PRODUCED BY:

Front Line Genomics

MULTI-OMICS: THE FULL PICTURE

A COMPREHENSIVE REPORT ON ALL THINGS MULTI-OMICS:
PAINTING A CLEARER PICTURE OF BIOLOGY, HEALTH AND DISEASE.

SPONSORED BY

canopy BIOSCIENCES A BRUKER COMPANY

mission bio

nanoString

Novogene

NVIDIA

CONTRIBUTORS

John Quackenbush
Professor, Department of Computational Biology and Bioinformatics
Harvard T.H. Chan School of Public Health

Miao-Ping Chien
Assistant Professor, Department of Molecular Genetics,
Erasmus University Medical Center
Principal Investigator,
Oncode Institute

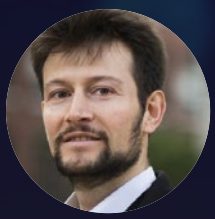
Stephanie Byrum
Associate Professor, Department of Biochemistry and Molecular Biology
University of Arkansas at Little Rock

Lihua (Julie) Zhu
Professor, Department of Molecular, Cell and Cancer Biology
University of Massachusetts Chan Medical School

Jianguo (Jeff) Xia
Assistant Professor, Department of Bioinformatics and Big Data Analysis
McGill University

Rong Fan
Associate Professor, Department of Biomedical Engineering
Yale University

Koichi Takahashi
Associate Professor, Department of Leukemia, Division of Cancer Medicine
The University of Texas MD Anderson Cancer Center



Nikolai Slavov
Associate Professor, Department of Bioengineering
Northeastern University

Andy Sharrocks
Professor, Division of Molecular and Cellular Function
University of Manchester

Andrew Smith
Assistant Professor
University of Milano-Bicocca

David Ruau
Head of Strategic Alliances, Drug Discovery AI
NVIDIA

Luciano Martelotto
Associate Professor, Single-cell and Spatial-Omics Lab, Adelaide Centre of Epigenetics
University of Adelaide

Kerstin Meyer
Principal Staff Scientist
Wellcome Sanger Institute

Anna Wilbrey-Clark
Staff Scientist
Wellcome Sanger Institute

Mathew Chamberlain
Principal Scientist
Janssen

Alex Tamburino
Director, Spatial and Multiomics Single-Cell Sequencing Lead
Merck Research Labs

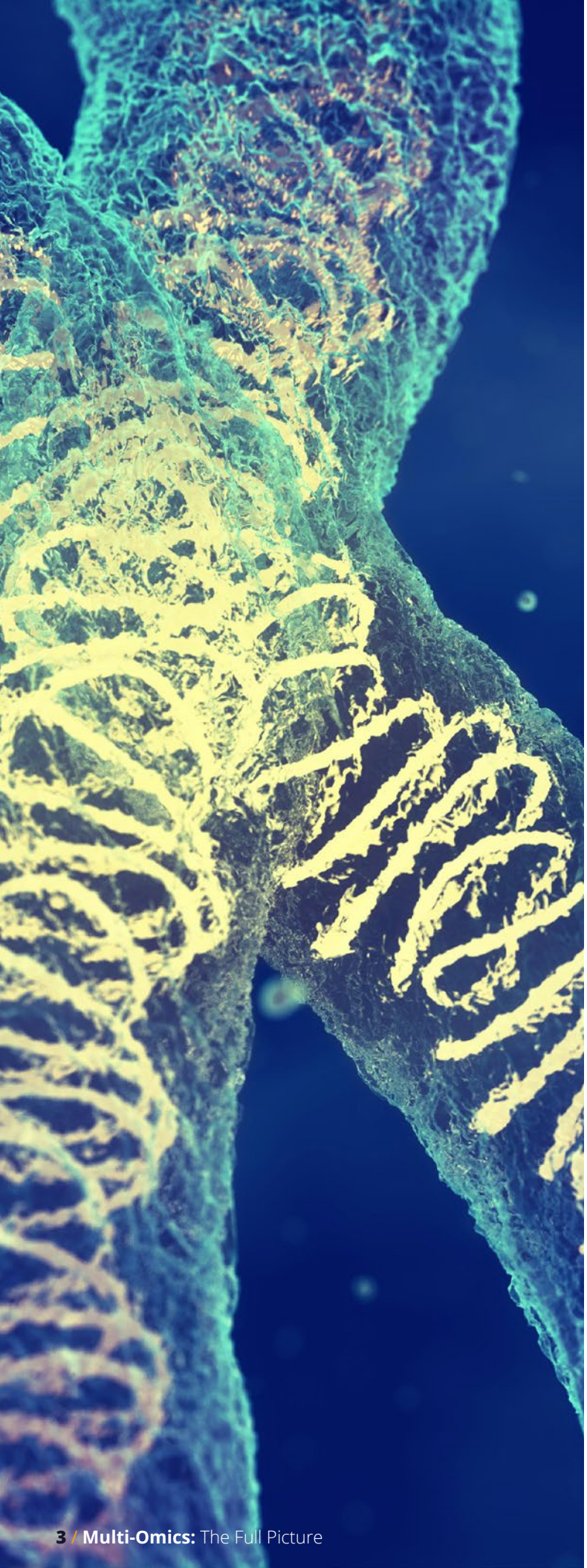
Rebecca Mathew
Principal Scientist
Merck Research Labs

Jeffrey Moffitt
Assistant Professor, Department of Microbiology
Harvard Medical School

Marshall Summar
Director, Rare Disease Institute Laboratory
Children's National Hospital (Washington D.C.)

Wanze Chen
Principal Investigator
Shenzhen Institute of Advanced Technology

Eleanor Matthews
PhD student, MRC iCASE DTP
University of Manchester



FOREWORD

Multi-omics is the combining of different "omes" – the genome, epigenome, transcriptome and proteome. Studying each layer in isolation can only colour in part of the picture. By bringing all these different layers of biological insight together, we can begin to paint a more complete picture of human biology and disease.

Over the last few years, rapid developments in this space have thrust multi-omics into the research spotlight and brought significant breakthroughs across a number of scientific disciplines. Therefore, we decided now was the time to release our most comprehensive report yet: "Multi-Omics: The Full Picture".

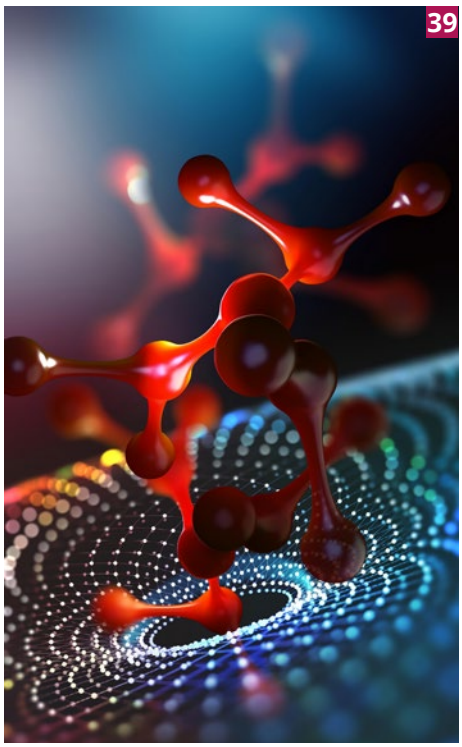
That said, it would be near impossible to cover every aspect of multi-omics in detail here. Instead, we have tried to focus on the topics that you, our audience, find most pressing. We start by exploring how each "omic" technique contributes a unique layer of biological insight (as well as their specific considerations and challenges) with advice and comments from the experts in this space. We then highlight recent advances, such as single-cell and spatial omics, and discuss what these technologies bring to the table. Later, we cover case studies that showcase just how powerful multi-omics can be. We also discuss how to integrate these different layers of analysis together. In our discussion roundtable with top developers and researchers, we go through the specific challenges in data integration and bioinformatics, as well as the emerging role of machine learning and AI in this space. Finally, we discuss innovative approaches that allow us to bring the next dimension – time – into the mix and discuss what the future has in store for this field.

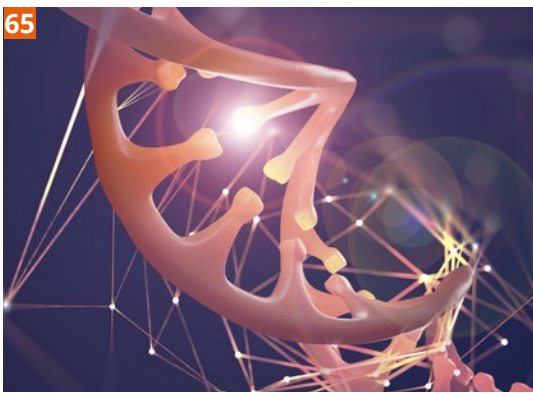
This report will not be an exhaustive list of different approaches for each step in a multi-omics workflow – as this field is constantly evolving, it would soon become outdated. Instead, we focus on bringing you expert advice from our contributors (as well as the rationale, the considerations, and the challenges involved) in a bid to show you why multi-omics is such a powerful tool in our journey to understanding human health and disease.

A huge thank you to all our contributors as well as our sponsors (Canopy Bioscience, Mission Bio, NanoString, Novogene and NVIDIA) for their time, advice and insights on all things multi-omics.

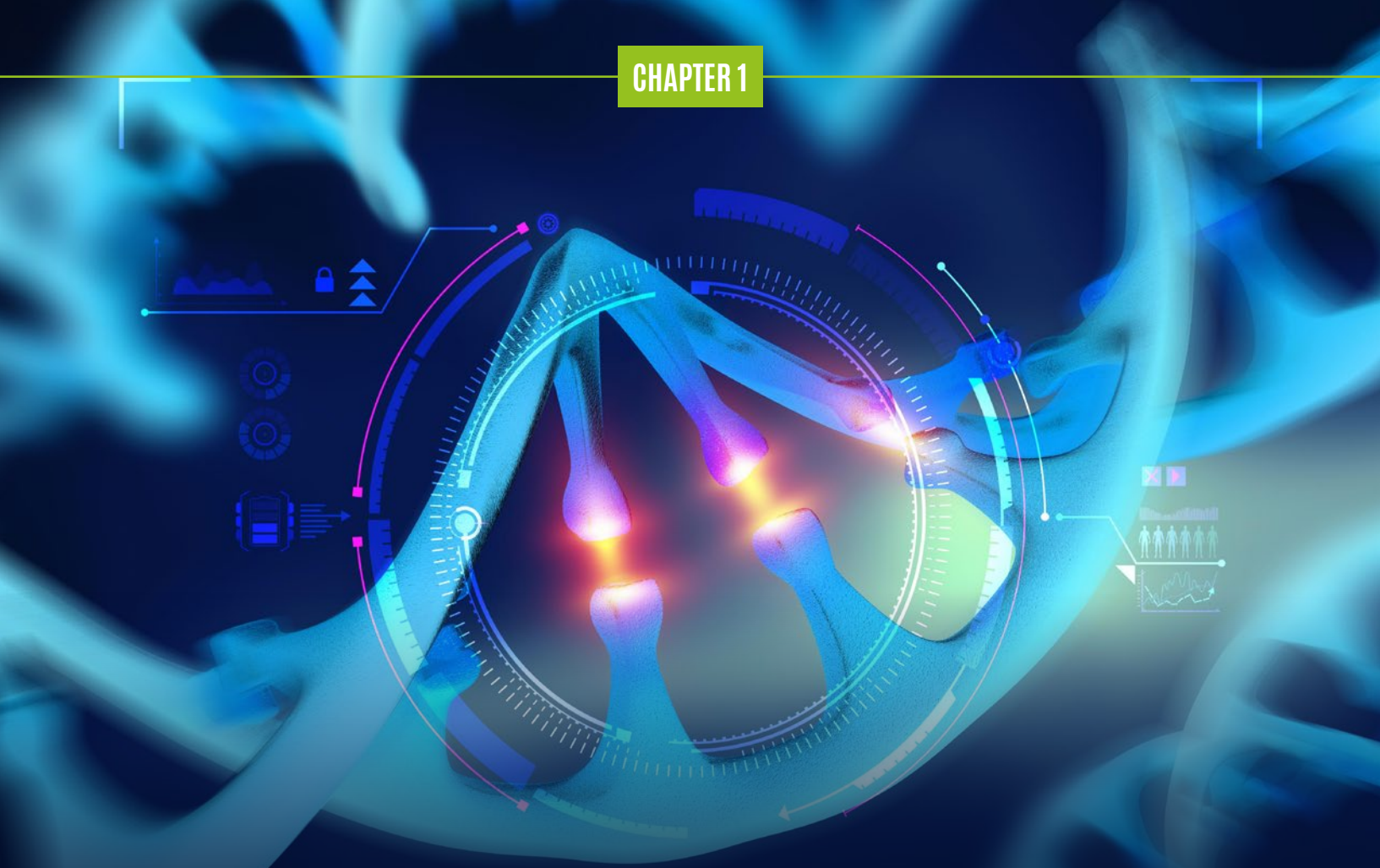
Miyako Rogers
Science Writer
Front Line Genomics

CONTENTS





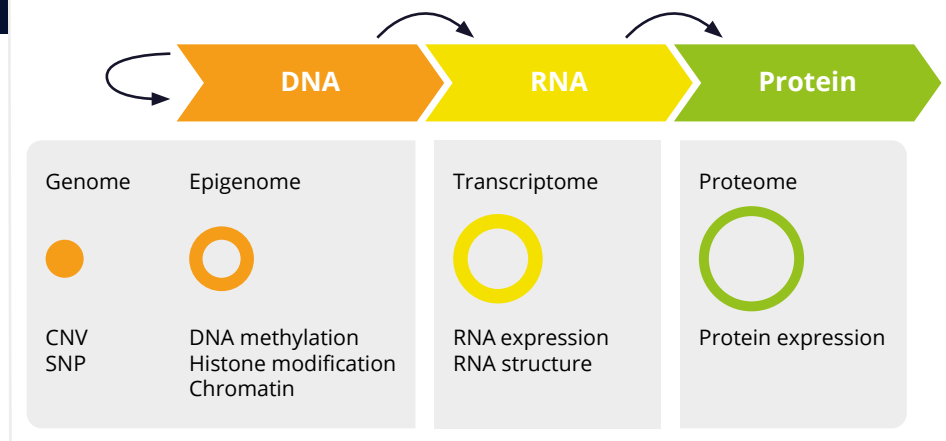
CHAPTER 1



THE DIFFERENT OMICS TECHNIQUES

A MULTI-OMICS APPROACH ALLOWS RESEARCHERS TO UNTANGLE DISEASE MECHANISMS, DEFINE DISEASE SUBTYPES, IDENTIFY POTENTIAL TARGETS FOR DRUGS, AND MUCH MORE.

FIGURE 1: SHOWING THE DIFFERENT OMES AND THE BIOLOGICAL LAYER THEY REPRESENT[1]



The simultaneous study of each "omic" can provide a more accurate, holistic, and representative understanding of the complex molecular mechanisms that underpin our biology.

However, each "omic" comes with its own set of considerations – and you can't have an onion without all the layers! So, in this chapter, we will look at each 'omic' in detail – the what, the why, the how and the challenges.

GENOMICS

LET'S START WITH GENOMICS – THE STUDY OF OUR FIRST LAYER, THE GENOME. WHILST NOT THE NEWEST, SHINIEST "OMIC" ON THE BLOCK, OVER THE PAST 15 YEARS THE TECHNOLOGICAL AND SCIENTIFIC ADVANCEMENTS MADE IN GENOMICS HAVE BEEN STAGGERING.

Fundamentally, genomics investigates the structure, function, mapping, evolution and editing of information coded in our (and other species) genomes. That includes single nucleotide variants (SNVs), indels, insertions, deletions, copy number variations (CNVs), duplications, inversions... the list goes on. In the past decade, genomics has allowed us to predict, diagnose and treat diseases in a more unbiased and precise way than we ever could before. And in research, genomics has revealed the genes or mutations involved in thousands of different phenotypes, biological processes and diseases. This has allowed us to identify new biomarkers, new drug targets and so much more. [2]

Progress in the genomics space has been rapid – in 2009, it cost about 10 million USD to sequence a genome. Today, it can cost a mere 100 bucks. [3] But, as always, **the insights you gather from genomics depends on the approach you take.** Read on and arm yourself with the latest information, insights, and expert tips and tricks from the genomics space.

"PROGRESS IN THE GENOMICS SPACE HAS BEEN RAPID – IN 2009, IT COST ABOUT 10 MILLION USD TO SEQUENCE A GENOME. TODAY, IT CAN COST A MERE 100 BUCKS."

THINGS TO CONSIDER: THE RIGHT SEQUENCING METHOD

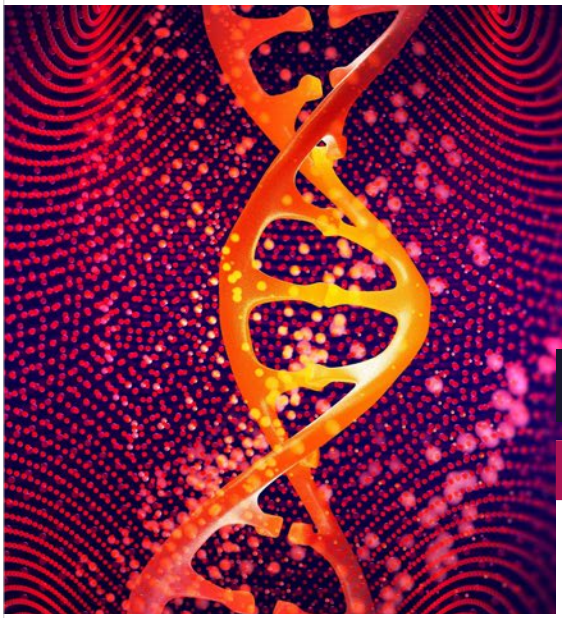


TABLE 1: SUMMARY OF THE ADVANTAGES AND DISADVANTAGES OF SHORT-READ AND LONG-READ SEQUENCING [4]

	Short-read	Londgread
Advantages	High sequence accuracy Scalable (high throughput data generation) Low cost Able to sequence fragmented DNA	Lack of amplification Easier library preparation De novo sequencing Start with larger DNA fragments Allows for unbiased discovery of novel mutations, etc.
Disadvantages	Only capable of reads between 200-300 bases long Not able to resolve structural variants or distinguish highly homologous genomic regions Not suitable for analysis of sequences that contain large numbers of repetitive sequence elements, transcript isoforms or methylation signatures	Signals obtained from individual fragments may be weak Lower accuracy overall More expensive

Selecting the right sequencing method is vital in any genomics analysis – and **this all depends on your biological question.** Today's sequencing methods can be split into short-read or long-read technologies, the advantages and disadvantages of which are summarised in Table 1. Short-read uses next-generation sequencing (NGS), also known as second-generation massively parallel sequencing, and is dominated by Illumina. Long-read sequencing, sometimes called third-generation sequencing, uses technology developed by Oxford Nanopore Technologies and Pacific Biosciences. [4]

Short-read (NGS) has the advantage of speed, scalability, lower cost and higher accuracy. Long-read has the advantage of de novo genome assembly and full-length isoform sequencing. **There is also the possibility of integrating both approaches together, to get the best of both worlds.** [4] There are many different NGS approaches now available on the market. For a comprehensive list, check out The Sequencing Buyers' Guide.

Koichi Takahashi
Associate Professor,
Department of Leukemia,
Division of Cancer Medicine,
The University of Texas MD Anderson Cancer Center:

Targeted sequencing may compromise discovery potential because you're looking for what's already known to be there, or has a high chance of being there. However, it is very useful in a clinical setting – in translational research, where we have some specific clinical question we want to answer. So if you're research is focused more on scientific discovery, a targeted sequencing approach might not be the application you want. **It really depends on the question that you are asking.**

However, whilst genomic data has been used in many studies and has led to countless scientific discoveries, there are limitations to only looking at this one layer – particularly when it comes to cancer.

BEYOND CANCER GENOMICS: THE LIMITATIONS OF A GENOMICS-ONLY APPROACH

CANCER IS OFTEN CHARACTERISED AS AN INHERENTLY GENETIC DISEASE. WHILE THIS IS ACCURATE, **LOOKING AT GENOMICS ALONE DOESN'T TELL THE WHOLE STORY.** WE SPOKE TO **KOICHI TAKAHASHI**, ASSOCIATE PROFESSOR, DEPARTMENT OF LEUKEMIA, DIVISION OF CANCER MEDICINE, **THE UNIVERSITY OF TEXAS MD ANDERSON CANCER CENTER** ABOUT HIS WORK TO CHARACTERISE LEUKAEMIA AND THE LIMITATIONS OF A GENOMICS-ONLY APPROACH WHEN INVESTIGATING CANCER.

FLG: You've done a lot of work characterising the heterogeneity of cancer and investigating drug resistance in cancer. How can multi-omics, compared to just genomics, better capture that heterogeneity, and also give us a deeper understanding of the mechanisms underlying drug resistance?

Koichi Takahashi: Heterogeneity can be understood based on what you look for. If you only look for genetics, then you can understand the genetic heterogeneity, but heterogeneity can be driven by other factors, such as phenotypic heterogeneity or epigenetic heterogeneity. If you don't look for it, you can't really understand it. So multi-omics analysis combining genomics with proteomics, transcriptomics, or epigenomics gives multiple layers of information which definitely increases the chance of you understanding the heterogeneity of cancer better. Moreover, only looking at one layer doesn't necessarily tell you about the mechanisms of drug resistance. **Cancer is essentially a genetic disease, so a lot of times genetics is the target of analysis. However, sometimes the resistance mechanism, or the reason why disease relapses or progresses, cannot be explained by genetics alone.**

"IF WE LOOK AT MECHANISMS OF DRUG RESISTANCE – BECAUSE THESE TUMOURS ARE SO FLEXIBLE IN CREATING DIFFERENT GENETIC SUBCLONES, IF OUR DRUG THERAPEUTIC STRATEGY IS DRIVEN BY GENETICS, DESIGNED BASED ON A GENETIC ABNORMALITY, IT'S ALMOST DESTINED TO FAIL, BECAUSE CANCER CELLS HAVE THE CAPACITY TO ACHIEVE THE SAME PHENOTYPE THROUGH MANY DIFFERENT GENETIC PATHWAYS."

FLG: You recently spoke at the Tri-Omics Summit USA. In your talk, you spoke a little about convergent evolution. Could you explain convergent evolution to us, how it contributes to drug resistance and how a genomics-only approach limits our understanding of this process?

Koichi Takahashi: Convergent evolution is an interesting evolutionary process in cancer development because at the genetic level, there is a lot of heterogeneity there, meaning that each clone, each subclone, has different combinations of mutations. However, most likely at the transcriptomic or phenotype level, they converge into one theme. So yes, there is genetic heterogeneity, but for the cancer as a whole, the phenotype, and what each subclone is trying to achieve is essentially the same. Often times, convergent evolution is an indication of the presence of strong selective pressure. Thus, this type of evolution is often seen in drug-resistance situation.

If we look at mechanisms of drug resistance – because these tumours are so flexible in creating different genetic subclones, if our drug therapeutic strategy is driven by genetics, designed based on a genetic abnormality, it's almost destined to fail, because cancer cells have the capacity to achieve the same phenotype through many different genetic pathways. If your drug only targets one specific genetic abnormality, the cancer has an inherent way to escape that mechanism. So convergent evolution is a source of drug resistance and I think it is key for the next generation of therapeutic development because the drugs we develop currently are increasingly targeted to specific genetic abnormalities. For instance, KRAS G12C targeting drug.

FLG: You gave KRAS G12C inhibitor as an example of a drug that was developed to target a specific gene.

Koichi Takahashi: KRAS G12C inhibitor is a really good example of cancer developing resistance against a drug that was developed to target a specific genetic abnormality. Cancer has so many other ways to upregulate RAS/MAPK pathways. So if you inhibit G12C mutation, then they generate other types of KRAS mutations or mutations in RAS/MAPK pathway genes to escape from the therapeutic effect.

FLG: It's essentially an arms race!

Koichi Takahashi: Exactly, especially if the target mutation is a late mutation, or the mutation at the sub-clonal level. The cancer has a tendency to create another subclone, with the same phenotypic characteristics.

FLG: Following on from what you just said, could a multi-omics approach be better for disease stratification/subtyping as well?

Koichi Takahashi: Yes, absolutely. With disease, and the way we understand things, we always want to group things together, right? We want to categorise things into one diagnosis because it is convenient. Let's say patient A has one disease and patient B also has a disease that looks very similar to the one with patient A, we can group them together and diagnose as the same disease and treat the same. **However, the truth of the matter is one individual's disease A is actually different from another individual's disease B, and with multi-omics we can analyse and start to dissect the unique features of each individual's disease. Yes, their clinical diagnosis may have the same name, but it actually presents as a totally different disease.** Subtyping in this way is really the driver for the development of personalised medicine. So multi-omics analysis will definitely be helpful in accelerating personalised medicine.

Novogene

NGS solutions for multi-omics research

Novogene is a leading provider of genomic services and solutions with cutting edge Next Generation Sequencing (NGS) and bioinformatics expertise. With one of the largest sequencing capacities in the world, we utilise our **deep scientific knowledge, first-class customer service and unsurpassed data quality** to help clients succeed on their multi-omics journey. DNA is the foundation of life and the foundational building block of multi-omics research, and with extensive experience in transcriptomic, epigenomic and genomic sequencing services, Novogene is here to assist our customers in laying those foundations.

Transcriptomics

Analysis of all RNA types to provide a global map of transcript levels and interactions

Genomics

Comprehensive genome coverage for identification of SVs, SNPs, CVs and InDels

Epigenomics

WGBS, ChIP-seq & RIP-seq for methylation and protein interaction analysis

illumina®

Oxford NANOPORE Technologies

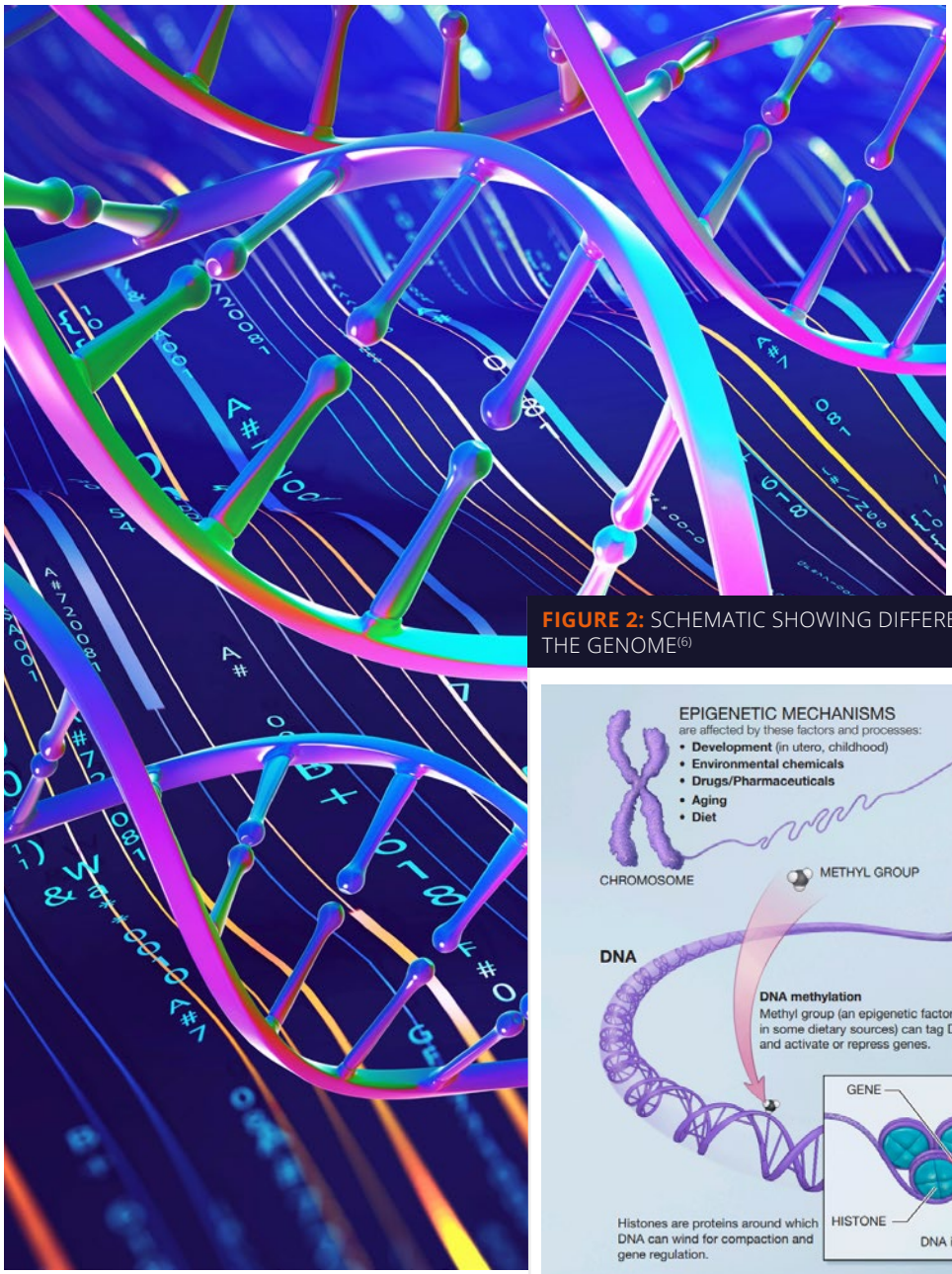
PacBio

Search ‘Novogene Europe’

Find out more: novogene.com

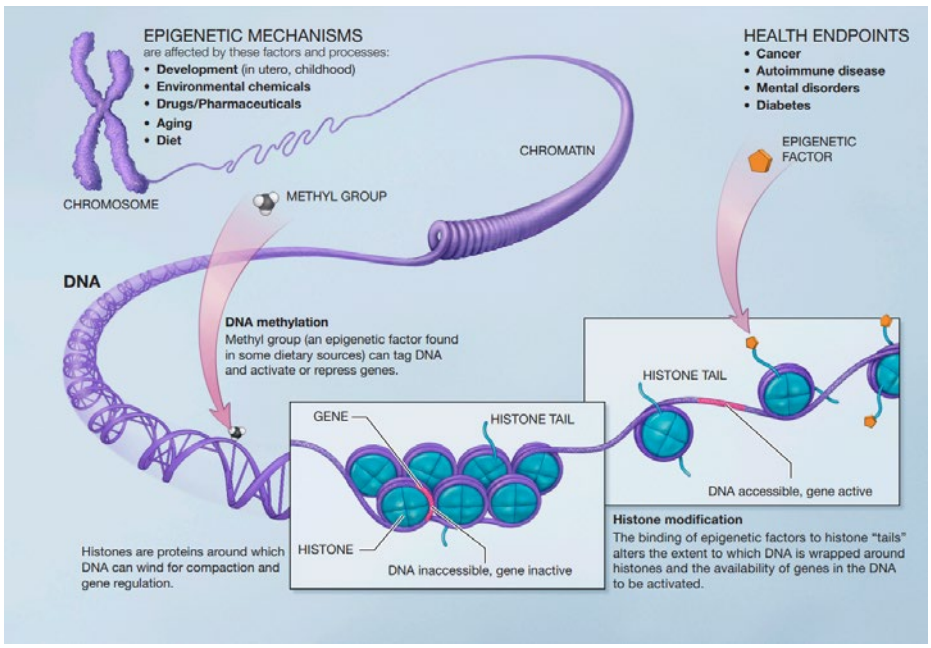
EPIGENOMICS

NOW WE MOVE ON TO THE LAYER "UPON" THE GENOME, APTLY NAMED WITH THE SUITABLE GREEK PREFIX – THE EPIGENOME. THE EPIGENETICS REVOLUTION WAS IN ITS HEYDAY IN THE 2010S, BUT ADVANCES IN SPATIAL OMICS HAVE MADE THE ROLE OF THE EPIGENOME ALL THE MORE RELEVANT.



Epigenomics investigates modifications of DNA or DNA-associated proteins, such as DNA methylation, chromatin interactions and histone modifications. Epigenetic regulation of DNA can determine cell fate and function, and the epigenome can change based on the environment. What's more, these DNA alterations can be passed on. These changes can act as markers for cancer, metabolic syndromes, cardiovascular disease and more. **They can be tissue-specific, cell-specific and even more specific than that – down to subcellular compartments – and changes can occur during both healthy and disease states.** [5]

FIGURE 2: SCHEMATIC SHOWING DIFFERENT FEATURES OF EPIGENETIC REGULATION OF THE GENOME[6]



THINGS TO CONSIDER: DIFFERENT APPROACHES

"DEPENDING ON THE BIOLOGICAL QUESTION YOU ARE INVESTIGATING, WHICH APPROACH YOU CHOOSE VARIES."

One of the first things to consider is the type of epigenomic analysis to conduct. Depending on the biological question you are investigating, which one you choose varies. There are several types of epigenomic analysis, with some taking advantage of the developments in next-generation sequencing.[5] These analyses can investigate different features of the epigenome:

Methylation sequencing: Cytosine methylation affects gene expression and chromatin remodelling and can be used to investigate the methylation status of the genome with single-nucleotide resolution.

- WGBS-Seq and RRBS-Seq (bisulfite dependent) for site-specific studies
- TAPS and EM-Seq (bisulfite free) for low-resolution, large-scale studies
- TAPS: Tab-Seq, Tab/OxBS Array
- MeDIP-Seq and MBD-Seq (affinity enrichment) for high-resolution, whole-genome studies
- MRE-Seq (endonuclease digestion) for high-resolution, whole-genome studies

ChIP-seq: Chromatin immunoprecipitation sequencing combines immunoprecipitation assays with sequencing to identify genome-wide DNA binding sites for transcription factors and other proteins. NGS.

ATAC-seq: Assay for transposase-accessible chromatin with sequencing to determine chromatin accessibility across the genome. Helps uncover how chromatin packaging and other factors affect gene expression. It can be used for nucleosome mapping, transcription factor binding analysis, novel enhancer identification, exploration of disease-relevant regulatory mechanisms, cell type-specific regulation analysis, and biomarker discovery.

HiC/3C/Capture-C: Analyses chromatin interactions. Hi-C extends 3C-Seq to map chromatin contacts genome-wide, and it has also been applied to studying in situ chromatin interactions. Capture-C to the 3C method with pull-down of the biotinylated fragments with magnetic beads.[5]

Andy Sharrock
Professor, Division of Molecular and Cellular Function
University of Manchester:

The DNA is encapsulated in chromatin, and you need to then remodel the chromatin to get changes in gene expression. So, you're revealing enhancer elements, promoter elements and other regulatory elements in the DNA to allow access to the machine that turns on gene expression. Equally, you can do the opposite; you can close down chromatin and closing down chromatin will then shut down gene expression, and that will typically shut down particular tumour suppressors that usually stop tumorigenesis. Shutting those down allows cancer to occur. You can also open up things, which tend to be oncogenic processes and the oncogenic proteins that are produced for these processes.

Integration of epigenomic data can be challenging, especially when there are no overlapping genes, as this is the simplest way to confirm gene expression. However, gene superposition methods can be avoided if needed by using direct and indirect functional analyses. For this, an interactome network needs to be developed to allow us to understand the direct relationships between genome and epigenome. The downside is that this networking method requires a reference database, so it is not suitable for rare diseases and species.[5]

Again, like with genomics, epigenetics doesn't involve processes that are as dynamic as the other omics. However here, spatial context is very important – this is why the development of spatial ATAC-seq is so revolutionary, which you can read more about in Chapter 3.

WHY STUDY EPIGENETICS?
EPIGENETICS AND CANCER

SO, WHAT INSIGHTS CAN EPIGENETICS GIVE YOU THAT THE OTHER OMICS CAN'T? WE ASKED **ANDY SHARROCKS**, PROFESSOR, DIVISION OF MOLECULAR AND CELLULAR FUNCTION, **UNIVERSITY OF MANCHESTER** WHO INVESTIGATES THE EPIGENETICS OF OESOPHAGEAL ADENOCARCINOMA, WHY STUDYING EPIGENETICS IN CANCER IS SO USEFUL.

FLG: Why investigate the epigenetics of oesophageal adenocarcinoma in particular? How does this disease differ from other forms of cancer?

Andy Sharrocks: Esophageal adenocarcinoma is a particularly deadly disease with high incidence and very poor survival, and part of the reason for this is because we don't understand the molecular pathways involved as much as we do in many other cancers. One of the issues with esophageal adenocarcinoma is it is a highly mutated cancer type, and yet there are no recurrent mutations. P53 is highly commonly mutated, as in many cancers, but the mutation with the next highest incidence is probably the receptor tyrosine kinase – up to 20 to 30%. After that, the usual common drivers aren't there in this cancer.

