my PhD from the Chinese Academy of Sciences in China. In my PhD work, I mainly focused on how to use mass spectrometry for biology studies, and I developed methods for mass spectrometry data processing and analysis. After I got my Ph.D., I came to Stanford in 2018 because I wanted to use these methods for precision medicine and biomarker discovery. I came here because Michael Snyder is a big name in the field of precision medicine.

In the lab, we have a lot of people using different methods and different omics data for precision medicine. This means we have different multi-omics data e.g., transcriptomics, proteomics etc. and I have used the different omics data for my projects here. I've focused on two biological diseases. The first one is a pregnancy-related disease - pre-term birth, which is a very serious problem for a lot of pregnant women. The second disease is ageing because ageing is a risk factor for a lot of diseases, e.g., cancer and diabetes. I focus on these two diseases, and the methods I've used in the lab integrate multi-omics data.

I have also worked on the microbiome, its integration and how it impacts human health. There are a lot of studies showing that the microbiome leads to risk for diabetes and risk for cancer. But how does the microbiome affect human health? We know that the gut microbiome can send chemicals to the brain. So, they produce some small compounds or peptides, and these small compounds and peptides can go to the brain and other body sites and affect human health. Ideally, we would want to profile all the metabolites and small compounds produced by their microbiome, but the data analysis for the mass spectrum is very challenging. So, I've been researching how we can integrate the microbiome and the metabolome. We need to construct a network between the microbiome, these compounds and how they impact health.

***FLG:** Could you just describe what multi-omics marker sampling looks like, how does it work? What do you capture? And just a little bit about the Nature Biomedical Engineering study?*

Xiaotao: That paper was just published in Nature Biomedical engineering. This multi-omics micro-sampling is another project I'm working on and I'm still working on this approach, because our paper is a methodology paper, so we just described the method and its potential use for precision medicine and for precision nutrition. But we didn't use this method for diagnostics. So, now we are working on this project, specifically for diagnosis.

Early diagnosis is very important for human health. In the US, people go to hospital once a year, or maybe once every two years, so not at a very high frequency. Or people only go when they have the symptoms of the disease, and then they go to the hospital and have a physical exam. If they're already in the late stage of disease, this is not great for their prognosis and their health. Even though studies have found earlier diagnostic biomarkers for diseases, these are not currently used in hospitals. This is because you need a method for testing that uses an abundant sample. With this in mind, I think blood samples are the most important material for diagnosis. In the hospital, the traditional method for blood collection is intravenous collection. You need a nurse or a doctor to collect these samples and you need to go to the hospital, we can't collect blood samples by ourselves in the home.

We wanted to know if it was possible to use a microsample of the blood, which could be collected by people at home. The microsample is very easy to collect and it's not painful. The first issue is that the microsamples are tiny blood samples, so we wanted to know if we can actually sample a lot of molecules and get enough biological information from the tiny blood sample. So, we used a different method to extract the molecules from the microsamples. Namely, we used mass spectrometry to measure the molecules from the microsamples. Mass spectrometry is a very high sensitivity instrument to measure low abundance markers from samples. And we can use this instrument to measure more than 1000 molecules from the microsamples, including proteins, metabolites and lipids. So, we built this multi-omics microsample platform and we wanted to know whether this platform could be useful for precision medicine or precision nutrition.

So, we performed two case studies in this paper. The first case study showed that we can use this method to measure people's blood response to the Ensure diet shake. For the second study, we collected a microsample almost hourly over seven days. This is a very high sampling frequency. If you use the traditional methods, you can't do that, because you can only produce samples maybe weekly or monthly. By using this measure, it's very easy to sample hourly. I think, in the future, this method can be used for diagnosis. So, people can just collect samples by themselves at home and send the sample to the lab. We can get then get the results, and if you have some known biomarkers for diseases, we will let you know. And you can go to hospital and get the full physical exam, right. I think this is this is a future application of this method.

***FLG:* It's definitely an exciting method. When you're analysing a micro sample, do you lose some of the depth i.e., the number and the different types of the amounts of molecules you can analyse from a micro sample compared to a normally derived intravenous sample?**

Xiaotao: So, in the paper, we compare the data from the microsamples and from the traditionally collected sample. To be honest, the microsample is so small, perhaps only 10 microliters, so we don't see the same amount of information or molecules from the microsamples as the traditionally collected samples. However, considering it is a microsample, we do see nearly all of the molecules that we can collect in data from traditional samples.

Another thing people may be concerned about is whether the molecules/information from the microsample are robust. We also compared the data from microsamples and the data from traditional samples and we found that the correlation between them is very high. We measured the metabolites and the lipids and the correlation between them was almost 0.9 - a very high correlation between them. So, this method can get highly robust data from microsamples.

> "MASS SPECTROMETRY IS A VERY HIGH SENSITIVITY INSTRUMENT TO MEASURE LOW ABUNDANCE MARKERS FROM SAMPLES. AND WE CAN USE THIS INSTRUMENT TO MEASURE MORE THAN 1000 MOLECULES FROM THE MICROSAMPLES, INCLUDING PROTEINS, METABOLITES AND LIPIDS."

