Report from Dagstuhl Seminar 24491

Deep Learning for RNA Regulation and Multidimensional Transcriptomics

Annalisa Marsico*[1], Uwe Ohler*[2], Igor Ulitsky*[3], Kathi Zarnack*[4], and Charlotte Capitanchik†[5]

1 **Helmholtz Zentrum München, DE.** annalisa.marsico@helmholtz-muenchen.de
2 **Max-Delbrück-Centrum – Berlin, DE.** uwe.ohler@mdc-berlin.de
3 **Weizmann Institute – Rehovot, IL.** igor.ulitsky@weizmann.ac.il
4 **Universität Würzburg, DE.** kathi.zarnack@uni-wuerzburg.de
5 **The Francis Crick Institute – London, GB.** charlotte.capitanchik@crick.ac.uk

Abstract

The Dagstuhl Seminar 24491 "Deep Learning for RNA Regulation and Multidimensional Transcriptomics" convened experts from computer science, computational biology, and experimental research to explore the intersection of artificial intelligence and RNA biology. The seminar facilitated discussions on the latest computational methods and experimental approaches that are reshaping our understanding of RNA-mediated gene regulation. With the rapid growth of transcriptomics data, deep learning methods are becoming essential tools for extracting insights from complex datasets, ranging from primary sequence information to intricate cellular dynamics.

A key theme of the seminar was the exploration of non-coding RNAs, including long non-coding RNAs (lncRNAs) and microRNAs, which play pivotal roles in regulating gene expression. High-throughput methods to profile these RNAs, combined with deep learning algorithms, are enabling the identification of novel regulatory mechanisms and the prediction of their cellular functions. The discussion underscored the challenges in classifying lncRNAs, deciphering their sequence features, and understanding their functional interactions.

The seminar also addressed the integration of deep learning in modeling RNA regulatory networks. Participants presented cutting-edge models for predicting RNA modifications, RNA–protein interactions, and the effects of genetic variants on RNA metabolism. Special attention was given to the interpretability of machine learning models, as understanding the biological significance of predictions remains a critical challenge. Advances in single-cell and spatial transcriptomics were highlighted as key drivers of future breakthroughs, offering unprecedented resolution of cellular heterogeneity and regulatory processes.

Another major focus was the role of deep learning in RNA-based therapeutic development. Discussions included the use of machine learning for designing RNA sequences in synthetic biology applications, predicting the efficacy of antisense oligonucleotides (ASOs), and identifying cancer-specific neoantigens. These applications demonstrate the potential of AI to accelerate the discovery of novel RNA-targeted therapies and improve precision medicine approaches.

In addition, the seminar emphasized the importance of community-driven initiatives to improve benchmarking, data curation, and collaborative model development. Participants highlighted the need for standardized datasets, transparent evaluation metrics, and shared computational resources to foster reproducibility and innovation. The discussions underscored the necessity of cross-disciplinary collaboration to ensure that machine learning methods address biologically meaningful questions and produce actionable insights.

Overall, the seminar illustrated how deep learning is transforming RNA biology by uncovering new layers of gene regulation and facilitating therapeutic discoveries. Moving forward, continued interdisciplinary collaboration and the development of scalable, interpretable models will be essential to unlock the full potential of AI in decoding RNA functions and advancing biomedical research.

* Editor / Organizer
† Editorial Assistant / Collector

Deep Learning for RNA Regulation and Multidimensional Transcriptomics
Dagstuhl Reports, Vol. 14, Issue 12, pp. 1–27
Editors: Annalisa Marsico, Uwe Ohler, Igor Ulitsky, and Kathi Zarnack

Dagstuhl Reports
Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

Seminar December 1–6, 2024 – https://www.dagstuhl.de/24491
2012 ACM Subject Classification Applied computing → Computational transcriptomics; Applied computing → Molecular sequence analysis; Computing methodologies → Artificial intelligence
Keywords and phrases deep learning, epitranscriptomics, rna, single-cell transcriptomics
Digital Object Identifier 10.4230/DagRep.14.12.1

1 Executive Summary

Kathi Zarnack (Universität Würzburg, DE)
Annalisa Marsico (Helmholtz Zentrum München, DE)
Uwe Ohler (Max-Delbrück-Centrum – Berlin, DE)
Igor Ulitsky (Weizmann Institute – Rehovot, IL)

The Dagstuhl Seminar 24491 "Deep Learning for RNA Regulation and Multidimensional Transcriptomics" brought together an interdisciplinary group of computer scientists, computational biologists, and experimentalists to discuss current challenges and emerging opportunities at the intersection of RNA biology and artificial intelligence. Across a broad range of sessions, participants showcased recent advances in deep learning models, RNA sequencing methods, and systems biology approaches, revealing novel insights into the epitranscriptome, RNA structure and function, RNA–protein interactions, regulatory mechanisms, and disease biology.