Signature common mutations like BRAF in melanoma, APC in colon cancer, or oncogenic fusions in leukaemia, are just not there, so getting targeted treatments becomes difficult. At the pathway level, you can see things slightly differently. Receptor tyrosine kinase pathways, for example, if you take different components and look at mutations within those, then you see a much higher prevalence so maybe 60 to 65% of all tumours would have mutations in this pathway. But again, the common DNA mutations just aren't there.

FLG: Focusing on epigenetics specifically, what insights and advantages could studying epigenetics have in our understanding of this and other cancers?

Andy Sharrocks: There are very few common DNA mutations between different types of cancer. **The next obvious thing to look at is epigenetics.** DNA is not naked in a cell, it's enclosed and encased in chromatin which controls the availability of regulatory elements that in turn control gene expression. So equally, **epigenetic changes may cause a change in cell phenotype,** which gives you a cancer phenotype. The other thing about epigenetics is it can also give you an idea of where the cells come from in the first place, i.e., **the cellular origin of cancers.** That's something we try and pursue in our research, trying to understand the **basic wiring of cancer cells and where that rewiring originates in the first place.**

FLG: Do you anticipate epigenetic insights into this cancer and other cancers to translate into patients in the future? Will these studies help cancer patients in the clinic in the near future or do you anticipate that future being far off?

Andy Sharrocks: Well, there's multiple answers to that question. What epigenetics can give you is an insight into tumours which **other types of analysis don't particularly give you.** While transcriptomics, DNA mutation analysis and typical genomics, for example, can give you one answer, epigenomics can reveal new things. In a way, **that's giving you new pathways, new therapeutic targets and potentially new diagnostic targets.** There are techniques which come and go; we can look at circulating tumour DNA from patients, for example, and you can then observe fragmentation patterns. From that, you can infer the chromatin state of the patient. This can help you begin to understand the pathways that are changed in patients. Using this as a diagnostic tool, you can A) tell if the patients have cancer, so it's a non-invasive biological tool from blood. B) you can then begin to understand what those changes might be. So yes, there are potential future uses of epigenetics and revealing new things, but there is also the potential for new diagnostic approaches through non-invasive blood sampling like liquid biopsies.

FLG: You mentioned diagnostics, but do you think these epigenetic insights could also be used to inform precision medicine treatments?

Andy Sharrocks: Again, that's possible. One of the beauties of doing epigenetics of the sort that we do is we use open chromatin accessibility assays. **What that does is reveal not just individual genes, but also programs.** So, if you get a transcriptional regulatory program regulated by a particular transcription factor, it gives you an insight with which to intervene in that pathway. Obviously, targeting transcription factors isn't as easy as targeting signalling pathways, but people are doing it which means it can be done. You just have to think a little bit out of the box to be able to do it.

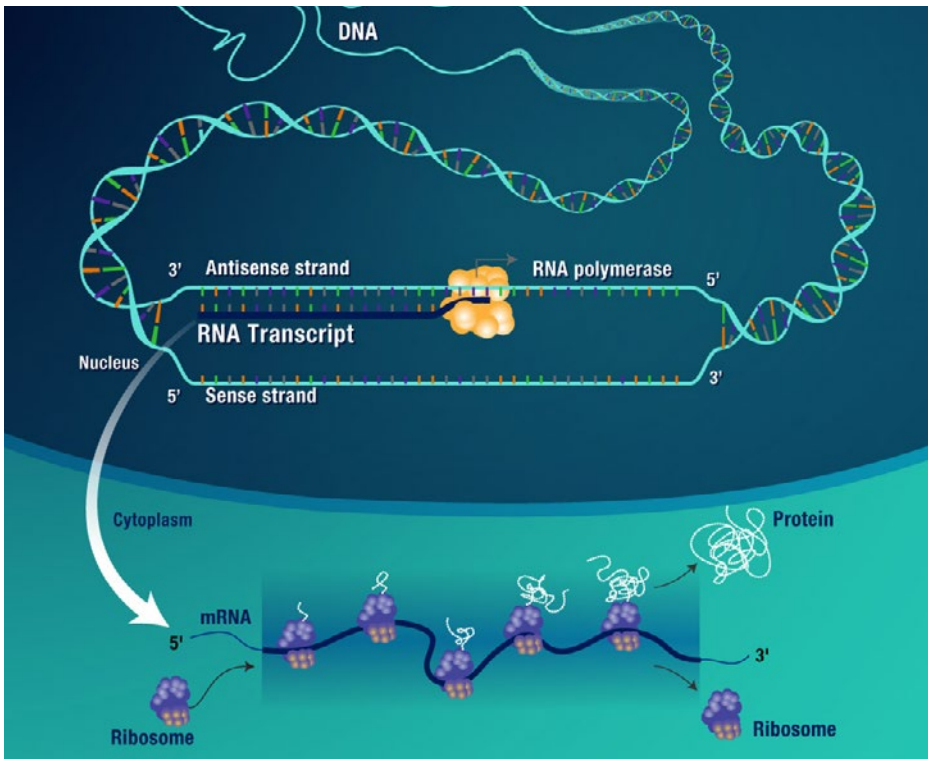
TRANSCRIPTOMICS

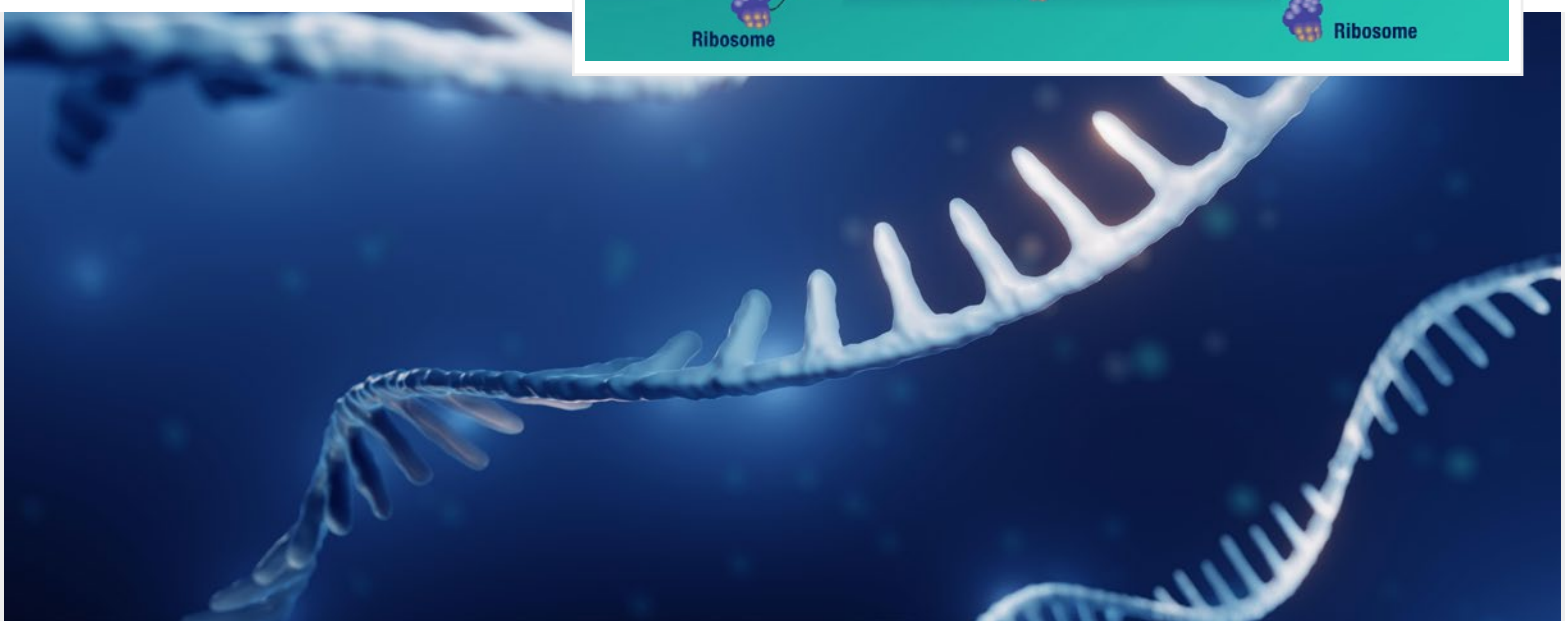
AND WHAT HAPPENS TO THOSE GENES? THEY DON'T JUST SIT THERE GATHERING DUST, THEY GET TRANSCRIBED! THIS LEADS NICELY ON TO OUR NEXT OMIC – TRANSCRIPTOMICS. AND NO, IT DOESN'T JUST EXCLUSIVELY APPEAR WITH THE WORD SPATIAL IN FRONT – THOUGH WE WILL BE COVERING THAT IN OUR NEXT CHAPTER.

Transcriptomics involves investigating RNA transcripts that are produced by the genome and how these transcripts are altered in response to regulatory processes. **It's the bridge between genotype and phenotype – the link between the genes and the proteins.** Sandwiched nicely in the middle, it can tell us a lot about our biology. [7]

RNA-seq analysis has become a staple in many labs now. However, the development of NGS has allowed for high-throughput RNA-seq analysis of targeted regions. Moreover, the development of long-read sequencing has now allowed researchers to conduct whole transcriptome analyses. This can help discover new RNA transcripts that have previously eluded analysis, as well as allow for better genome annotation and the identification of long non-coding RNAs or fusion transcripts. [7]

FIGURE 3: FIGURE ILLUSTRATING TRANSCRIPTION OF GENES INTO RNA[8]





THINGS TO CONSIDER: HIGH DIMENSIONALITY

RNA-seq is the method of choice for analysing the transcriptomes of disease states, biological processes and more. RNA-seq has a broad dynamic range, has very sensitive and accurate measurements of fold changes in gene expression and can be applied across a wide range of species. However, there is a lack of standardisation between sequencing platforms and read depth, which can compromise the reproducibility of this analysis. Whole transcriptome analysis captures both known and novel features, allows researchers to identify biomarkers across the broadest range of transcripts and enables a more comprehensive understanding of phenotypes of interest. [7]

Compared to the genome and epigenome, the **transcriptome is much more dynamic and highly dimensional.** Therefore, spatial context is very important when studying the transcriptome, which is likely one of the reasons why spatial transcriptomics was the first to emerge and become popularized. In the context of many diseases and normal biological processes **the neighbouring cells and the surrounding environment can alter the transcriptome be that directly or indirectly.** [9] To learn more about spatial, read on to Chapter 3.

Transcriptomic events can change rapidly over time, so making sure things **line up temporally** when integrating the transcriptome with the other omics can be a **major challenge.** [10] This is why it's especially important to consider your integration approach at the very beginning of your study. We cover advances in multi-omics that have enabled us to better profile biological events as they change over time in Chapter 7.

"SPATIAL CONTEXT IS VERY IMPORTANT WHEN STUDYING THE TRANSCRIPTOME, WHICH IS LIKELY ONE OF THE REASONS WHY SPATIAL TRANSCRIPTOMICS WAS THE FIRST TO EMERGE AND BECOME POPULARIZED."

HEAR FROM THE EXPERTS: TACKLING DYNAMIC DATA

WE ASKED OUR EXPERTS HOW THEY GO ABOUT INTEGRATING HIGHLY DIMENSIONAL AND DYNAMIC DATA, SUCH AS TRANSCRIPTOMIC DATA. THEY TOLD US WHY ENSURING EVENTS LINE UP TEMPORALLY IS SO IMPORTANT, AND HOW THEY TACKLE THIS CHALLENGE.

JOHN QUACKENBUSH
Professor, Department of Computational Biology and Bioinformatics
Harvard T.H. Chan School of Public Health

LIHUA (JULIE) ZHU
Professor, Department of Molecular, Cell and Cancer Biology, **University of Massachusetts Chan Medical School**

ANDREW SMITH
Assistant Professor
University of Milano-Bicocca

ANDY SHARROCKS
Professor, Division of Molecular and Cellular Function
University of Manchester

RONG FAN
Associate Professor, Department of Biomedical Engineering
Yale University

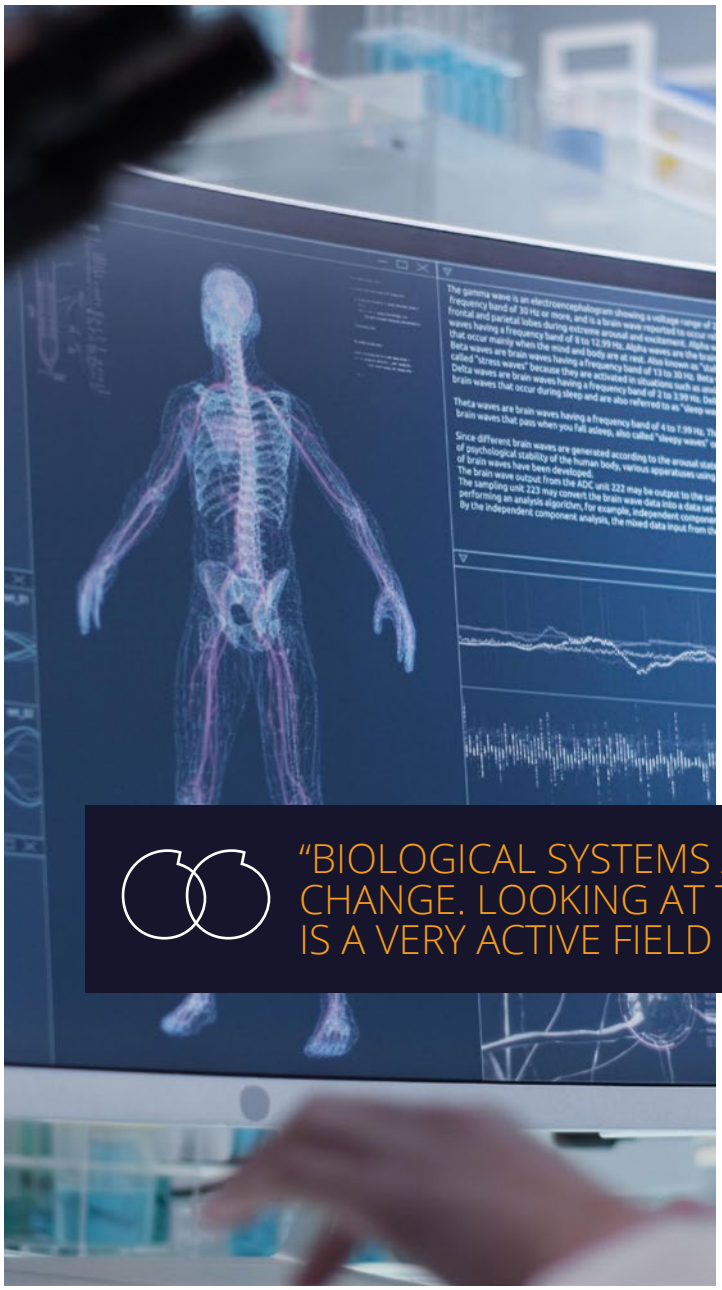
John Quackenbush: One of the most important things is to **try to get these multi-omics datasets on the same sample.** So, you know, if I'm looking at gene expression in me and protein expression in you and trying to draw some parallels - at the surface, it doesn't make a lot of sense. But if I had gene and protein expression data on me and gene and protein expression on you and 10 other people or 100 other people, then I could start to look at relationships between RNA and protein in a biologically meaningful way. Part of being able to do this analysis comes back to asking the right biological questions, selecting the right biological data, and then using analytical methods that respect the limitations of the data set that exists between them.

Trying to incorporate protein and gene expression data out of the box does become really difficult and challenging. One major reason is a **batch effect.** They exist on very different scales in terms of the abundance levels that we see and interpret as either RNA or protein expression but the level of protein can't be compared directly to the level of RNA. And so, as we try to build models and analyse the data, our starting point is going to be understanding the limitations of those datasets.

Lihua (Julie) Zhu: There's definitely a danger in misaligning data, temporally speaking. For example, within different phases of cell cycle, gene expression changes. I think it all goes back to experimental design and making sure you don't compare apples to oranges. **Metadata annotation can really help with this.** I think it's also really challenging when the corresponding data isn't available – and you may be tempted to integrate some suboptimal data. Caution definitely needs to be taken.

Andrew Smith: In our lab, we use tissue which has been formalin-fixed as part of clinical routine, most commonly during the post-surgical phase. Therefore, once the pathological tissue has been excised, it is fixed immediately in formalin and, in theory, our biomolecules of interest are preserved so that they closely represent the natural, molecular, state of the tissue. Naturally, as much as possible, we try to ensure that all the samples are fixed at the corresponding moment, and for the same length of time, but when working within a clinical context, it can be sometimes challenging to achieve this.

Based on our experience, however, when working with the inherent variability that is to be expected with clinical tissue, these slight variations in fixation time do not appear to insert significant bias or artefacts within our molecular imaging data. Naturally, within the first few moments post-excision, the less stable molecules may degrade to some degree, **but this is part of the challenge of working with clinical tissue and both the experimental and bio statistical approach has to be sufficiently robust to account for these small variations that may occur within the tissue processing workflow.** It can be challenging, and is a key consideration when designing the experimental approach, but can be accounted for accordingly in my opinion.



Andy Sharrocks: In cancer, looking at things temporally can help us understand how things go from A to B into an intermediate state, especially if you're looking at signalling events. In oesophageal cancer, for example, we have the receptor tyrosine kinase signalling pathways upregulated and they're obviously signalling through to chromatin and gene expression. **One of the problems with looking at signalling is that it is dynamic, again we end up looking at end points relative to start points.**

There are new technologies like **LIVE-seq** that allow us to look at gene expression changes in real-time. You're able to extract part of the cytoplasm from individual cells and analyse that. **It's quite an exciting technology because you're able to sample the same cell multiple times.** At the moment, most single-cell technologies are not able to do that – we have to take snapshots. You can do that through multi-omics, in the sense that you can use it to connect things together in single cells. However, I think the new technologies where you're able to look at what's going on in a single cell over time are particularly exciting. You can't do that with chromatin changes at the moment and I can't envisage a way of currently doing it. But if we could do that, it'd be really exciting.

Rong Fan: **Biological systems are dynamic, they are in constant change.** Looking at time course, and event ordering, is a very active field of research in the omics space. There are several approaches that allow you to get some temporal information. One is computational approaches such as chromatin velocity and **RNA velocity.** If you're able to do single-cell unbiased base by base sequencing of whole transcriptome, you can look at relative ratios of spiced and unspliced RNAs to investigate how cells change from one state to another. So, **you can see how the cell differentiates from one type to another with data from one snapshot.** There have also been recent experimental approaches that allow you to generate time-dependent data such as **LIVE-seq.** So what I will say is that this is a very active field of research and **multi-omics is a very powerful tool and a very viable solution to accurately address this time course question.**

"BIOLOGICAL SYSTEMS ARE DYNAMIC, THEY ARE IN CONSTANT CHANGE. LOOKING AT TIME COURSE, AND EVENT ORDERING, IS A VERY ACTIVE FIELD OF RESEARCH IN THE OMICS SPACE."

To learn more about RNA velocity and LIVE-seq, as well as other computational and experimental approaches which tackle the question of time, check out Chapter 6.

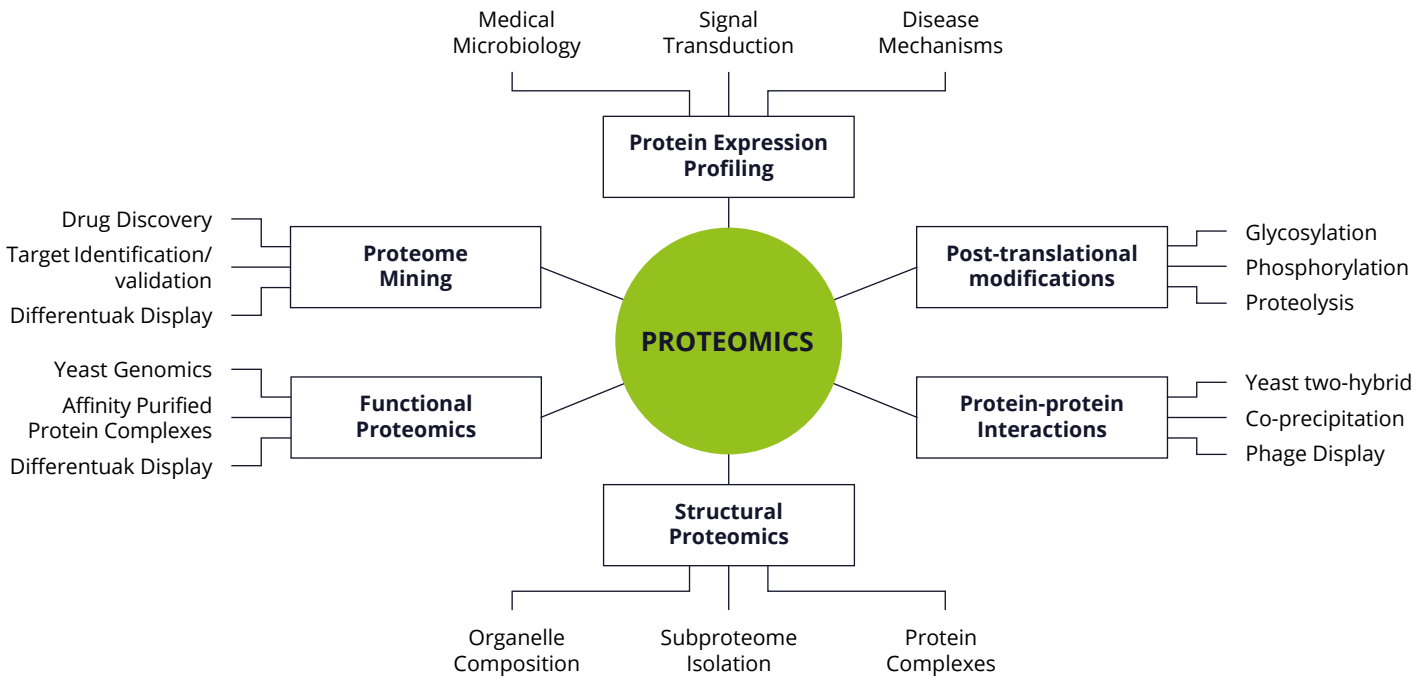
PROTEOMICS

ONCE THOSE GENES GET TRANSCRIBED, THE PROTEINS ARE PRODUCED. THIS IS WHERE PROTEOMICS COMES IN. PROTEOMICS IS PERHAPS THE MOST DYNAMIC AND SPECIFIED OMIC OF THEM ALL. TIME IS IMPORTANT. SPACE IS IMPORTANT. AND AS A RESULT, PROTEOMICS REPRESENTS THE FINAL FRONTIER OF SPATIAL, SINGLE-CELL, AND MULTI-OMICS ANALYSIS IN GENERAL.

So, what can you use proteomics for?

- **Functional annotation:** By linking genomic data to proteomic data, you can not only confirm the existence of a particular gene but see how it affects human health and disease.
- **Protein expression:** Understanding the mechanisms of several biological processes. Analysing the changes that take place in a cell and how they impact the disease and biology.
- **Protein localisation:** Proteins in the wrong place can lead to very wrong outcomes. Understanding the trafficking of proteins and creating a 3D map of the cell can lead to unparalleled insights.
- **Interactomics:** Cell growth, death, and homeostasis involve signals travelling as protein interacts with protein.
- **Protein structure:** Looking at protein structure can help us understand how they bind, interact and function.
- Researchers can also study **post-translational modifications** such as phosphorylation, acetylation, ubiquitination, nitrosylation, and glycosylation. These modifications are involved in maintaining cellular structure and function. [11]

FIGURE 4: SCHEMATIC SHOWING THE DIFFERENT INSIGHTS AND STUDIES THAT FALL UNDER THE PROTEOMICS UMBRELLA REFERENCE [11]



THINGS TO CONSIDER: WHAT IS MASS SPEC?

There's no doubt that proteomics is important. But it's not particularly approachable. A niche that's cultivated its own community and designated experts for some time, proteomics can often appear too difficult to integrate into your research. Moreover, whilst the other omics deal with sequencing, **mass spectrometry is the technology of choice here.**

Mass spectrometry measures the **mass-to-charge ratio of ions to identify and quantify molecules** in simple and complex mixtures. Methods can generally be segregated into either bottom-up or top-down. Simply put, bottom-up means that proteins are digested by proteolytic enzymes before analysed by mass spec. This approach has been around longer and is the most widely used approach. Top-down involves the characterisation of intact proteins. This allows for almost 100% sequence coverage and the characterisation of proteoforms. However, top-down is far more expensive and less efficient. [12]

> "WHILST THE OTHER OMICS DEAL WITH SEQUENCING, MASS SPECTROMETRY IS THE TECHNOLOGY OF CHOICE HERE."

Much like the transcriptome, the proteome is also highly dimensional and dynamic. Therefore, the considerations we listed above for transcriptomics also need to be taken into account when integrating proteomic data. Moreover, spatial context is again very important here, and recent developments in spatial proteomics have allowed researchers to see how neighbouring cells and the surrounding environment impacts protein expression, localisation and function. [13] To learn more about spatial, check out Chapter 3.

However, there are some considerations specific to proteomics – namely because of the use of mass spectrometry.

HEAR FROM THE EXPERTS: PROTEOMICS-SPECIFIC CHALLENGES

AS PROTEOMICS INVOLVES MASS SPECTROMETRY, THERE ARE SEVERAL CONSIDERATIONS YOU NEED TO MAKE IN A PROTEOMIC STUDY THAT YOU DON'T NEED TO ACCOUNT FOR WITH THE OTHER OMICS. WE WENT TO OUR EXPERTS TO ASK THEM WHAT **CRUCIAL POINTS RESEARCHERS NEED TO BEAR IN MIND WHEN WORKING WITH PROTEINS**, AND COVERED TOPICS SUCH AS **SEQUENCING DEPTH, DYNAMIC RANGE, DEPLETION AND BIAS.**

STEPHANIE BYRUM
Associate Professor, Department of Biochemistry and Molecular Biology
University of Arkansas at Little Rock

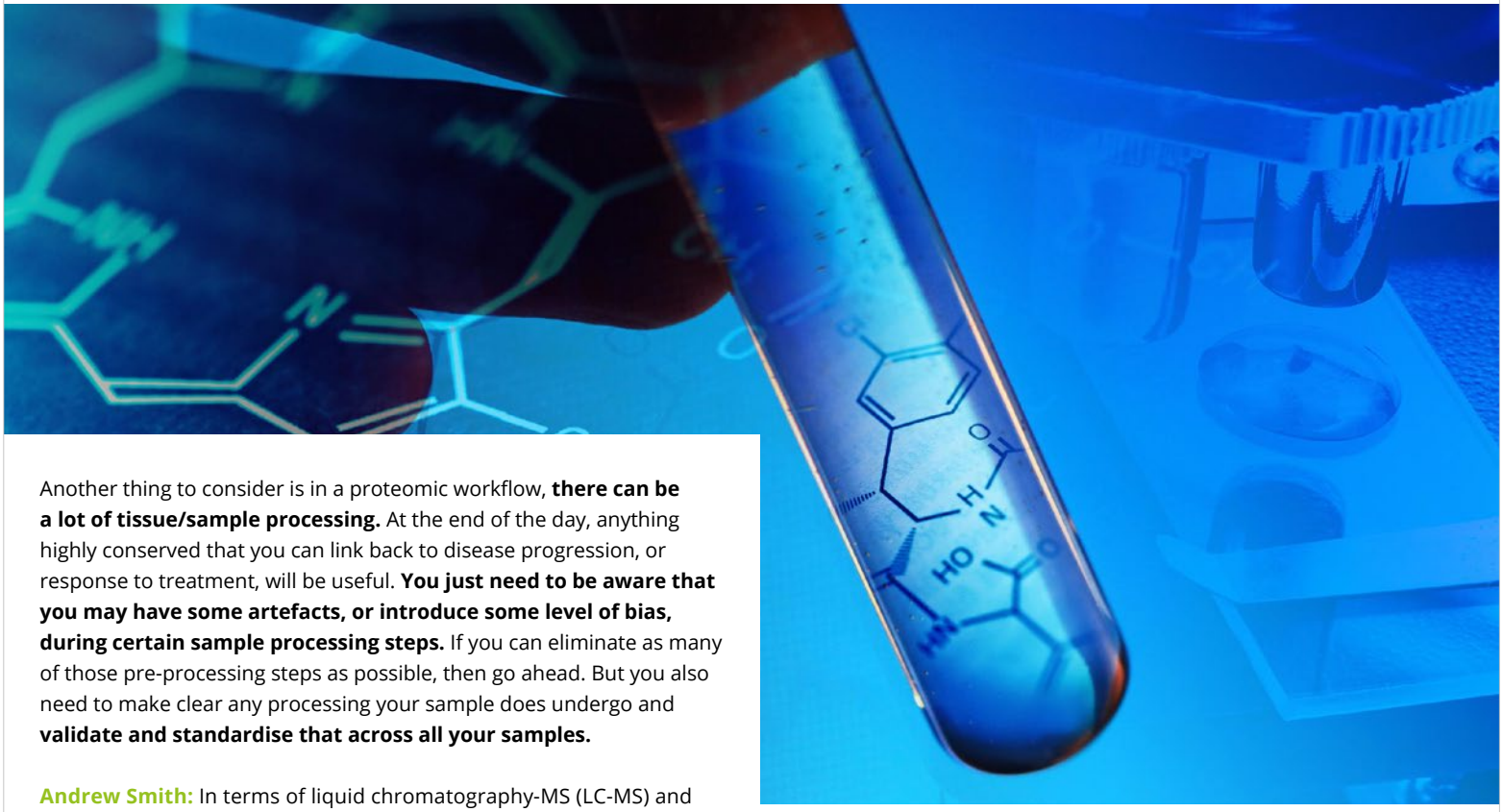
RONG FAN
Associate Professor, Department of Biomedical Engineering
Yale University

ANDREW SMITH
Assistant Professor
University of Milano-Bicocca

Stephanie Byrum: Well, one problem with including data types such as proteomics is **sequencing depth.** With RNA-seq, you can get 16,000 genes that you can identify, but the proteome is limited to 10,000 just based on the technology. **So, you're already automatically limiting your dataset by whatever your lowest amount of input is.** Until the sequencers catch up and can get us more and more data, we may have to do some in silico kind of prediction models or something.

It is only in the **last couple of years that mass spectrometers have gotten to the level to allow us to do integration at the protein level.** So, I think you're going to see a big explosion, we already are, with people trying to understand protein data and apply new methods to that. **But mass specs are very different from sequencing DNA and RNA because you're not actually sequencing the full protein.** You're looking at its mass-to-charge ratios, and then you have to match that back to libraries, spectral libraries, to come up with your sequence. So, the other complication with proteomics is modifications. You can have **a lot of data come out of the raw data files and the spectra, but if you're not searching for modifications, you're not searching for the right things.** We're only using a subset of the available data.

Rong Fan: **Measuring proteins gives us that direct evidence – it's the product of all the other biological layers.** But the proteome is much bigger, it's huge compared to the transcriptome. But it's very hard to look at the proteome in an **unbiased manner.** Earlier this year we released a preprint in BioRxiv showing we can spatially profile 300 proteins in conjunction with whole transcriptome in one analysis. It's a bit of a debate amongst scientists, but generally with mass spec, if you can detect more than 500 proteins, that's considered fairly unbiased and anything below that is considered targeted. Even in our case, with 300 proteins, it's pretty close to being considered "large-scale" proteomics. **So, integrating that with the other omics, where you get a lot more data, can be a challenge as things may not match up quite as nicely as you would like.**



Another thing to consider is in a proteomic workflow, **there can be a lot of tissue/sample processing.** At the end of the day, anything highly conserved that you can link back to disease progression, or response to treatment, will be useful. **You just need to be aware that you may have some artefacts, or introduce some level of bias, during certain sample processing steps.** If you can eliminate as many of those pre-processing steps as possible, then go ahead. But you also need to make clear any processing your sample does undergo and **validate and standardise that across all your samples.**

Andrew Smith: In terms of liquid chromatography-MS (LC-MS) and MS-based proteomics, yes, there are some specific challenges that can be faced. Steps such as depletion are particularly important for certain biological samples, including bio fluids such as plasma or urine, given that the proteins are present with a large degree of dynamic range. For example, urine contains a substantial abundance of albumin which must often be depleted in order to detect the lower abundant proteins which may be more functionally relevant. This is one particular characteristic of proteomics that should be considered, especially when employing biological fluids. However, with tissue-based or single-cell proteomics, this is somewhat less challenging because the dynamic range of those proteins residing in tissue is markedly reduced and thus, in the majority of cases, it is not necessary for this depletion step to be performed.

I believe that one of the biggest challenges that we have faced recently in terms of proteomics, particularly single-cell proteomics, is related to instrument sensitivity. We now have MS-based spatial approaches which are able to individuate single cells of a particular phenotype and, in these instances, we truly require instrumentation which has sufficient sensitivity. Taking advantage of instruments which possess parallel accumulation–serial fragmentation (PASEF) technology, we are able to enhance the efficiency of the proteomics investigation and ensure that a larger proportion of the detected molecules can be dissociated and more comprehensively identified/quantified, even with limited protein quantities. I think this is a prime example of how the community and vendors have identified a challenge within the field and then adapted the technology accordingly to render this type of research more feasible.

References:

1. Hu Y, An Q, Sheu K, Trejo B, Fan S, Guo Y. "**Single-cell Multi-Omics Technology: Methodology and Application.**" Front Cell Dev Biol. 2018;6. doi:10.3389/fcell.2018.00028
2. McGuire, Amy L et al. "**The road ahead in genetics and genomics.**" Nature reviews. Genetics vol. 21,10 2020: 581-596. doi:10.1038/s41576-020-0272-6
3. LeMieux, Julianna "**Ultima Genomics Bursts Onto NGS Scene Targeting the $100 Genome**" Acccessed 23/11/2022 https://www.genengnews.com/topics/omics/ultima-genomics-bursts-onto-ngs-scene-targeting-the-100-genome/
4. Amarasinghe, S.L., Su, S., Dong, X. et al. "**Opportunities and challenges in long-read sequencing data analysis.**" Genome Biol 21, 30 2020. doi:10.1186/s13059-020-1935-5
5. Mehrmohamadi, Mahya et al. "**A Comparative Overview of Epigenomic Profiling Methods.**" Frontiers in cell and developmental biology vol. 9 714687. 22 Jul. 2021, doi:10.3389/fcell.2021.714687
6. "**A Scientific Illustration of How Epigenetic Mechanisms Can Affect Health**" Accessed 21/11/22 National Insitutes of Health https://commonfund.nih.gov/epigenomics/figure
7. Lowe, Rohan et al. "**Transcriptomics technologies.**" PLoS computational biology vol. 13,5 e1005457. 18 May. 2017, doi:10.1371/journal.pcbi.1005457
8. "**Transcriptome Fact Sheet**" Accessed 21/11/2022 https://www.genome.gov/about-genomics/fact-sheets/Transcriptome-Fact-Sheet
9. Williams, Cameron G et al. "**An introduction to spatial transcriptomics for biomedical research.**" Genome medicine vol. 14,1 68. 27 Jun. 2022, doi:10.1186/s13073-022-01075-1
10. Gorin, Gennady et al. "**RNA velocity unraveled.**" PLoS computational biology vol. 18,9 e1010492. 12 Sep. 2022, doi:10.1371/journal.pcbi.1010492
11. Graves, Paul R, and Timothy A J Haystead. "**Molecular biologist's guide to proteomics.**" Microbiology and molecular biology reviews: MMBR vol. 66,1 (2002): 39-63; table of contents. doi:10.1128/MMBR.66.1.39-63.2002
12. Chait, Brian T. "**Chemistry. Mass spectrometry: bottom-up or top-down?**" Science (New York, N.Y.) vol. 314,5796 2006: 65-6. doi:10.1126/science.1133987
13. Lundberg, Emma, and Georg H H Borner. "**Spatial proteomics: a powerful discovery tool for cell biology.**" Nature reviews. Molecular cell biology vol. 20,5 2019: 285-302. doi:10.1038/s41580-018-0094-y

CHAPTER 2

SPOTLIGHT ON SINGLE-CELL AND SPATIAL OMICS

SINGLE-CELL ANALYSIS HAS ALLOWED RESEARCHERS TO STUDY THE INNER WORKINGS OF A CELL AT NEVER-BEFORE-SEEN RESOLUTION AND REVEAL THE FULL COMPLEXITY OF CELLULAR DIVERSITY.