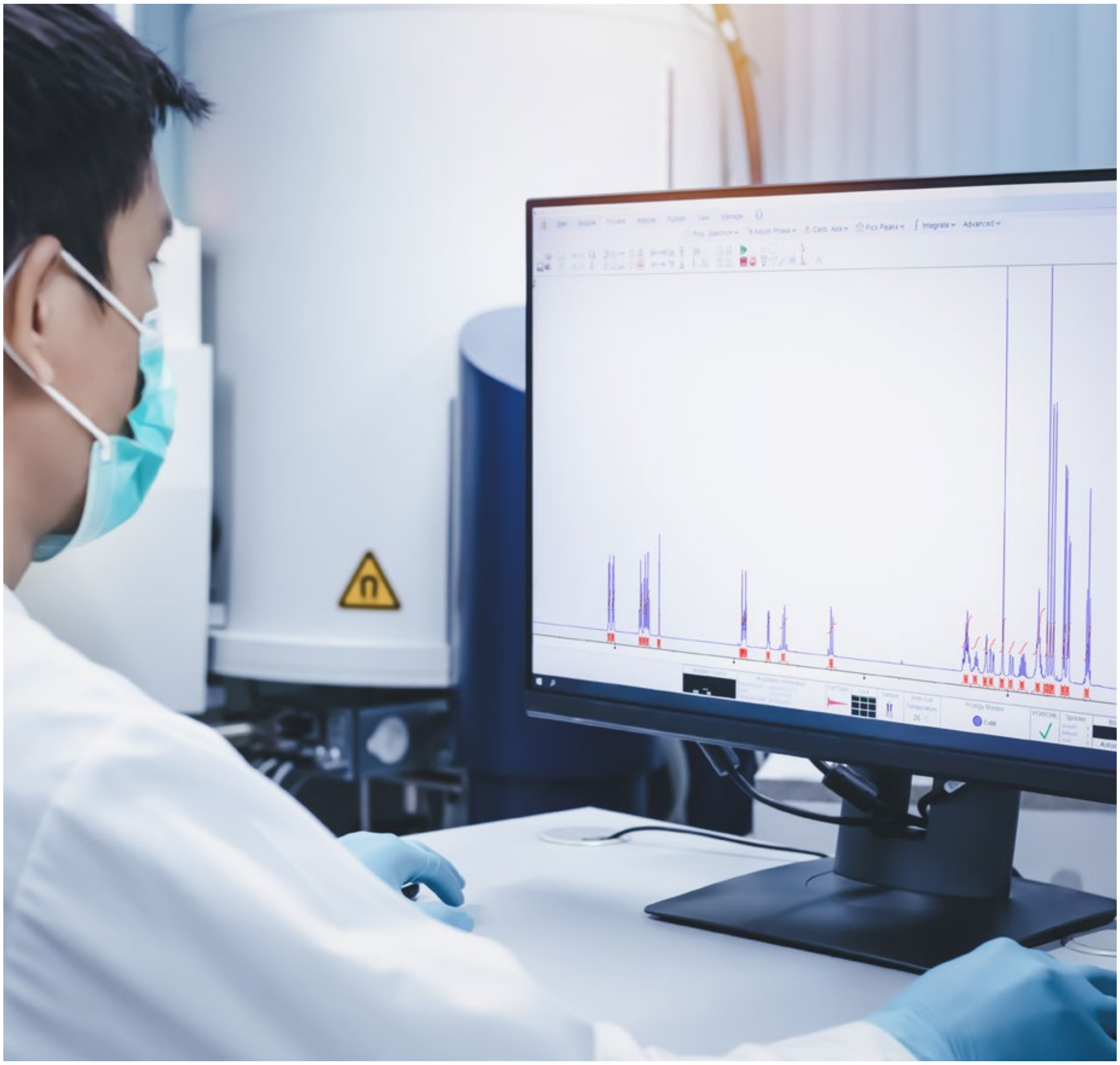
Multi-omics in drug discovery

While we can use multi-omics to better understand disease, identify biomarkers and better diagnose patients, a crucial application of this holistic approach is for drug and treatment discovery.

Drug discovery is a challenging task and traditional methods often fall short. Single-target drugs that modulate the activity of a specific gene or protein may not be effective in the diseases discussed above due to complex molecular pathways and heterogenous disease presentation.

Multi-omics presents the natural alternative for drug discoverers to assess multiple molecular pathways from the same sample and build a compelling case to earn funding and support for promising drug targets[35].

We recently caught up with **Mathew Chamberlain**, Principal Scientist at Johnson & Johnson to get his take on the value that multi-omics has brought to drug discovery.



INTERVIEW: MATHEW CHAMBERLAIN

PRINCIPAL SCIENTIST JOHNSON & JOHNSON INNOVATIVE MEDICINE

Mathew: I specialise in the pharmaceutical industry in early drug discovery. I've worked in single-cell omics for about six years now, mostly for drug discovery efforts, often across autoimmune diseases, and sometimes immune-mediated diseases like oncology.

***FLG:* What is the value of multi-omics methods for drug target discovery? And for the clinical applications that you're working on?**

Mathew: At a high level, for a lot of the diseases we study, some patients have the same molecular phenotypes, but different symptoms, or they have different symptoms, but the same phenotypes. This is very confusing. How do you understand disease if you don't have a strong understanding of the molecular basis for it? And it raises questions about whether or not there are multiple different diseases within one disease. Are we aggregating diseases into a composite condition?

Since the advent of omics about 25 years ago, it really allowed us to start looking at patient data for the first time, in a high-throughput way. Multi-omics is now a very high throughput way of analysing samples from patients, and this is very valuable tissue that we're talking about here. You have potentially thousands of dollars a vial because it takes a lot to get a biopsy sample of some tissues, it's very hard to do. So, you really want to get the most you can possibly get out of a vial. And when it comes down to our team, we are asking ourselves 'What experiments would you recommend for getting the most out of your clinical sample?' And the answer is single-cell multi-omics. It's the most bang for your buck.

> "HOW DO YOU UNDERSTAND DISEASE IF YOU DON'T HAVE A STRONG UNDERSTANDING OF THE MOLECULAR BASIS FOR IT? AND IT RAISES QUESTIONS ABOUT WHETHER OR NOT THERE ARE MULTIPLE DIFFERENT DISEASES WITHIN ONE DISEASE. ARE WE AGGREGATING DISEASES INTO A COMPOSITE CONDITION?"

***FLG:* And with these multi-omics methods in drug discovery, are you hoping that they're going to lead to more targets or a much more refined set of drug targets. Or both?**

Mathew: The main opportunities are positioning drugs into patient populations in a smarter way. For example, there are TNF inhibitors that are the number one bestselling drug for the past 20 years in pharmaceuticals and it is special because it's approved across about 20 different disease indications. It then opens up a question; a lot of these other targeted therapies, could they be applied in other disease indications as well? Maybe you have a drug that's going to work in one disease area, but maybe it'll actually work in other ones too. You just don't know about it yet. And single-cell multi-omics lets you explore that idea.

It also lets you explore completely novel ideas. For example, no-one's really studied the interactions between nonimmune and immune cells in a specific tissue. Single-cell multi-omics allows you to completely open it up and there might be some new targets in here. It might also let you segment clinical trials better. Sometimes, you have a result in a clinical trial that almost worked, and it looks like it might help some patients but not others. Often, we don't really know why it helps some people and not others. So, it would be lovely to take a deep dive into the molecular status of disease and you might derive two different types of patients at the molecular level that you didn't appreciate you had before.