A central theme of the seminar was the exploration of transcriptome complexity. This extends well beyond coding sequences, as non-coding RNAs, such as long non-coding RNAs (lncRNAs), microRNAs and enhancer RNAs, play essential roles in modulating gene expression. Toward the molecular code of lncRNA function, several speakers presented high-throughput approaches to study the spatiotemporal behavior of lncRNAs and dissect their functions in cis and trans. In addition, several presentations emphasized how translating ribosome footprints illuminate new regulatory events in non-coding regions, while new deep learning models decipher the interplay between motif composition and translation outcomes.

Another major focus was the promise of AI-driven analyses of vast repositories of RNA-sequencing data. Participants showcased deep learning strategies to identify regulators of splicing, alternative polyadenylation, RNA stability, and translation. From the single-cell perspective, novel methods–particularly those combining metabolic labeling or spatial transcriptomics with deep learning–offered unprecedented resolution into gene regulatory cascades in development and disease. These high-throughput approaches highlighted how combinatorial readouts of RNA modifications, transcript isoforms, and protein–RNA binding events can be turned into powerful predictive frameworks through deep neural networks.

At the level of RNA–protein interactions, discussions centered on data from experiments like individual-nucleotide resolution UV crosslinking and immunoprecipitation (iCLIP) and multiplexed profiling platforms that revealed how RBPs coordinate transcript processing and decay. Deep learning architectures such as RBPNet and panRBPNet showcased how to predict RBP binding sites directly from primary sequence, integrating background noise models and improved interpretability. Complementarily, methods that quantify the uncertainty of predictions or predict changes in splicing and translational efficiency broaden the scope and robustness of computational pipelines.

Disease contexts, including metabolic liver disorders, neuroblastoma, and other malignancies, highlighted the biomedical importance of precisely quantifying transcriptomic complexity and RNA-regulatory events. By capturing single-cell and subcellular heterogeneity, these advanced molecular methods can inform novel therapeutic strategies. Participants presented case studies on the roles of specific lncRNAs in cellular pathways, microproteins in cardiomyocytes, and the impact of RNA modifications on stability and translation.

Complementary to the scientific presentations, the Dagstuhl Seminar included several panel discussions and dropout sessions that carved out twelve critical research tasks for RNA-focused machine learning research. These tasks encompass understanding RNA molecules by identifying functional lncRNAs and their mechanisms of action, mapping chemical modifications and their effects, predicting RNA secondary and tertiary structures, and representing RNA conservation and functional regions. Furthermore, the challenge of understanding interactions and complexes was addressed, including context-aware prediction of regulatory targets, modeling RNA-protein complex formation including condensates, and constructing combinatorial maps of cellular compartments and their RNA-regulatory networks. In addition, the discussions highlighted several domain-specific challenges for machine learning in RNA biology as biological data is inherently complex, often messy, and continually evolving, making data processing and curation particularly demanding. Benchmarking emerged as a crucial aspect of model evaluation as participants stressed the need for unbiased datasets, developing benchmarks for non-model organisms, and organizing community challenges to foster competition and innovation. Future directions emphasized were improving data curation and infrastructure, collaborative model development, transfer learning for cross-species applications, and creating practical, efficient models accessible to the academic community.

In sum, the Dagstuhl Seminar provided a rich forum for bridging computational and experimental frontiers in RNA biology. Emerging deep learning algorithms are revealing complex layers of RNA regulation, while large-scale, high-resolution data–spanning ribosome profiling, single-cell labeling, nanopore direct RNA sequencing, and other transcriptomic techniques–continuously expand our understanding. Altogether, the meeting underscored the need for continued collaboration between computational method developers and bench biologists to realize the transformative potential of deep learning in decoding RNA-based regulation and its application to human health.

2 Table of Contents

3 Overview of Talks

3.1 Machine-Learning in the Context of Bioinformatics Research

Rolf Backofen (Universität Freiburg, DE)

It is becoming increasingly clear that a RNA-binding proteins (RBPs) are key elements in regulating the cell's transcriptome. CLIP-seq is one of the major tools to determine binding sites but suffers from high false negative rate due its dependency on expression levels, hindering the large-scale use of public CLIP-data. We will show in several examples how use of raw public CLIP data can lead to false biological reasoning and how advanced machine learning approach can overcome this problem. I will further discuss our results from Nature paper, showing that the human RNA helicase DHX9 predominately binds to IRAlu elements and thus suppresses the negative effect of Alu inflation in transcripts. I will then proceed by discussing distance measures for RBP-binding sites, and how the AI-models can be used to determine properties of binding sites of RBPs.