Single-cell techniques started with transcriptomics, but in subsequent years other omics have been added into the mix. In particular, single-cell proteomics has seen the most recent developments in technology and application.[1]

Anna Wilbrey-Clark
Staff Scientist
Wellcome Sanger Institute:

It's really amazing just how much you can discover with single-cell (and single-nuclei) sequencing. As a natural sceptic, I initially thought most of the data we collected [for the Human Cell Atlas project] was going to be noise. But actually, there was a lot of useful data, which allowed us to identify some very rare novel cell types.

Projects like the Human Cell Atlas have utilised current advances in single-cell analysis to reveal a previously unrecognised heterogeneity of cell types and **defined new cell states that are associated with diseases from cancer to liver disease, Alzheimer's and heart disease**.[2]

Rebecca Mathew
Principal Scientist
Merck Research Labs:

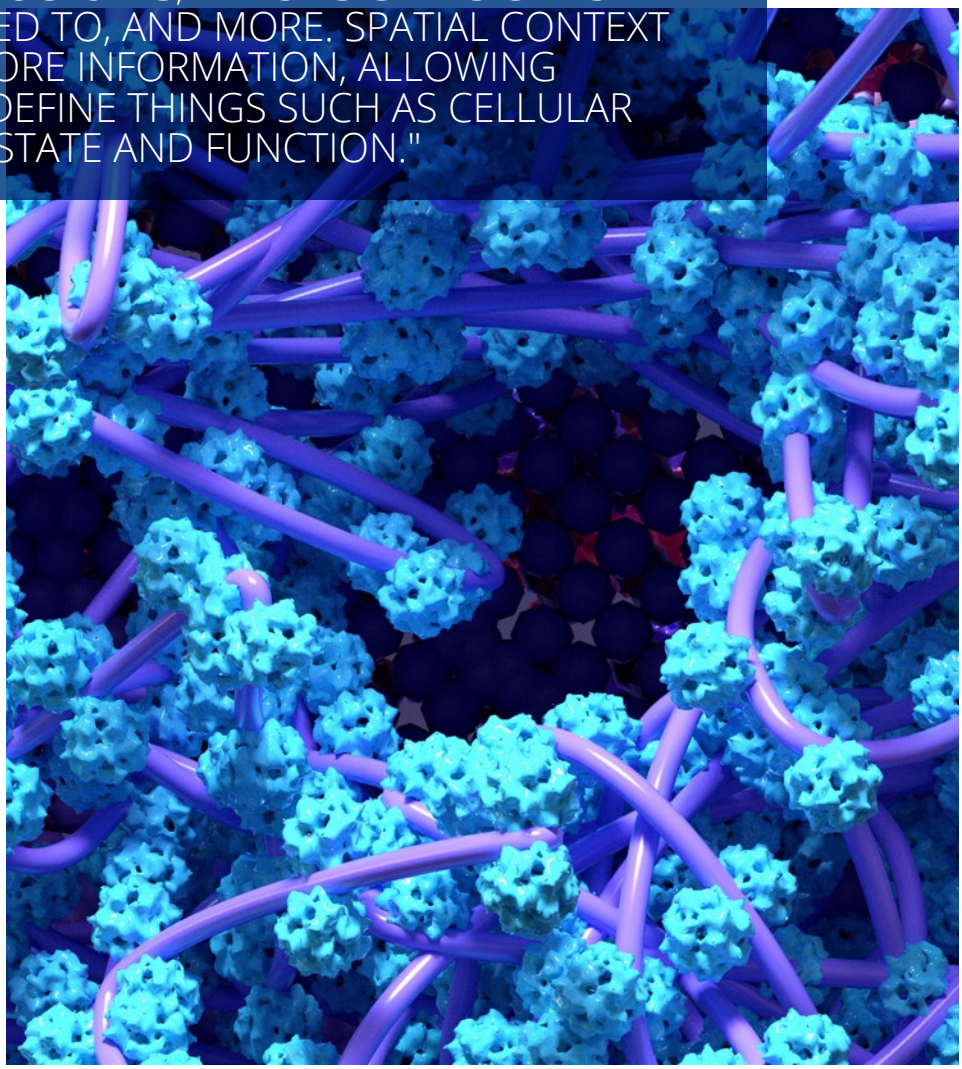
To capture the significance of this type of technology in the neuroimmunology space specifically, **there have been tremendous discoveries in characterising the sub-states of these cells as they interact and associate** with both acute and chronic stressors. There's been the discovery of disease-associated microglia cell populations using single-cell sequencing technologies, which has transformed our understanding of the responsiveness of the cell type to pathologies that exist in Alzheimer's disease, for example.

However, one crucial step in the single-cell analysis workflow is dissociation – **breaking down tissues to prepare them for analysis.** This breakdown of tissue means the **spatial context is lost**, as well as potentially changing features due to stress, cell death or cell aggregation.[1]

Alex Tamburino
Director, Spatial and Multiomics Single-Cell Sequencing Lead
Merck Research Labs:

Single-cell is very powerful. **It has enabled us to achieve unbiased whole transcriptomic profiling from single cells.** It's unlocked a lot of biology in low abundance or rare cell types. **The major drawback of single-cell is the need to dissociate tissues into individual cells before profiling, divorcing the cells from each other and the disease pathology.** The field, including ourselves, is investing into spatial transcriptomics because we can **profile cellular microenvironments and individual cells while retaining spatial information.** This enables us to understand how cellular neighbourhoods and disease pathology impact cell abundances, cell states and gene expression.

"NOW RESEARCHERS CAN SEE NEIGHBOURING CELLS, NONCELLULAR STRUCTURES, WHICH SIGNALS CELLS MAY HAVE BEEN EXPOSED TO, AND MORE. SPATIAL CONTEXT ALSO PROVIDES MORE INFORMATION, ALLOWING RESEARCHERS TO DEFINE THINGS SUCH AS CELLULAR PHENOTYPE, CELL STATE AND FUNCTION."



This is where spatial omics come in. New tools have allowed researchers to map the whole genome, epigenome, transcriptome, proteome, – and many other "omes" of hundreds of thousands of cells **while preserving morphological and spatial context.** [3]

Anna Wilbrey-Clark
Staff Scientist
Wellcome Sanger Institute:

I love that with spatial technologies you are able to look at the single-cell data, and the whole transcriptome, in the tissue, in situ. It's good to be able to see the cells in their context. When you dissociate tissues, you not only lose that context, but the cell populations do change. **I think we can learn so much more from looking at the tissues and seeing where those cells are, how they interact, what cells are next to each other, and so on.**

Now researchers can see **neighbouring cells, non-cellular structures, which signals cells may have been exposed to, and more.** Spatial context also provides more information, allowing researchers to define things such as cellular phenotype, cell state and function. **This is why spatial-multi-omics was named one of the seven technologies to watch in 2022** in a Nature article earlier this year. [4]

Jeffrey Moffitt
Assistant Professor, Department of Microbiology, **Harvard Medical School,** Investigator, Program in Cellular in Molecular Medicine, **Boston Children's Hospital,** Associate Member, **Broad Institute:**

In many ways, spatial biology as a term captures many things. But perhaps one way to think about this emerging field is that it is the merger of spatial technologies like microscopy – techniques that allow one to image where different cells or molecules are found within tissues and also determine properties like cell morphology or the organization of molecules within cells - with genomics technologies - techniques that allow one to probe the complexity of gene expression within tissues as a whole, with techniques like bulk RNA-sequencing, or within single cells, with techniques such as single-cell RNA sequencing.

The success of genomics in biology is clear, and these technologies have provided tremendous insight into a wide range of biological questions. The same can be said of microscopy. **What is exciting about the new suite of technologies that are captured in this term 'spatial biology' is that they promise to provide the same set of insights but simultaneously for the same sample. We will make microscopy-style measurements but with genomic-scale information.**

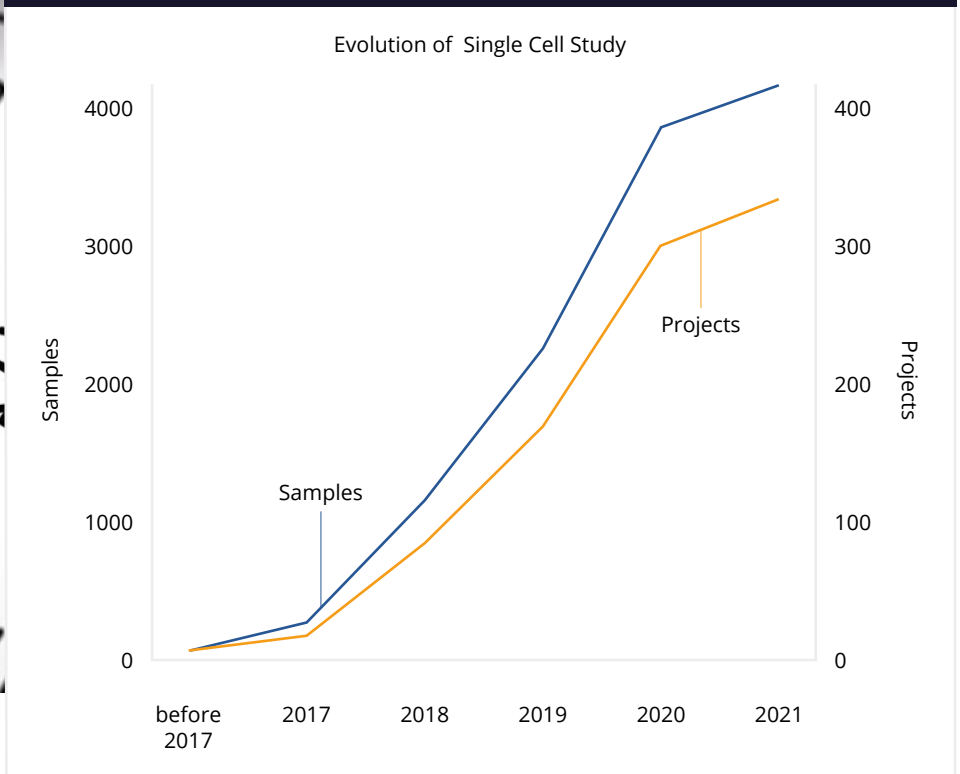
Read on to learn more about single-cell analysis, spatial-omics, and how we can bring the best of single-cell and spatial together.

SINGLE-CELL

SINGLE-CELL SEQUENCING TECHNOLOGIES HAVE ALLOWED RESEARCHERS TO PROFILE INDIVIDUAL CELLS AT UNPARALLELED RESOLUTION.

The enhanced resolution of single-cell sequencing technologies has allowed researchers to study the cellular, genomic, epigenomic, transcriptomic and proteomic heterogeneity in different contexts, both in disease as well as normal biological processes.[1] Consequently, the adoption of this technique has become wildly popular – as you can see in Figure 1 below, the number of projects using single-cell sequencing techniques has exploded in the past few years.

FIGURE 1: THE GROWTH IN THE NUMBER OF SINGLE-CELL PROJECTS AND SAMPLES [5]



HEAR FROM THE EXPERTS: WILL SPATIAL REPLACE SINGLE-CELL?

HOWEVER, AS SINGLE-CELL HAS RISEN IN POPULARITY AND BECOME UTILISED MORE AND MORE, IT MAY APPEAR TO SOME THAT SINGLE-CELL SEEMS TO HAVE REACHED A PLATEAU, ESPECIALLY WHEN YOU COMPARE IT TO THE BUZZ AROUND SPATIAL OMICS. AS RAPID ADVANCES IN SPATIAL HAVE ALLOWED FOR HIGHER AND HIGHER RESOLUTION, HAS IT USURPED THE APPLICATION OF SINGLE-CELL? WE ASKED EXPERTS IN THE FIELD THE SIMPLE QUESTION:
WILL SPATIAL REPLACE SINGLE-CELL?

MIAO-PING CHIEN
Assistant Professor, Department of Molecular Genetics, **Erasmus University Medical Center,** Principal Investigator, **Oncode Institute**

KOICHI TAKAHASHI
Associate Professor, Department of Leukemia, Division of Cancer Medicine, **The University of Texas MD Anderson Cancer Center**

RONG FAN
Associate Professor Department of Biomedical Engineering **Yale University**

NIKOLAI SLAVOV
Associate Professor, Department of Bioengineering **Northeastern University**

ANDREW SMITH
Assistant Professor **University** of **Milano-Bicocca**

KERSTIN MEYER
Principal Staff Scientist **Wellcome Sanger Institute**

ALEX TAMBURINO
Director, Spatial and Multiomics Single-Cell Sequencing Lead **Merck Research Labs**

Miao-Ping Chien: Many people have developed great single-cell-omics techniques, so I would say it's more accessible now. However, compared to bulk cell sequencing, it's still not that accessible. **I think the next frontier in terms of single-cell omics is to reach levels of accessibility and commonality similar to bulk cell sequencing.** Not only for RNA sequencing, but also for single-cell genomic sequencing, epigenomic sequencing and hopefully single-cell proteomics as well. I think the next step, after single-omics techniques becoming more accessible, will be to have **easy accessibility for multiple 'omics' at a single-cell level.**

Another important part will be analysis. There are a lot of well-developed analysis algorithms out there in the community, **but it's still not that standardized yet.** So, this is another thing I think people will work on in the coming years. The last key development, I think, will be to **integrate these single-cell-omics methods with spatial omics.**

Koichi Takahashi: Single-cell analysis has evolved so much over the past few years, and the capability is getting so good that the field has become a bit saturated. But there's still some room for innovation, particularly for multi-omics. In terms of single-cell and spatial, **I envision both will start being used for different purposes, as well as being used in conjunction with one another and complementing each other. I think that there is a potential for single-cell analysis to move into the clinic** – the question there is can these single-cell platforms enter the clinic to help diagnose or monitor disease response better than existing methods? To tell the truth, single-cell analysis has been in clinical use already. We actually use flow cytometry, which is a single-cell protein analysis platform, for a diagnosis and treatment response assessment purpose. So why not single-cell genomics?

Rong Fan: I hear this question all the time! I don't think spatial will replace single-cell. I think that each has its own unique advantage. I will say that in some of our own work and publications **we often use both spatial and single-cell approaches, and integrate them together.** We've found that we can always learn a lot more by integrating the data from both approaches together, so I think in the future more people will start doing that.

I also think that we are **going to start to get some more large-scale single-cell reference data.** At the moment, if you are sequencing something novel, you may have to generate your own single-cell data. Reference single-cell sequencing data would be more useful because they are often **larger datasets, with information from millions of cells,** and you can't achieve that on your own. You won't be able to get that **granular, detailed, information that you can get by integrating your spatial omics data with such a large-scale single-cell reference dataset.**

Nikolai Slavov: I'd love to see the technology being so **accessible and so inexpensive, that every biological project that can benefit from doing single-cell protein measurements, can do them as easily and perhaps more cheaply than currently.** Not only making it cheaper and easier, but extending the reach of what we can measure because **measuring proteins expands the scope of our analysis, but it's clearly insufficient.** We want to be able to measure protein interactions, protein modifications, protein localisation in the cell, protein activities, and protein conformations... **all of these very important layers of biological activities and regulation**, and our technologies can be extended to make these measurements accessible because they currently cannot do all of these things. We have not yet done it. But I see no reason why this cannot be developed.

I think it's going to be a very exciting path of technology development towards achieving this. Once we have that technology, I think we'll be in a much better position to **catalyse a more mechanistic approach to single-cell biology.** Not only to identify different clusters of cells and to describe differences in cell states, in different pathophysiological conditions, but to be able to measure the molecular processes that **ultimately underpin those different stages and contribute to either health or disease.**

Andrew Smith: No, I don't think spatial will replace single-cell, **I think they will become complementary to one another.** Considering that single-cell omics provides you with molecular depth **that mass spec imaging approaches alone are unable to provide,** whilst the imaging aspect provides spatial context regarding a structured cellular network, I feel that these two approaches will effectively go hand-in-hand and complement one another.

Kerstin Meyer: I think they will start to answer different questions. **In the clinic we can easily obtain blood from patients, and in a sense when working with blood we are we are back to single-cell – the cells are already dissociated.** The tech that gives you the same insights at a **lower cost – that's the future.** Routine sequencing of very large numbers of cohorts – **we can follow responses to drug treatments or population genetics at single-cell resolution.**

However, there are still challenges single-cell will need to overcome other than lowering cost for clinical implementation.

Alex Tamburino: Generally speaking, clinical implementation for single-cell is challenging as you are **collecting samples at one location and then profiling them and then doing the experiments at another location. Maintaining sample quality and preserving that sample for longer periods of time and across sites has been challenging.** There's been a lot of advances in order to improve that, but it's always going to be compared to the quality you can get doing profiling in the same lab.

> "I THINK THAT SINGLE-CELL WILL START TO MOVE INTO THE CLINIC – THE QUESTION THERE IS CAN THESE SINGLE-CELL PLATFORMS ENTER THE CLINIC TO HELP DIAGNOSE OR MONITOR DISEASE RESPONSE BETTER THAN EXISTING METHODS?"

THINGS TO CONSIDER: SAMPLE PREP IS CRUCIAL IN SINGLE-CELL STUDIES

SAMPLE PREP IS CRUCIAL IN SINGLE-CELL STUDIES, BUT THIS STAGE IS OFTEN OVERLOOKED OR LITTLE DISCUSSED. WE SPOKE TO SOME EXPERTS IN THE SINGLE-CELL SPACE AND ASKED THEM ABOUT THE IMPORTANCE OF CORRECT SAMPLE PREP AND DISSOCIATION IN SINGLE-CELL STUDIES.

LUCIANO MARTELOTTO
Associate Professor, University of Adelaide
Single-cell and Spatial-Omics Lab
Adelaide Centre of Epigenetics

KERSTIN MEYER
Principal Staff Scientist
Wellcome Sanger Institute

ANNA WILBREY-CLARK
Staff Scientist
Wellcome Sanger Institute

Luciano Martelotto: Something people usually forget is the **importance of sample preparation, and how to take care of the samples.**

Kerstin Meyer: Sample preparation is really **critical.** Our approach to sample prep has evolved in recent times – initially we tried to process our samples as soon as possible. What we've now realised is that storing samples at 4°C preserves them to some extent. More recently with single-nucleus sequencing we can now use frozen samples – **this changes the considerations involved.**

Anna Wilbrey-Clark: Different enzymes release different cell types. Most do the job at 37°C but cold enzymes can work at 4°C. In terms of stability – 24 to 72 hours after storage are the results the same? The way we do our investigations has changed, initially when we received tissue samples, we dissociated them, loaded them on 10x Genomics and looked at what the total population of the cells were. **Now we are doing a lot more cell sorting and enriching specific populations of cells.** This takes a long time.

Kerstin Meyer: Keeping the samples at 4°C and using cold enzymes could mean that **during enriching stages the cells are more stable.**

Luciano Martelotto: You also need to be careful the cell-surface markers are not removed. How do you decide which enzymes to choose for your sample prep? Do you check published work or trial and error? This is especially important for precious samples. Sorting can be a great help – but if cells are too fragile it may actually make things worse.

Anna Wilbrey-Clark: It's tricky. The first thing to do is check the literature. In terms of testing, we can trial with pig tissue, but ultimately you do have to trial some human samples with the different enzymes and see what works best. It is very precious material – and naturally, that makes people nervous to work with it. But at the end of the day, the human material is going to be different to the animal.

Kerstin Meyer: Every different organ system also appears to have its own favourite set of enzymes. However, for the Human Cell Atlas, to allow us to compare different cell types, we needed to use the same enzymes across a whole range of organs. But we're sort of coming around to realising that this isn't necessary because all the dissociated cells have a stress signature - and this doesn't seem to vary between different enzymes for sample prep. So now we are going back and specialising the sample prep to optimize it for each enzyme. Having some diversity in the way you assess your sample is also a good thing.

Luciano Martelotto: Think about what you are investigating; things vary a lot cell to cell. It's not the same working with brain vs breast cancer vs liver, heart, and so on. The way we need to consider that is to do the research before we go in and test things out. It also depends on the type of cell you want to recover. We also need to bear in mind that past studies do not have the resolution we have now, and we also now define a cell by transcriptional state. What we see now is not the same thing they saw in the past. All this information is important when considering how we go about choosing our sample prep.

"THINK ABOUT WHAT YOU ARE INVESTIGATING; THINGS VARY A LOT CELL TO CELL. IT'S NOT THE SAME WORKING WITH BRAIN VS BREAST CANCER VS LIVER, HEART, AND SO ON."

ADVANCES IN SINGLE-CELL PROTEOMICS: SPOTLIGHT ON SCOPE2

IN THE PREVIOUS CHAPTER, WE DISCUSSED SOME OF THE CHALLENGES TO CONSIDER WITH PROTEOMICS – NAMELY SEQUENCING DEPTH. SINGLE-CELL PROTEOMICS IS A RAPIDLY DEVELOPING FIELD, AND INNOVATIVE APPROACHES HAVE ALLOWED RESEARCHERS TO PROFILE PROTEOMICS AT HIGHER DEPTH AND RESOLUTION THAN EVER BEFORE. WE RECENTLY SPOKE TO **NIKOLAI SLAVOV,** ALLEN DISTINGUISHED INVESTIGATOR AND ASSOCIATE PROFESSOR, DEPARTMENT OF BIOENGINEERING, **NORTHEASTERN UNIVERSITY.** HIS WORK FOCUSES ON SINGLE-CELL PROTEOMICS AND THE ROLE OF PROTEINS IN PATHOLOGY AND HEALTH. IN THIS INTERVIEW, NIKOLAI DISCUSSES THE METHODS HE'S DEVELOPED FOR HIGH THROUGHPUT SINGLE-CELL PROTEOMICS BY MASS SPECTROMETRY.

FLG: What drew you to focus on single-cell proteomics?

Nikolai Slavov: It is well appreciated and understood that **most biological functions are performed by proteins. But our inability to analyse proteins in depth at high-resolution has resulted in biomedical research being focused primarily on analysing DNA and RNA molecules.**

My background was previously focused on transcriptomics. I did my PhD in systems biology with David Watson's group where we used the technology of the day, DNA microarrays. But as we were studying that, often my data would suggest that the interesting biological processes were taking place or happening at the level of protein synthesis or protein degradation, and we were limited in our ability to analyse those.

Around 2011, I realised that existing mass spectrometry technologies can be applied to analyse proteins with **much higher throughput, and better quantitative accuracy and sensitivity than had previously been achieved.** This particular opportunity wasn't being realised by anybody, as far as I could tell. This made me think that I could help accelerate the development of this area of biomedical science. I saw this as a very big opportunity.

I decided to give it a try to see if we could actually develop single-cell proteomics tech. It seemed obvious to me (and everybody else) that if we were able to achieve cheap, quantitative protein analysis of individual mammalian cells, with their own mini applications, the significance of this research would never be in doubt. What, at the time, was much more in doubt and controversial, was the feasibility. Many of the leading experts in the field at the time believed that this was not possible and they were very sceptical. I was a newcomer, without a background in that particular technology. Not surprisingly, I did not convince the leaders in mass spectrometry overnight that we can do something that they believed was impossible. But I thought it was worthwhile giving it a try.

FLG: You have developed methods for high throughput single-cell proteomics by mass spectrometry and you're using them to quantify proteome heterogeneity during cell differentiation. It's an emerging field, there's been a variety of methods proposed by different groups. So, what makes your SCoPE2 project stand out from the crowd?

Nikolai Slavov: With SCoPE2, and other methods that we have tried to develop, we emphasise its accessibility from the very beginning.

Let's take a step back and get a perspective on methods, how they have been developed and their differences. Looking from the outside, it is easy to see multiple different names being used and different methods. The field appears to be quite crowded, but the methods that are being used fall into only a couple of categories. And the methods within a category are quite similar to each other.

One approach is to analyse one cell at a time. In the jargon of mass spectrometry, we call this 'label free', because cells are not labelled, and we analyse only one cell at a time. This approach has been attractive to a number of colleagues. **From my perspective, a major limitation of these approaches has been their limited throughput.** Because mass spec instruments are quite expensive and mass spec time is expensive. If we only analyse one cell per hour, that is going to have **limited scope and biological applications.**

Our strategy has been to use labels, which are often called '**multiplex approaches**'. Instead of analysing one cell, we can **barcode proteins from different single cells with single-cell specific barcodes, and then we can analyse a dozen or more single cells at the same time.**

At the end of the experiment, we can tell which protein came from which single-cell and therefore do the single-cell quantification. This is one specific aspect that we introduced using isobaric mass tags, and until recently, all of the multiplex methods that anybody has used is the technology that we introduced using isobaric mass tags. More recently, we have introduced a different kind of multiplexing, which will be published in Nature Biotechnology, using non-isobaric mass tags. This has a different set of advantages, which we'll discuss later on. **But one aspect of our first approach, SCoPE-MS, and its second generation, SCoPE2, is the use of multiplexing, which allows increased throughput.**

Another aspect at the time that was absolutely crucial, was the use of an isobaric carrier. In addition to the single cells, we barcoded a small bulk sample of cells, which allowed us to reduce losses of single cells and various surfaces that the labelled cells interact with. This allowed us to increase our ability to assign amino acid sequences to the peptides, and that was key for being able to quantify proteins in single cells in our very first attempt. At this point, new technologies developed by my laboratory and other laboratories have made the use of isobaric carrier less essential. **But at the time, this was the first demonstration, to my knowledge, of being able to quantify hundreds of proteins across hundreds of single cells.** That was very much enabled by using this approach of the isobaric carrier. In fact, it was enabled by using old equipment. We did not have access to cutting-edge, state-of-the-art mass spectrometry equipment. In fact, we did not have any equipment. I had to collaborate with a good friend of mine at Harvard, who made a key contribution by providing access to equipment that I did not have at Northeastern.

FLG: Have other labs and researchers adopted SCoPE2?

Nikolai Slavov: This whole approach of multiplexing, the isobaric carrier and other aspects that were introduced, made it very easy to implement in other laboratories. What has been a guiding philosophy for us, is not only to develop the best methods that we can possibly develop to give us the most accurate and highest throughput methods… all of that is great. But we have a significant constraint in everything we do. **That constraint is that what we do should be reproducible in other laboratories; others should be able to do it.** Ideally, we make it as easy as possible for others **by using equipment that is widely available, that is relatively inexpensive.** They're not cheap, nothing is very cheap, but at least we can use equipment that is much cheaper; in some cases, orders of magnitude cheaper, than alternative equipment. That has been a very important guiding principle for us, trying to make the technology accessible, make it as inexpensive as possible, make it high throughput.

I think that these are distinctive aspects that have allowed many laboratories to adopt and start using SCoPE2. I know of a number of mass spectrometry facilities, both in the US and Europe, that have successfully implemented SCoPE2. There are several other methods using multiplexing, sometimes with slightly different names, but they're really variations on the same approach that SCoPE-MS introduced. I mentioned other methods that exist – they are label-free approaches. Another set of approaches we recently introduced was by doing data-independent acquisition, combined with multiplexing, non-isobaric labels are used in that context. I'm very enthusiastic about the potential of these approaches to inherit many of the advantages of SCoPE2, in terms of being accessible, being relatively inexpensive etc. **But they have the potential to provide even deeper proteome coverage and substantially higher throughput.**

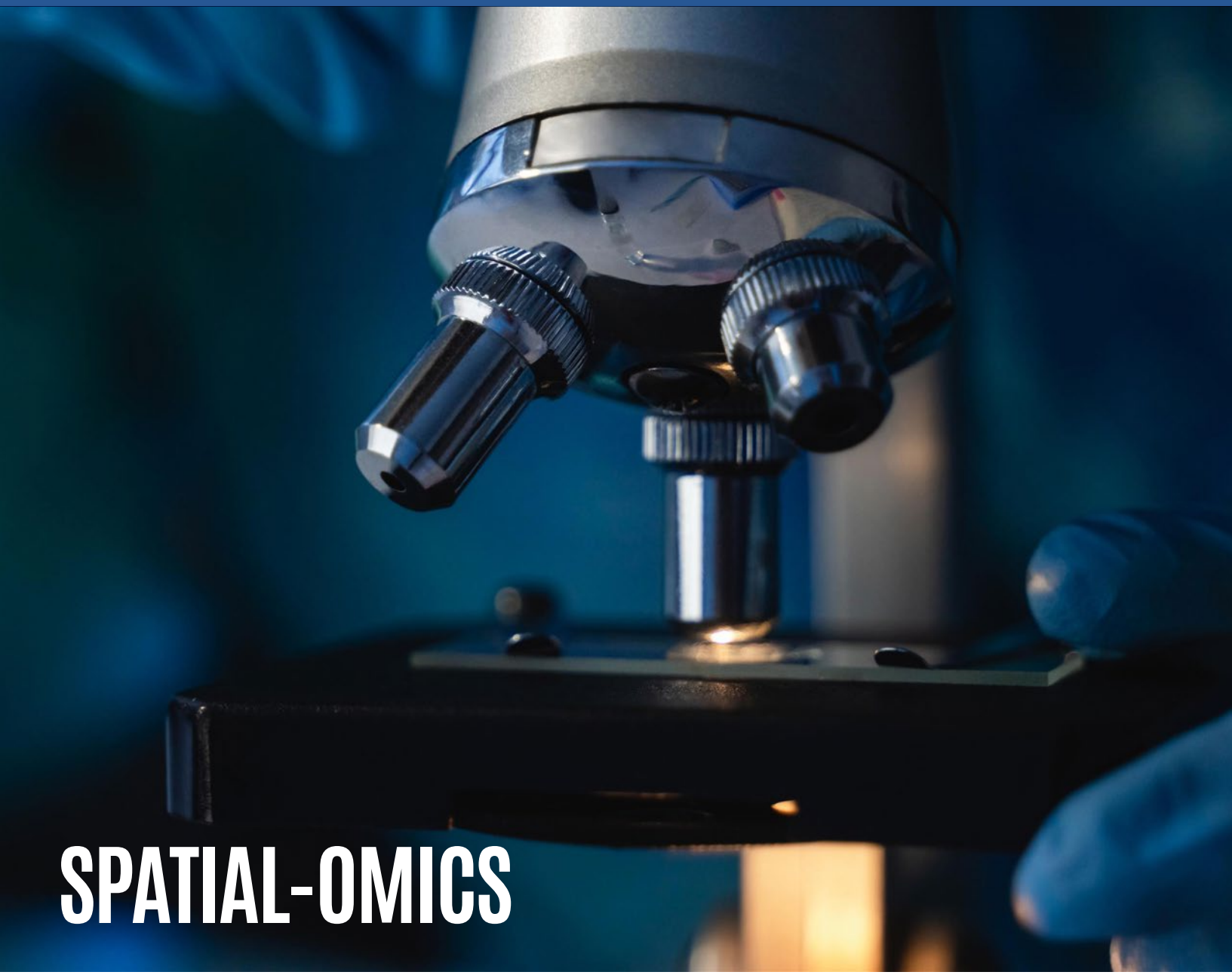
FLG: You've talked about the benefits and potentials there. What about some of the challenges? What are some of the key challenges in scaling up single-cell analysis to the proteome?

Nikolai Slavov: Some of the challenges are very similar to the challenges of mass spectrometry proteomics. **Any kind of mass spectrometry proteomics analysis is not as widely integrated with biomedical research as DNA or RNA sequencing methods.** The reasons for that are numerous. Some of them are technological, I think a lot of them are societal and policy-based. It's a level of understanding of the technology by colleagues who drive biomedical research, it's funding from various governments, institutions, and so on. All of these problems that have generally made mass spectrometry proteomics less accessible and less integrated with biomedical research, are also applied to single-cell protein analysis.

Fortunately, these problems are not unsolvable. They certainly have solutions. They're not simple. I cannot solve them overnight. But we try to help, certainly at the level of education, **we're very passionate about doing our best to explain the technology in an accessible manner to the wider community.** Some of this has to do with articulating a compelling vision and justifying funding to develop standard operating procedures. **We also articulate the problems in biology and medicine that really need proteomics, and why we should invest in doing the protein analysis as opposed to focusing on the more accessible transcriptomic and genomic analysis.** In terms of adoption of SCoPE2 in existing facilities and laboratories that can already do protein analysis well, I think all of these laboratories and facilities that can do quantitative proteomics should be able to implement SCopPE2, so there are no major additional bottlenecks.

There is one disadvantage, compared to single-cell RNA sequencing, which has a high throughput. 10x Genomics has made it possible to analyse in the order of 10,000 single cells in a relatively convenient way, in a single sample. This is even more challenging to deal with. SCoPE2 throughput is much more comparable to the multi well plate-based approaches such as CEL-Seq, SMART-Seq, and so on. And, to some extent, this reflects the current state of the field. **It's not a limitation of mass spectrometry proteomics or single-cell analysis. It is simply the level of throughput that current technologies have achieved.**

In fact, with the new multiplexed data-independent acquisition framework that we've introduced, we believe that we can get to analysing the proteomes of five thousand single cells per day, per single instrument, and potentially scale that even further. [6,7] **The opportunity certainly exists to increase throughput substantially; though at the moment, the current situation is relatively weak when compared to the more mature techniques such as single-cell RNA sequencing.**



SPATIAL-OMICS

SPATIAL HAS ALLOWED RESEARCHERS TO PROFILE THE VARIOUS "OMES" OF CELLS WHILE PRESERVING THE TISSUE MORPHOLOGY – THUS ALLOWING SPATIAL CONTEXT TO BE PRESERVED.

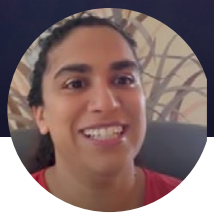
As we've already mentioned, developments in the spatial omics field have been rapid, skyrocketing past expectations. The development of this technology means that spatial context is preserved, and researchers can now profile cells and tissue in their morphological context and understand the influence of their local environment and surrounding cells. [3]

Luciano Martelotto
Associate Professor
University of Adelaide Single-cell and Spatial-omics Lab, Adelaide Centre of Epigenetics:

If we go back to 5 years ago – if you described spatial omics to me, I would have responded, "What? What are you even talking about? That sounds like science fiction". The future is now – and if you want to get into the spatial world, now is the time.

THE IMPORTANCE OF SPATIAL CONTEXT

BUT WHY IS RETAINING SPATIAL CONTEXT SO IMPORTANT? WELL, WE'VE ALREADY MENTIONED A FEW REASONS WHY – BUT READ ON TO HEAR IT FROM THE EXPERTS THEMSELVES.



REBECCA MATHEW
Principal Scientist
Merck Research Labs

MATHEW CHAMBERLAIN
Principal Scientist
Janssen

ANDREW SMITH
Assistant Professor
University of Milano-Bicocca

JEFFREY MOFFITT
Assistant Professor, Department of Microbiology
Harvard Medical School

Rebecca Mathew: There's a tremendous impact that spatial technologies have on us being able to **understand the interplay between different cell types in intact tissue.** We propose or hypothesise in sophisticated tissues, like the brain, where there's a great degree of cellular heterogeneity, that different cell types may interact and contribute to disease progression or pathology in that microenvironment. **Spatial transcriptomics allows us to see that and test our hypotheses rather than speculate.**

Mathew Chamberlain: If cells couldn't communicate with each other inside tissue, you wouldn't have a lot of diseases, but you'd have a lot of other problems. With many of these neurological and immunological diseases, we're moving past the idea that there's one gene that can causally drive one disease. **As a result, you have to think about diseases as systems with components that interact.** Spatial data is very helpful for identifying different nearby immune cells and understanding how they interact with other cells. Some of the earlier data didn't quite have the resolution we needed for that analysis. The more recent spatial data seems to have much better resolution. We're getting closer.

Andrew Smith: **It's a very apt way of understanding how these biological molecules, and therefore processes, are altered within complex pathological tissue.** It allows us to visualise how certain cells are distributed and the molecular environment in which they are found, which is highly important within pathological tissue. Pathological tissue is known to be very dynamic and cell state may be governed or driven by the types of cells which surround it. If we think of a tumour immune environment, for example, and study the phenotypes of these tumour cells, we realise that the phenotype of the tumour cells present may also be governed by the lymphocytes surrounding them. Cytotoxic lymphocytes and macrophage can impact the function of the other array of cells found within the tumour microenvironment, and how they're communicating with one another is really important information to understand what's going on in the pathological tissue. **I think it's one aspect to take consider a cell in isolation, but it can become an entirely different story if you're then visualising how its interacting with its surroundings.**

Jeffrey Moffitt: Let me make an analogy. Imagine that you want to understand how a car engine works – this is a complex machine comprising many different parts. Having a catalogue of the parts of the engine – pistons, spark plugs, etc. – would provide tremendous insight into how an engine functions and what might be the origin of problems when the engine breaks. Yet, imagine you were given the parts of the engine but no understanding of how to assemble them. Clearly, your ability to understand how the function of these different parts gives rise to the function of the engine would be limited.