***FLG:* You recently presented some work on multi-omics data integration for new target discovering in Crohn's disease, could you just give us a brief summary of that work?**

Mathew: So, if you're going to present new target discovery from single-cell data, it's really going to help to integrate data from multiple sources. With Crohn's disease, there have been recent publications from Aviv Regev, Ramnik Xavier and from some other papers on gut inflammation and some internal data using spatial, CITE-seq and single-cell. Across these there are hundreds of thousands of cells. In our work, I combined GWAS information (genetics) from published sources with bulk sequencing data from tissue biopsies from Crohn's from another published source. I combined that with single-cell data using ligand receptor analysis methods - CellPhoneDB from Sarah Teichmann's group - and created an integrated view. Here, you could look at ligands and receptors, which are enriched in disease samples in single-cell data, and are associated with genetic evidence and then differentially expressed in tissue biopsies and bulk sequencing.

This whole big data package is the kind of thing you have to put together in pharma companies to motivate people to do some validation experiments. Because, in general, if you have one piece of evidence, people will think, 'Well, I don't know if this is supported by genetics, I don't know if this is a single-cell sequencing related artefact or only observable using those assays'. Or in bulk sequencing, 'I don't know if it's going to occur, and at the single-cell level, what it really looks like'. So, when you combine all three different data types together into a single data package, that new target discovery method is typically convincing, even for seasoned drug discovery experts. It's not easy to convince a roomful of people who have worked on drugs that had been approved by the FDA to work on new targets, that's why it's risen to this level of interest.

AI and machine learning presents the other part of this toolkit, which is very promising for drug discovery. Machine learning can be used to find patterns in these complex datasets and find interactions between omics layers, which create diseases states[36,37]. An example of this is the DeepInsight-3D approach[38,39,] which was released this year, and uses deep learning and multi-omics data to predict patient-specific anticancer drug responses.

Furthermore, these models can be used to reposition drugs for new targets[40] or to predict the effect of a drug target and how it would effect a system in silico. This includes tools such as ChemCPA[41,42], which can predict perturbation effects of unseen drug combinations.

Chapter 7 references

1. van Karnebeek, C.D. *et al.* **The role of the clinician in the multi-omics era: are you ready?** *Journal of Inherited Metabolic Disease* **41**, 571-582 (2018).
2. Clark, C., Rabl, M., Dayon, L. & Popp, J. **The promise of multi-omics approaches to discover biological alterations with clinical relevance in Alzheimer's disease.** *Frontiers in Aging Neuroscience* **14**, 1065904 (2022).
3. Kodam, P., Sai Swaroop, R., Pradhan, S.S., Sivaramakrishnan, V. & Vadrevu, R. **Integrated multi-omics analysis of Alzheimer's disease shows molecular signatures associated with disease progression and potential therapeutic targets.** *Scientific Reports* **13**, 3695 (2023).
4. Khullar, S. & Wang, D. **Predicting brain-regional gene regulatory networks from multi-omics for Alzheimer's disease phenotypes and Covid-19 severity.** *Human Molecular Genetics* **32**, 1797-1813 (2023).
5. Sathyanarayanan, A. et al. **Multi-omics data integration methods and their applications in psychiatric disorders.** *European Neuropsychopharmacology* **69**, 26-46 (2023).
6. Kharb, S. & Joshi, A. **Multi-omics and machine learning for the prevention and management of female reproductive health.** *Frontiers in Endocrinology* **14**, 358 (2023).
7. Maitre, L. *et al.* **Multi-omics signatures of the human early life exposome.** *Nature Communications* **13**, 7024 (2022).
8. Ma, A., Xin, G. & Ma, Q. T**he use of single-cell multi-omics in immuno-oncology.** *Nature Communications* **13**, 2728 (2022).
9. Akhoundova, D. & Rubin, M.A. **Clinical application of advanced multi-omics tumor profiling: Shaping precision oncology of the future.** *Cancer Cell* **40**, 920-938 (2022).
10. Liu, Q. & Song, K. **ProgCAE: a deep learning-based method that integrates multi-omics data to predict cancer subtypes.** *Briefings in Bioinformatics,* bbad196 (2023).
11. Yang, S. *et al.* **Integrated Multi-Omics Landscape of Liver Metastases.** *Gastroenterology* 164, 407-423. e17 (2023).
12. He, D.-n. *et al.* **Multi-omics analysis reveals a molecular landscape of the early recurrence and early metastasis in pan-cancer.** Frontiers in Genetics **14**, 1061364 (2023).
13. Chen, G., Zhang, J., Fu, Q., Taly, V. & Tan, F. **Integrative analysis of multi-omics data for liquid biopsy.** *British Journal of Cancer* **128**, 505-518 (2023).
14. Zou, Y., Zhao, Z. & Song, Y. A**n overview of multiomics: a powerful tool applied in cancer molecular subtyping for cancer therapy.** *Malignancy Spectrum* (2023).
15. Murai, H. *et al.* **Multiomics identifies the link between intratumor steatosis and the exhausted tumor immune microenvironment in hepatocellular carcinoma.** *Hepatology* **77**, 77-91 (2023).
16. Hsieh, W.-C. *et al.* **Spatial multi-omics analyses of the tumor immune microenvironment.** *Journal of Biomedical Science* **29**, 96 (2022).
17. Chai, H. *et al.* **Integrating multi-omics data through deep learning for accurate cancer prognosis prediction.** *Computers in biology and medicine* **134**, 104481 (2021).
18. Stephenson, E. *et al.* **Single-cell multi-omics analysis of the immune response in COVID-19.** *Nature Medicine* **27**, 904-916 (2021).
19. Wang, K. *et al.* **Sequential multi-omics analysis identifies clinical phenotypes and predictive biomarkers for long COVID.** *Cell Reports Medicine.*
20. Zhang, B. et al. **Multimodal single-cell datasets characterize antigen-specific CD8+ T cells across SARS-CoV-2 vaccination and infection.** Nature Immunology **24**, 1725-1734 (2023).
21. Diray-Arce, J. *et al.* **Multi-omic longitudinal study reveals immune correlates of clinical course among hospitalized COVID-19 patients.** *Cell Reports Medicine* **4** (2023).
22. Vasaikar, S.V. *et al.* **A comprehensive platform for analyzing longitudinal multi-omics data.** *Nature Communications* **14**, 1684 (2023).
23. Lu, T., Wang, Y. & Guo, T. **Multi-omics in COVID-19: Seeing the unseen but overlooked in the clinic.** *Cell Reports Medicine* **3**(2022).
24. Guo, M. *et al.* **Multi-omics for COVID-19: driving development of therapeutics and vaccines.** *National Science Review* **10**(2023).
25. Zhu, Z. *et al.* **A comprehensive review of the analysis and integration of omics data for SARS-CoV-2 and COVID-19.** *Briefings in Bioinformatics* **23**(2021).
26. Ma, J., Deng, Y., Zhang, M. & Yu, J. **The role of multi-omics in the diagnosis of COVID-19 and the prediction of new therapeutic targets.** *Virulence* **13**, 1101-1110 (2022).
27. Lunke, S. *et al.* **Integrated multi-omics for rapid rare disease diagnosis on a national scale.** *Nature Medicine* **29**, 1681-1691 (2023).
28. Marić, I. *et al.* **Early prediction and longitudinal modeling of preeclampsia from multiomics.** *Patterns* **3**(2022).
29. Jehan, F. *et al.* **Multiomics Characterization of Preterm Birth in Low- and Middle-Income Countries.** *JAMA Network Open* **3**, e2029655-e2029655 (2020).
30. Espinosa, C.A. *et al.* **Multiomic signals associated with maternal epidemiological factors contributing to preterm birth in low- and middle-income countries.** *Science Advances* **9**, eade7692 (2023).
31. Babu, M. & Snyder, M. **Multi-Omics Profiling for Health.** *Molecular & Cellular Proteomics* **22**, 100561 (2023).
32. Kellogg, R.A., Dunn, J. & Snyder, M.P. **Personal omics for precision health.** *Circulation research* **122**, 1169-1171 (2018).
33. Schüssler-Fiorenza Rose, S.M. *et al.* **A longitudinal big data approach for precision health.** *Nature medicine* **25**, 792-804 (2019).
34. Shen, X. *et al.* **Multi-omics microsampling for the profiling of lifestyle-associated changes in health.** *Nature Biomedical Engineering* (2023).
35. Zielinski, J.M., Luke, J.J., Guglietta, S. & Krieg, C. **High Throughput Multi-Omics Approaches for Clinical Trial Evaluation and Drug Discovery.** *Frontiers in Immunology* **12** (2021).
36. Feldner-Busztin, D. *et al.* **Dealing with dimensionality: the application of machine learning to multi-omics data.** *Bioinformatics* **39** (2023).
37. Pammi, M., Aghaeepour, N. & Neu, J. **Multiomics, artificial intelligence, and precision medicine in perinatology.** *Pediatr Res* **93**, 308-315 (2023).
38. Sharma, A., Lysenko, A., Boroevich, K.A. & Tsunoda, T. **DeepInsight-3D architecture for anti-cancer drug response prediction with deep-learning on multi-omics.** *Scientific Reports* **13**, 2483 (2023).
39. Cai, Z., Poulos, R.C., Liu, J. & Zhong, Q. M**achine learning for multi-omics data integration in cancer.** *Iscience* (2022).
40. Cong, Y. & Endo, T. **Multi-omics and artificial intelligence-guided drug repositioning: Prospects, challenges, and lessons learned from COVID-19.** *OMICS: A Journal of Integrative Biology* **26**, 361-371 (2022).
41. Lotfollahi, M. *et al.* **Predicting cellular responses to complex perturbations in high-throughput screens.** *Mol Syst Biol* **19**, e11517 (2023).
42. Hetzel, L., Boehm, S., Kilbertus, N., Günnemann, S. & Theis, F. **Predicting cellular responses to novel drug perturbations at a single-cell resolution.** *Advances in Neural Information Processing Systems* **35**, 26711-26722 (2022).