3.2 How to deal with complexity of eukaryotic translation?

Pavel Baranov (University College Cork, IE)

Joint work of Jack A. S. Tierney, Michał Świrski, Håkon Tjeldnes, Anmol Kiran, GionMattia Carancini, Stephen Kiniry, Audrey Michel, Jonathan M. Mudge, Joanna Kufel, Nicola Whiffin, Eivind Valen, Pavel V. Baranov

Main reference Jack A. S. Tierney, Michał Świrski, Håkon Tjeldnes, Jonathan M. Mudge, Joanna Kufel, Nicola Whiffin, Eivind Valen, Pavel V. Baranov: "Ribosome Decision Graphs for the Representation of Eukaryotic RNA Translation Complexity", bioRxiv, Cold Spring Harbor Laboratory, 2023.

URL https://doi.org/10.1101/2023.11.10.566564

We present RiboSeq.Org, a suite of tools for processing, analyzing, and visualizing publicly available ribosome profiling (RiboSeq) data. RiboSeq provides genome-wide, nucleotide-resolution mapping of ribosome positions on cellular RNA. Our analysis of these data reveals widespread use of multiple translation initiation sites across human cytoplasmic RNAs, including both mRNAs and non-coding RNAs. This translational complexity arises from leaky scanning and reinitiation downstream of short translated regions (translons). The interdependence of these translation events regulates protein synthesis through their precise organizational hierarchy. To analyze these relationships and effects of sequence variations on translation, we developed Ribosome Decision Graphs (RDG), a novel framework for abstractly representing translons organisation within individual RNA molecules. RDGs combine all possible ribosome trajectories through an mRNA into a single graph, with initiation sites and reinitiation-enabling stop codons serving as branch points. We demonstrate RDG utility in analyzing condition-specific RiboSeq data and pathogenic effects of 5' leader variants in human mRNAs. The framework accommodates unconventional translation mechanisms such as frameshifting, readthrough, and selenocysteine insertion and it can be extended to yet undiscovered RNA decoding phenomena.

3.3 Towards building RBP networks using all available RNA-seq data

Charlotte Capitanchik (The Francis Crick Institute – London, GB)

Joint work of Charlotte Capitanchik, Guillermo E. Parada Gonzalez, Benjamin J. Blencowe, Jernej Ule

The ever-increasing volume of publicly available short-read RNA sequencing datasets presents an opportunity to study RNA processing signatures (e.g. splicing, alternative polyadenylation) at an unprecedented scale. Given computed RNA signatures in a novel cellular context, one could hypothetically search across hundreds of thousands of samples to identify those with similar signatures. From here, gene regulatory networks or investigation of sample metadata could identify novel regulators. Here, I present correlation of RNA-binding protein gene expression profiles with specific exon programs across 150,000 human RNA-seq datasets. I demonstrate that gene-gene expression correlations can show very different relationships than gene-splicing correlations and that this information can be used to explore novel RBP interactions through the example of KHSRP and PTBP1 which act antagonistically. Further, when interesting sample subsets are identified, large language models can be used to summarise and interpret thousands of rows of sample metadata, which can be sparse and weakly structured.

3.4 Small ORFs in the heart & How translation initiations factors shape the proteome

Christoph Dieterich (Universitätsklinikum Heidelberg, DE)

Joint work of Christoph Dieterich, Jeroen Krijgsveld, Toman Bortecen
Main reference Brandon Malone, Ilian Atanassov, Florian Aeschimann, Xinping Li, Helge Großhans, Christoph Dieterich: "Bayesian prediction of RNA translation from ribosome profiling", Nucleic Acids Research, Vol. 45(6), pp. 2960–2972, 2017.
URL https://doi.org/10.1093/nar/gkw1350
Main reference Shirin Doroudgar, Christoph Hofmann, Etienne Boileau, Brandon Malone, Eva Riechert, Agnieszka A. Gorska, Tobias Jakobi, Clara Sandmann, Lonny Jürgensen, Vivien Kmietczyk, Ellen Malovrh, Jana Burghaus, Mandy Rettel, Frank Stein, Fereshteh Younesi, Ulrike A. Friedrich, Victoria Mauz, Johannes Backs, Günter Kramer, Hugo A. Katus, Christoph Dieterich, Mirko Völkers: "Monitoring Cell-Type–Specific Gene Expression Using Ribosome Profiling In Vivo During Cardiac Hemodynamic Stress", Circulation Research, Vol. 125(4), pp. 431–448, 2019.
URL https://doi.org/10.1161/CIRCRESAHA.119.314817
Main reference Maja Bencun, Laura Spreyer, Etienne Boileau, Jessica Eschenbach, Norbert Frey, Christoph Dieterich, Mirko Völkers: "A novel uORF regulates folliculin to promote cell growth and lysosomal biogenesis during cardiac stres", Sci Rep. 15(1):3319. PMID: 39865126; PMCID: PMC11770079, 2025 Jan 27
URL https://doi.org/10.1038/s41598-025-87107-3