The promise of spatial context is the ability to understand how these different parts fit together and how their individual functions collective give rise to the function of the whole—an understanding of how to assemble the engine from these parts. To stretch the analogy, single-cell RNA sequencing and single-nucleus sequencing have done a phenomenal job of giving us parts catalogues for a wide range of biological systems. **We know what the cell types and states are. But until we understand how these cells are assembled, it will be challenging to understand how the behaviour and function of individual cells cooperatively gives rise to the function of the tissue as a whole.** The promise of spatial biology techniques is the ability to both discover the parts of biological systems—cell types—and how they fit together.

THINGS TO CONSIDER: FILTERING NOISE AND HANDLING BIG DATA

ALEX TAMBURINO
Director, Spatial and Multiomics Single-Cell Sequencing Lead
Merck Research Labs

ANDREW SMITH
Assistant Professor
University of Milano-Bicocca

RONG FAN
Associate Professor, Department of Biomedical Engineering**Yale University**

The high-resolution and dimensionality means spatial omics studies are very data rich. We asked researchers how to tackle big data problems, as well as other challenges specific to spatial omics.

Alex Tamburino: The biggest challenge for everyone, not just pharma companies but everyone in the field, **is the complexity of these experiments.** Spatial transcriptomics experiments and analyses are built on many years of advances in molecular biology, next-generation sequencing, high-dimensional data analysis, microscopy and image analysis. **Those are all fields unto themselves. Spatial transcriptomics experiments require expertise in all these domains, and researchers must be able to extract all the relevant information from these multi-omics and multi-level analyses.** Having a team of experts who can work together and have fluency in the different domains and have the expertise in their own specific domain is essential to implement these technologies and utilise them to their full potential.

Andrew Smith: As of today, in order to render our work more user-friendly, many of the commercial software packages allow you to perform pre-processing of the data, including noise removal, in a streamlined manner. This enables more reliable data to be obtained and filters out some redundant information that only serves to increase the data load. Of course, from time to time, a manual check with an "expert eye" is required but, in my opinion, we are trending toward more and more sophisticated automated pre-processing workflows which can only be beneficial if the findings of molecular imaging are to be translated into clinical utility.

Rong Fang: For us and many working in the spatial omics field, **registration can be very difficult as there is a lot of data** (registration makes sure information, such as spatial information, is properly "registered" to the corresponding omics data or different spatial image data). This is very important with spatial data collected from multiple tissue samples, but even data from one tissue sample, or from tissue section adjacent to each to each other – cells in these serial sections are still not the same cell. **So, registering the spatial mapping data is very important.**

SPATIAL MULTI-OMICS MAPPING THE IMMUNOLOGICAL DEVELOPMENT OF THE LYMPH NODE

IN THIS CASE STUDY, PRESENTED BY OUR SPONSOR NANOSTRING, RESEARCHERS USED SPATIAL MULTI-OMICS PLATFORMS GEOMX® DIGITAL SPATIAL PROFILER (DSP) AND THE COSMX™ SPATIAL MOLECULAR IMAGER TO MAP THE IMMUNOLOGICAL DEVELOPMENT OF THE LYMPH NODE.

The natural physiological processes regulating immune cell maturation, and dissemination across the body are critical in informing the way in which we view and understand responses to therapeutic intervention from conditionings including neurological disease, auto-immune conditions, and cancer. Single-cell atlases mapping immune cells provide hints to these aspects of immunology but lack essential spatial-temporal relationships between cells. With the advent of spatial multi-omics we can resolve RNA and protein molecules simultaneously in situ, enabling direct insight into the dynamics occurring as immune cells mature and migrate through tissue.

SPATIAL EXPLORATION

We profiled lymph node samples using complementary spatial multi-omics platforms: the GeoMx® Digital Spatial Profiler (DSP) and the CosMx™ Spatial Molecular Imager (SMI). With GeoMx DSP, we profiled whole transcriptomes from 5 patients focusing on key structures within the lymph node including the germinal center, mantle zones, medulla, and paracortex. To complement the structural profiling, we captured multi-omic RNA and protein expression profiles at sub-cellular resolution capturing 1000 genes and 63 proteins across serial sections covering >100mm2 and >1.4 million cells/section (Figure 1).

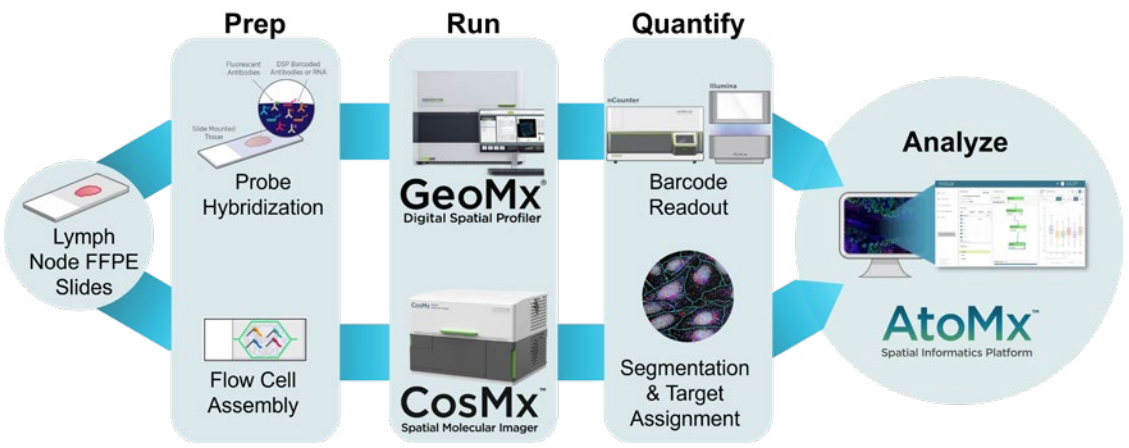


FIGURE 1: LYMPH NODE SAMPLES WERE ANALYZED USING BOTH THE GEOMX® DIGITAL SPATIAL PROFILER AND THE COSMX™ SPATIAL MOLECULAR IMAGER FROM STANDARD HISTOLOGICAL SLIDES. DATA ANALYSIS WAS PERFORMED WITH ATOMX™ SPATIAL INFORMATICS PLATFORM, NANOSTRING'S CLOUD COMPUTING PLATFORM FOR SPATIAL MULTIOMIC DATA ANALYSIS.

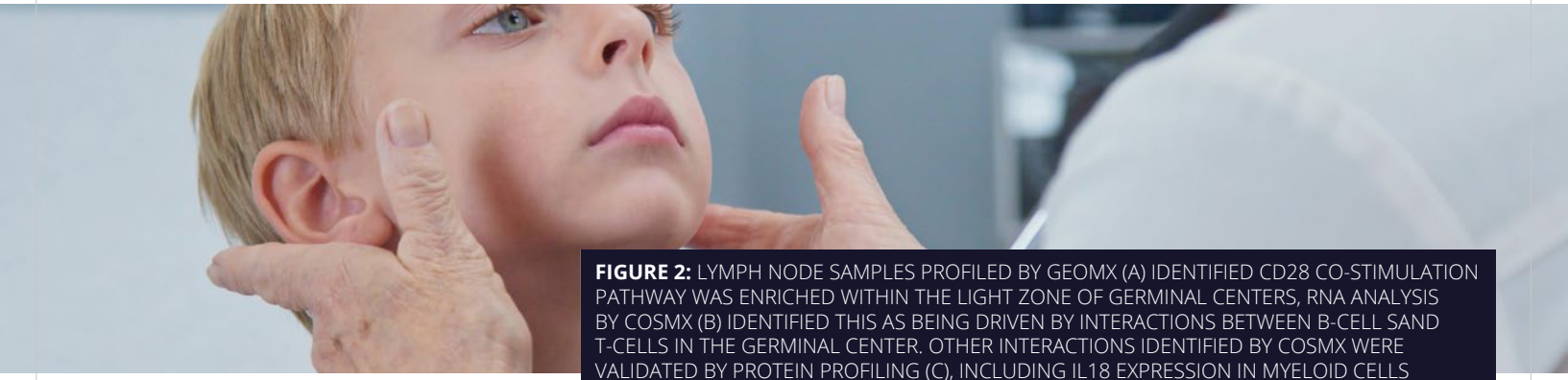
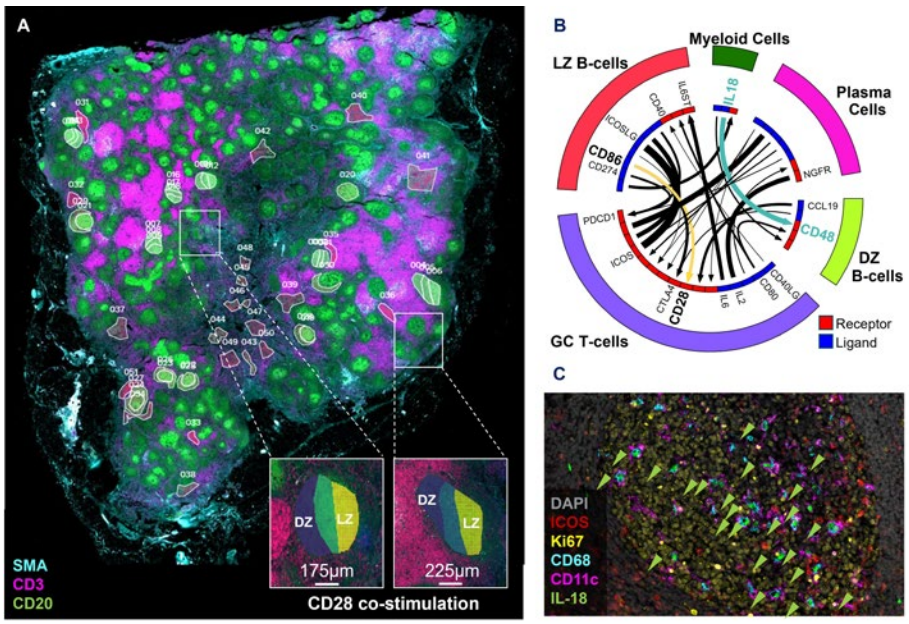


FIGURE 2: LYMPH NODE SAMPLES PROFILED BY GEOMX (A) IDENTIFIED CD28 CO-STIMULATION PATHWAY WAS ENRICHED WITHIN THE LIGHT ZONE OF GERMINAL CENTERS, RNA ANALYSIS BY COSMX (B) IDENTIFIED THIS AS BEING DRIVEN BY INTERACTIONS BETWEEN B-CELL SAND T-CELLS IN THE GERMINAL CENTER. OTHER INTERACTIONS IDENTIFIED BY COSMX WERE VALIDATED BY PROTEIN PROFILING (C), INCLUDING IL18 EXPRESSION IN MYELOID CELLS (GREEN ARROWS) SIGNALING WITHIN THE GERMINAL CENTER.



A NEW ROADMAP FOR IMMUNOLOGICAL RESPONSE

Across structures profiled with GeoMx Whole Transcriptome Atlas (WTA) we identified over 2,500 genes associated with distinct functional regions within the lymph node, and hundreds significant signaling pathways. These findings are can be explored in an interactive Minerva story[1], which will guide you through the in-depth profiling performed on these samples and the resulting spatially-defined gene expression and pathway patterns related to lymph node biology. Similarly, profiling with CosMx Human Universal Cell Characterization panel identified 27 cell types, 6 of which were not captured in dissociated single-cell reference studies.

BRIDGING SPATIAL SCALES

By integrating the results from GeoMx DSP and CosMx SMI, we identified over 600 pathways enriched across the dark and light zones of the germinal center or at their interface, as well as over 100 ligand-receptor interactions driving such pathways. For example, co-stimulation of CD28 was identified within the light zone of the germinal center by GeoMx DSP and CosMx SMI confirmed that CD86 ligands within light zone B cells were significantly colocalized with the CD28 receptors of the TfH cells of the germinal center (Figure 2A and B).

HIGH PLEX SPATIAL MULTI-OMICS

However, not all interaction can be confirmed by profiling RNA alone. Integrative analysis leveraging high-plex protein profiling on a single slide was able to confirm results from both GeoMx DSP and CosMx RNA profiling which suggested that IL18 signaling between macrophages and B cells within the germinal center may represent a critical cross-talk between these two cell populations (Figure 2C).

A SPATIAL ECOSYSTEM

This study is one of the first of many spatial multi-omics analyses which can drive novel understanding of well-profiled tissues. The GeoMx DSP and CosMx SMI platforms provided an unprecedented view into both the cellular interactions and structural underpinnings of the lymph node, and immunological processes and timescales happening as cells transit throughout this tissue. These studies shed light on novel interactions across key immunological interfaces, which may lead to better understanding of the mechanisms of many disease states.

Case Study Reference:

1. Spatial Organ Atlas, a tissue atlas of lymph development. https://nanostring.com/products/geomx-digital-spatial-profiler/spatial-organ-atlas/human-lymph-node/

BRINGING TOGETHER THE BEST OF SINGLE-CELL AND SPATIAL

IN AN IDEAL WORLD, WE WOULD HARNESS THE RESOLUTION AND INSIGHTS OF SINGLE-CELL AND PRESERVE THE SPATIAL INFORMATION AND TISSUE STRUCTURE/MORPHOLOGY. WELL, THE LATEST ADVANCEMENTS AND RESEARCH IN THE SCIENTIFIC COMMUNITY ALLOW US TO DO JUST THAT. WE RECENTLY INTERVIEWED **MIAO-PING CHIEN** ASSISTANT PROFESSOR, DEPARTMENT OF MOLECULAR GENETICS, **ERASMUS UNIVERSITY MEDICAL CENTER** AND A PRINCIPAL INVESTIGATOR AT THE **ONCODE INSTITUTE**. HER LAB HAS DEVELOPED NOVEL SPATIALLY RESOLVED SINGLE-CELL APPROACHES. IN THIS INTERVIEW, WE ASKED HER TO TELL US MORE ABOUT HER WORK, AND THE MICROSCOPY-BASED FUNCTIONAL AND SPATIAL SINGLE-CELL SEQUENCING APPROACH SHE AND HER TEAM DEVELOPED TO ANALYSE THERAPY-RESISTANT CANCER CELLS.

FLG: Why were you drawn to studying therapy-resistant cancer cells?

Miao-Ping Chien: That's a good question. We're interested in those rare and aggressive cancer cells that are responsible for tumour metastasis or therapy resistance because we know at this stage, cancers are not really curable. Part of the reason we cannot completely cure cancers is that after the treatment, there are always small subpopulations of cancer cells that survive; we give these surviving cell populations a collective term, aggressive cancer cells. These cells survive after treatment and over a period of time, they can relapse and regrow, sometimes becoming even more aggressive than before. So, that's the reason we're interested in studying those small populations of resistant cells.

FLG: Could you tell us about the challenges in identifying and analysing these aggressive cancer cells?

Miao-Ping Chien: So, people in this field are aware of the existence of these types of cells, but in the past few decades, the majority of techniques we use to study these cells mainly involved looking at specific markers. If we know the markers of cells, we can use techniques like cell sorting, for example, where you use an antibody with a probe to target certain cells and then those cells can be separated using the FACS machine, for instance, and this works quite well to a certain extent. People also use these techniques to isolate resistant cells. B**ut the thing is, it's not thorough enough because not every cell will express those markers.** So purely using those markers to identify or study those cells is simply not enough.

Because of this, the approach we are using in our group is basically to look at these cells not based on their markers, but on their behaviour. So, we profile, study and analyse individual cells' behaviour, and from this data we can then identify the subpopulations that are potentially more aggressive, after which we further isolate and profile these cells. The challenge here is really to **capture the dynamic information of cell behaviour.** And collecting this type of information is very challenging using currently available technologies. So that, to me, is one of the main challenges, but it has been tackled by the technology developed in our group.

FLG: Like you said, your lab developed a variety of multidisciplinary approaches to investigate these rare, aggressive cancer cells. Could you give us a quick overview of the different tools that you have developed?

Miao-Ping Chien: Yes, of course. As mentioned, our group is multidisciplinary, and we operate within four different research pillars. One is advanced imaging; we actually built our own custom-built microscope which allows us to screen really large quantities of cells. We're talking about tens, hundreds of thousands of cells in one single field of view, which is very impressive. Importantly, despite being able to see large numbers of cells, **we don't lose any spatial resolution.** This allows us to look at individual cell information. So, we can see many cells, but don't lose the single-cell or sub-cellular information. This kind of combination is really hard to achieve using commercial microscopes, and that's part of the reason we have our own setup.

In this setup we also implemented a device so that no matter what kind of cells you identify, you can also pinpoint and photo-label those cells, and then the cells can be isolated accordingly. That is part of the reason we have this capability to image and screen a bunch of cells, and from there we're able to identify and isolate certain subpopulations of cells. That's the first tool we use in our group, advanced imaging.

The second part of our process is, once we create a lot of big imaging data, we also need to have a really robust and advanced image analysis algorithm because, as I also mentioned earlier, we are interested in the behaviours of individual cells.

We need to have an algorithm that can help us analyse individual cells or the imaging data we acquired on the fly, and this is also very challenging. For that, we also scripted our own image analysis to be able to instantly profile the data and instantly export individual cells' features, and we can then use that information for further cell isolation.

So that's the second part, and the third part will be single-cell technology development. We use quite a lot of single-cell RNA sequencing, and we started to use single-cell genomic sequencing and proteomic profiling as well. Those techniques have been developed quite well in the field, and we kind of adapt it to fit with our technique, our pipelines. So that's the third tool we use, and the last one is bioinformatic analysis. We create quite a lot of data, and this data is also quite different than the standard single-cell sequencing you will normally get. Because of this, we also need to adapt the algorithms a little bit and use them to identify potential targets or hidden driving mechanisms that cannot be identified using current analysis methods. Those are basically the four different types of tools we use.

FLG: A lot of multi-omics data integration there, a lot of focus on single-cell data and spatial resolution as well as combining that with image analysis. I'm assuming a lot of Machine Learning and AI is involved, too. Could you tell us a little bit about the microscopy-based part of your own work and elaborate on the functional and spatial single-cell sequencing technique?

Miao-Ping Chien: First, we developed what we call microscopy-based functional single-cell sequencing. Through this, **we can profile or sequence individual cells based on their functional features observed under a microscope, and that's why we call it functional single-cell sequencing.** We're able to isolate cells based on any functional feature visualised under a microscope. Say we screened 10,000 cells under a single field of view, from there we'd identify maybe 100 cells that display aggressive migration. We'd then want to isolate those 100 cells and can use the photolabeling technique I mentioned earlier to label and isolate the cells of interest for downstream single-cell sequencing. The way we do that is: you screen a bunch of cells and then **perform real-time image analysis.** You then identify the cells you want followed by photolabeling and cell isolation. And I've used the example of aggressive migration here, but you can do this with any feature you can imagine. We've isolated cells before based on their abnormal DNA damage response to radiation, where we looked at DNA damage responses of individual cells after radiation. From there, you can also see heterogeneous populations. We can also look at abnormal cell division, because under the microscope you can see how cells divide during mitosis their mitosis. From observing those regular and irregular mitosis features, you can already see which cells are abnormal and which cells are normal, and so we can then isolate the cells we want to study.

You can also see which immune cells interact with cancer cells, which do not, which can kill cancer cells, which cannot, etc. You can observe these interactions under a microscope and can then selectively separate those cells. I mentioned the techniques to profile cells based on their functional features, and another consideration is the fact that cells are located in different spatial locations. **That's how we expand our technology to spatial omics, because under this setup, you can also see where those cells are located and which area they're in, as well as what cells they interact with.** So, we can also immediately identify this information, and this can also be used. That's why our technique can also be applied to spatial profiling.

FLG: Thank you for elaborating on that. That's really interesting. You also use Machine Learning and AI for your analysis, so could you expand on the Machine Learning and AI approach you use, as well as the advances in Machine Learning that excite you most and whether or not you think it's going to be applied more and more for data integration or multi-omics?

Miao-Ping Chien: That's a very good question. AI has birthed some really useful techniques, but as you also mentioned, it has its own challenges as well. We do apply AI in our research, and earlier, I talked about four different pillars in our research group. We currently implement AI not only for image analysis, but also for bioinformatic analysis. In terms of image analysis, when we curate large image data, we need to instantly process them to recognise, detect and track cells. This needs to be done quite accurately to avoid contamination. So, that's part of the reason why we implement AI, to improve our detection precision. So we do impart cell segmentation techniques using AI. The second aspect of AI in our research is about bioinformatics. **Many people also use AI or Machine Learning to dig out useful information from sequencing data and try to see if they can identify some hidden markers or target genes.**

We do that too, but one of the differences for our research is that we train our algorithms differently. AI is very good at identifying genes that are shared or distinct across different conditions or across different samples, but in our technique, we have the annotation information. We know which ones are and aren't aggressive. So by training the algorithm and providing these annotations, we can more straightforwardly identify the genes that are distinct, and uniquely expressed in aggressive cells and not expressed in non-aggressive cancer cells. This is the way we apply AI, because we have the annotation of each cell. When you don't have this annotation, you can still use AI to try to extract information, but that will be quite challenging and messy **because any variation between samples will complicate the training and lead to unreliable outcomes.** We do see some promising results using AI, but the reason it can work for us is because we have this very clear annotation at the beginning. So that's another way we apply AI in bioinformatic analysis.

The third one we have recently implemented is, as I mentioned earlier, looking at individual cells' behaviours. Quite often we will rely on recognisable features like migration, morphology and location, and now we are also training AI to observe images and to identify the cells that, for example, haven't displayed aggressive features yet and so have not been identified as such, but they are destined to become aggressive cells. We're training this programme to be able to identify the cells at an earlier time before they become aggressive cells, and the benefit of this is we can get more comprehensive information about how this cell will eventually drive aggressive features. This kind of thing is the third application of using AI in our group.

FLG: So obviously, with this AI and your analysis, you're looking at cells that develop their features over time, and you're looking at different behavioural changes that occur over time. You mentioned that you look at everything in real-time. Do you think more studies will start to use this real-time imaging instead of snapshot imaging? What advantages does real-time have over snapshot, and what challenges are there in other studies that don't do real-time data analysis?

Miao-Ping Chien: That's a great question. One of the things I mentioned earlier was that in the past, people studied those aggressive or therapy-resistant cells based on markers, and by doing that you're gathering snapshot, static information. Although this information is already very powerful, what we do is to offer additional information because some cells just don't have these universal markers, and some have no markers at all. So we started to look at the behaviour changes, and that requires real-time imaging. What people can get out of this is additional information, in addition to what we have studied based on the static information, is to have more comprehensive understanding about the driving mechanisms of those aggressive cells, and that's the main advantage I think it provides.

The challenging part is the real-time imaging, so after acquiring this large image data, one of the challenging elements is to process the data in a real-time fashion. We really need this because it would allow us to immediately identify cells after data acquisition. **That process is very important, otherwise the target cells that you're interested in might already migrate away from the original location, making it difficult to separate and isolate them.** We have developed this algorithm and hopefully will release it soon publicly, because we just submitted it. In the paper, it details how you can implement the algorithm in your experiments and setups. Regarding real-time image analysis, the second part will be to do with the fact that from the data, we need to be able to extract cellular features defined at the beginning of the analysis. That information will be further used for downstream cell isolation. I started to see more and more people using this kind of information for cell selection in their own studies. In terms of the real-time image analysis, these are the challenges I see and hopefully, in the near future, our soon-to-be-published paper will help the community.

FLG: Obviously, there have also been advances in spatial omics or increasing resolution, and there's been a lot of work done in that area. When it comes to spatial omics, how do you think we could still push that frontier further, and where are there going to be further advances?

Miao-Ping Chien: It definitely hasn't plateaued yet. The field of spatial omics is really changing very rapidly - every six months to every year, you'll see big jumps happen. Personally, I'm actually very impressed by the progress in this field. I think the ultimate goal people want to have for spatial omics is to have what you can do with current single-cell omics. What are the benefits of single-cell omics? We would like to have in-depth sequencing profiles of tens of thousands of cells, and to have tens of thousands of genes per cell, and to have single-cell resolution. **This is what the current state-of-the art single-cell omics techniques can reach, but not for spatial omics yet.**

Obviously, spatial omics techniques consist of spatial information, but the combination of these three properties that I just talked about, a large quantity of cells with in-depth sequencing profiles and with single-cell resolution, doesn't exist in current spatial omics methods yet. That's basically what this field is going towards. So if you look at an individual element, like having single-cell resolution? Yes, there are some spatial omics techniques providing single-cell resolution, but they can only profile, hundreds of genes, or maybe a maximum of 1000 genes, and that's it. So it's still different, but I have no doubt that in the near future people will be able to reach that goal, including my group. **In our research, the technique we are developing is basically to combine these three aspects without losing their spatial information.**

FLG: Thank you very much. Going back to your labs, when you're working on your analysis and isolating these aggressive cancer cells, do you still miss some of these rare aggressive cancer cells? If so, how are you trying to isolate those hard to spot aggressive cancer cells?

Miao-Ping Chien: Great question. During this whole isolation process, we definitely lose some cells for sure. As long as we have a minimum of 50 to 100 good quality cells after the whole process including screening, isolation and profiling, it will be sufficient. Even though we lose some isolation, profiling and analysis, it would be sufficient. The cells with phenotypes that we cannot yet observe under a microscope will be missed, but that is part of our plan for AI implementation that I talked about earlier. We want to develop methods so that they're not only based on recognisable features seen under a microscope; we also want to have a programme that can detect the cells before they display these aggressive features. With a tool like this, we can also identify, isolate and sequence those rare and to-be-aggressive cells. So that's something else we're also developing and implementing.

References:

1. Stuart, Tim, and Rahul Satija. "**Integrative single-cell analysis.**" Nature reviews. Genetics vol. 20,5 2019: 257-272. doi:10.1038/s41576-019-0093-7
2. "**Human Cell Atlas**" Accessed 23/11/2022 https://www.sanger.ac.uk/collaboration/human-cell-atlas/
3. Lewis, Sabrina M et al. "**Spatial omics and multiplexed imaging to explore cancer biology.**" Nature methods vol. 18,9 2021: 997-1012. doi:10.1038/s41592-021-01203-6
4. Eisenstein, Michael. "**Seven technologies to watch in 2022**." Nature vol. 601,7894 2022: 658-661. doi:10.1038/d41586-022-00163-x
5. Li, Mengwei et al. "**DISCO: a database of Deeply Integrated human Single-Cell Omics data.**" Nucleic acids research vol. 50,D1 2022: D596-D602. doi:10.1093/nar/gkab1020
6. "**Framework for multiplicative scaling of single-cell proteomics.**" Nature biotechnology, 18 Jul. 2022, doi:10.1038/s41587-022-01411-1
7. Derks, Jason et al. "**Increasing the throughput of sensitive proteomics by plexDIA.**" Nature biotechnology, 14 Jul. 2022, doi:10.1038/s41587-022-01389-w

Sponsored by

CHAPTER 3

THE MULTI-OMICS APPROACH

MULTI-OMICS, AT ITS ESSENCE INVOLVES BRINGING THE MULTIPLE "OMICS" TOGETHER, SO WE CAN GET A CLEARER AND MORE COMPREHENSIVE PICTURE OF BIOLOGICAL PROCESSES, DISEASE PATHOLOGY, IDENTIFY MORE ROBUST DRUG TARGETS AND BIOMARKERS, AND MORE. **THIS IS THE MULTI-OMICS APPROACH: LINKING GENOME TO EPIGENOME, TRANSCRIPTOME TO PROTEOME, AND BRIDGING THE GAP BETWEEN GENOTYPE AND PHENOTYPE.**

But how do we go about integrating the data together? How do we achieve "The multi-omics approach"? In this chapter, we will explore some case-studies which integrate different "omes" together and show how the whole is greater than the sum of its parts.

Andy Sharrocks
Professor, Division of Molecular and Cellular Function
University of Manchester:

I think the clue's in the name with the 'multi' part of things. **If you want to understand any biological process, if you look at multiple facets of the same cell, you're able to then uncover deeper insights into what's going on in that cell.** That's where multi-omics comes in, because you're able to look at different aspects. For example, you can look at the chromatin level and the gene expression level, and you can integrate those together to understand how chromatin changes lead to gene expression. Then there's the protein level downstream, and that's what actually gives your cell its ultimate phenotype. Multi-omics can help to understand how that phenotype comes from gene expression profiles.

Rong Fan
Associate Professor, Department of Biomedical Engineering
Yale University:

A multi-omics approach is **truly essential to get a clear picture of what's going on. Every layer, each omic, tells you a different story.** To get the complete story, a complete answer to your question, multi-omics is the approach you need.

GENOMICS AND TRANSCRIPTOMICS

GENOMICS AND TRANSCRIPTOMICS CAN BE INTEGRATED TO PRIORITISE DIFFERENT VARIANTS, ANALYSE THE FUNCTION OF GENES, UNCOVER MECHANISMS OF DISEASE, POWER DRUG TARGET IDENTIFICATION, AND FUEL BIOMARKER DISCOVERY. THE FOLLOWING CASE STUDY DETAILS HOW GENOMICS AND TRANSCRIPTOMICS CAN BE USED TOGETHER TO ANALYSE THE FUNCTIONS OF GENES AND UNCOVER DISEASE MECHANISMS.

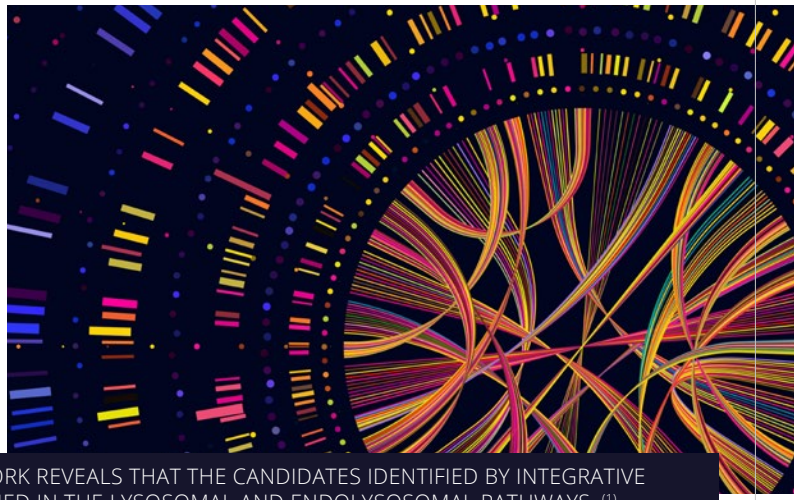
In a recent study published in the journal Human Molecular Genetics, researchers developed a multi-omics approach to discover and validate genes in Parkinson's Disease. [1] **This study highlights the limits of a genomics-only approach and how multi-omics can help fill in the gaps and deepen our understanding of not just Parkinson's Disease, but other complex trait disorders.**

Genome-wide association studies (GWAS) have been the primary method for analysing and comparing different genomes to pick out variants that are associated with disease. GWAS are very effective in identifying specific associations, but further studies in animal models are required to determine their functional relevance and move past correlation to causation. In more complex disorders such as Parkinson's Disease, GWAS can also be inefficient and cumbersome, not to mention time and labour intensive.

This is where Transcriptome-wide association studies (TWAS) (a relatively new technique) can come in handy. **TWAS integrates data from GWAS with gene expression datasets to identify gene-trait associations.** By identifying the associations between gene expression levels and the pathogenesis of the disease, **TWAS can give us a clearer picture of the disease rather than just revealing the associated variants.**

Using a multi-step analysis, the team identified 160 candidate genes. Together with neuronal dysfunction assays and computational analyses, the team whittled down the list to 50 risk genes and 14 potentially protective genes.

In a single screening, the team created a new method which **combined genomics and transcriptomics insights – far more efficient than the sole implementation of each "omic."** This new method of integrated genomic analysis can be used in other complex disorders where GWAS alone doesn't shed enough light. [1]



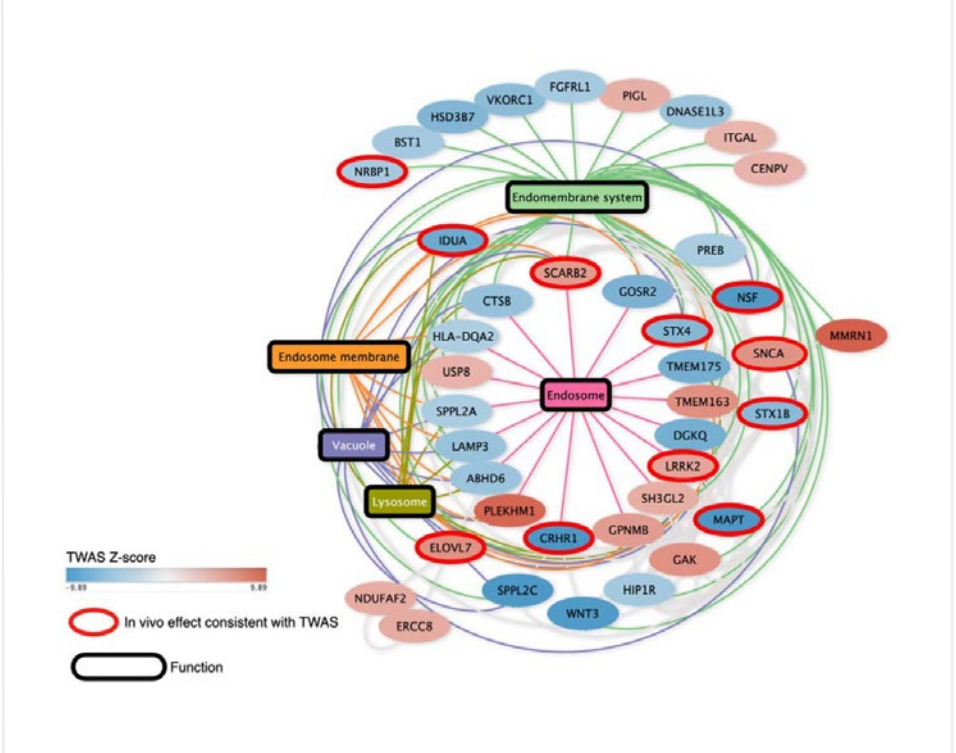


FIGURE 1: PPI NETWORK REVEALS THAT THE CANDIDATES IDENTIFIED BY INTEGRATIVE ANALYSIS ARE ENRICHED IN THE LYSOSOMAL AND ENDOLYSOSOMAL PATHWAYS. [1]

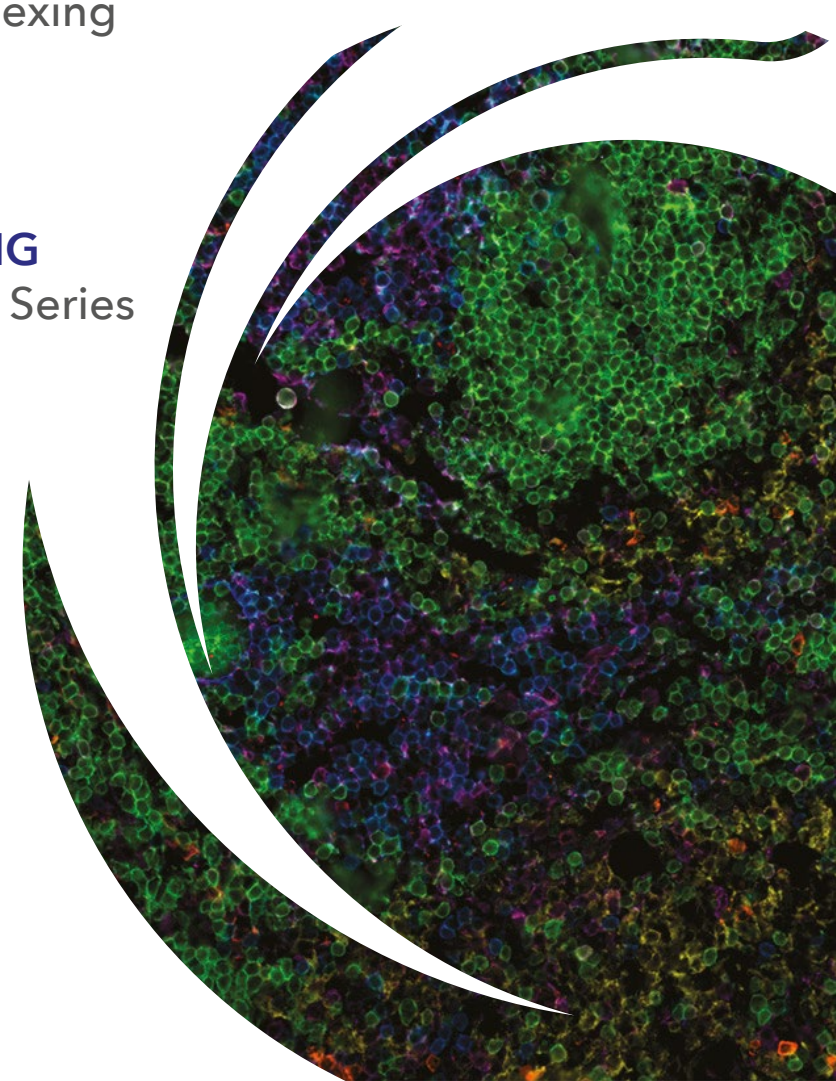
canopy
BIOSCIENCES® A BRUKER COMPANY

See the power of spatial biology in your lab. Or ours.