CHAPTER 8

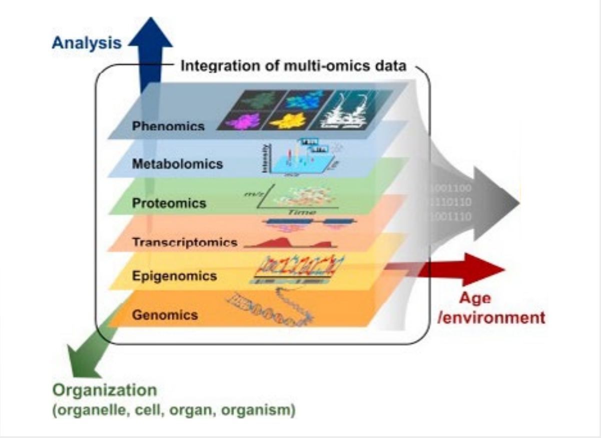
INSIGHTS FROM THE MULTI-OMICS-VERSE

IN THIS FINAL CHAPTER, WE WANTED TO EXPLORE THE REMAINING CHALLENGES IN MULTI-OMICS AND IDENTIFY WAYS THESE CHALLENGES CAN BE ADDRESSED. SO, WE ASKED OUR CONTRIBUTORS THE SAME QUESTION; WHAT ARE THE GREATEST CHALLENGES FACING THEIR FIELD AND WHAT COULD BE DONE ABOUT THEM? OUR FINAL CHAPTER WILL HIGHLIGHT THESE CHALLENGES AND LOOK AT THE WAYS THESE CHALLENGES COULD BE ADDRESSED AS WE LOOK AHEAD INTO 2024.

Across this playbook, we have highlighted many of the advancing areas of single-cell and spatial multi-omics. Due to the inherent complexity of integrating multiple modalities, the progress of building reliable and standardized practices is slow but steady.

In the list of 11 grand challenges for single-cell data science, published in early 2020[2], integrating single-cell data across samples and modalities was challenge number 10. In late 2023/early 2024, many new approaches to integration have been realized (see Chapter 2). However, data integration is still present in the list of top challenges facing multi-omics experiments. But what are the other challenges holding the progress of multi-omics back? Is integration still near the top?

Image Credit: Kim, et al. [1]



We asked a range of contributors **what they think is the biggest challenge still facing the multi-omics field, and what is being done or what should be done to tackle the issue.**

The nature of multi-omics data

The sparseness and lack of correspondence between data was highlighted as a challenge by several contributors, followed by their ideas for how this is still being tackled.

MIRJANA EFREMOVA

Group Leader
Barts Cancer Institute

FLG: What do you think is the biggest challenge still facing the multi-omics field? And what is being done to tackle the issue, or what should be done to tackle the issue?

Mirjana: *A challenge for the application of many current integration methods is still the requirement of feature correspondence between different modalities (for example, gene scores are often calculated from scATAC-seq data). Paired data (where we measure different modalities from the same cell) can be very informative in integrating different omics layers from unpaired data because it can provide ground truth data and inform associations between features of different modalities. Therefore, it should be used whenever available to improve multimodal integration of unpaired data.*

PAU BADIA I MOMPEL

PhD Candidate, Saez-Rodriguez Group
Heidelberg University

FLG: What do you think is the biggest challenge still facing the multi-omics field?

Pau: *Sparsity. The more multimodal assays we add, the sparsity increases, it becomes harder to try to profile both technologies, and also the actual cost of these technologies increases. I mean, there are some alternatives that are coming up, that are more open source such as ISAAC-seq. But I would say the actual monetary cost is still quite restrictive, especially for a lot of labs without access to huge funds, but they can still contribute to the field.*

FLG: Is there any way this challenge is currently being addressed or could be addressed?

Pau: *Aggregation. Pseudo-bulking or performing meta cells. Pseudo-bulking is just the matter of aggregating all the counts, for example, in a given cell type into one single profile. But sometimes if you don't have enough true replicates, that's restrictive. A middle ground would be to try to identify which groups of cells are behaving more or less the same, and then pull them together into these meta cell profiles. So, you'd still have some kind of granularity but at least those profiles are richer than the original.*

FLG: Would you say that you lose some of the value of doing it by single cell in the first place by pseudo-bulking?

Pau: *Not necessarily. Why do we do single-cell in the first place? Theoretically, it is because we want to know what happens in each individual cell, but due to sparsity, that's not possible. I think right now, the value of single-cell is that you can unbiasedly profile 1,000s of cells without having to worry about 'I need to FACS sort these', so you can computationally separate the signals between cell types in an easier way than having to do it in the in the lab.*

Data Integration

As expected, data integration was still front and centre for several of our contributors.

IAIN MACAULAY

Technical Development Group Leader
Earlham Institute

FLG: What do you think is the biggest challenge still facing multi omics?

Iain: *I think the biggest challenge is probably in data integration, and understanding what it means. You can generate so many different types of data, but the challenge is understanding how they all fit together. From the G&T-seq paper[3], which is now a million years old, even that data is not fully analysed. We've just gotten to a certain point where we had to stop but we haven't done a huge amount of single nucleotide analysis; we know we can do it, but we'll have to write new pipelines to get some of that data out and integrate it. It's amazing how easy it is to generate that much data. You could generate a PhD's worth of data in a few weeks if you've got all the methods set up, but you will never finish analysing it, so I think data integration and data visualisation is the biggest challenge.*

FLG: What do you think can be done for data integration? What approach could we try? Is it machine learning? Is it community-based efforts?

Iain: *It's really hard, because you've got so many methods emerging, and each method has been analysed, and the pipeline is developed by that lab's computational team or postdoc. The standardisation of methods only emerges when enough people are generating enough data using a specific method. So, I think some things are going to emerge out of people doing 10x Multiome, because lots of people are doing that. Things like G&T-seq is probably too niche.*

I think it would be good if there was more funding for computational science experts to come together and do this kind of work. It would be great if there was just more funding for a multi-omics facility that you can engage. For certain methods, I think general principles will start to emerge, such as how do you integrate epigenetic data with transcriptome data, regardless of method? But with these more bespoke methods, you just don't get the traction of having a bunch of people really caring enough to develop a nice pipeline for it.

SUHAS VASAIKAR

Principal Scientist, Clinical Biomarker and Diagnostics
Seattle Genetics (Seagen)

FLG: What do you think is the biggest challenge still facing the multi-omics field and what is being done to tackle the issue, or what should be done to tackle the issue?

Suhas: *One of the biggest challenges still facing the multi-omics field is the integration of different omics data types in a meaningful and interpretable way. While there has been significant progress in developing computational methods for integrating omics data, there are still challenges in dealing with the complexity and heterogeneity of the data.*
To tackle this challenge, there are several approaches that can be taken.

- *One approach is to* ***develop more advanced machine learning algorithms*** *that can handle the complexity of multi-omics data and identify meaningful biological patterns.*
- *Another approach is to* ***develop more standardized protocols for collecting and processing multi-omics data,*** *which can help to reduce variability and improve data quality.*
- *Additionally, there is a need for* ***more collaboration and interdisciplinary research in the field of multi-omics.*** *Integrating data from different omics platforms requires expertise in multiple fields, including biology, statistics, computer science and data visualization. By bringing together researchers from different disciplines, we can develop more comprehensive and effective approaches for integrating multi-omics data.*
- *Finally, there is a need for* ***more open data sharing and collaboration in the multi-omics field.*** *Sharing data and methods can help to accelerate research and enable more effective integration of omics data. Open data platforms, such as the Genomic Data Commons and the Cancer Genome Atlas, are already playing a critical role in advancing multi-omics research.*

In summary, more advanced machine learning, collaboration and interdisciplinary research and open data sharing are some of the ways to tackle the issue of developing challenging computational methods for integrating omics data.

XIAOTAO SHEN

Postdoctoral Research Fellow, Snyder Lab
Stanford University

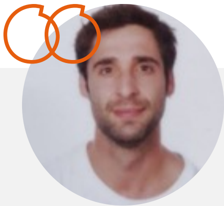
FLG: What do you think is the biggest challenge still facing the multi omics field? And is there anything being done to tackle that challenge?

Xiaotao: *For me, the most challenging issue is integration. Especially knowledge-based integration, which is very truly difficult. For example, now we are working on the projects of integration between microbiome and metabolites. We know that microbiomes can produce some compounds and small peptides, and then these small compounds and peptides can impact human health. But how can we connect them and measure all the metabolites from the microbiome?*

Currently, some researchers culture the bacteria and then measure the media to see what compound the bacteria can produce. However, this method is very difficult for all of the microbiome. As we know, there are hundreds and thousands of bacteria from the human microbiome, so we can't culture all of them. So, I think in the future, we could try to integrate them using the knowledge and data-driven network to get the whole interaction network between microbiome and metabolites

Data accessibility and model validation

Linked to the previous comments, accessibility of high quality data to validate experiments, methods and AI models is becoming an essential challenge to address.



RICARD ARGELAGUET

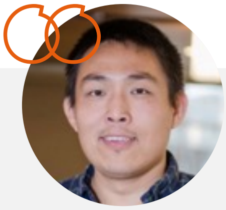
Senior Research Scientist
Altos Labs

FLG: What do you think is the biggest challenge still facing the multi-omics field? And what is being done to tackle it?

Ricard: *Since the next step for integration is on very diverse and large data sources, one of the main challenges is data accessibility. Machine learning scientists are not going to be reprocessing all of these datasets from scratch. They need a curated effort to try and get all of these different datasets together.*

Some effort was done on our SingleCellMultiModal package. But there are a lot more datasets than they were three years ago and there are extremely good efforts. For example, by the John Zuckerberg Foundation on trying to make those data accessible. And this is one of the main challenges that is being tackled by the big consortia.