Instruments & CRO Services for Spatial & Single-Cell Omics

Access cutting-edge spatial biology and multi-omic platforms:

- **SPATIAL PROTEOMICS**
  ChipCytometry™ Spatial Multiplexing
- **SPATIAL TRANSCRIPTOMICS**
  GeoMx® Digital Spatial Profiling
- **SINGLE-CELL RNA-SEQUENCING**
  10x Genomics® Chromomium X Series

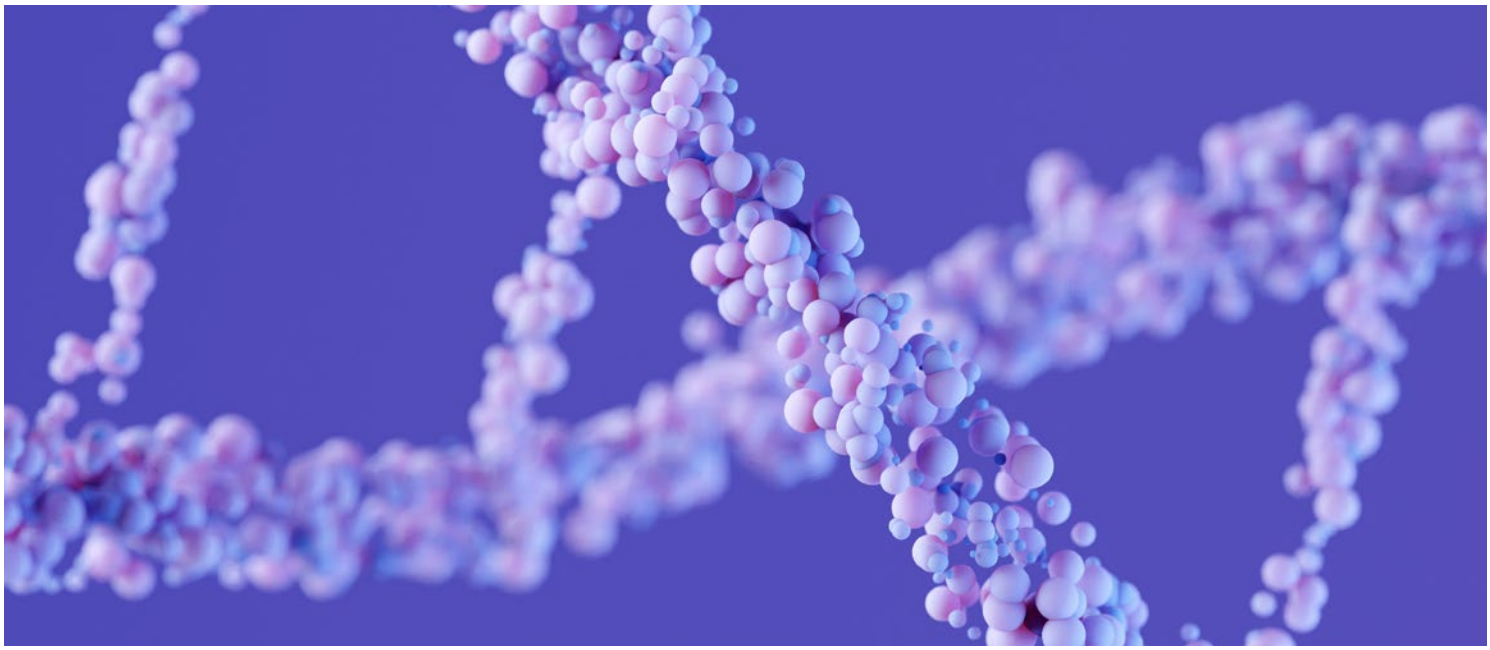


Take your clinical trial to the next level with our CRO Services Team.

Canopy Biosciences.com

EPIGENOMICS AND TRANSCRIPTOMICS

EPIGENOMICS AND TRANSCRIPTOMICS CAN TIE GENE REGULATION TO GENE EXPRESSION, REVEALING PATTERNS IN THE DATA AND HELPING TO DECIPHER COMPLEX PATHWAYS AND DISEASE MECHANISMS. IN THE FOLLOWING CASE-STUDY, BY STUDYING BOTH THE EPIGENOME AND TRANSCRIPTOME, RESEARCHERS COULD DERIVE NEW INSIGHTS INTO BIOLOGICAL PROCESSES AND DISEASE PATHOLOGY.



In a recent study published in Nature, researchers integrated epigenomic and transcriptomic data on biological processes such as injury, repair and remodelling to create a **first-of-its-kind multi-omics map of late-stage myocardial infarction.**[2] Not only did the team **elucidate and comprehensively explore the pathology of this disease,** but by creating this map they have produced an invaluable resource for future researchers to further investigate myocardial infarction, and perhaps even develop new therapies.

The researchers analysed single-cell gene expression, chromatin accessibility and spatial transcriptomic profiles of human hearts from 23 individuals. By analysing the genetic expression of features associated with inflammatory and fibrotic remodelling events, they were able to reveal new insights into remodelling and repair processes, **characterising the differences between healthy functioning parts of the heart and ischaemic and damaged tissue. They then investigated how these remodelling events can create changes in the vasculature and architecture of cardiac tissue.**

The study identified potential gene regulatory mechanisms in cardiac cells and fibroblasts. Differences were defined between different cell states and subtypes of cardiac, endothelial, myeloid cells and fibroblasts. They also revealed a border zone, which separates injured and healthy cardiac tissue. In this border zone, the team conducted further analysis and showed that there was significant upregulation of specific genes such as ANKRD1, a known mediator of cardiac cell responses to stress. Remodelling of tissue in late-stage myocardial infarction was found to be driven by fibrosis and upregulation of specific genes appears to be involved in this process.

The insights from this study alone further understanding of late-stage myocardial infarction by leaps and bounds. But the map created from the data in this study will be a publicly available resource, meaning future researchers can use this atlas of comprehensive information to do further analyses into the transcriptome and epigenome and find out even more about how they regulate and affect the pathology of this disease.[2]

GENOMICS, EPIGENOMICS AND TRANSCRIPTOMICS

THE COMBINATION OF THE SEQUENCING DATA FROM GENOMICS, EPIGENOMICS AND TRANSCRIPTOMICS CAN HELP US UNDERSTAND THE MECHANISMS CONTROLLING SPECIFIC PHENOTYPES, UNCOVER NEW REGULATORY ELEMENTS, HELP IDENTIFY CANDIDATE GENES, BIOMARKERS AND THERAPEUTIC TARGETS. IN THE FOLLOWING CASE STUDY, RESEARCHERS INTEGRATED THESE 3 LAYERS TO BETTER UNDERSTAND, STRATIFY, AND SUBTYPE LYMPHOCYTIC LEUKAEMIA, A COMPLEX DISEASE.

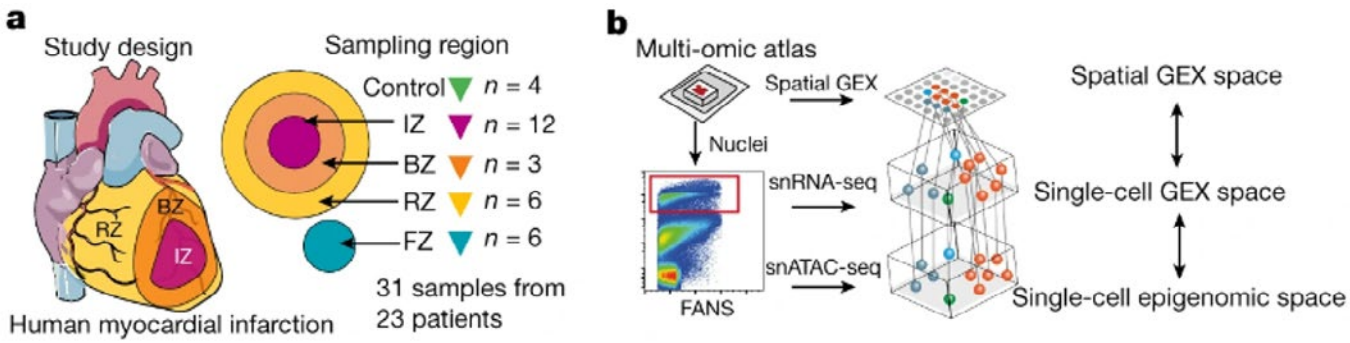
In a recent study, published in Nature, researchers have constructed a comprehensive and **high-resolution map of the landscape of genetic changes in chronic lymphocytic leukaemia** (CLL), a cancer that exists in diverse forms and can have various causes. [3]

This map is the first of its kind to comprehensively characterize the **genome, epigenome and transcriptome of CLL.** Previous studies have only provided **a few puzzle pieces of a CLL map**, and each study has been limited in the patients included, using fragmented, limited, or incomplete data.

The current study integrated genomic, epigenomic and transcriptomic data from over 1,000 patients to identify 202 candidate genes (109 of which were novel). **The team characterized the genes with distinct genomic characteristics and prognoses, as well as their expression patterns, allowing them to subtype CLL to allow for more tailored precision medicine treatments.** The clinical outcomes of the patients were also associated **with a combination of genomic, transcriptomic and epigenomic features. The insights from this comprehensive analysis could allow for better prognosis for patients.**

"We are releasing a CLL map 'portal' that is based on the CLL map and will be an interactive website for translational researchers to use as a resource for further investigation – such as learning more about the different drivers and subtypes of CLL," says Getz, one of the authors of the study. This interactive map could become a potentially invaluable resource for other researchers and clinicians. [3]

FIGURE 2: A) SCHEMATIC OF SAMPLING REGIONS TAKEN FROM DONOR HEARTS. B) SCHEMATIC SHOWING DIFFERENT MULTI-OMIC TOOLS USED TO BUILD THE MAP [3]



GENOMICS AND PROTEOMICS

THE COMBINATION OF GENOMICS AND PROTEOMICS CAN BE VERY EFFECTIVE AS IT ALLOWS YOU TO LINK GENOTYPE DIRECTLY TO PHENOTYPE. THIS APPROACH CAN ELUCIDATE AND CHARACTERISE BIOLOGICAL PROCESSES, HELP US UNTANGLE DISEASE-DRIVING MECHANISMS, AND INFORM THE DEVELOPMENT OF THERAPEUTICS. IN THE FOLLOWING CASE-STUDY, RESEARCHERS COMBINED GENOMICS AND PROTEOMICS TO SHED LIGHT ON CERTAIN PROCESSES INVOLVING THE IMMUNE SYSTEM, HOW IMMUNE CELL NETWORKS ARE ALTERED IN SPECIFIC DISEASES, AS WELL AS HELP DEVELOP THERAPEUTIC APPROACHES.

Immune cells constantly travel around the human body, forming connections and communicating with each other in a complex and dynamic network. This constant communication is vital for maintaining our immune system and fighting off disease, but it is also implicated in the development of auto-immune diseases, such as multiple sclerosis.

In a recent study, researchers developed an **interactome map, detailing the network of connections that make up our immune system by integrating protein-protein interaction data with single-cell genomic datasets of different human tissues.** [4] By systematically mapping these connections, researchers can now gain **an unprecedented level of information and understanding about different biological processes and diseases.**

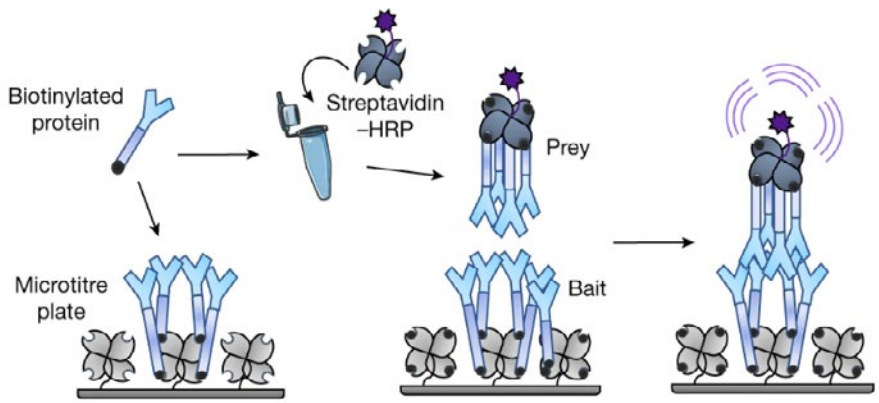
To build this map, the researchers first tested protein-protein interactions using a SAVEXIS (scalable arrayed multi-valent extracellular interaction screen) – a type of high-throughput screening they developed for this purpose. They then independently checked each newly discovered protein-protein interaction to provide information about the biophysical characteristics of each connection. **This protein interactome was then integrated with single-cell genomic datasets of different human tissues, creating a multi-organ map of interactions.**

Finally, the researchers assigned functions to different connections by targeted stimulation of specific proteins in human immune cells and then analysed the proteins with multiplex high-content microscopy. In this way, the researchers created a comprehensive map of connections in the immune system. The scale in creating a map of this size and detail should not be understated; there are hundreds of distinct surface proteins on each immune cell and protein-protein interactions are often transient, hence the need to develop SAVEXIS.

This map systematically documents and describes the intracellular wiring of the immune system, from cell-cell connections down to the biophysical properties of surface proteins. In this study alone, researchers identified several potential therapeutic targets. For example, they identified major histocompatibility complexes HLA-E and HLA-F as ligands for immune checkpoint receptor VISTA (V-domain immunoglobulin suppressor of T cell activation). They also highlighted the SLITRK4 pathway in lymphocyte responses as a pathway that should require further investigation.

However, this map has future applications beyond the scope of this study. The methods and strategies used to develop this map, such as SAVEXIS, could be used in future research to map other cellular networks in the human body. **Furthermore, by disentangling the intracellular wiring of the immune system, this resource could prove invaluable for future research and the development of immunotherapies.**

FIGURE 3: SCHEMATIC SHOWING HOW SAVEXIS WORKS - HIGH-THROUGHPUT SCREENING FORPROTEIN BINDING INTERACTIONS BETWEEN RECOMBINANT EXTRACELLULAR DOMAINS [4]



TRANSCRIPTOMICS AND PROTEOMICS

THE COMBINATION OF TRANSCRIPTOMICS AND PROTEOMICS IS POWERFUL AS IT CAN TIE NEW DISCOVERIES BACK TO KNOWN MARKERS AND CLINICAL OUTCOMES, GIVING INSIGHTS INTO HOW GENE EXPRESSION AFFECTS PROTEIN FUNCTION AND PHENOTYPE. IN THIS CASE STUDY, RESEARCHERS COMBINED TRANSCRIPTOMICS AND PROTEOMICS TO UNRAVEL SPECIFIC CELLULAR PROCESSES, REVEALING THE PREVIOUSLY OVERLOOKED ROLE THE NUCLEOLUS PLAYS IN RNA DECAY.

A recent study in Nature **used proteomics and transcriptomics to demonstrate the role that the nucleolus plays in regulating RNA turnover of pro-inflammatory genes during infection.** [5] The study was comprehensive, investigating mechanisms of action and elucidating the role of the protein nucleolin (NCL) and the Rrp6-exosome complex in this process, furthering our understanding of the molecular pathology of inflammation-associated diseases. The findings could potentially improve therapies for cancer, autoimmune disease, and sepsis.

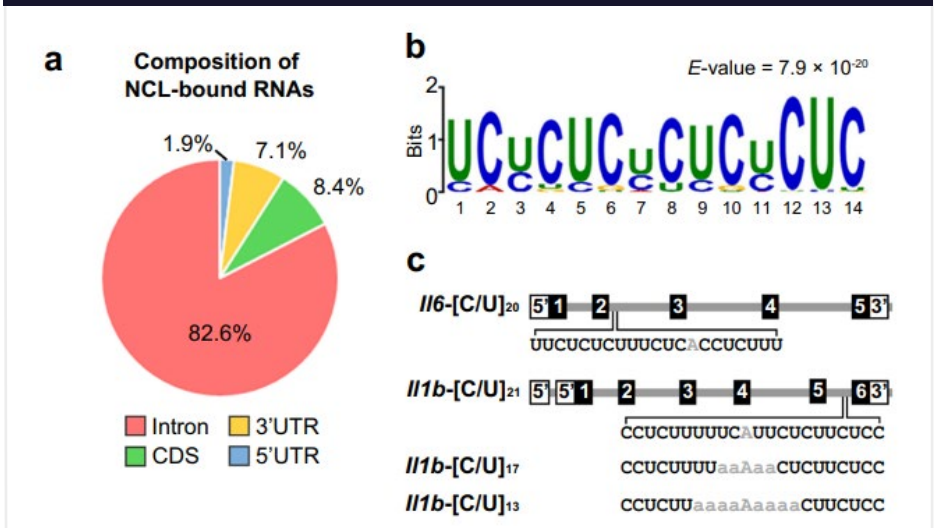
Analysis of RNA-seq data of fractions extracted from cells infected with lipopolysaccharides (LPS) showed that not only do inflammatory genes have higher intronic read densities than non-inflammatory genes, but their mRNAs are highly enriched in nucleoli during infection. What does this suggest? Well, introns are known to increase transcript levels by enhancing transcription, nuclear export, and the efficiency of mRNA translation. This means inflammatory genes are likely highly expressed in nucleoli during infection.

To explore how these inflammatory mRNAs are enriched at the nucleolus, the team screened for potential RNA-binding proteins by analysing **nucleolar proteome datasets** to identify NCL. These datasets also showed that not only does NCL reside primarily in the nucleoli, but it is able to translocate to different subcellular locations in response to alterations in the cellular environment. The researchers then sequenced the mRNAs and analysed the sequences to identify NCL binding sites. Sure enough, NCL depletion caused an increase in inflammatory mRNAs at the nucleolus, and this effect was reversed by overexpressing NCL again.

The team went one step further and investigated how NCL guides inflammatory mRNAs to the nucleolus. **Mass spectrometry analysis** showed that a significant number of polypeptides, including the nucleus-specific exosome component Rrp6 and RNA helicases (DDX5 and DHX36, DEAD-Box Helicase 5 and DEAH-Box Helicase 36) were associated with NCL in response to LPS exposure. Rrp6 is known to facilitate RNA decay in the nucleus by binding to core exosomes, so the team decided to explore the role of Rrp6 in decay of NCL-bound inflammatory pre-mRNAs. They discovered that NCL recruits the Rrp6 complex.

This study is a fantastic example of how **transcriptomics and proteomics can be used in combination to comprehensively elucidate the previously unknown role of a subcellular compartment in a biological process and disease mechanism.** [5]

FIGURE 4: DIFFERENT ANALYSES PERFORMED IN STUDY. A) PIE CHART PRESENTING PERCENTAGES OF NUCLEOLAR RNA READS CAPABLE OF BINDING NCL MAPPED TO THE INDICATED FEATURES BASED ON PAR-CLIP ANALYSIS. B) IDENTIFICATION OF A NCL-BINDING CONSENSUS MOTIF ANALYZED BY PAR-CLIP. C) SCHEMATIC REPRESENTATION OF WILD-TYPE OR MUTANT IL1B OR IL6 RNA PROBES. PAR-CLIP (PHOTOACTIVATABLE RIBONUCLEOSIDE-ENHANCED CROSSLINKING AND IMMUNOPRECIPITATION) IS METHOD FOR IDENTIFYING THE BINDING SITES OF RNA-BINDING PROTEINS.



Spatial Biology Without Limits

Meet CellScape™

The Next Generation of ChipCytometry™ Instrumentation

CellScape is an end-to-end solution for highly multiplexed spatial omics. Combining an advanced, purpose-built imaging system with easy-to-use fluidics for walk-away automation, the CellScape system accelerates exploration in the rapidly evolving field of spatial biology.

CanopyBiosciences.com/CellScape

canopy
BIOSCIENCES® A BRUKER COMPANY

PROTEOMICS, METABOLOMICS, AND MORE

IN THIS REPORT, WE'VE FOCUSSED A LOT ON SEQUENCING, AND A GENETICS-FIRST APPROACH. HOWEVER, MULTI-OMICS CAN GO BEYOND THE PROTEOME, AND USING MASS SPEC WE CAN ALSO LOOK AT THE METABOLOME, LIPIDOME, N-GLYCOME, AND MORE. MOST SPATIAL-OMICS STUDIES USE FFPE TISSUE SECTIONS – HOWEVER, ONE PROBLEM WITH FIXING AND PRESERVING TISSUE IN THIS WAY IS THAT THE USE OF PARAFFIN WAX AND ORGANIC SOLVENTS DEPLETE MANY LIPID SPECIES. THE LIPIDOME CAN OFFER CRUCIAL INSIGHTS INTO THE PATHOLOGICAL STATUS OF A TISSUE, SO LOSING THIS INFORMATION CAN BE DETRIMENTAL. HOWEVER, RECENT STUDIES USING ADVANCED SPECTROSCOPY HAVE SHOWN THAT SOME SOLVENT-RESISTANT LIPID SPECIES ARE MAINTAINED. **ANDREW SMITH, ASSISTANT PROFESSOR, UNIVERSITY OF MILANO-BICOCCA,** RECENTLY PUBLISHED A PAPER WHERE HE DESCRIBED A NOVEL WORKFLOW FOR SPATIAL MULTI-OMICS OF LIPIDS, N-GLYCANS, AND TRYPTIC PEPTIDES ON A SINGLE FFPE TISSUE SECTION[6]. WE SAT DOWN TO TALK TO HIM ABOUT HIS WORK.

FLG: Could you tell me a little about yourself and the work you do please?

Andrew Smith: My general area of expertise involves the use of mass spectrometry imaging, a powerful technique which allows you to visualise the distribution of various biomolecules in pathological tissue. Our unit is flanked by a large clinical centre and this definitely facilitates the type of work that we do. As a result, a large portion of our research work involves the use of MS-based imaging approach to determine how certain biomolecules are altered in various pathologies and how this information could be exploited to obtain findings of clinical relevance.

My principal line of research throughout the last few years has focused primarily on glomerular diseases of the kidney. However, I have also recently expanded my line of research to focus on a number of different oncological contexts. For the most part, I've focused on the spatial proteome but, in recent years, and with the help of my colleagues, I have also developed protocols which have enabled, or facilitated, the possibility of **performing spatial lipidomics in formalin-fixed paraffin embedded tissue,** which is of course the gold standard tissue type in clinical centres.

FLG: You recently published a paper titled spatial multi-omics of lipids and glycans and tryptic peptides on a single FFPE tissue section, where you describe the novel workflow you developed to enable mapping of lipids and in glycans and peptides on mouse brains and clear cell renal carcinoma tissue. Could you tell us what's novel about this approach?

Andrew Smith: Matrix-assisted laser desorption mass spectrometry (MALDI-MS) based imaging has been around for a number of years now, and there are several previous works that have demonstrated the possibility to map multiple biomolecules on a single tissue section, both fresh frozen tissue and FFPE tissue. **Now, for the first time, we've also been able to map lipids in FFPE tissue sections as part of a sequential, multi-omics workflow.** This was really one of the most challenging aspects that we've had to overcome, because quite commonly, in FFPE tissue, a large number of lipids are depleted during the fixation and embedding process. So our approach has enabled us to maintain this spatial lipidome - I think that's one of the novel aspects. Another novel aspect is related to how we analysed and integrated the various datasets.

So, does integrating all of this data together increase your capacity to characterise the various regions within the tissue? **It's very nice to visualise the distribution of various biomolecules given that each molecular class provides a complementary piece of information to add to the pathological jigsaw.** Moreover, there are many disease cases where considering the proteome alone does not provide you with sufficient depth, or coverage, to obtain diagnostic, prognostic, or predictive information about that tissue. By integrating different layers of biological insight together, you can better stratify patients compared to using a single omics level alone.

With this new technique, I tend to think of (because I live in Italy) the Roman deity Janus which has 2 faces. With this technique there are two uses which I hope will give this approach further legs: You can either use it on the whole tissue for patient stratification, in a tissue typing manner, or use it as a guide to individuate altered cells, which can then be excised for more in-depth investigations.

FLG: You mentioned the challenge with FFPE tissue sections, but what are some of the other problems and challenges that you had to overcome when developing this workflow?

Andrew Smith: One of the biggest challenges was related to **maintaining the spatial localisation of the biomolecules** following the multiple analytical steps and deciphering when the most appropriate moment was to perform antigen retrieval. This was a question we posed in a paper published about two years ago, that also incorporated an antigen retrieval step to help liberate some of the membrane lipids that become trapped during the formalin fixation process. However, in this sequential workflow, we found that it [antigen retrieval] had to be performed following spatial lipidomics to ensure that both the N-glycans and proteins did not delocalise to a significant degree. So, a major challenge was **organising the protocol in a way that ensured molecular localisation was maintained at all stages.**

FLG: How did you maintain that spatial localisation throughout your experiments to make sure you mapped the proteomic data or the lipid data accurately?

Andrew Smith: There are a variety of different analytical steps which can result in bioanalyte delocalisation. One of these steps enzymatic deposition. In order to liberate N-glycan structures from their linked proteins, you need to use an enzyme such as PNGase F. Subsequently, you use proteolytic enzymes, such as trypsin, which are selected based on the protein coverage you desire. Given that we required our analytes to remain in situ, the deposition of these enzymes should be performed to ensure that only small droplets of the enzyme accumulate on tissue, but at the same time ensuring that there is sufficient water content for efficient action of the enzyme. This is a difficult balance, and a compromise between the two aspects has to be struck. This was the first challenge, but it's a challenge associated with MALD-MSI analysis of FFPE tissue in general.

The second aspect is related to the deposition of the MALDI matrix. For MALDI-MS, an organic matrix is required and is responsible for extracting your biomolecules from the cells of the tissue, incorporating

them within a network of co-crystals, as well as assisting in the ionisation process. Once more, depending on the nature of the biomolecules present within your tissue, a balance between the amount, or density, of the organic matrix and the size of the resulting co-crystals must be struck and again required optimisation in our instance. Analyte extraction that is too reserved can lead to limited sensitivity, whilst on the other side, too much can result in analyte delocalisation.

Finally, you may also encounter similar issues as a result of some of the tissue washing steps. In particular, we noticed that the MS-imaging of N-glycans require the tissue to be sufficiently rehydrated in order to ensure the efficacy of the digestive enzyme (PNGaseF). So again, it was about finding the right balance to have sufficient tissue rehydration without promoting further analyte delocalisation.

> "MUCH WORK IS BEING PERFORMED WITHIN THE MS-IMAGING COMMUNITY TO ENSURE THAT PROTOCOLS AND METHODS CAN BE BROADLY APPLIED BY MULTIPLE CENTRES, PRODUCING ROBUST AND REPRODUCIBLE DATA."

"IF YOU HAVE DEVELOPED AN APPROACH THAT IS HIGHLY SPECIFIC, AND YOU GET ANOTHER RESEARCH GROUP TO TRY IT WITH ANOTHER TYPE OF SAMPLE OR EVEN A VERY SIMILAR SAMPLE, IT SOMETIMES JUST DOESN'T WORK WHATSOEVER AND THE RESULTS CAN BE VASTLY DIFFERENT."

FLG: Could you tell us a little bit more about the N-glycome and the lipidome, why it's important to study them, and what unique biological insights they can give us?

Andrew Smith: The N-glycome, especially in tumour biology, has really come to the fore in the last 5 to 10 years. This is because inside cells, we have glycol-transferase enzymes, which are responsible for the synthesis and production of these N-glycans. **Under certain pathological instances, modifications in gene expression can change the activity of these enzymes, which has a downstream effect on N-glycan structures.** And in certain types of cells, this may mean that you have different N-glycan receptors, which are expressed on the surface of the cell. **This changes their functionality or their phenotypic state.** Therefore, the N-glycan receptors which are expressed on the surface of the cell can be very good indicators of the phenotypic nature of that particular cell. **It is becoming ever more evident that there seems to be a number of interactions between these different tumour cell phenotypes based on the N-glycan receptors present on the cell surface and the other immune cells which are circulating in the tumour microenvironment as well.** This again highlights the relevance of using a spatial approach that considers the different cell types that communicate with each other within this complex landscape.

Lipids play a variety of different roles in human biology. They may affect the structure of a cell, serve as an energy source, and can also function as signalling molecules. In clear cell renal carcinoma (CCRC), the type of tumour that was used as proof of concept in this study, is characterised by lipid accumulation within the cells, resulting in their "clear" appearance. In fact, this accumulation of lipids is also associated with chemo-resistance in certain cancers, as well as CCRC specifically. **So, the lipidome, the N-glycome, the proteome, are all providing complementary pieces of information, and by integrating them together you can have a greater molecular understanding of the processes that drive pathology.**

FLG: What tissues do you anticipate your workflow being used on?

Andrew Smith: In theory, there shouldn't be any particular limitations regarding the type of tissue to which it can be applied. For example, workflows for MALDI imaging of tryptic peptides can, and have been, be applied for a variety of different tissue types, as is evident within the extensive literature. However, as touched upon previously, the workflow, in particular the enzyme and MALDI matrix deposition, should be optimised for that specific tissue type.

FLG: Most of the time researchers produce a workflow that is specific to the experiment they're doing. Whereas yours, as you said, it's not as limited because theoretically, you should be able to use it for other tissues. Do you think that when we're developing approaches we need to make sure that they're more generalisable, and they're applicable in different contexts so that we can actually collect more data that can be compared, rather than working in isolation?

Andrew Smith: I certainly think this would help. If you have developed an approach that is highly specific, and you get another research group to try it with another type of sample or even a very similar sample, it just doesn't work whatsoever and all the results are vastly different. It would really be beneficial to take a base protocol and only have to modify or optimise it slightly for that tissue type, and then be able to incorporate it within an existing workflow.

I think we're quite fortunate in this aspect because, with the technology we use, the general broad workflow between different tissue types does remain the same – only a slight optimisation in terms of enzyme deposition and matrix deposition is required. Of course, instrumental parameters need to be tweaked depending on the type of tissue as well, as you can't guarantee that the dynamic range that you have in kidney tissue is going to be the same as what you have in the brain for example. However, much work is being performed within the MS-imaging community to ensure that protocols and methods can be broadly applied by multiple centres, producing robust and reproducible data. This is of course very promising if the final goal is to push the technology towards clinical utility.

References:

1. Li, Jiayang et al. **"Integration of transcriptome-wide association study with neuronal dysfunction assays provides functional genomics evidence for Parkinson's disease genes."** Human molecular genetics, ddac230. 29 Sep. 2022, doi:10.1093/hmg/ddac230
2. Kuppe, Christoph et al. **"Spatial multi-omic map of human myocardial infarction."** Nature vol. 608,7924 2022: 766-777. doi:10.1038/s41586-022-05060-x
3. Knisbacher, B.A., Lin, Z., Hahn, C.K. et al. **"Molecular map of chronic lymphocytic leukaemia and its impact on outcome."** Nat Genet 2022. doi:10.1038/s41588-022-01140-w
4. Shilts, Jarrod et al. **"A physical wiring diagram for the human immune system."** Nature vol. 608,7922 2022: 397-404. doi:10.1038/s41586-022-05028-x
5. Lee, T.A., Han, H., Polash, A. et al. **"The nucleolus is the site for inflammatory RNA decay during infection."** Nat Commun 13, 5203 2022. doi:10.1038/s41467-022-32856-2
6. Denti, Vanna et al. **"Spatial Multiomics of Lipids, N-Glycans, and Tryptic Peptides on a Single FFPE Tissue Section."** Journal of proteome research vol. 21,11 2022: 2798-2809. doi:10.1021/acs.jproteome.2c00601

CHAPTER 4

DATA INTEGRATION AND BIOINFORMATICS

DATA INTEGRATION IS THE PROCESS OF COMBINING DIFFERENT OMICS DATASETS, ALLOWING RESEARCHERS TO STACK THE MULTIPLE LAYERS OF BIOLOGICAL INSIGHT TOGETHER TO GET THE WHOLE PICTURE. **INTEGRATION IS AT THE CORE OF THE MULTI-OMICS APPROACH – HOWEVER, THIS STAGE IS OFTEN CITED AS THE MOST CHALLENGING.** [1]

The optimal data integration strategy or approach depends on several factors. Firstly, the biological question being addressed has an impact. Different approaches can be broadly split into 3 categories: disease subtyping, disease insights and biomarker prediction. Another factor is the data: data type, quality, size and resolution can impact how the data should be analysed, interrogated and integrated. The third is the experiment itself – the animal, and even the tissue type, can impact which tool or package to use. [1]

We've already covered how rapidly technology in the multi-omics space is advancing, and the bioinformatic and computational sector is no exception. In fact, the pace of new packages, tools, and software releases can be overwhelming. Being surrounded by so many choices can often be more confusing than helpful. Add to this the fact that computational biology is a highly specialised and complicated field and **picking the right approach can be a daunting task.**

Furthermore, recent advances in areas such as long-read sequencing, single-cell and spatial have allowed for higher resolution and consequently more data to be collected. Handling big data and filtering out the noise is particularly challenging. Digesting and visualising big data in a reproducible and robust way is essential. [2]

In this chapter, we have included a discussion roundtable with top researchers who have developed tools and packages for the integration of multi-omics data, so you know what limitations, considerations and challenges you may face when you embark on a multi-omics study. To close things off, we will be looking at different data integration tools with Lihua (Julie) Zhu, who has developed numerous different tools and packages for multi-omics data integration. In the next chapter, we will discuss the emerging role of AI and Machine learning in data integration, and how the incorporation of these technologies has allowed researchers to tackle big data at scale.

mission bio

Ask the questions
you could never ask before

What if I could precisely distinguish leukemia from clonal hematopoiesis?

The Tapestri® single-cell multi-omics (scMRD) platform integrates mutational and immunophenotypic signatures of the same cell, enabling measurable residual disease detection with unparalleled resolution, sensitivity, and specificity.

Mission Bio is now offering a Tapestri scMRD early access program for acute myeloid leukemia.

Learn more at go.missionbio.com/scMRD-AML-EA

MISSIONBIO.COM

ROUNDTABLE DISCUSSION: DATA INTEGRATION TIPS, CHALLENGES AND THINGS TO CONSIDER

WE SPOKE TO SEVERAL TOP RESEARCHERS WHO HAVE DEVELOPED TOOLS AND PACKAGES IN THE MULTI-OMICS SPACE ABOUT THE THINGS RESEARCHERS NEED TO CONSIDER AND THE KEY CHALLENGES IN DATA INTEGRATION.

STEPHANIE BYRUM
Associate Professor, Department of Biochemistry and Molecular Biology
University of Arkansas at Little Rock

DAVID RUAU
Head of Strategic Alliances, Drug Discovery AI
NVIDIA

JOHN QUACKENBUSH
Professor, Department of Computational Biology and Bioinformatics
Harvard T.H. Chan School of Public Health

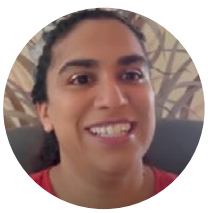
MIAO-PING CHIEN
Assistant Professor, Department of Molecular Genetics, Erasmus University Medical Center, Principal Investigator
Oncode Institute

JIANGUO (JEFF) XIA
Assistant Professor, Department of Bioinformatics and Big Data Analysis
McGill University

LIHUA (JULIE) ZHU
Professor, Department of Molecular, Cell and Cancer Biology
University of Massachusetts Chan Medical School

NIKOLAI SLAVOV
Associate Professor, Department of Bioengineering
Northeastern University

MATHEW CHAMBERLAIN
Principal Scientist
Janssen



REBECCA MATHEW
Principal Scientist
Merck Research Labs

MARSHALL SUMMAR
Director, Rare Disease Institute Laboratory
Children's National Hospital (Washington D.C.)

FLG: What are some of the considerations you need to make when developing your data integration approach?

Stephanie Byrum: Your data integration approach is very dependent on how you set up your experimental design at the beginning. You want to make sure that you're comparing apples to apples and not different cells. You want to make sure you haven't done different sequencing at different times and introduced all **these different batching events**.

Ideally, it would be best to think about your data integration approach **at the very start of your study.** Then you can get your data from the same samples, extract out your DNA, RNA and protein, and really store that well. You limit the technical variability, and you limit the degradation of the molecules – this really helps prevent batching. The more you can think about it at the beginning, the better off you are.