Another important challenge, which I think is important to always keep in mind is, why are we developing those models? Are we trying to teach a new deep learning model for the sake of teaching a new deep learning model? We need to think of what biological questions we are not currently able to answer. And which question would we be able to answer by developing these models. So, I think that's why it's very important to keep this field interdisciplinary and have machine learning scientists, machine learning engineers, bioinformaticians and biologists all working together. We then disconnect this knowledge generation process.



QIN MA

Professor, Department of Biomedical Informatics
The Ohio State University

FLG: What do you think is the biggest challenge still facing the multi omics fields? And what is being done to tackle the issue, or what should be done to tackle the issue?

Qin: *In single-cell, I think the data noise level is very high. There's a real issue with potential false positives and inaccurate predictions. So, validation will be a very critical thing. Either validating one prediction using multiple datasets, so that you select a phenomenon to determine whether it's causality or only observation. You can then check this in a lot of other datasets and see whether this is a real insight. Or you can try different methodologies. We can use the different tools generated by various other labs that are targeting the same thing, to see whether we can identify something similar. If not, that will be a question mark. If yes, that will be strong evidence that this may be a real insight. The last validation is experimental validation. Can we really make it happen in mice, and then in humans? And then we can go to clinical trial.*

The second challenge is interpretation in machine learning and deep learning. The algorithms are doing too well to be in a black box. Currently, we don't know what's happening with this very rigorous and smart AI. Also, if we were not in a biomedical or human health field, people wouldn't care. If AI is earning you more money by handling your investments, you won't care why, and you'll give your money to AI. In science, we are eager to know what happened, we want to know the mechanism so we can learn from it. Why can AI do a better job than traditional methods? Why can AI remove that noise? These things will help us to shape the future directions in science.

Perturbations for GRNs

Validating models was a challenge on the minds of several other contributors, but this time with the perspective of perturbation experiments to validate gene network assumptions.

SAMANTHA MORRIS

Associate Professor of Development Biology and Genetics
Washington University School of Medicine in St. Louis

FLG: What do you think is the biggest challenge still facing like your field? And is there any way that multi-omics could help tackle it?

Sam: *One of the limitations of CellOracle is that we can only simulate the perturbation of one transcription factor at a time. So, one of the next challenges is perturbation of multiple factors at the same time, which is challenging. When multiple transcription factors interact together, particularly in an overexpression scenario, how do you predict the behaviour of those factors? We need to better understand how they impact gene expression, and then how they impact cell identity, potentially creating new cell types that we haven't observed before. That is a huge challenge in the field that I'm really excited to see people make progress on.*

SUSHMITA ROY

Professor, Department of Biostatistics and Medical Informatics, **University of Wisconsin-Madison**
Faculty, **Wisconsin Institute of Discovery**

FLG: What do you think is the biggest challenge still facing your field – multi-omics and gene regulation? And what is being done to tackle the issue, or what should be done to tackle the issue?

Sushmita: *I would say that the problems we work with are fundamentally unsupervised. Most of the time we don't know what the truth is, and typically we have very little 'truth' that we can use to actually benchmark our models. One thing that we really need more of are these high throughput perturbation experiments like Perturb-seq. This would mean we could really test our model's predictions in a high-throughput way, because our models make 1000s of predictions.*

Currently, people can do one regulator knockout and get five edges/five targets that they can test, but that's not really assessing things in a high-throughput way. So, incorporating perturbations into the models to infer causal gene regulatory networks, I would say, is really the direction that we should be heading in. Going forward, I think multi-omics will be beneficial because we want to learn these integrative multi-layer networks. So, we would want perturbation approaches that not only measure gene expression, but also, we want a 'multimodal Perturb-seq' or whatever you want to call it. Ultimately, we need to build predictive, interpretable models of gene expression.

Transcriptomic and epigenomic methods

Other contributors visualized ways to improve current transcriptomic and epigenomic methods.

RUI CHEN

Professor of Molecular and Human Genetics
Baylor College of Medicine

FLG: What do you think is the biggest challenge still facing multi omics? And what is being done to tackle the issue, or what should be done to tackle the issue?

Rui: *I think for the RNA/ATAC ones, if you want to profile them at the same time - true multiome from single cells - I think that the biggest hurdle is the difficulty in getting high quality data from both modalities at the same time. It is very sensitive to the time of treatment. And it's not always the same, particularly when you deal with human samples, a well-controlled mouse animal is probably much easier. So, I think the protocol optimization is still very finicky.*

I think getting the true multiome is critical, because otherwise you need to use computational methods to co-embed. Then you introduce error. However, the current methods of co-embedding are quite accurate, even for very dynamic data. If you're looking at extremely dynamic data, then it's probably best to get a good single-cell multiome to avoid the misalignment. If your system is not extremely dynamic, e.g., adult tissue or even developmental tissue, I think you can just provide them separately and format and put them together.

BINGJIE ZHANG

Postdoctoral Research Fellow, Satija Lab
New York Genome Center

FLG: What do you think is the biggest challenge still facing the multi-omics field, and what is being done or should be done to tackle that issue?

Bingjie: *Setting aside multi-omics, we are still at the early stage with the single modality technologies. Currently, most single-cell studies are heavily focused on transcriptome and chromatin accessibility, likely because 10x Genomics provides commercial kits that are user-friendly.*

On one hand, I hope to see more innovative methods that explore modalities beyond RNA and ATAC. For example, we don't even have a good method to profile the transcriptional factor binding sites. More sensitive and easy-to-implement methods targeting histone modifications, intracellular proteins or DNA methylation would also be beneficial.

On the other hand, it's important to consider how to make these novel methods easily accessible to labs that are eager to apply them. Our lab has made a real effort in this regard. We have initiated a technology sharing program. Anyone interested in the method we developed can sign up for free 'starter kits', which essentially include everything needed to implement the method in their own lab. We now have launched the starter kits for NTT-seq, CaRPool-seq and Phospho-seq. While it's currently only available to CEGS groups, we are keen to extend this to more labs in the future.

Newness and costs of the technology

Some contributors looked at the wider problem for the whole field, namely the cost, accessibility and lack of experience with this plethora of technology.

ANDREA CORSINOTTI

Single-cell Multi-omics Facility Manager, Centre for Regenerative Medicine, Institute for Regeneration and Repair
University of Edinburgh

FLG: What do you think is the biggest challenge still facing the kind of single-cell multi-omics community?