David Ruau: It's important to think about the end goal and start by using a toolkit that will scale and help you deploy. In omics research, the data and tools are very diverse, which means a researcher will likely use many of them one after the other. **This is a good reason to orchestrate workflows with tools like workflow description language (WDL), to make the process more efficient.** Using an accelerated library can complete research faster, without having to wait days for computation to finish before running the next project.

John Quackenbush: Any model we build is going to be an approximation based on our knowledge and understanding of the world around us, the world we are trying to explore. A lot of the models my colleagues and I build are gene regulatory network models. We recognise that they are incomplete, that the data we're using to build the models is noisy, that we are sampling a relatively small population and trying to understand 1000s, or 10,000s, of interactions between genes. Moreover, the models we build are of a process that's fundamentally dynamic, but the kind of information that we have is from a sample that's at a specific point in time. So the models we are trying to make with the static data we have are at best approximations of these dynamic regulatory processes. We try to learn the structure of the regulatory network, and then iteratively optimise the network given the data. What we always try to come back to is asking ourselves, well, does the model actually inform our understanding of the systems we are studying?

We are constantly checking to ensure that we are adding data in a principled way and improving our understanding of these regulatory processes so that we can validate our networks. Invariably, if we do this well, we find that our ability to predict things really does improve over time. **The fundamental thing to bear in mind is data by itself isn't necessarily useful, but more data collected on the right samples in the right way, and then integrated in a way that respects our knowledge of biology, can really lead us to a better understanding of the processes we are studying.**

The thing I always tell my students and colleagues is that in doing science, **the starting point is asking the right questions.** The first step is making sure the technology you choose is appropriate for the question you're trying to address. What is the right tool to generate the data to answer those questions? The other thing is to understand the limitations of the data that we can generate. How noisy is it? How reproducible is it? What are the assumptions that are going to go into this analysis? What are the right tools to make best use of the data?

Miao-Ping Chien: The most important challenge in the integration of multi-omics data is to link data from different sources in a way that is biologically meaningful. I would also consider the data integration approach right at the beginning when designing your experiment. For example, subjecting all the samples to sequencing at once, instead of sequencing them sequentially can avoid the concern of the batch effect.

Another example is making sure you have enough replicates. **It is always wise to have enough technical and biological replicates in order to draw a robust conclusion.** When only one biological sample is accessible, then having technical replicates (2-3 times) is advised. If we only have one repeat, technical biases or noise will sometimes be identified as biological findings; this can be avoided by running multiple repeats. Also, when deciding which method to use, carefully evaluate the benchmark methods and discuss them with experienced researchers in the field.

Jianguo (Jeff) Xia: My team is really passionate about omics data analysis and integration, especially about making multi-omics data more accessible to users. **We are motivated to assist our scientist or researcher's mind with a more intuitive computational framework.** This requires taking into consideration of both the backgrounds of different researchers and the context of the data. **A lot of contextual data, metadata, like gender, age, treatment, time, location and other information are important during data analysis.** We need to bring the data analysis closer to the domain experts of the biological processes or diseases that the multi-omics data aim to address. There are a lot of new technologies in computing and visualization that can help the process. Part of my research aims to leverage these new technologies to empower researchers.

"THE MOST IMPORTANT CHALLENGE IN THE INTEGRATION OF MULTI-OMICS DATA IS TO LINK DATA FROM DIFFERENT SOURCES IN A WAY THAT IS BIOLOGICALLY MEANINGFUL."

Lihua (Julie) Zhu: I think one major thing to consider with data integration is the **batch effect.** You can have all this data, be it Chip-seq data, ATAC-seq data, you name it, but **often this data isn't collected at the same time, or from the same cell, or the same sample, sometimes not even from the same lab.** There are computational approaches that can account for this, but they have their own limitations. **It's a major challenge and something researchers should be very careful about.**

FLG: Many researchers struggle finding the right tool for the right question, especially as there are so many different tools, packages, etc., being released all the time – it's hard to stay up to date. What advice do you have for researchers who may not be computational experts in how they approach data integration?

Stephanie Byrum: It's very overwhelming. And most of the time, biologists probably do not have that domain expertise. **That's where I think your core facilities come into play.** I direct our bioinformatics core, and we're trying to initiate workshops, training, and we have a lot of people interested in single-cell RNA-seq, so we are trying to put together a workshop on that. The idea is to really try to incorporate an education around where your samples come from, how you prepare your data, but then the actual downstream analysis part we will probably do as a fee for service in the core with the experts that know how to do it. We already have experts that run mass spectrometers – you're going to have experts on the analysis side too.

At the back end of that, I like to meet with the students and try to incorporate the training aspects as well. We also have special courses that go over omics platforms – like with our graduate school courses – and so there's this whole training and education component that has to happen. But if somebody is going to use this technology just once for one study, they don't necessarily need to be an expert in everything. **Sometimes you get PIs who ask their students to analyse everything, but they're asking them to get multiple PhDs – it becomes unrealistic.**

That's something that we're trying to consider when developing models too: How do you take something and not make it so specific to your experiment, but generalise that model so that you can pull-down other datasets from TCGA or any other sequencing core and still use the same algorithms to get information out of it? **I think that becomes a problem – when we develop tools, we'll make them very, very specific to the data that it was developed with.** I'm trying really hard to not do that. It's a challenge for sure. But by utilising our own datasets, we can build and test with the simpler set of experiments. We know the ins and outs, we know what's validated, we can go back to the wet-lab, validate the biology, but then also apply it to publicly available datasets and see how it's going to perform with those that have also been validated, but have been sequenced by other labs.

Jianguo (Jeff) Xia: I think with multi-omics data, **data integration is an iterative, ongoing process.** Researchers doing multi-omics must keep open-minded and **not to just blindly use the tools to get a result. Otherwise, they are jumping to conclusions** rather than letting data speak for themselves, which could miss important discoveries enabled by multi-omics data. Most people who do multi-omics usually have experience doing a single omics, so it should be natural to use an iterative approach. For example, if they're familiar with RNA-seq, then start from the RNA-seq data, do a thorough analysis and use their favourite tools, and then start thinking about the big picture informed by the RNA-seq result. Gradually add another layer and start thinking about whether it's making sense or not. Basically, have an incremental approach – start from your comfort zone and gradually become more adventurous.

However, you cannot have mathematicians and computer scientists happily in their own domain, give them the tools and hope that it will be useful to clinicians. People need to work together and there needs to be that dialogue; they need to communicate. I don't believe clinicians need to learn advanced stats or Machine Learning algorithms. But there needs to be that happy middle ground, the computer scientists need to learn a bit about disease and the clinicians need to learn a little bit about why you're doing this or the concept behind it. In the end, everybody learns a bit more and everybody will converge on a space where it seems accessible to everyone. We are already moving towards this direction because of cloud computing which we use tremendously.

David Ruau: Tools and packages in the -omics spaces are often research outputs, and seeing many groups work on tools around a shared technology (e.g., single cell sequencing analysis) is a good indicator that the technology is ripe for production. The diversity in tooling, however, often highlights that there is still a lot of unexplored territory, with varying degrees of uncertainty. If you do look at the spaces where multiple tools or packages exist, you will discover that there are some common denominators – data standards or visual representation approaches – that are cross-tool, which could be the starting point to a data harmonisation campaign.

Miao-Ping Chien: It's a very good question. I would first suggest reading papers and reviews that outline detailed comparisons between different state-of-the-art and benchmark methods. For example, papers, for example a recent review by Lahnemann et al., detail 11 grand challenges that researchers may encounter and need to be aware of when running single-cell experiments and analyses. [3] It also guides researchers on how to select the right tools for different scenarios. Another practical piece of advice would be to join (local or international) communities. For example, in the Netherlands, we have a Dutch single-cell network community, where many newly developed methods or tools are presented and discussed - people can also brainstorm together through this platform.

FLG: The high resolution of new technologies such as long-read, single-cell and spatial mean that researchers can now collect more data than ever. But converting that raw data to informative data and filtering out noise can be very difficult. Do you have any advice to give on handling big data, and finding those unique insights?

Nikolai Slavov: I tend to see data analysis as being a very exciting and productive component of what we do. It is true that we generate gigabytes, sometimes terabytes, of data, and those need to be analysed. Fortunately, we have access to clusters that make this analysis quite doable.

I would say that data analysis in terms of the volume of the data is certainly not the bottleneck. **Current algorithms are not extracting as much from the data as I believe we can extract, so one can say that they're limiting – this would represent the glass being half empty. But the half-full part of the glass is the opportunity to advance those algorithms so that we can interpret a lot more from the data.** We think that's a very exciting opportunity. Big data, data interpretation, data analysis, Machine Learning – they're certainly very important parts of the field. I think we'll see many of the advances ahead of us be driven, or at least aided, by improvements in data analysis – and you're certainly not limited by computational power.

"YOU CANNOT HAVE MATHEMATICIANS AND COMPUTER SCIENTISTS HAPPILY IN THEIR OWN DOMAIN, GIVE THEM THE TOOLS AND HOPE THAT IT WILL BE USEFUL TO CLINICIANS. PEOPLE NEED TO WORK TOGETHER AND THERE NEEDS TO BE THAT DIALOGUE; THEY NEED TO COMMUNICATE.

Miao-Ping Chien: It's hard to answer this question in a few sentences, because it depends on what biological insights researchers would like to extract from the big data and what data sources researchers have. I would recommend reading reviews such as those by Angerer et al., which highlights the inherent opportunities and challenges in the context of big data analytics. [4] Another growing attention in terms of finding unique insights from big data is the use of deep learning. Read on to Chapter 6 to learn about the AI and ML.

David Ruau: Each case is problem dependent. One option is to go down the exploratory route and throw every model you have at the problem, but this can be quite time-consuming. Usually, most of the data is noise. The real advantage is knowing how to filter effectively if you know what you're looking for.

John Quackenbush: Handling big data is hard! Big data gives us an opportunity, but it's one in which we have to be very thoughtful about the ways in which we take this information and use it to make discoveries. **I often tell people the biggest challenge in being a scientist is asking the right question.** Then the next challenge, of course, is having the right data to be able to answer the question. I'm not afraid of big datasets; I welcome having them. We can do a lot more with them if they're big than we can if they're small. Technology such as single-cell sequencing is extraordinarily exciting, as we can start to understand the diversity of biological processes, even in a single individual and their collection of diverse cells. **People often ask me; would you rather have single-cell data on 10,000 cells or whole genome data on 1000 individuals? And my answer is: it depends on the question I want to ask.** There are some questions where the individual cell data is not too informative, but there are other places where whole genome profiles on a population are not very informative. It depends on the question, but more data gives us the opportunity to ask and answer more questions.

However, with all these new technologies and the data they produce, we are running into the same problems we saw years ago with DNA microarray. It's so interesting because there's almost nothing new under the sun. Every time there's a new technology, scientists using it almost have to relearn the same lessons we've learned before. When we started doing experiments with DNA microarrays, everybody complained the data was noisy, everybody complained the data was not reproducible. Everybody tried to analyze the smallest number of samples possible because it was expensive and justifying small numbers because microarrays were so much better than the previous technology. **But once we get beyond those bad assumptions, we have to go back to the fundamental basics of doing good experiments, we have to use technologies that we optimise, we have to follow good protocols, we have to check for batch effects.** We want to do statistical analysis and tests on our data instead of just kind of eyeballing them, and we really want to do our due diligence – analysing the dataset to make sure that what we're measuring is good, and that the conclusions we draw from them are reliable.

I mentioned DNA microarrays because they had become fairly reliable when a new technology broke onto the scene, which was RNA sequencing.

I remember going to meetings and hearing people say, well, with microarrays we needed lots of replicates, because the data was noisy, but with RNA-seq we can only do two or three samples because we're counting RNAs so it's a much more reliable technology. And you know, the funny thing is, there are advantages and disadvantages to RNA-seq, but it's still noisy and it has its own sources of noise. **Having the new technology never meant that you could draw more meaningful insights by looking at fewer samples—because you hope that any noise is going to be small so that the signal should emerge from the noise.** But at the end of the day, our ability to detect signals really depends on the number of samples we examine and how big the noise is. At the end of the day, you have to look at each dataset individually, clean them as well as you can, and respect them for their limitations. And if you do, you can learn something from nearly every dataset.

Mathew Chamberlain: There are many technical and computational challenges associated with a new data type and new technology – as it's all very new. Defining computational workflows with bulk sequencing microarray has been around for 10-15, even 20 years. **However, in single-cell, best practices are still being established.**

Rebecca Mathew: There is such a large amount of data that comes out of these individual experiments. **Reaching a stopping point where you're ready to interpret the data and stop refining the analysis pipelines can be a challenge as well. There's often a desire to do both; with these large datasets, you want to be able to come back to interpretation, especially on the biologist's side.** That's what we're eager to see; interpretation from these experiments. That can be a challenge as well.

Stephanie Byrum: Higher resolution has definitely impacted my approach to data integration. Long-read is definitely useful if you're looking at those repetitive regions, and if you're trying to find other variants, structural variation is a big thing. We are also looking at a lot of epigenetics. You want to see the chromatin structure along with those long reads. With long read technology, they have higher error rates, which you can fix with short-read sequencing. So again, that's another layer of integration. You could actually utilise both long-read and short-read to call your variants and get structural information as well as SNPs. It all has to come back to experimental design, what's your biological question? And so whatever question you're trying to answer dictates which technologies we're going to utilise and what we're trying to get out of it. Sometimes I think we try to put too many factors in the same experiment, and then you lose some of your power. And you have a harder time interpreting those results. **So, you don't want to make it overly complex. You want to have - here's my question, and then how are we going to answer that question?**

"CURRENT ALGORITHMS ARE NOT EXTRACTING AS MUCH FROM THE DATA AS I BELIEVE WE CAN EXTRACT, SO ONE CAN SAY THAT THEY'RE LIMITING – THIS WOULD REPRESENT THE GLASS BEING HALF EMPTY. BUT THE HALF-FULL PART OF THE GLASS IS THE OPPORTUNITY TO ADVANCE THOSE ALGORITHMS SO THAT WE CAN INTERPRET A LOT MORE FROM THE DATA."

Another thing to consider is replicates. The power analysis researchers have to do for grants is not the right power analysis. It's all coming from classical statistics and not necessarily Machine Learning or regression-type models. So, I feel like from a grant perspective, the way we do power analysis has got to shift. A t-test isn't appropriate but we have to put it in there. There needs to be a better way to evaluate the power of these models. There are some power calculators for DNA methylation, and some other tools because you're trying to get to what's the depth of the molecules present in your data that you're expecting. Are you expecting 2000 proteins because it's blood serum? Are you expecting 10,000 proteins, so the level of protein depth can change the power of your model? And so that could help us with the reproducibility of statistics. **We're kind of still stuck in the old power analyses a bit. I think we are finally getting there with the push into big data - I can see a turning point there.**

Jianguo (Jeff) Xia: That's a great question. Big data, multi-omics data, is still expensive at the moment, so you must have a very careful plan. **I think high resolution, temporal multi-omics data** is great for precision medicine. **Before we venture into this, we need to have a good knowledge of reference baseline and a refined hypothesis.** A lot of statistical thinking about variance and replication must be interpreted with the context of the reference baseline because if we see something different, we want to know whether this change is meaningful, functional or just noise. This is challenging at the moment due to lack of such baseline.

Lihua (Julie) Zhu: I think when handling big data, the pre-experimental design process is really important. You have to make sure you have enough power – otherwise when you are differentiating between those true biological insights and noise, or artefacts, it's hard to validate and compare if you don't have enough replicates and sequencing depth. You want to make sure you have high replicability, sufficient sequencing depth, and low variability, and the signal of interest should rise above the noise if you have enough power.

FLG: Multi-omics data is very heterogeneous, and the lack of standardisation between many different datasets is also a big challenge. What are your thoughts on this?

Stephanie Byrum: I think standardisation is starting to happen already. The FAIR principles have just come out, and the NIH is initiating a new data management sharing policy going into effect in January. So, I think more and more people are having to think about these issues and metadata is key. A lot of times they only want to give you some of the information upfront, but you need to have sample metadata sheets that are consistent across projects and have information on the parameters of the instrument for example. For proteomics data, more information is required upfront – you have to say this is the instrument we used, this is the pipeline that we ran, here's the methods for the project. I don't necessarily have to do that with some of the DNA datasets. **The more we communicate about the workflow and the parameters, and conduct provenance tracking, the better. So as you're building out the pipeline, you're building the workflow and recording each of those steps, noting all the parameters that were used.** But with the data we currently have, hopefully whatever molecules overcome the biology are going to be robust no matter what we do. So those are usually the molecules I'm looking for – no matter what algorithm I throw at it, it doesn't go away.

To tackle this problem, we developed a new proteogenomic pipeline, because we can't always identify genetic mutations in our protein datasets – **often with multi-omics the annotation doesn't link up.** So you have all these different databases that are updated at different times. So one ID may not match another in a different molecular-type database. Annotation is really our slowest step with projects that we're working on. The idea behind the tool is that instead of having to curate databases for every individual project, we actually make the tool do it for us. You can select whatever reference genome from the genomic level that you're working on, whether that's the HG 19, or the 38, whatever version it is for your dataset. You can select that and then merge that with the protein datasets and have standard curated IDs and annotation. **We want to set up data standards for integration, alongside the FAIR standards and those types of things.**

It's a challenge for sure because things are constantly updating and it can get all over the place, but we're trying to standardise more of the annotation workflow with this part of the pipeline.

That will bring together variants, so if you run a variant caller from genetic mutations, we can also include those isoforms in the curated database. So, you have your known curation information, but then you also have the novel components that are going to be specific to whatever experiment you're running. And so that's kind of the idea behind that one. That's really an upstream pipeline that we are developing.

The downstream part of that is now that we have the datasets, and we've got curated annotation, and they're consistent, how do we do the multi-omics integration? And so that'll be your Machine Learning type of algorithms from the multi-omics layers. We're playing around with some Machine Learning algorithms to try to find the features that are important at each molecular data type. But most of those are going to be based on correlation-type workflows, such as mix-omics. I've used that one a lot. So that's kind of my overall lab focus right now.

Miao-Ping Chien: This is indeed an important, and yet hard-to-address issue. However, people are aware of this and are paying more attention to newly generated data. This is particularly important when generating big cohort data; things like standardizing data type, format, and nomenclatures of metadata fields are points of attention.

David Ruau: In a recent paper, Lipkova et al. provide a summary of an emerging AI approach to solve heterogeneous data integration. Multimodal AI models are able to learn patterns within and across data modalities.[5] *To learn more about AI and ML in multi-omics data integration, read on to Chapter 5*

Jianguo (Jeff) Xia: Yes, this is a great question, especially for multi-omics datasets. Part of my research is on how to analyse different omics data, how to make data accessible, and working on developing some tools to address the issue. **But as we all know - garbage in, garbage out.** There is an urgent need for standardized multi-omics data repositories for tool development and benchmarking studies.

FLG: When you are performing experiments and collecting data, how do you assess the quality of the data? How do you know when you should go back and repeat an experiment?

Stephanie Byrum: Yeah, pre-processing, QC is very, very critical upfront. We do a lot of tests, like with the proteomics side of it, I have a tool called proteiNorm.[6] And what that does is it takes the input raw data and evaluates eight different normalisation methods that are common methods utilised for proteomics data. **When evaluating variability and correlation among your replicates in your samples – you want your replicates to have low variability and high correlation.** So trying to apply some actual statistical evaluation to the data upfront, because a lot of times you can detect a protein or a sample that didn't sequence well. Overall, it's going to have low-intensity values and you may have to throw it out. It's the same with RNA, DNA, any of those. There are different tools for QC, we use a lot to just evaluate quality. You need to have the appropriate read depth. That's the other point when you're doing multi-omics integration. Each technology has its different limitations, and it's different QC pre-processing steps that are specific to that platform. So, when I'm doing my integration, I'm actually starting from the pre-processed normalised data and not necessarily the raw data. You have different things to account for in different platforms.

> "WE'RE KIND OF STILL STUCK IN THE OLD POWER ANALYSES A BIT. I THINK WE ARE FINALLY GETTING THERE WITH THE PUSH INTO BIG DATA - I CAN SEE A TURNING POINT THERE."

Lihua (Julie) Zhu: The quality of input data is really important. We developed ATACseqQC for quality assessment of ATAC-seq data, which is for genome-wide profiling of chromatin accessibility. It allows users to quickly assess whether their ATAC-seq experiment was successful. This allows researchers to make a quick decision about the quality of their ATAC-seq data, save time and improve a study by allowing researchers to make an evidence-based decision on whether or not to redo an experiment with an improved protocol, or simply sequence more reads from the same library.

FLG: While making tools easier to use and more accessible can be a good thing, domain expertise is also important – especially when choosing the right approach or strategy. If you don't really understand the intricacies of the data integration approach, that spells problems because you may not be doing data analysis properly, and this is particularly problematic when trying to ensure results are reproducible. What are your thoughts on this?

John Quackenbush: I think it's a two-way street. The best relationships are partnerships where the partners have some understanding of each other. Having somebody in a wet lab try to pick up computational tools and analyse data is often fraught with challenges. When I was at Dana-Farber I used to run the Centre for Cancer Computational Bbiology and we offered classes in RNA-seq analysis for investigators. What we discovered is you can teach them how to use the tools, which is great. But once they sit down with real datasets, they run into all these different problems because there's so much cleaning and so much dataset-dependent work that you often have to do. So we evolved to try and provide a high-level view of what you could do with the tools and then tell people that if you really want to use them, come and talk to us and we'll help you work through it. I don't think there's necessarily a paternalistic view, but it reflects the importance of that partnership, where the biologically grounded members of a team rely on the computational experts.

But you also see the flip side. People with biological datasets sometimes throw them over the fence to somebody computational or a biostatistician who then analyses the data and they kind of toss it back. **The problem with that is it's easy to find things that are statistically significant but not biologically significant or meaningful, especially if you don't understand the biology of the system, or the limitations of the datasets that have been collected. And you need to have that two-way dialogue, that two-way partnership to make the relationship effective.**

Stephanie Byrum: **So, there's a give and take between making things easy, but then also not understanding what you're doing.** And I think that's one of the things the NIH is trying to initiate with the STRIDES programme. They've got some Jupyter notebooks that they're going to put into these training modules. And they're trying to develop different training modules based on expertise, but it's difficult. So, I created one for proteomics, but with limited, very simple input. Here's some data, this is what we're looking at QC, but kept it very broad, very simple, because there's a lot to learn. There's a lot of intricacies about how you interpret the data that comes off a mass spectrometer because it matters if you're running it with data-dependent acquisition data, independent acquisition, how you're setting up the instrument parameters, the interpretation... **I think sometimes you don't know what you don't know.**

David Ruau: Especially where humans are concerned, any research where their health, disease, or treatments are involved does require deep domain expertise on top of technical expertise, data collection, and extensive data science work on multiple modalities/fronts. **Healthcare and life sciences is one of the few fields where data science has a direct impact when we don't know the answer a priori.** This is different to image recognition, for example, where we compare machines to humans, or astronomy which may not directly impact human health. **Therefore, we do need heterogeneous models, experts and experiments to create checks and balances.** Reproducibility in life sciences is a tough problem to solve, because we work under pressure and on limited data for any one particular task. It's only when we do multi-modal work that we create systems to check the correctness of systems, even if they are reproducible.

Mathew Chamberlain: Computationally, it makes a lot of sense to take a look at the field and align with some of the major software packages that are out there. **There's this wonderful open-source community with thousands of developers creating and contributing to computational pipelines.** If you just adopt the pipelines that people are writing right now, it's like hiring hundreds of developers for free. **It's such a beautiful example of open-source science.** I wouldn't get intimidated. There's a strong and wonderful community of scientists out there that try to make this seemingly complex technology and data just very accessible. Dive in.

Jianguo (Jeff) Xia: Multi-omics data analysis is still fast evolving, and many new methods are proposed every day. Many methods are based on multivariate statistics which is challenging for most researchers. **My group has made significant efforts to help reduce this barrier by integrating statistics with powerful visualization – we call it "visual analytics".** For instance, our recent tool OmicsAnalyst, allows researchers to explore various multi-omics dimensionality reduction methods within interactive 2D / 3D scatters plots with various advanced support for highlighting and in-situ functional analysis.[7]

However, I do think there is some danger in lowering the barriers. **As more people enter the multi-omics data analysis field without proper training, they can often be overconfident in their results, without paying close attention to either statistics or biology.** I see a major effort in future is cross-disciplinary training and education to make sure that these methods and tools are used properly, and results are meaningful and reproducible.

Lihua (Julie) Zhu: I think reproducibility is a very important issue. **Part of the problem with big data is, if you interrogate it enough you can always get *something* out of it.** However, is that something biologically relevant? Is it a real signal or just an artefact? This is why having enough biological replicates and applying the proper tools with the right parameter settings are so important, and someone who doesn't have good domain knowledge, data science knowledge, that's something they may overlook. You just have to make sure you are properly trained. It's just like a car – you need pass a test and get a licence to know how to drive it properly!

"AS MORE PEOPLE ENTER THE MULTI-OMICS DATA ANALYSIS FIELD WITHOUT PROPER TRAINING, THEY CAN OFTEN BE OVERCONFIDENT IN THEIR RESULTS, WITHOUT PAYING CLOSE ATTENTION TO EITHER STATISTICS OR BIOLOGY."

FLG: Do you think that making things open-source and being transparent about data integration/analysis strategies is important, particularly for reproducibility?

Jianguo (Jeff) Xia: I am an open science advocate. **We should always try our best when communicating results, but we also need to communicate exactly *what* we have done.** It definitely will take more time to communicate that but if you follow a protocol and you run your own machine, you will probably get a different result because of multiple factors in addition to the protocol. With regards to data reproducibility, if you have the same code and the same input, you should get the same result. The practice should be incorporated for all bioinformatics tool development.

Miao-Ping Chien: Yes, absolutely. This is a trend and it is now also a common request (by journals) to deposit the used data and analyses to publicly accessible sites when publishing papers.

David Ruau: Absolutely – a lot of the solution stack can be made openly accessible. However, there are rightful considerations about patient privacy, or intellectual property for custom solutions. There are pathways that exist to make sure as much data as possible is accessible and reproducible.

Nikolai Slavov: I think this is an example of a win-win strategy. I think that open research is very beneficial for the community, but also very beneficial for groups who practice it. One way that benefited us very significantly in the early days in establishing credibility in this emerging field, where we, the newcomers, proposed that we can do something that the established leaders in the field couldn't do, or claimed wasn't possible to do. Of course, that resulted in a lot of scepticism. Part of what helps overcome the scepticism is that colleagues from other laboratories downloaded our data, or repeated our analysis, and obtained results that were qualitatively identical, for all practical purposes, to what we had done. While reproducibility is not the same as accuracy, this ability to reproduce our results landed a very large degree of credibility to what we had done, and was very, very healthy for the field.

Another example of this benefiting us is, I think it sets the bar high for all the students and postdocs in the group. If they make the work easy for others to reproduce, it also becomes very easy for them to introduce new data to their pipelines and easily revise their papers in the process of peer review. Which is, unfortunately, not the standard in the community. It may take more time to begin with to establish your producible pipeline of data analysis. But in the long term, it actually saves you time because, for impactful papers, one has to revise the figures multiple times. If everything is set up in a way that allows these revisions to be done simply and quickly in the long term, it enables you to benefit from constructive feedback from reviewers, and also allows us to incorporate new data for others to build upon it. Ultimately, that's why we do science in my view.

John Quackenbush: One of the things that I believe very strongly in, and my entire research group does as well, is the **fundamental importance of reproducible research. The whole scientific process is predicated on the idea that any theory, any model, any explanation that we have for a process we observe in nature and in biological systems should be testable and falsifiable.** So, if I tell you we have a particular model, and I'm getting an answer by applying that model in software to a dataset, the lowest barrier is ensuring that you or anyone else listening to this can take my model and my data and run it and get the same answer. When I say that, you might think that's absurd, everybody should be able to do it. **But the sad thing is that I can point you to countless examples in the past 20 years where people have taken datasets and taken published methods and run them and not gotten the same answer.**

So, that's the lowest bar, but then you want to be able to reproduce an analysis and understand what the method is doing. That all requires access to the software code, the underlying source code, so you can actually read it and understand what the method doing and whether if there's an error in the method. I will not say the code that we have written over time has always been error-free; sometimes we go back and discover errors, which is good. Sometimes people discover errors and tell us, which is good, and mostly, they're minor. But having access to code is fundamental to your research being part of the scientific process.

Our group has been committed to open-source software development for a pretty long period of time, and one of the things we do is to write all our code in a variety of different languages. We tend to write all of our methods in R, which is a statistical programming language. We also write in Python, which is widely used, especially in machine learning applications, MATLAB, which is a proprietary software system, but you can publish the code, and then C, which is just a generic programming language.

Each one of them has its own advantages, but one of the things that we try to do in building these methods is to ensure that people have access to software and code so that they can reproduce our analysis and so they can also apply our methods to their own data.

FLG: Multi-omics data often just shows correlation, not causation. How do we go from just making suggestions, to actually proving a functional link?

Stephanie Byrum: I have a colleague that works on causal inference – so we would really like to integrate in not only the multi-omics data, but also that causal inference model. **So it's not just correlation, it's causation. And then hopefully, with the causation, we can actually get to - these are the things that we need to go back to the wet lab and validate and really hone in on the molecules that are causing the disease.**

Miao-Ping Chien: I guess this can be partially answered by a proper design of the experiments. If possible, collecting samples and data from multiple time points (with/without treatment) and computing trajectory analysis might be a way to address this demand.

David Ruau: The causation link is established over time, and as evidence accumulates. This is a method that has been in place for a long time, where a researcher would prove that a molecule is having an effect using clinical trials. The next frontier in establishing evidence will be done through laboratory test automation, coupled with AI-assisted experimental suggestions.

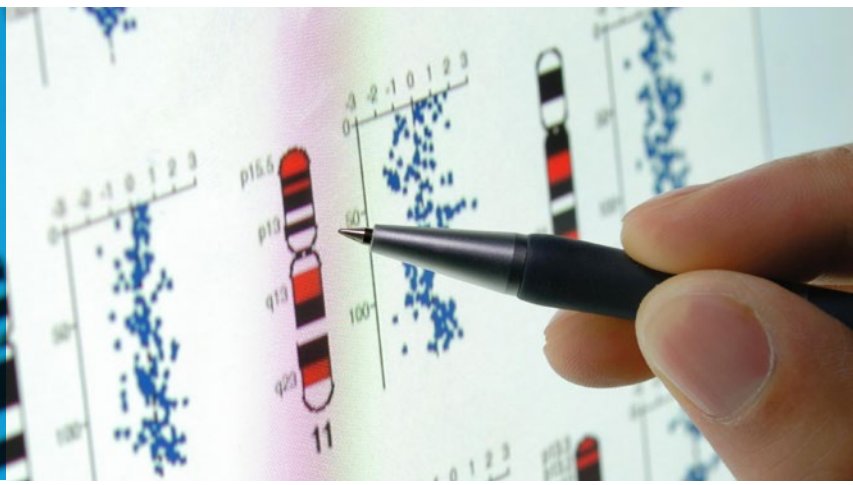
John Quackenbush: The best way to try to combine data in a meaningful way is to apply the lessons of an idea from computer science called the "no free lunch" theorem and to build models that rely on our biological knowledge about the processes we're studying. For our work, we try to learn gene regulatory models from data—which means we model the links between genes and the things that regulate them using multi-omic data. One of my pet peeves is hearing people talk about gene regulatory network models when all they're doing is looking at correlation. I get really frustrated when people tell me they're building multi-omics models and all they're doing is looking at each individual data types by themselves, finding significant things in each data type, and doing a big Venn diagram overlap to guess what might be going on.

I won't ever tell you that you can't learn something from doing that. **But in the end, you always learn so much more if you take these different data types and try to model their interactions based on what we know about how the corresponding cellular elements work together and interact with each other.** So I get very excited about all these multi-omics datasets. We try to ensure that we have each data type for each individual. We try to be very careful about making sure the data are good quality, and make sure that we adjust for things like batches, and then as we bring them together into models we really try to respect the biological associations between the various factors we measure. What we've found is that by doing that, we gain much deeper insight than by using methods naïve to how elements in the cell interact.

Marshall Summar: If I sequenced both of us, I would find about 10 million variations between us. Let's say one of us has a disease and the other does not; which of those 10 million variations caused the disease? There's a bit of an interpretive dance involved. If it's something we've seen before, that's a lot easier. You know, the delta f508 mutation in the CFTR gene is known to cause cystic fibrosis. That one's pretty easy. But what if you find a change in that gene that no one has seen before?

We're going to hear the term variants of unknown significance thrown around a lot. If I sequence anyone, I'm going to find some serious changes in some genes. But are they relevant to the clinical picture in front of us? You still have to analyse the data and see if there's a link between the changes you find. I guess one way of putting it is, we've got a giant monster-size puzzle with lots of pieces, and we're starting to fill in some of the edge pieces. We've got to fill in a whole lot more before you can really see the impact of some of the cheaper sequencing. It's the sequence analysis because we still require a knowledgeable human intermediary to say, "Okay, this makes sense that this variation might link to this." But the most common answer in human sequence analysis right now is "maybe". What you get is a "maybe this variation may have caused this, there's a high probability" It's one of the reasons we like to sequence families as much as possible, as opposed to an individual. Could this come from the parents? Are there any siblings with the same change who don't have anything or may have the same thing? So, we're still kind of feeling our way through. I think that will take a few years because it's a complex problem. But every day, we fill in a little bit more of the puzzle.

"THE WHOLE SCIENTIFIC PROCESS IS PREDICATED ON THE IDEA THAT ANY THEORY, ANY MODEL, ANY EXPLANATION THAT WE HAVE FOR A PROCESS WE OBSERVE IN NATURE AND BIOLOGICAL SYSTEMS SHOULD BE TESTABLE AND FALSIFIABLE."



mission bio

Ask the questions
you could never ask before

What if you could pinpoint which clones are destined for relapse?

The Tapestri® single-cell multi-omics (scMRD) platform integrates mutational and immunophenotypic signatures of the same cell, enabling measurable residual disease detection with unparalleled resolution, sensitivity, and specificity.

Mission Bio is now offering a Tapestri scMRD early access program for acute myeloid leukemia.

Learn more at go.missionbio.com/scMRD-AML-EA

MISSIONBIO.COM

DEVELOPING TOOLS FOR DATA INTEGRATION: SPOTLIGHT ON LIHUA (JULIE) ZHU

WE SPOKE TO **JULIE ZHU**, WHO, ALONGSIDE HER TEAM, HAS DEVELOPED SEVERAL TOOLS AND PACKAGES FOR INTEGRATION OF MULTI-OMICS DATA. BELOW WE DESCRIBE SOME OF THE TOOLS HER TEAM HAS DEVELOPED AND HIGHLIGHT HOW THESE TOOLS HELP MAKE DATA INTEGRATION MORE ACCESSIBLE AND EASY-TO-USE FOR RESEARCHERS IN THIS SPACE. IF YOU'RE NOT A COMPUTATIONAL EXPERT, THERE'S NO NEED TO FEAR – JULIE AND HER TEAM DESIGN THEIR TOOLS WITH YOU IN MIND.

Lihua (Julie) Zhu Professor, Department of Molecular, Cell and Cancer Biology, University of Massachusetts Chan Medical School: All the tools we have developed are to meet the specific needs of our community, its all needs-based and organically grown. The aim is to make data integration, analysis, and visualisation run more efficiently and reproducibly.

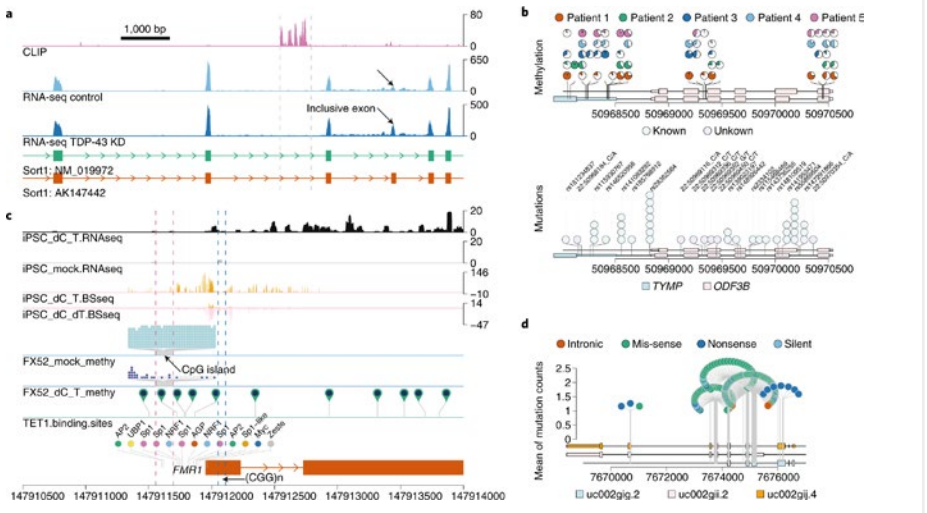
Several of these tools and packages are available on Bioconductor, a project that aims to "develop, support, and disseminate free open-source software that facilitates rigorous and reproducible analysis of data from current and emerging biological assays." [9] Bioconductor uses the R statistical programming language and host a dedicated community of developers and researchers committed to collaboration, transparency, and accessibility. Many of the other packages and tools that are included in this report are also available through Bioconductor.

One of the tools developed includes trackViewer, (published in Nature Methods) which is for the integration and visualisation of multi-omics data. Several genome browsers have been developed, but the majority of these tools do not have an easy programming interface that can be plugged into a pipeline. As well as this, unlike other genome browsers, trackViewer can perform specialised plots such as lollipop plots (also known as needle plots) for methylation, mutation, and SNP data. In the image below you can see how trackViewer allows for the integrative visualisation of multi-omics data.

Lihua (Julie) Zhu Professor, Department of Molecular, Cell and Cancer Biology, University of Massachusetts Chan Medical School: Looking at the image below (A), by simultaneously visualizing CLIP-seq and RNA-seq data together with the gene model, you can clearly see how a binding site may regulate the inclusion of an exon. You can then go ahead and knock that out to see what's going to happen – so the integrative visualization of multi-omics data can better inform your hypothesis going forward.

FIGURE 1: EXAMPLE OF HOW TRACKVIEWER CAN BE USED FOR INTEGRATIVE VISUALISATION OF MULTI-OMICS DATA.

a) Visualisation of RNA-seq and CLIP-seq data along with gene models of Sort1, with arrows and dashed lines representing regions of interest. b) Lollipop plots of methylation data with mutation data for genes TYMP and ODF3B from multiple individuals. The white part of the circle represents the methylated percentage, the coloured the unmethylated percentage. In the mutation plot the different colours depict different mutation events, and the number of circles indicates the number of events. c) Visualisation of coverage tracks from multiple datasets, together with lollipop plots of methylation data and binding sites of TET1 and several other transcription factors, and the gene model of the FMR1 promoter. d) Visualisation of dense mutation data for TP53 in a dandelion plot. [10]



In (B) you can see the different patients by colour-coding the patients, with the circle representing the percentage methylation alongside the mutation data. Simultaneous visualization of methylation data and mutation data in such a concise way helps to uncover the correlation between methylation status and mutation status.

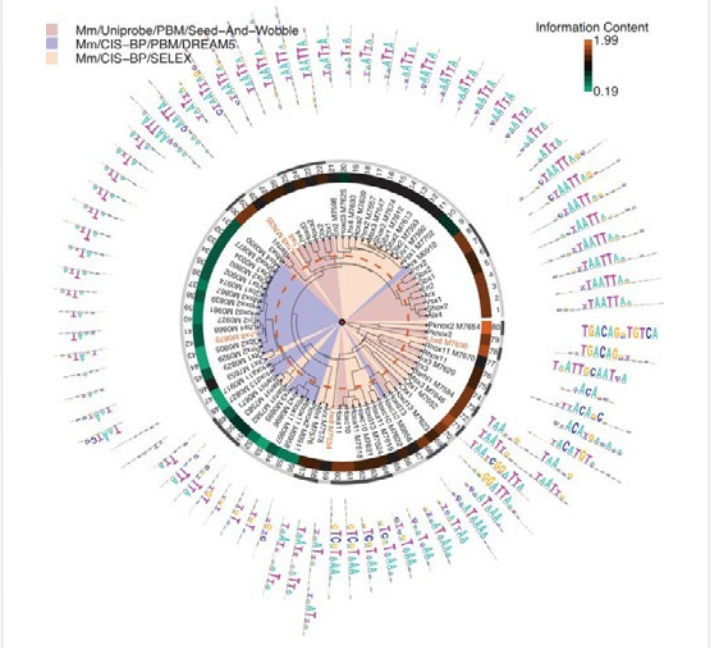
In (C) there are even more different data types integrated together: We can see transcription factor binding sites, promoter regions, repeats, RNA-seq data, methylation data, and more – so you can see a lot of data presented in one graph. Looking at this graph you can see how abnormal methylation may have contributed to changes in transcription factor binding, which causes changes in gene expression, and so on – in this way you can inform your hypothesis by looking at the data visualised in a lot of different ways.

As you can see it's a very nice visualisation tool, and easy to use for biologists with the web interface and the 'browseTracks' function. The pictures you produce are perfect for publication. With the 'browseTracks' function, users can generate interactive figures—that is, figures one can easily customize the features of by clicking, dragging, and typing, and it's ideal for people who don't want to do a lot of programming as well, so it's very accessible.

motifStack (published in *Nature methods*) can be considered as a tool for downstream analysis of all multi-omics data. [11] For example, let's say you have DNA-seq data, RNA-seq data, ATAC-seq data, ChIP-seq data, etc, and you want to identify some motifs. To do so, you can perform a motif enrichment analysis, and once some motifs are identified, the next step is to see how the motifs are all connected. How they compare between different platforms, different motif identification algorithms, different experiments, different databases, how to identify outliers, closely related transcription factors, etc. motifStack helps visualize that data so you can see what's going on and make those connections, as you can see in Figure 2 .

FIGURE 2: MOTIFS FOR A SET OF MOUSE HD TFS PRESENT IN THREE DIFFERENT DATASETS ARE DEPICTED AS A RADIAL PHYLOGENIC TREE USING MOTIFSTACK.

Tree branches are coloured to highlight the source of each motif. The inner ring is coloured to indicate the information content (IC) of the motifs. The alternating light and dark grey colours in the second ring delineate different motif clusters. [11]



Lihua (Julie) Zhu Professor, Department of Molecular, Cell and Cancer Biology, University of Massachusetts Chan Medical School: Figure 2 is one way of visualising multi-omics data using motifStack, by clustering and aligning data based on sequence motifs, which can be transcription factor binding sites, or mRNA splicing signals, or a functional region of a protein domain. So, it's one graph, but it's very information rich, as you can see how different data relates to each other and different motifs. Without aligning the motifs and visualizing different motifs in such a concise way as in motifStack, it would be very hard to visualise the connection among so many motifs.

ChIPpeakAnno is for integrated analysis of ChIP-seq and any experimental data resulted in genomic ranges. ChIPpeakAnno was the first batch annotation tool for ChIP-seq data and is one of the top downloaded bioconductor packages and is highly cited and been used extensively – despite having been released about 12 years ago, it has stood the test of time. With ChIPpeakAnno, you can annotate peaks to genes and enhancers, perform pathway enrichment analysis, overlap analysis (e.g., replicates, of different transcription factors, or different omic profiles), and more.

Lihua (Julie) Zhu Professor, Department of Molecular, Cell and Cancer Biology, University of Massachusetts Chan Medical School: ChIPpeakAnno was developed in such a way that if you follow the instructions, there's just one line of code for each type of analysis, and ChIPpeakAnno then processes everything inside of a workflow, on the back-end. So, you can just enter your data, set the parameters, and you're done. You can do batch annotation, peak boundary analysis, overlap analysis, GO and pathway enrichment analysis, all of this in ChIPpeakAnno.

And you can visualise the data easily as well. You can visualise enriched GO for different biological processes, cellular locations, or gene functions. You can visualise the overlap of two different datasets or maybe for different studies of the same transcription factors. Or maybe they are for different transcription factors, maybe they're co-factors you want to analyse or two different replicates. You can also easily see how the signal density around TSS (Transcription Start Sites) looks different for different transcription factors. For example, maybe one transcription factor didn't work out or maybe it is just not related to promoter binding. You can look at a simple plot for different transcription factors, or you can use colour-coding to order the data by rank. So, this is ChIPpeakAnno, which has proved to have a lasting impact.

Julie and her team have also developed some **web applications.** Their most recent is OneStopRNAseq which is for comprehensive and efficient analyses of RNA-seq data. [13] As we've already discussed in our discussion roundtable, many biologists struggle finding the right tool for the right question, as there are so many different tools being released all the time. This is one of the major reasons Julie and her team developed OneStopRNAseq, as it's simple to use with many different types of analysis all in one place. The back end is a workflow, and the front end is a web application, and you can click and choose different types/approaches for analysis and integration at each step.

Lihua (Julie) Zhu Professor, Department of Molecular, Cell and Cancer Biology, University of Massachusetts Chan Medical School: The reason why we developed OneStopRNAseq is because RNA-seq has become so prevalent. I think at almost every biological research laboratory, someone is doing RNA-seq at some point, so it's really becoming a staple technology. However, although there are already a lot of tools out there, researchers have to chain different tools together, and you also have to install different versions, updates, etc. To analyze datasets from public repositories such as GEO, you have to download the data and then verify for yourself that the data downloaded is intact. We developed OneStopRNAseq to simplify the process, thereby democratising the analysis of RNA-seq data including those from GEO.

With OneStopRNAseq, all you have to do is enter the GEO number or Dropbox links to sequence files, alignment files, gene-expression-count tables, or rank files with the corresponding metadata, select the genome and types of analyses you want to do. Once the analyses are complete, you will receive a link for downloading all the analyses that have been done for you, saving you a lot of time. It's gaining in popularity – we only published in 2020 but we have already been getting quite a few citations.

OneStopRNAseq supports read quality assessments (QA), read alignment, post-alignment RNA-seq-specific QA, count summarization, differential gene expression (DGE), and differential alternative splicing (DAS) analyses. It also supports differential transposable element expression (DTE) analysis, allele-specific gene expression (ASE) quantification, GO terms and KEGG pathway overrepresentation analysis, and MSigDB-based gene-set enrichment analysis (GSEA). In addition, it generates many standard plots, such as volcano plots to visualize enriched or depleted genes, and heatmaps.

It's great for reproducibility as well as robustness. If anything changes, such as a new version, or an update to the software, that might affect the analysis results. We don't want researchers to just put in their methods that they used OneStopRNAseq because that doesn't tell you which analyses are used within the workflow. We provide the method write-up, giving credit to all the packages with versions and parameter setting used in the workflow. That way the researcher can copy the method write-up and tweak it a little bit for their paper, and it also means other researchers can repeat the analyses easily

scATACpipe (single cell ATAC pipeline) is a workflow for analysing and visualising scATAC-seq data. [14] While there are other tools and pipelines for analysis of scATAC-seq data, scATACpipe is unique in that it is an end-to-end analysis of scATACseq data, and is easy-to-use, scalable, reproducible, and comprehensive. scATACpipe can perform extensive quality assessment, pre-processing, dimension reduction, clustering, peak calling, differential accessibility inference, integration with scRNA-seq data, transcription factor activity and footprinting analysis, co-accessibility inference, and cell trajectory inference.

Lihua (Julie) Zhu Professor, Department of Molecular, Cell and Cancer Biology, University of Massachusetts Chan Medical School: Development has been so rapid and so quick, especially for single cell technologies. Similar to OneStopRNAseq, our motivation behind developing this scATACpipe was to eliminate the need for users to chain different tools together to analyse and visualise their scATAC-seq data. Another problem is often, when an error occurs, you have to restart the whole analysis. With scATACpipe, if there is an error, you can just change specific parameters - it's not going to redo the entire analysis especially the time-consuming parts, such as mapping or filtering which may have already completed successfully. Therefore, you only need to focus on the high level analyses and parameter tunning. It significantly eliminates errors and resource waste and has an interface which helps users fine-tune their parameters easily.

References:

1. Subramanian, Indhupriya et al. **"Multi-omics Data Integration, Interpretation, and Its Application."** Bioinformatics and biology insights vol. 14 1177932219899051. 31 Jan. 2020, doi:10.1177/1177932219899051
2. Vahabi, Nasim, and George Michailidis. **"Unsupervised Multi-Omics Data Integration Methods: A Comprehensive Review."** Frontiers in genetics vol. 13 854752. 22 Mar. 2022, doi:10.3389/fgene.2022.854752
3. Lähnemann, David et al. **"Eleven grand challenges in single-cell data science."** Genome biology vol. 21,1 31. 7 Feb. 2020, doi:10.1186/s13059-020-1926-6
4. Angerer, Philipp & Simon, Lukas & Tritschler, Sophie & Wolf, F. & Fischer, David & Theis, Fabian. **"Single cells make big data: New challenges and opportunities in transcriptomics."** Current Opinion in Systems Biology. 2017 4. 10.1016/j.coisb.2017.07.004.
5. Lipkova, Jana et al. **"Artificial intelligence for multimodal data integration in oncology."** Cancer cell vol. 40,10 (2022): 1095-1110. doi:10.1016/j.ccell.2022.09.012
6. Graw, Stefan et al. **"proteiNorm - A User-Friendly Tool for Normalization and Analysis of TMT and Label-Free Protein Quantification."** ACS omega vol. 5,40 25625-25633. 30 Sep. 2020, doi:10.1021/acsomega.0c02564
7. Zhou, Guangyan et al. **"OmicsAnalyst: a comprehensive web-based platform for visual analytics of multi-omics data."** Nucleic acids research vol. 49,W1 2021: W476-W482. doi:10.1093/nar/gkab394
8. Gómez, David, and Alfonso Rojas. **"An Empirical Overview of the No Free Lunch Theorem and Its Effect on Real-World Machine learning Classification."** Neural computation vol. 28,1 2016: 216-28. doi:10.1162/NECO_a_00793
9. Reimers, Mark, and Vincent J Carey. **"Bioconductor: an open source framework for bioinformatics and computational biology."** Methods in enzymology vol. 411 2006: 119-34. doi:10.1016/S0076-6879(06)11008-3
10. Ou, Jianhong, and Lihua Julie Zhu. **"trackViewer: a Bioconductor package for interactive and integrative visualization of multi-omics data."** Nature methods vol. 16,6 2019: 453-454. doi:10.1038/s41592-019-0430-y
11. Ou, Jianhong et al. **"motifStack for the analysis of transcription factor binding site evolution."** Nature methods vol. 15,1 2018: 8-9. doi:10.1038/nmeth.4555
12. Zhu, Lihua J et al. **"ChIPpeakAnno: a Bioconductor package to annotate ChIP-seq and ChIP-chip data."** BMC bioinformatics vol. 11 237. 11 May. 2010, doi:10.1186/1471-2105-11-237
13. Li, Rui et al. **"OneStopRNAseq: A Web Application for Comprehensive and Efficient Analyses of RNA-Seq Data."** Genes vol. 11,10 1165. 2 Oct. 2020, doi:10.3390/genes11101165
14. Hu, Kai et al. **"scATACpipe: A nextflow pipeline for comprehensive and reproducible analyses of single cell ATAC-seq data."** Frontiers in cell and developmental biology vol. 10 981859 2022. doi:10.3389/fcell.2022.981859

Sponsored by

NVIDIA

CHAPTER 5

MACHINE LEARNING AND AI

MACHINE LEARNING (ML) AND AI APPROACHES ARE BECOMING INCREASINGLY POPULAR IN SCIENTIFIC RESEARCH. IN THE PREVIOUS CHAPTER, WE DISCUSSED THE CHALLENGES OF HANDLING BIG DATA, AND MANY RESEARCHERS IN OUR ROUNDTABLE DISCUSSION IDENTIFIED AI AND ML APPROACHES AS BEING POTENTIAL SOLUTIONS.

"THE BIG PROBLEM WITH ML AND AI IS THAT RESEARCHERS ARE OFTEN TRYING TO USE THESE METHODS WITHOUT UNDERSTANDING WHAT THEIR LIMITATIONS ARE."

However, ML or AI should not be considered a magic bullet – as with any technique, each has their own limitations and challenges. Moreover, a lot of these approaches are not even that novel – in fact the buzz-wordy nature of these terms means that they are often used for relatively basic and old models like Random Forest, which was developed back in 1995. That being said, there's a lot of innovation and development in the AI/ML space – and having some background knowledge may help you identify what is truly fresh and ground-breaking.

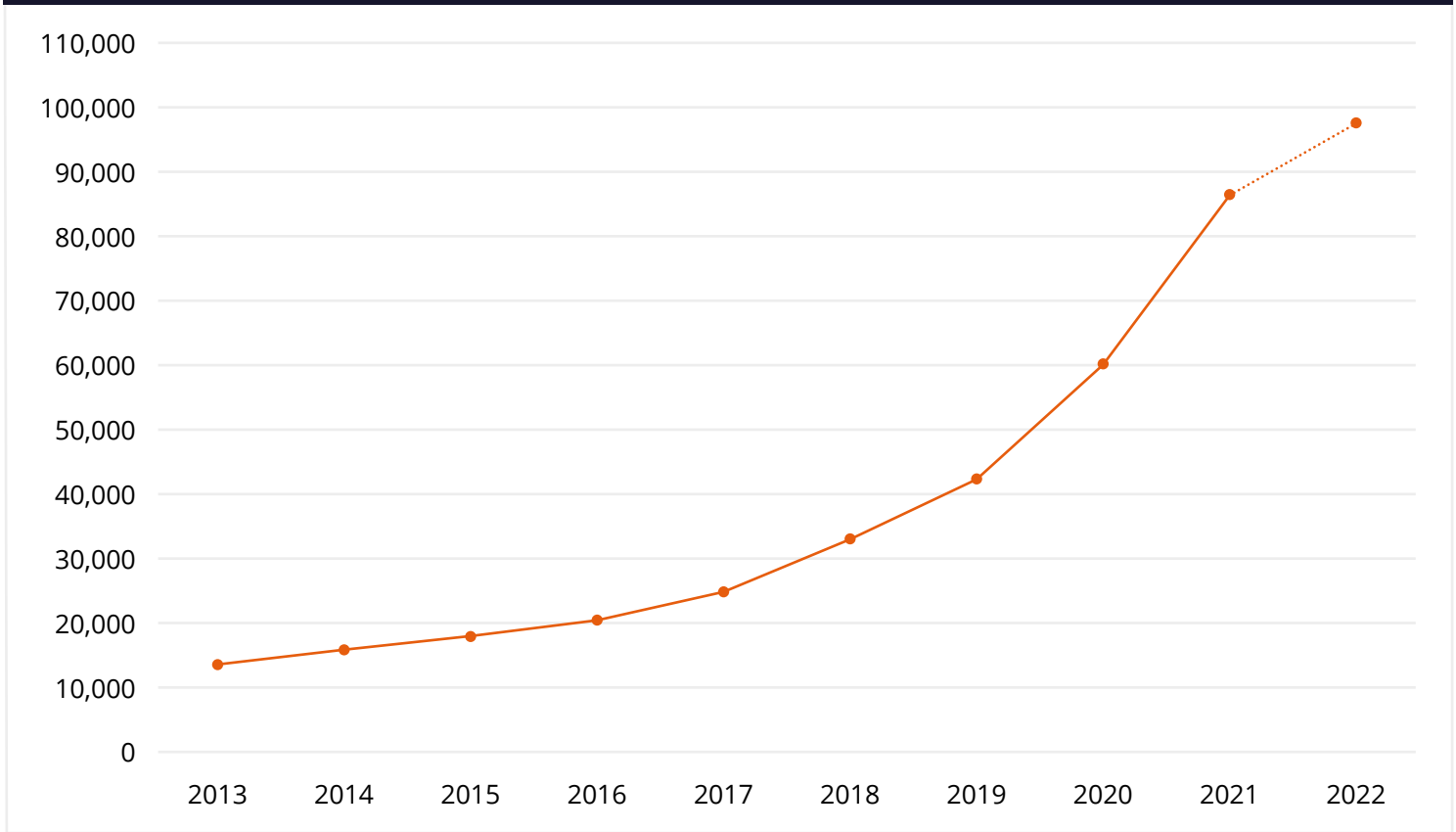
Jianguo (Jeff) Xia
Assistant Professor, Department of Bioinformatics and Big Data Analysis
McGill University:

It is very clear that ML and AI will be used much more for data integration in the near future. It's just unstoppable from my point of view, but we need to have a realistic view of what the data and technology can offer. Many people may want to use these new techniques but might not understand their limitations.

Stephanie Byrum
Associate Professor, Department of Biochemistry and Molecular Biology
University of Arkansas at Little Rock:

It's kind of funny because honestly, a lot of the ML we are doing is quite old, not really all that new. For example, PCA is an ML algorithm, but it's really a basic model. Random Forest, support vector machines, all of these algorithms have been around for a while. There's been an interesting shift in language, it seems to be more of a shift in terminology than it is a change in the methods. However, I think a novel application of ML is in some of the causal network inference models.

FIGURE 1: GRAPH SHOWING THE RAPID INCREASE PUBLICATIONS IN BIOLOGICAL SCIENCES USING THE TERM "MACHINE-LEARNING[1]



So what is ML? What is AI? And how can we use them for multi-omics studies? We asked John Quackenbush this question, and here is what he had to say.

John Quackenbush
Professor, Department of Computational Biology and Bioinformatics
Harvard T.H. Chan School of Public Health:

The big problem with ML and AI is that researchers are often trying to use these methods without understanding what their limitations are. First, you have to understand what people mean when they say ML or AI. It's sort of trendy to call almost everything ML or AI. I can tell you that our gene regulatory network models are based on an AI and ML approach, and some people would classify them that way. One of the things you often see called ML is Random Forest analysis, which we used in studies over 10 years ago. Back them, we never thought we would be calling it ML. But it is in the sense that you are taking a computational method, running it on a big dataset, you're coming up with weights, predictions, and building a model. So yeah, you are using machines to learn the parameters in the model. And the truth is that Random Forests, trained properly, can be very useful for making predictions on the right data.

If you want to try ML and AI approaches, you can't just grab a method and assume it will work. You need to careful to pick the right method and then use it with the right data to answer the right question. But that's also true of statistical models – I can run a t-test on any dataset, but the utility of that t-test is going to be determined by how big the sample size is, and the distribution of errors in the data. Most of the time, when we analyse data, we make the assumption that the errors are normally distributed, and that's swept under the rug. But if you have a dataset that doesn't align with those assumptions, even though you can apply a t-test, and **you can get an answer, that answer** isn't necessarily correct because you are using the wrong tool.

This is the same principle for ML and AI. What we are currently seeing with the application of ML and AI is the same kind of mistakes people made with statistical techniques. **We have to think carefully about if we are picking the right method for the right analysis, and whether or not we have the right data to use that method.** Furthermore, to really get good information out of these ML and AI approaches, there's got to be that connection and partnership between the biological domain experts and the computational domain experts so that we can do a better job of discovering something meaningful.

There are clearly some limitations for AI and ML. But before we cover those, let's take a closer look at how useful ML can be for integrating multi-omics data.

USING MACHINE LEARNING FOR MULTI-OMICS IN CANCER

IN A RECENT STUDY, PUBLISHED IN FRONTIERS IN GENETICS, A TEAM OF RESEARCHERS USED A MACHINE LEARNING APPROACH BASED ON MULTI-KERNEL LEARNING TO INTEGRATE DATA ACROSS MULTIPLE OMICS PROFILES AND SUBTYPE HEPATOCELLULAR CARCINOMA (HCC) INTO DISTINCT GROUPS[2].

"IF YOU WANT TO TRY ML AND AI APPROACHES, YOU NEED TO CAREFUL TO PICK THE RIGHT METHOD AND USE IT WITH THE RIGHT DATA TO ANSWER THE RIGHT QUESTION."

Multi-kernel learning is one example of a ML approach **to integrate diverse datasets such as specific omics data.** rMKL-LPP (regularised multiple learning with locality preserving projections) is an approach that reduces dimensionality and integrates data simultaneously. In this study, rMKL-LPP was used to integrate mRNA, miRNA and DNA methylation data from 287 HCC patients.

Researchers found that 2 subtypes had significantly higher mortality rates, with the high-risk group found to be 3.37 times more likely to die within the first 3 years than the low-risk group. Further investigation into the distinct subtypes allowed researchers to elucidate the underlying biological processes that lead to this significant difference in mortality rates. They highlighted 6 pathways that were significantly different between the 2 groups: Hypoxia, MAPK, EGFR, NF-kbeta, and TNFalpha pathways were found to be significantly more active in the high-risk group, whilst in the low-risk group VEGF pathway activity was found to be higher.

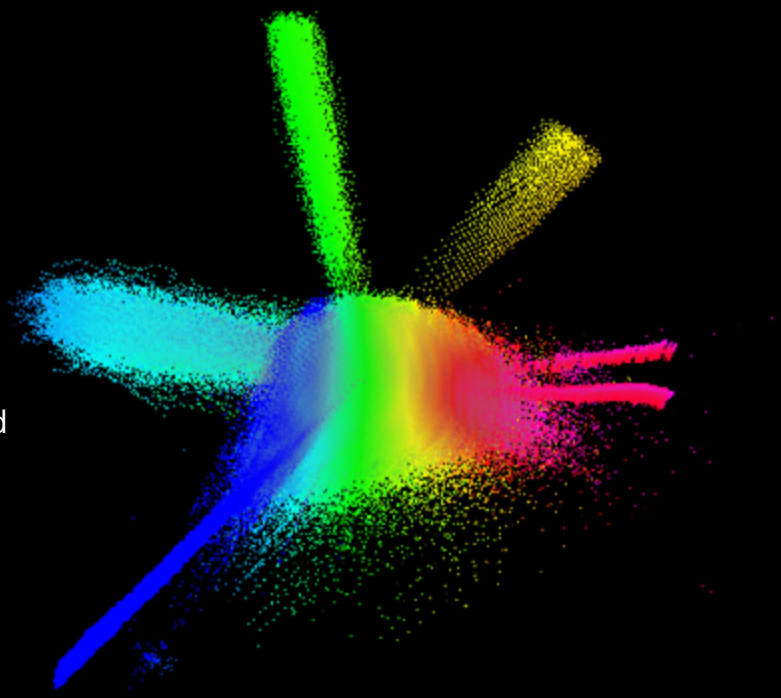
Researchers also found through immune cell infiltration analyses that 9 immune cell seemed to have different concentrations between the 2 subtypes. Myeloid cells, T-cells (particularly CD8+ T-cells) and dendritic cells were found in significantly higher concentrations in the high-risk group compared to the low-risk group. Furthermore, the cytotoxicity score of these immune cells was found to be significantly higher than in the low-risk group. Altogether, the high cytotoxicity of these immune cells, the higher concentration of these immune cells, as well as the increased activity of pathways associated with inflammation suggests the high-risk group exhibits an enhanced inflammatory response which could be causing the increased mortality rates. Targeting these specific pathways, as well as these tumour-infiltrating cells, could prove to be an effective precision medicine strategy for HCC.

Researchers also conducted weighted gene co-expression network analysis and identified various gene modules that may impact prognosis. These genes were found to be involved in biological processes that may enhance the development of liver cancer, helping elucidate the genetic causes of the mechanisms which underlie HCC progression. CDK1, CDCA8, TACC3 and NCAPG were significantly associated with poor HCC outcome and could be potential biomarkers used for prognosis.

The study authors said “the selected potential pathogenic genes, pathways and tumour-infiltrating immune cells can be used as references to control related gene expression or interfere with their target signal transduction pathways to provide potential opportunities for the treatment of HCC. Our findings may bring novel insights into the subtypes of HCC and promote the realization of precision medicine.”

This study is a great example of **using ML as a tool to account for differences in the data whilst integrating it together** – differences such as when the sample was collected and when the analysis was performed. It allowed the researchers to get the **disparate datasets, combine them together, and mine it for unique insights into differences in HCC between patients.** Moreover, data integration and analysis were done **in tandem with a biological understanding of the disease – and this allowed the researchers to really investigate the underlying pathways and mechanisms that cause the differences in phenotype** – namely the differences in mortality rate.[2]

NVIDIA Expands Large Language Models to Biology



Leading pharma companies, biotech startups and pioneering biology researchers are developing AI applications with the NVIDIA BioNeMo LLM service and framework to generate, predict and understand biomolecular data.

As scientists probe for new insights about DNA, proteins and other building blocks of life, the NVIDIA BioNeMo framework will accelerate their research.

NVIDIA BioNeMo is a framework for training and deploying large biomolecular language models at supercomputing scale — helping scientists better understand disease and find therapies for patients. The large language model (LLM) framework will support chemistry, protein, DNA and RNA data formats.

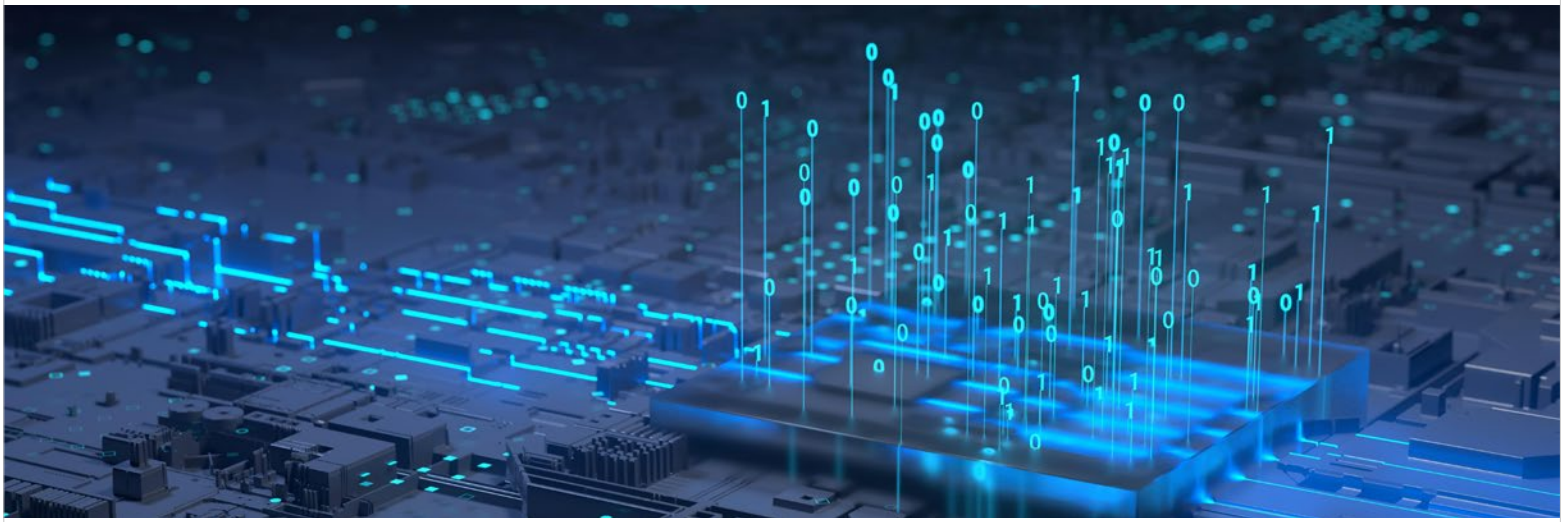
Just as AI is learning to understand human languages with LLMs, it's also learning the languages of biology and chemistry. By making it easier to train massive neural networks on biomolecular data, NVIDIA BioNeMo helps researchers discover new patterns and insights in biological sequences — insights that researchers can connect to biological properties or functions, and even human health conditions.

Learn how AI and accelerated computing are improving every stage of drug discovery with faster, more accurate insights. High-performance computing (HPC) applications, pretrained AI models, and domain-specific application frameworks are powering accelerated genomics applications, protein structure determination, virtual drug screening, medical imaging, natural language processing, and more.

nvidia.com/bionemo

Learn More

THINGS TO CONSIDER: CHALLENGES IN ML AND AI



As we've touched on already, like with every approach, AI and ML have their own limitations. **So, when it comes to selecting an AI approach, what are some of the key things to take into account?** Here we've included 5 major considerations - many thanks to **John Quackenbush** and **Julie Zhu** for helping us to identify these. Read on to learn more about data shift, underspecification, overfitting vs underfitting, data leakage, and black box models.

DATA SHIFT

Data shift occurs when there is a **mismatch between the data an AI or ML model was trained and tested on and the data it encounters in the "real world."** Essentially, training fails to produce a good ML model because the training and testing data does not match other datasets and is not generalisable.

As you can imagine, this can be a real issue when dealing with multi-omics data. Lack of standardisation between datasets, data collected under very specific experimental conditions, data collected at different times, by different people, under different environments – **all these factors can mean that our ML model may have data shift issues.** Making sure the data you trained your ML approach on fits other data you may later test it on, is vital to ensure that your results are reliable, accurate, robust and reproducible. Assessing the data you use to train and test your model, and taking into account the limitations of that data can help avoid data shift, as well as computational techniques such as anchor regression. [3]

UNDERSPECIFICATION

Underspecification is a problem that is also observed in statistics. **Essentially, even if a training process can produce a model that performs well on the test data, that model can still be flawed.** This is because with ML models, the training process can produce many different models that all work on your test data, but these models differ in small, seemingly unimportant ways. These differences can be attributed to many things, such as the way training data is selected or represented, the number of training runs, and so on. In neural network models, random values which are given to the nodes before training even starts can cause these differences.

These small, sometimes random differences appear arbitrary, especially because they don't affect how a model performs on the testing data. **But when applied to other datasets, it can end up causing unanticipated problems.** To illustrate why underspecification is cause for concern researchers used the same training processes to produce multiple ML models, which all seemingly performed well on the test data. They then ran these models through stress tests, which revealed distinct differences between the models. [4]

To avoid underspecification issues, one option is to introduce an additional stage to the training and testing process, where you produce many ML models instead of just one. These different models can then be tested again on another set of test data, made to compete against each other, and whichever performs best can be selected. [4]

Another thing to bear in mind is how strongly the training set influences the model that you build.

John Quackenbush
Professor, Department of Computational Biology and Bioinformatics
Harvard T.H. Chan School of Public Health:

How his team investigated this problem: "We started to turn the standard paradigm of training on one set and testing on a number of others on its head – we instead trained the model on one training set, saw how it performed on a test set. And then trained the model on another training set, and tested it again on the original test set. We did this to evaluate how stable our model was, independent of how we trained the model. That's a really subtly interesting question, because if the model is robust and stable, it should give me the same classifications independent of on which dataset it was trained. That's the model I would have much more confidence in. Because at the end of the day, what I want to know is that my method gives me the same answer independent of where the original training data came from."

OVERFITTING VS UNDERFITTING

Overfitting is when a statistical or ML model **fits too exactly against its training data – and as a result, when the model is tested against unseen data, it cannot perform accurately.** Basically, if the model is trained on the sample data for too long, or if the model is too complex, **it can start to memorize the noise or artefacts within the dataset.** Consequently, it is unable to generalize to new "real world" data. To combat overfitting, researchers can look at the training data and the test data – if the training data has a low error rate, and the test data has a high error rate, overfitting is likely an issue.

Underfitting is, predictably, the opposite of overfitting. One may reasonably assume that to avoid overfitting, you should spend less time training your model – this is known as "early stopping," – reducing the complexity of the model. **However, pausing too early may cause the model to miss or exclude important features, leading to underfitting.** This means the model, like with overfitting, is unable to generalize to new "real world" data.

The more a model learns, the more its bias reduces. However, if it trains for too long, the variance increases. **The ideal situation is to find a balance between bias and variance and hit a sweet spot where the model can perform well on new unseen data.** [5]

DATA LEAKAGE

Data leakage is a major problem in ML when developing **predictive models.** The goal of a predictive ML model is to make accurate predictions on new unseen data. **When information from the data a model is trained on includes data that it is later tested on, the model has effectively already seen the answers, and its predictions seem much better than they really are.** In other words, if the data from the training set "leaks" into the testing data, it causes results to be unreproducible.