Andrea: *I would say cost. We need to do more and more of these experiments with more cells, more samples, more meaningful data sets. The technologies work, there is constant development, and I am confident that the technologies will keep developing and technical solutions will become available in a scalable, commercial way for everyone. The limitation that remains is the cost. It's not only the cost of the experiment but the cost of the single-cell reagents, the cost of sequencing, because single-cell experiments always require more sequencing than bulk experiments.*

This is a big limitation. By spending all this money on these aspects, the capacity to generate larger datasets will be affected. Hopefully costs will keep going down, and the more they do, the more we can make sense of what we are doing. All these questions about benchmarking, what technology works better, do we need an alternative? We will not find the answers until we have exhausted the technology. We need to have done so many experiments and we will know whether something works better than something else, but in order to do that, people need to be able to afford to do these experiments.

FLG: Is there anything that can be done? Or should be done to tackle the issue?

Andrea: *What should happen is people should use these technologies and share the data. It would be useful to know: 'Okay, we tried some different technologies, different reagents, different companies, and did some benchmarking, and now we can say that they are equivalent or even better than what was available before'. Having as much information as possible at this level is going to be critical in the near future, to bring down costs and to increase the ability of researchers to use these technologies.*

MATHEW CHAMBERLAIN

Principal Scientist
Johnson & Johnson Innovative Medicine

FLG: What do you think the big challenge is for your current field?

Mathew: *I think that the biggest challenge now is how new single cell technology is. The quality and richness of datasets will get much better with time, like bulk RNA-sequencing. That's related to the newness of the technology, also to the cost and feasibility in adopting single-cell software and methods.*

FLG: Is there anything that can be done to address the challenge?

Mathew: *I think that some of the cost reducing methods definitely help. For example, some of the multiplexing methods that various groups have come out with recently. Those certainly help. Every time I think about these questions, I just think about what happened with bulk sequencing 15 years ago, and then assume it's going to happen a bit faster this time, because we've essentially done it once before. In bulk sequencing, it took about 10 years for the field to coalesce on a pretty set standard of computational workflows that most people are reasonably happy with. In single-cell, we're basically starting that process now. The field is starting to coalesce a bit, which, together with reduced costs, will help standardize everything from sample collection to results.*

Some aspects of multi-omics are far behind this in single cell. I've seen a lot of spatial omics talks this year, which remind me of single-cell five years ago. In spatial right now, you have a few samples, and you're sort of observing something and it looks interesting and enticing. But, if you were any researcher in the field and had a dataset of 100 spatial samples today, before and after treatment, there's no computational workflow to even answer those questions right now. You'd have to design it yourself. That's where I see the field going, a lot of multi-omics integration workflows, where you're putting together different data types, and build out larger and larger data sets and atlases. I see the field going that way in the near future.

Then I see the field moving more into spatial and then increasingly as the sample sizes get larger, more tissue layers and more omics layers get added. At that point, the number of available workflows goes almost to zero pretty quick. Then you have to start the clock again when the fields coalesce on a standard workflow. You need a statistical analysis plan going into it but usually the omics data typically aren't an endpoint, they're just an observation that goes together with your trial.

That's why I mentioned cost and throughput as two things I would work on now, because then at least you'll get observational data. And we're starting to get that from trials that we run, other pharmaceuticals are doing that as well.

Integrating Spatial

Several contributors highlighted the need to bring spatial information into the analysis paradigm to measure spatially-mediated second order phenomena such as cell-cell interaction.

SHIRLEY GREENBAUM

Postdoctoral fellow, Department of Pathology, **Stanford University**
Resident, Department of Obstetrics and Gynaecology
Hadassah-Hebrew University Medical Center

FLG: What do you think is the biggest challenge still facing the multi-omics field? And what is being done to tackle the issue, or what should be done to tackle the issue?

Shirley: *The biggest challenge facing the multi-omics field, particularly in the context of the placenta, is the effective correlation of multi-omics data with spatial information. This challenge is especially crucial concerning the placenta, the focus of my work. During placentation, fetal trophoblasts invade maternal tissue, leading to regions where maternal and fetal cells are adjacent to one another. It becomes critically important to precisely associate multi-omics data with specific individual cells within this intricate spatial context.*

One promising approach to tackle this issue is the utilization of techniques like Multiplexed Ion Beam Imaging (MIBI). As mentioned, MIBI has the distinct advantage of allowing researchers to preserve tissue integrity during processing while simultaneously providing high-resolution spatial data. This enables the association of multi-omics data with the precise location within the tissue, thereby addressing the spatial correlation challenge.

ZONGMING MA

Professor, Department of Statistics and Data Science
Yale University

FLG: What do you think is the biggest challenge still remaining in the multi-omics field? And what is being done to tackle the issue, or what should be done to tackle the issue?

Zongming: *I think the biggest challenge at the moment is the neglect of second-order information. A lot of the multi-omics technologies give you a finer understanding of things on the first order. By this, I mean you can get a very precise differentiation of different cell states, trajectory analysis, so on and so forth. These have been enabled by multi-omics information. What would be more interesting is how you combine such understanding with spatial data objects and get an understanding of the second order. By second order, I mean cell-cell interaction, cell-cell communication and regulation of certain cells by neighbouring cells. You can then get a mechanistic understanding of how cells interact with each other by doing both spatial measurements, and also single-cell multi-omics measurements.*

Mapping metabolites

And last but not least, our metabolomics contributors highlighted some key challenges and potential solutions for identifying the nature and roles of metabolites as well as modelling this rapidly changing omic.

THEODORE ALEXANDROV

Team Leader, Structural and Computational Biology Unit
European Molecular Biology Laboratory

FLG: If you had to pick a big challenge for single-cell and spatial metabolomics, what would you say is the big one remaining?

Theo: *The next challenge for metabolomics is in data interpretation. In metabolomics, in terms of detection, we have amazing technologies already, in terms of hardware and instruments. We can do a lot of things, but in terms of interpretation, metabolism and metabolic pathways are inherently very complex. Metabolites are building blocks and energy sources, but the same molecules can play signalling roles.*

There are also many unknown roles of metabolites. Lately, the Rutter lab published a paper in Nature Cell Biology showing how metabolites have unstudied roles, in particular in controlling the activity of enzymes in their pathways and also enzymes throughout the whole metabolism. This kicked off the discussion of 'metabo-verse', a universe of all the small molecules with diverse functions and roles.