Data leakage is more of a problem with **complex datasets** – and we can agree most multi-omics data fits that description. One subtle form of data leakage to look out is temporal leakage, which is when training data includes points from later in time than the test data. This is a problem because the model has essentially seen the future, and improper metadata annotation or the batch effect can be an issue here. Ways to prevent data leakage include preparing the data well within cross-validation folds and reserving a validation dataset for final checks of any developed models. [6,7]

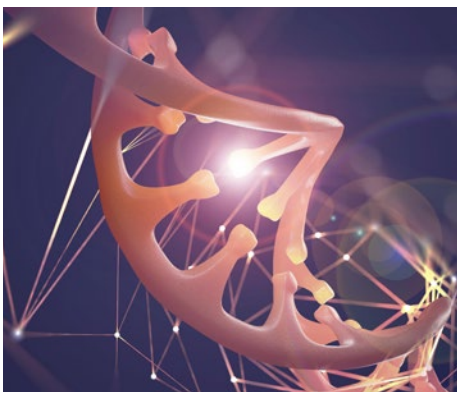
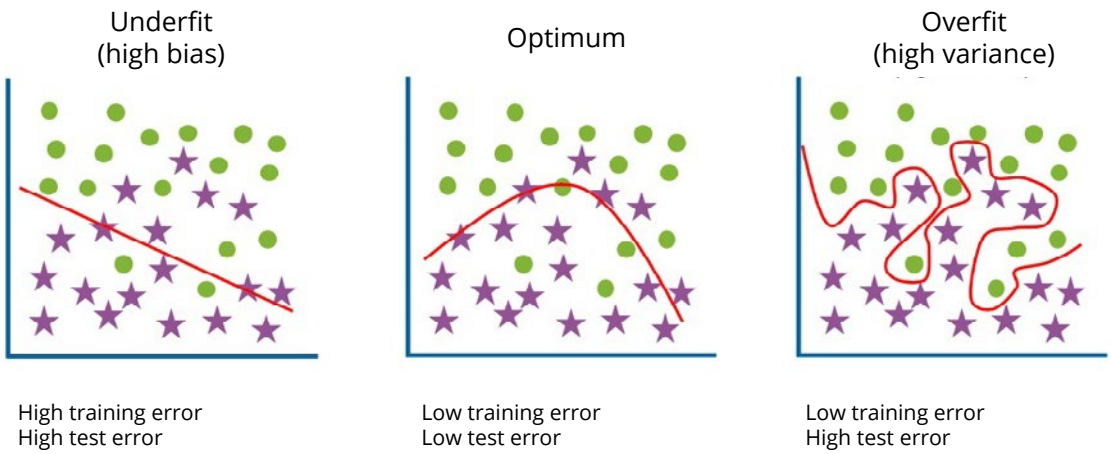


FIGURE 2: ILLUSTRATIONS TO SHOW THE EFFECTS AND TELL-TALE SIGNS OF UNDERFITTING, OVERFITTING, AND AN OPTIMUM ML MODEL. [4]



BLACK BOX MODELS

Some ML and AI models are referred to as "black box models," **where users and researchers know the inputs and the outputs, but do not know how the model actually works.** However, if we can't interpret the model, how can we falsify, test, and reproduce the results? Interpretable models, or explainable models, instead make clear how the model works. Often these models are also open-source, and all the code is made easily accessible and freely available.

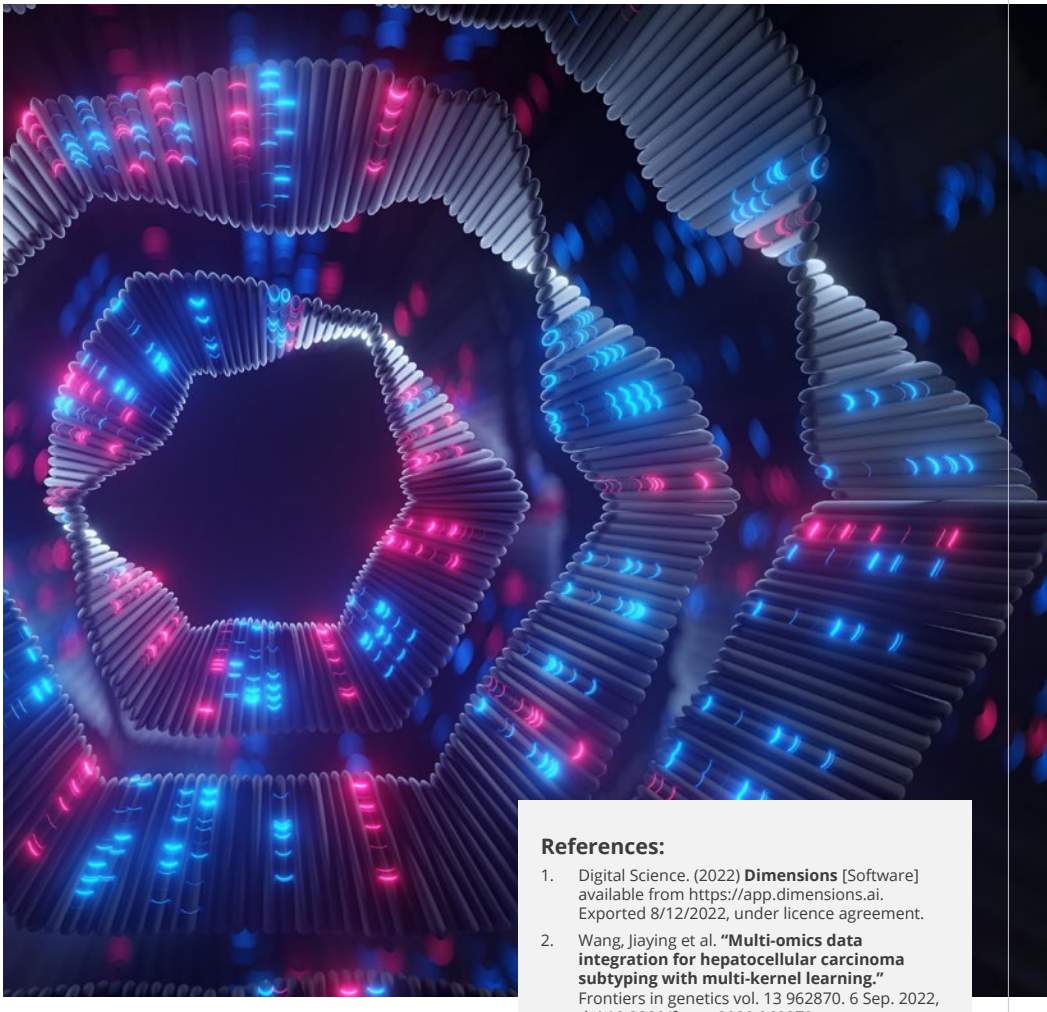
Moreover, black box models are created directly from data by an algorithm. In the context of multi-omics, and scientific research in general, **this can actually limit the utility of these models, as they do not incorporate domain knowledge.** In many scientific disciplines, such as systems biology, biological information is present in the form of graphs and networks – and this information can be incorporated in network-based algorithms to make them more versatile and applicable in many research areas. [8,9]

REPRODUCIBILITY

The reproducibility crisis has been a point of concern in the scientific community for some time. However, one would assume that as long as you make the data and code available, **it should be relatively easy to reproduce studies utilising ML and AI approaches.** Unfortunately, this doesn't seem to be the case, mainly due to the aforementioned problems.

John Quackenbush
Professor, Department of Computational Biology and Bioinformatics
Harvard T.H. Chan School of Public Health:

> What we've seen from all these papers on ML being published is that everybody is falling back into the trap of where we were 10, 15, and 20 years ago, producing analyses that aren't reproducible. Part of my crusade as a scientist is convincing people that we all have to be open to having our models falsified, and exposed for all the warts they may carry, especially if the methods are going to be applied in a clinical setting. That means we have to make the code available, we have to make the models available, and we have to make the underlying data available, so others can replicate our findings or show us that our findings are wrong so we can fix the mistakes.



Reproducibility and robustness of these ML and AI based approaches are crucial if we want to trust the findings from studies, **and if we want to cement the legitimacy and credibility of studies using ML and AI. Both can be incredibly useful tools, as long as we make sure we are using it in the right way.** *For more information about the reproducibility crisis, why don't you check out our recent blog-post: Reproducibility: The science communities' ticking timebomb. Can we still trust published research?*

References:

1. Digital Science. (2022) **Dimensions** [Software] available from https://app.dimensions.ai. Exported 8/12/2022, under licence agreement.
2. Wang, Jiaying et al. **"Multi-omics data integration for hepatocellular carcinoma subtyping with multi-kernel learning."** Frontiers in genetics vol. 13 962870. 6 Sep. 2022, doi:10.3389/fgene.2022.962870
3. Stewart, Matthew **"Understanding Dataset Shift"** Accessed 21/11/2022 https://towardsdatascience.com/understanding-dataset-shift-f2a5a262a766
4. Haven, Will Douglas **"The way we train AI is fundamentally flawed"** Accessed 21/22/2022 MIT Technology Review https://www.technologyreview.com/2020/11/18/1012234/training-Machine learning-broken-real-world-heath-nlp-computer-vision/
5. **"Overfitting"** Accessed 21/11/2022 IBM Cloud Education https://www.ibm.com/cloud/learn/overfitting#:~:text=Overfitting%20is%20a%20concept%20in,unseen%20data%2C%20defeating%20its%20purpose.
6. Brownlee, Jason **"Data Leakage in Machine learning"** Accessed 21/11/2022 https://machinelearningmastery.com/data-leakage-Machine learning/
7. Gibney, Elizabeth. **"Could Machine learning fuel a reproducibility crisis in science?."** Nature vol. 608,7922 2022: 250-251. doi:10.1038/d41586-022-02035-w
8. Pfeifer, Bastian et al. **"Multi-omics disease module detection with an explainable Greedy Decision Forest."** Scientific reports vol. 12,1 16857. 7 Oct. 2022, doi:10.1038/s41598-022-21417-8
9. Rudin, C., & Radin, J. **Why Are We Using Black Box Models in AI When We Don't Need To?** A Lesson From an Explainable AI Competition. 2019 Harvard Data Science Review, 1(2). doi:10.1162/99608f92.5a8a3a3d

CHAPTER 6

THE NEXT DIMENSION - TIME

WE HAVE ALREADY COVERED HOW IMPORTANT AND REVOLUTIONARY THE INTRODUCTION OF SPATIAL CONTEXT IS FOR MULTI-OMICS RESEARCH. **HOWEVER, TEMPORAL CONTEXT IS EQUALLY VITAL – BIOLOGICAL SYSTEMS ARE DYNAMIC, AND THINGS CAN CHANGE IN A MATTER OF SECONDS.** IN THE LITERATURE, AND ACCORDING TO OUR CONTRIBUTORS, ADDING THE NEXT DIMENSION – TIME – IS THE NEXT FRONTIER, NOT JUST IN THE MULTI-OMICS SPACE, BUT FOR SCIENTIFIC RESEARCH IN GENERAL.

In this chapter, we will look at some computational approaches that have allowed researchers to take data from snapshot experiments and interrogate the data in a way that elucidates the time course of events. We will also cover experimental advances which have allowed for real-time tracking and recording of biological events. A special thank you to Rong Fan for his advice and contributions to this chapter.

RNA VELOCITY

WHILE RNA-SEQ IS A POWERFUL TOOL THAT ALLOWS RESEARCHERS TO PROFILE THE GENE EXPRESSION OF A CELL (OR A TISSUE WHEN USED FOR BULK SEQUENCING), IT ONLY CAPTURES THE TRANSCRIPTOME AT **A STATIC SNAPSHOT IN TIME**. THIS CAN BE CHALLENGING WHEN INVESTIGATING PROCESSES THAT ARE DYNAMIC AND CONSTANTLY CHANGING – THOSE IN NORMAL HUMAN BIOLOGY SUCH AS EMBRYOGENESIS OR REGENERATION, OR IN THE CONTEXT OF DISEASE SUCH AS NEURODEGENERATION OR CANCER CELL EVOLUTION.

Cell differentiation can happen over hours to days – **a similar time scale to the typical half-life of mRNA.** The paper "RNA Velocity of Single Cells" (published in *Nature*) by La Manno et al. describes RNA velocity, which works by measuring the relative abundance of unspliced (nascent) mRNA and spliced (mature) RNA, thus estimating rates of gene splicing and degradation, and **therefore predicting the future state of an individual cell's transcriptome.** [2] Most single-cell sequencing protocols use oligo-dT primers to enrich for polyadenylated mRNA molecules, but when examining single-cell RNA-seq datasets, La Manno et al. found reads contained a substantial number of unspliced intronic sequences. Investigations (using metabolic labelling) showed that these molecules and their correlation with exonic counts may represent unspliced precursor mRNAs.

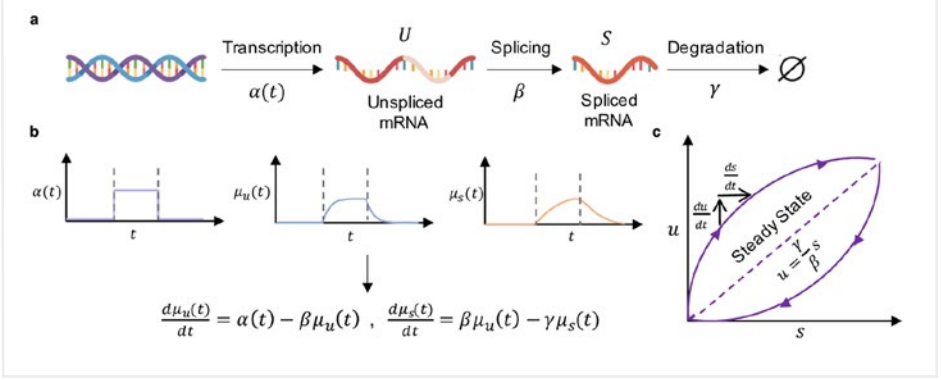
For RNA velocity, La Manno et al. assumed a simple model for transcriptional dynamics to quantify the time-dependent relationship between unspliced and spliced mRNA numbers.

This simple model assumes that the rate of spliced mRNA abundance is determined by the rate of transcription, the rate of splicing (production of spliced mRNA from unspliced mRNA) and mRNA degradation (see Figure 1).

During dynamic processes, the transcription rate escalates, causing a rapid increase in unspliced mRNA, followed by a subsequent increase in spliced mRNA, until a new steady state (equilibrium) is reached, and vice versa when transcription rates drop. When gene expression is induced, the number of unspliced mRNAs is in excess of the expected number, whilst during repression the opposite is true. Therefore, the balance of unspliced vs spliced mRNA is an indicator of the future abundance of spliced mRNA, and therefore of the future transcriptome of cell.

In this paper, La Manno et al. showed how RNA velocity worked in a number of ways. To demonstrate the ability of RNA velocity to predict transcriptional dynamics, the researchers analysed (SMART-seq25) data from mouse chromaffin cells. Chromaffin cells are the neuroendocrine cells of the adrenal medulla (in the brain), and during development a substantial percentage of chromaffin cells arise from Schwann cell precursors. The direction of differentiation can be validated easily by lineage tracking. RNA velocity estimates accurately recapitulated the transcriptional dynamics of individual cells, as well as the general movement of differentiating cells either towards the chromatin fate, or towards/away from the intermediate differentiation state. [2]

FIGURE 1: SHOWING MODEL USED FOR RNA VELOCITY ANALYSIS OF TRANSCRIPTION, SPLICING, AND DEGRADATION OVER TIME. ALPHA = TRANSCRIPTION, BETA = SPLICING, GAMMA = DEGRADATION, U = UNSPLICED, AND S = SPLICED. [1]



The computational biology field has harnessed the potential of RNA velocity and has adopted the method to **produce a number of new models based around the core concept of unspliced vs spliced mRNA levels. Some new methods incorporate the other omics,** including *protaccel* (which incorporates newly available protein data), as well as chromatin velocity and *MultiVelo* (which incorporate chromatin accessibility). Other new methods enhance RNA velocity – *scRegulosity* identifies local trends, *Velo-Predictor* incorporates Machine Learning, *dyngen* and *VeloSim* are used to simulate RNA velocity data, and *VeloViz* and evo-velocity can be used to construct velocity-inspired visualizations. [1]

MEFISTO

MULTI-OMICS FACTOR ANALYSIS (MOFA) IS A HIGHLY CITED AND FREQUENTLY USED MULTI-OMICS COMPUTATIONAL APPROACH THAT USES PRINCIPAL COMPONENT ANALYSIS (PCA) TO INTEGRATE DATA. NOW, THE TEAM BEHIND MOFA HAVE DEVELOPED A NEW APPROACH (CALLED MEFISTO) THAT IS **CAPABLE OF INTEGRATING TEMPORAL AND SPATIAL CONTEXT INTO THE ANALYSIS.**

MEFISTO is an "unsupervised approach to integrate multi-modal data with continuous structures among the samples, e.g., given by spatial or temporal relationships." [3] MEFISTO aims to reduce the dimensionality between samples taken at different time points and disentangle sources of variation caused by factors that change gradually, as well as other independent sources of variation. MEFISTO can also be used to interpolate/extrapolate to unseen time points or locations.

In a recent webinar Britta Velten Postdoctoral Fellow, DKFZ German Cancer Research Center, told us more about MEFISTO and the importance of preserving spatial and temporal context.

> **Britta Velten Postdoctoral Fellow, DKFZ German Cancer Research Center**: Existing methods for multi-omics data integration often don't account for other types of relationships between samples, such as spatial or temporal information. Temporal and spatial information can give us very important insights into molecular dynamics for example, or spatial factors that play a role in biological processes.
>
> One area in which this has become very popular is precision medicine, where longitudinal studies now follow up with patients over a period of time making various omics measurements at different time points – this can of course give us a lot of information about disease progression, disease onset and treatment outcomes that snapshot data cannot provide. Another area is in the field of developmental biology, where we are interested in the temporal axis to understand how transcriptional dynamics or translational processes are regulated along various developmental stages.
>
> With MEFISTO, we want to integrate all this data, including spatial and temporal data, perform dimensions reduction, and have an overview of the major sources of variation in the data, but also account for the temporal and spatial dependencies between samples that are naturally present in such datasets."

In other words, MEFISTO takes highly dimensional data that has measurements from multiple sources – different omics data, different sample groups, and different timepoints – and reduces them to a small number of factors in a time-aware manner, as you can see in Figure 2.

MEFISTO can identify temporal and spatial patterns and major sources of variation while accounting for heterogeneity across different groups of samples. This is particularly useful for studies with repeated spatial and temporal measurements, such as longitudinal studies involving many individuals, species or experimental conditions. MEFISTO can infer the extent to which spatio-temporal patterns are shared across groups. MEFISTO can also interpolate and extrapolate data – it can predict (infer) what will happen in the future, based on what it has seen [4].

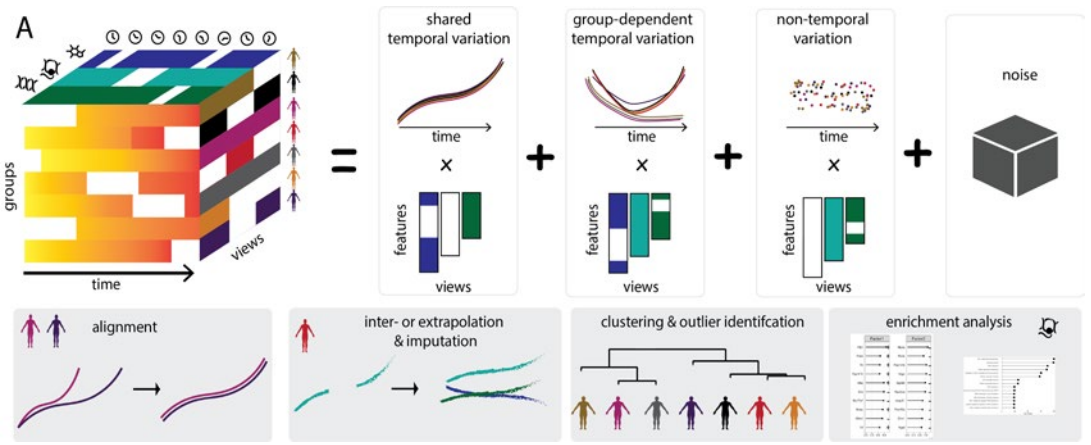


FIGURE 2: ILLUSTRATION OF MEFISTO FOR TIME-RESOLVED DATA – THE BOXES ON THE RIGHT ARE SCHEMATICS ILLUSTRATING THE TYPES OF TEMPORAL VARIATION SEEN ACROSS SAMPLE GROUPS, AND THE BOXES BELOW SHOW THE FEATURES OF MEFISTO.

RETRO-CASCORDER

AS YOU'VE ALREADY SEEN, COMPUTATIONAL METHODS USING INFERENTIAL STATISTICS AND MODELS CAN ESTIMATE OR PREDICT CHANGES IN TRANSCRIPTION OVER TIME. **HOWEVER, THESE ARE STILL ESTIMATES WHICH RELY ON CERTAIN ASSUMPTIONS.** MOLECULAR RECORDERS DO NOT RELY ON INFERENCE, ASSUMPTIONS, OR ESTIMATES – **INSTEAD, THEY CONTINUOUSLY RECORD GENE EXPRESSION CHANGES AND STORE THEM IN A PHYSICAL RECORD, USING DNA AS THEIR HARD DRIVE.** IN A RECENT PAPER PUBLISHED IN NATURE, BHATTARAI-KLINE ET AL., DEVELOPED SUCH A DEVICE, WHICH THEY CALLED "RETRO-CASCORDER". [3]

Retro-Cascorder makes sequential recordings of transcriptional events, logging receipts of gene expression using CRISPR-Cas integrases to incorporate retron barcodes (engineered RNA barcodes) into a cell's genome. **Once you sequence the genome, you uncover the history of gene expression in that cell.**

> **Shipman An author of this paper:** DNA is a flexible data storage medium in which you can really encode whatever you want. It's compact, it's flexible, it's got a nice code we can work with, it's stable. It's not something that you ever have to worry about falling apart, even over really long timescales. [5]

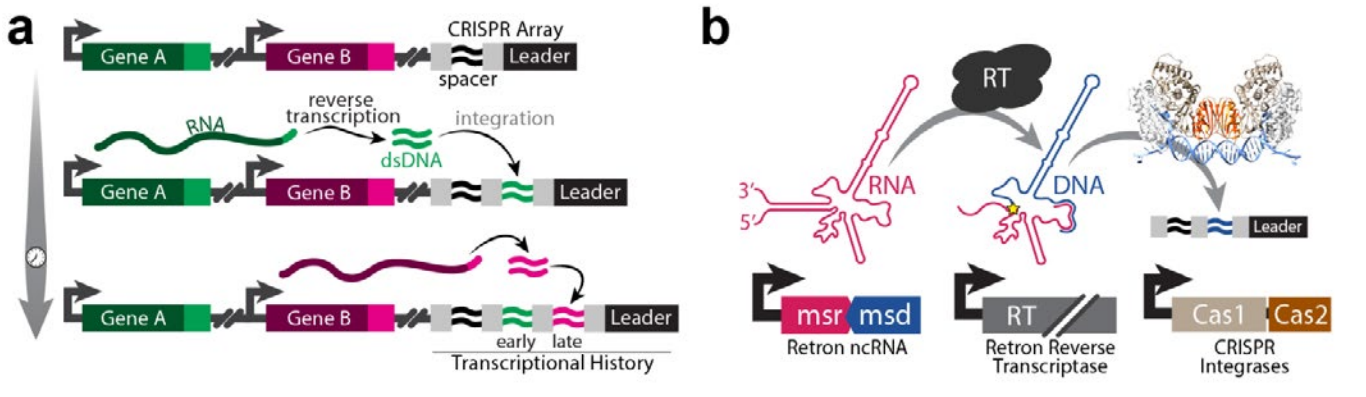
What makes Retro-Cascorder unique is that, unlike previous molecular recorders, **it can record more than one event at a time** – it achieves this by adding retrons to the gene of interest. To develop Retro-Cascorder, retrons were engineered to produce a specific tag sequence and then placed under the control of promoter sequences for specific genes of interest in E. coli. When the promoter is activated, the tag sequence is transcribed to RNA. This RNA sequence is then reverse-transcribed by the retron reverse-transcriptase contained within Retro-Cascorder, generating a DNA receipt. CRISPR integrases then integrate the DNA receipt into a CRISPR array, thus creating a record of transcription. CRISPR arrays contain spacer sequences and if another promoter is then activated, this process repeats, placing the new DNA receipt after the first spacer.

> **Santi Bhattarai-Kline An author of this paper:** That retron acts like a receipt that tells you the gene was just turned on [5]

To test Retro-Cascorder, the researchers tagged specific genes in E. coli, which are known to activate in the presence of specific chemicals. Over 48 hours, transcription events were recorded and the order of DNA receipts in the CRISPR array corresponded to the order in which the chemicals were applied. Thus, Retro-Cascorder records the temporal history of specific gene transcription events.

At the moment, Retro-Cascorder has only been used in bacteria and only shows the order in which genes are transcribed, not the time between each event. In the future, as this technology continues to be developed, Retro-Cascorder could also be used to record gene expression patterns during complex cell events, such as in immune cells during an inflammatory response or in cancer cells to unravel the process of tumour formation. [6,7]

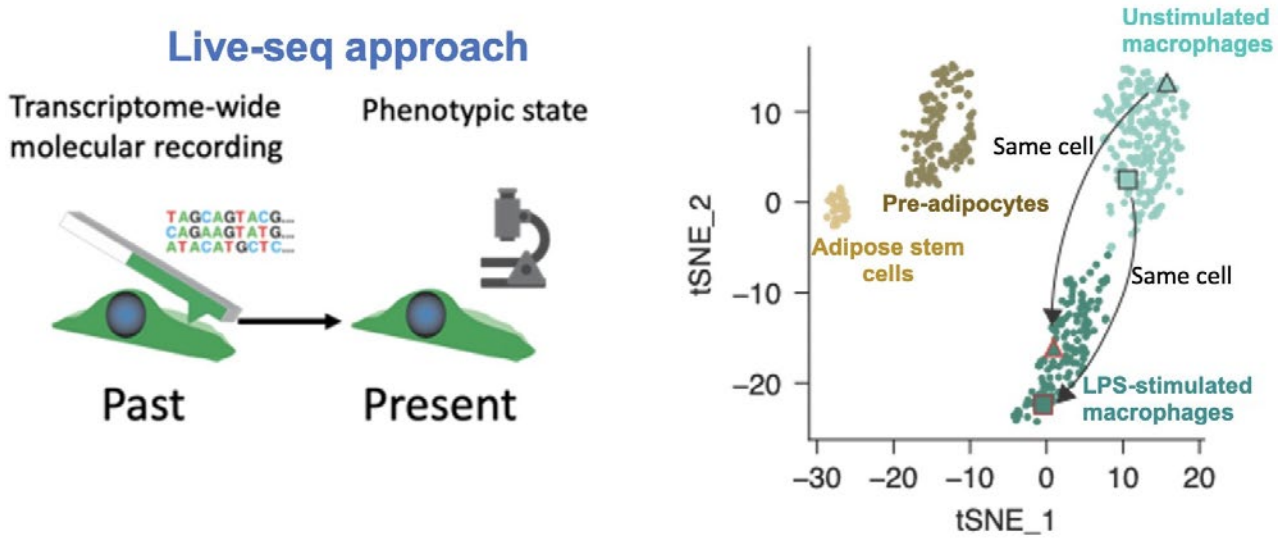
FIGURE 3: CAS1-CAS2 INTEGRATES RETRON RT-DNA. A: ILLUSTRATION OF RETROELEMENT-BASED TRANSCRIPTIONAL RECORDING INTO CRISPR ARRAYS. B: ILLUSTRATION OF BIOLOGICAL COMPONENTS CONTAINED IN THE RETRON-BASED RECORDER.



LIVE-SEQ

IN A PAPER RECENTLY PUBLISHED IN NATURE, CHEN ET AL. DESCRIBE LIVE-SEQ. LIVE-SEQ TAKES THINGS ONE STEP FURTHER, **BY ALLOWING FOR TRANSCRIPTOMIC PROFILING OF SINGLE CELLS WHILST THEY ARE STILL ALIVE AND FUNCTIONAL.** IN OTHER WORDS, **LIVE-SEQ CAN MONITOR THE ACTIVITY OF THOUSANDS OF GENES IN A SINGLE CELL AT DISCRETE TIMEPOINTS WHILE KEEPING THAT SAME CELL ALIVE TO SEE HOW THE TRANSCRIPTOME CHANGES, BUT ALSO HOW ITS PHENOTYPE AND BEHAVIOUR CHANGE.** [8]

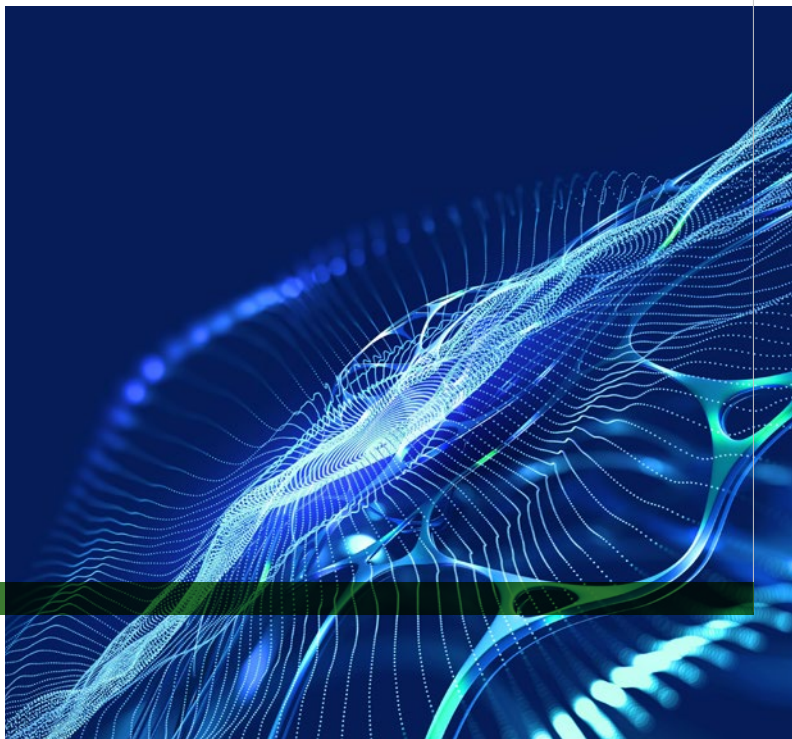
FIGURE 4: USING LIVE-SEQ TO RANK GENES INFLUENCING HETEROGENEITY OF MACROPHAGES' RESPONSE TO LPS.

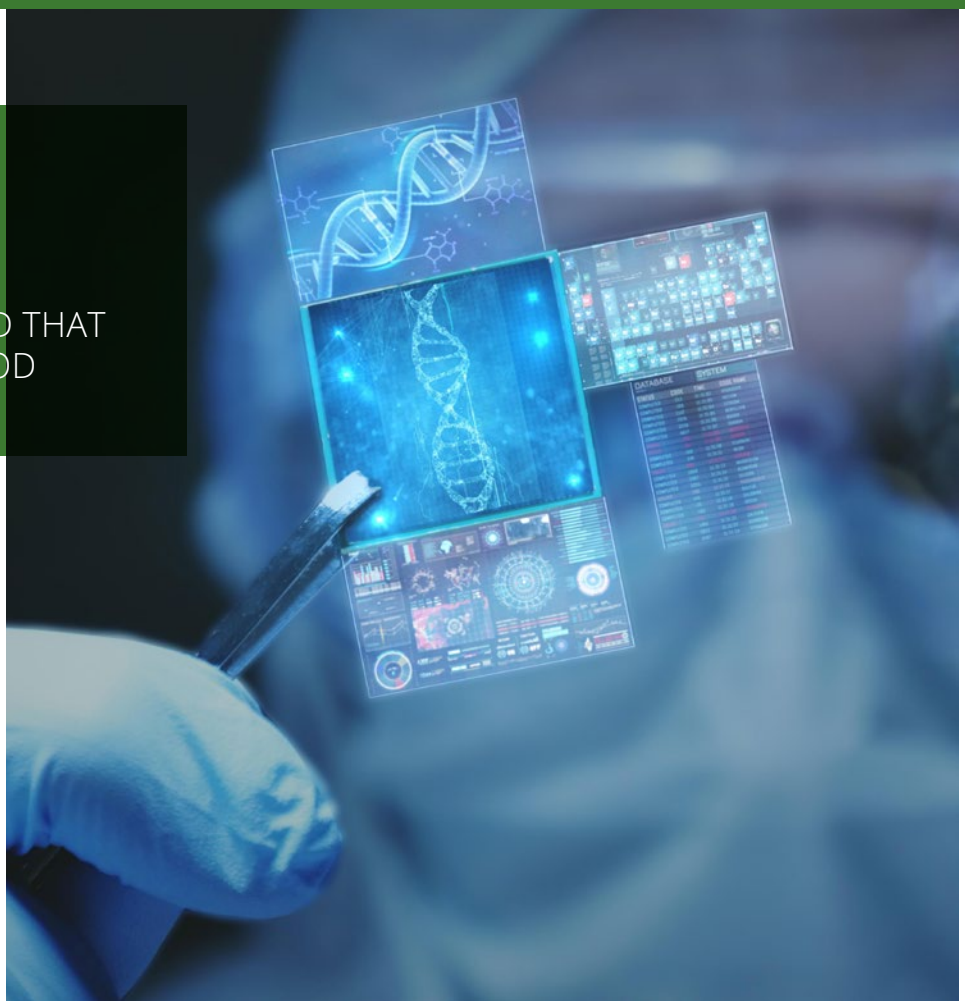


Bart Deplancke An author of this paper: With Live-seq, we can now uniquely address highly interesting and biomedically relevant questions, such as why certain cells differentiate and sister cells do not, or why certain cells are resistant to a cancer drug, while their sister cells are again not. [9]

LIVE-seq uses a cytoplasmic biopsy approach to extract RNA whilst preserving cell viability – in this way, researchers can link a cell's current transcriptomic state to its downstream molecular or phenotypic behaviour. This technique thus preserves both the spatial and temporal context of transcriptomic data.

Julia Vorholt An author of this paper: Live-seq can address a broad range of biological questions by transforming scRNA-seq from an endpoint to a temporal and spatial analysis approach. [9]





"LIVE-SEQ THUS ALLOWS FOR QUESTIONS TO BE ADDRESSED THAT NO OTHER SCRNA-SEQ METHOD DIRECTLY CAN."

The cytoplasmic biopsy approach is based on fluidic force microscopy (FluidFM), which has been previously shown to extract RNA from single cells while preserving their viability. However, only a limited amount of cytoplasmic mRNA can be extracted with FluidFM, so optimising the amount of mRNA extracted, minimizing mRNA degradation, and coupling this with a low-input RNA sequencing approach allowed for LIVE-seq to work.

The team behind LIVE-seq showed that it can accurately stratify diverse cell types and states, without introducing significant disturbances. They also used LIVE-seq to map the trajectory of macrophages before and after activation, as well as map the trajectory of adipose stromal cells before and after they become active. The researchers used LIVE-seq as a transcriptomic recorder as well, by tracking molecular events that are predictive of a macrophage's downstream phenotype, allowing them to predict how a macrophage would react to an immunological challenge [8].

Wanze Chen
Principal Investigator
Shenzhen Institute of Advanced Technologyand an author of this paper:

We were able to show that Live-seq is orthogonal to any other scRNA-seq approach today because it keeps cells alive and functional, while all other approaches cannot. This, in turn, enables the transcriptome of the cell to be captured prior to phenotyping, or to sequentially profile the transcriptome of a cell at different time points. Live-seq thus allows for questions to be addressed that no other scRNA-seq method directly can. These include, as illustrated in our study, how molecular and cellular heterogeneity is established, and what the actual (and not statistical) trajectory of cells are. We anticipate that Live-seq has the potential to transform scRNA-seq from its current end-point-type assay into a real-time analysis workflow.

References:

1. Gorin, Gennady et al. **"RNA velocity unraveled."** PLoS computational biology vol. 18,9 e1010492. 12 Sep. 2022, doi:10.1371/journal.pcbi.1010492
2. La Manno, Gioele et al. **"RNA velocity of single cells."** Nature vol. 560,7719 2018: 494-498. doi:10.1038/s41586-018-0414-6
3. **"MEFISTO - A Method for the Functional Integration of Spatial and Temporal Omics data"** Accessed 21/11/2022 https://biofam.github.io/MOFA2/MEFISTO.html
4. Velten, Britta et al. **"Identifying temporal and spatial patterns of variation from multimodal data using MEFISTO."** Nature methods vol. 19,2 2022: 179-186. doi:10.1038/s41592-021-01343-9
5. **"What If Cells Kept Receipts of Their Gene Expression?"** Accessed 21/11/2022 https://www.wired.com/story/what-if-cells-kept-receipts-of-their-gene-expression/
6. Bhattarai-Kline, Santi et al. **"Recording gene expression order in DNA by CRISPR addition of retron barcodes."** Nature vol. 608,7921 2022: 217-225. doi:10.1038/s41586-022-04994-6
7. **"New molecular device 'Retro-Cascorder' uses CRISPR to track and record gene expression over time"** Accessed 21/11/2022 https://frontlinegenomics.com/new-molecular-device-retro-cascorder-uses-crispr-to-track-and-record-gene-expression-over-time/
8. Chen, Wanze et al. **"Live-seq enables temporal transcriptomic recording of single cells."** Nature vol. 608,7924 2022: 733-740. doi:10.1038/s41586-022-05046-9
9. **"Live-seq: Sequencing a cell without killing it"** Accessed 21/11/2022, https://www.sciencedaily.com/releases/2022/08/220817114214.htm