However, we do not have the databases of roles and functions for metabolites, similar to how we have them for genes and for proteins. There is work to be done to create these catalogues of their functions and associations with cell types. I think these databases will eventually explain the roles and contain all the biochemical molecular functions of the molecules. That will be a big breakthrough, but this is our current challenge.

FLG: Can you see a way to address those problems you just highlighted with the data interpretation?

Theo: *First of all, we lack robust protocols, particularly in single-cell and spatial, that are accessible for scientists who are not from analytical chemistry labs. We'll have impact with this technology only when it will be accessible by biologists and by clinical scientists. Over the last years, it started happening with spatial metabolomics when a number of biology labs installed their first imaging mass spectrometer. However, we're not there yet. Once this happens, they will create bigger opportunities.*

For this to work, we need to have better instrumentation, we need to have robust protocols, which are relatively easy to execute, we need to have user friendly software, which is not for geniuses in mass spectrometry. On top of this, we need to have databases that this software can tap into to enhance data interpretation. And, in particular, we need to link to other omics, because metabolomics is not an ultimate tool, it is orthogonal and complementary to other omics.

Obviously, to have a biological or medical conclusion or a decision for drug development, one needs to use all the tools, and have metabolomics in your portfolio. There should be more to be done to link metabolomics with other omics tools that will help achieve much bigger impact.

INGELA LANEKOFF

Professor, Department of Chemistry-BMC
Uppsala University

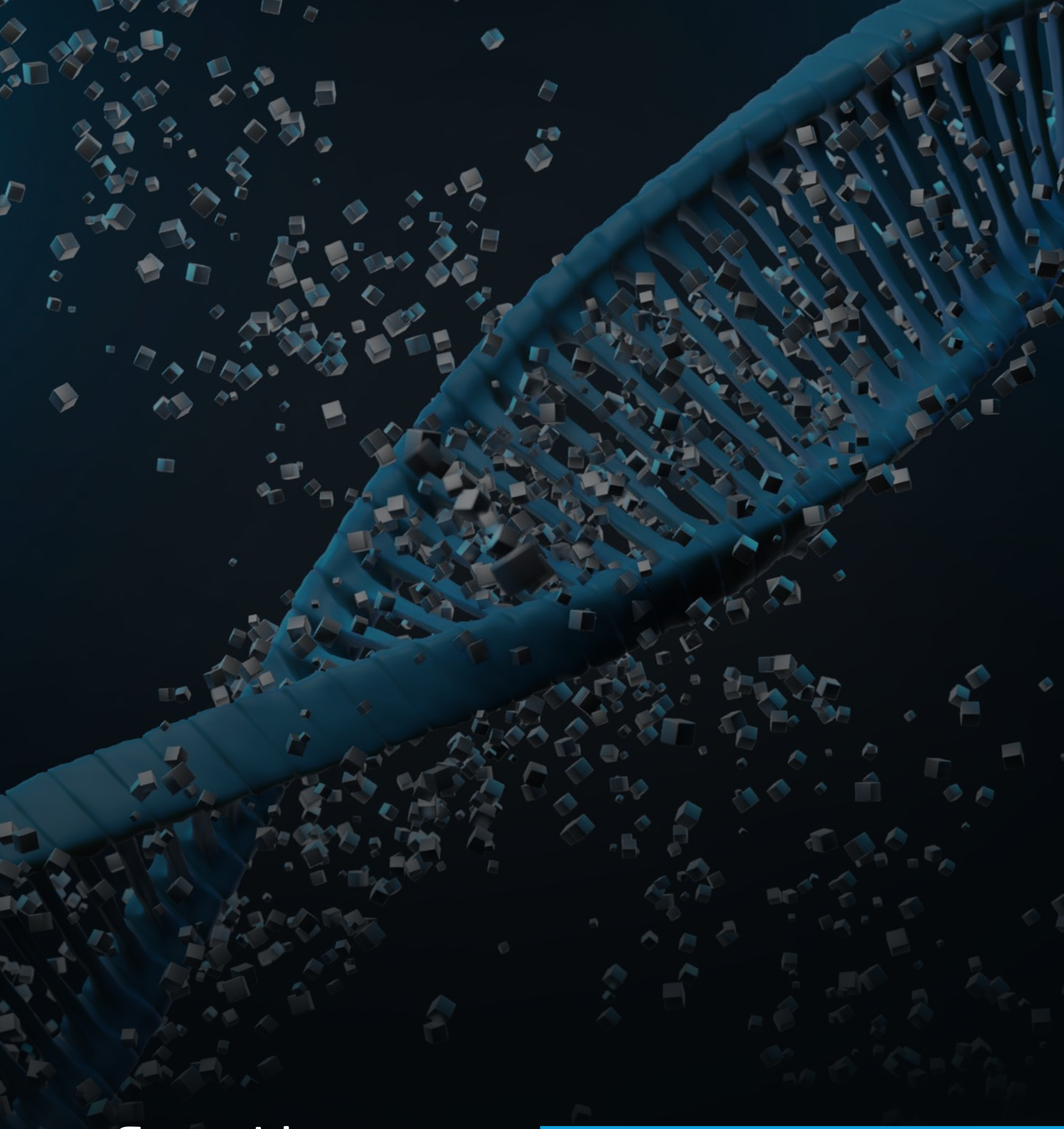
FLG: What do you think is the biggest challenge still facing the metabolomics field? And what is being done to tackle the issue or what should be done to tackle the issue?

Ingela: *I think that the biggest challenges of the metabolomics field are what also makes it so interesting, which is that the metabolome can rapidly alter to reveal the cellular status at that time. This is particularly challenging with sample collection of living biological material, where the sample will actually change if not immediately quenched by, for example, snap freezing in liquid nitrogen.*

The community is highly aware of importance of reproducibility in sampling and sample handling, and is continuing to establish protocols to reduce these types of artefacts in their metabolomics studies.

Chapter 8 references

1. Kim, J., Woo, H.R. & Nam, H.G. **Toward systems understanding of leaf senescence: an integrated multi-omics perspective on leaf senescence research.** *Molecular plant* **9**, 813-825 (2016).
2. Lähnemann, D. *et al.* **Eleven grand challenges in single-cell data science.** *Genome Biology* **21**, 31 (2020).
3. Macaulay, I.C. *et al.* **G&T-seq: parallel sequencing of single-cell genomes and transcriptomes.** *Nature methods* **12**, 519-522 (2015).



Front Line Genomics

frontlinegenomics.com

info@frontlinegenomics.com

@FLGenomics

@frontlinegenomics

Festival of Genomics: festivalofgenomics.com