Open Reading Frames (ORFs) occur randomly in RNA sequences. We use Ribosome Footprinting to uncover actively translated subsequences on RNA. We have mapped the entire translatome in the mouse heart using physiological and pathological stress models. Subsequently, we have focused on upstream ORFs, which may play a role in control of RNA translation of canonical ORFs (in cis or trans) or produce an independently acting microprotein. Candidate uORFs were selected based on positional and sequence conservation across human, mouse and rat. We have presented some results on one uORF of Folliculin (Flcn). Ablation of the uORF start codon leads to a five-fold increase in translation of the canoncial downstream ORF. A smaller increase could be attained with an antisense

oligo as well. Through an array of additional experiments, we could show that Flcn levels indirectly regulate lysosomal biogenesis by transcription regulation of autophagy-related genes. This phenomenon occurs under hypertrophic conditions when the uORF is more actively translated. In the second half of the presentation, we presented a genetic ablation screen of dozens of RNA translation initiation factors and discussed how they shape the proteome. We reported that half of the tested factors lead to a clear down-regulation of protein synthesis across thousands of loci, while not changing RNA abundance.

3.5 Heterogeneity-seq predicts causal factors affecting outcomes of molecular challenges

Florian Erhard (Universität Regensburg, DE)

A major limitation of standard single cell RNA-seq (scRNA-seq) is that each cell can only be measured once. As a consequence, any study comparing cells before and after a molecular challenge such as virus infection is purely correlative. Metabolic RNA labelling over short time scales combined with scRNA-seq can provide the prior and current state of each cell, breaking this limitation. Using statistical data analysis and causal inference techniques, we demonstrate that this approach we coined "Heterogeneity-seq" can identify causal factors that impact on molecular outcomes. By the same principle, we construct single cell trajectories to characterize the progression of individual cells over longer time scales. Heterogeneity-seq uncovered genes with an effect on drug treatment and novel pro- and antiviral host factors of cytomegalovirus infection.

3.6 Decoding the epitranscriptome at single-molecule resolution

Eduardo Eyras (Australian National University – Canberra, AU)

Joint work of Stefan Prodic, Akansha Srivastava, Agin Ravindran, Aditya J. Sethi, Gaby Santos-Rodriguez, Shafi Mahmud, Madhu Kanchi, Grazi Vieira, Arash Hajizadeh-Dastjerdi, Alice Cleynen, Rippei Hayashi, Robert Weatheritt, Nikolay Shirokikh, Eduardo Eyras

Main reference Pablo Acera Mateos, Aditya J Sethi, Agin Ravindran, Akansha Srivastava, Katrina Woodward, Shafi Mahmud, Madhu Kanchi, Marco Guarnacci, J Xu, Zaka Yuen, You Zhou, Alexandra Sneddon, William B. Hamilton, Jing Gao, Lora M. Starrs, Rippei Hayashi, Vihandha Wickramasinghe, Kathi Zarnack, Thomas Preiss, Gaetan Burgio, Nathalie Dehorter, Nikolay E. Shirokikh, Eduardo Eyras: "Prediction of m6A and m5C at single-molecule resolution reveals a transcriptome-wide co-occurrence of RNA modifications", Nat Commun 15, 3899 (2024)

URL https://doi.org/10.1038/s41467-024-47953-7

Messenger RNAs (mRNAs) are key molecules that express the genetic code to be translated into proteins that sustain life. mRNAs are composed of specific combinations of four nucleotides that define mRNA properties and the proteins they encode. These nucleotides can also be chemically modified in specific ways, defining the epitranscriptome, which modulates the properties and fate of mRNA molecules. We still lack complete knowledge of the modifications present in mRNA and their specific functions. To address this challenge, we have generated AI-based models to identify modifications in mRNA transcriptome-wide. These models use the ionic signals from nanopore sequencing devices to identify and label

different modified nucleotides at single-molecule resolution. We trained our AI models using a combination of controlled datasets specifically generated to contain known modifications and data from cellular RNAs with known modification sites. We applied our models to RNA from various species and experimental conditions to discover that mRNA modifications are positionally conserved across evolution. Further, we found that modifications occur in specific combinations of sites and modification types on mRNA molecules. Remarkably, we found evidence that modifications are deposited very early during mRNA transcription and before splicing catalysis, unlike what was previously assumed. Our AI models and analysis strategies provide an unprecedented opportunity to characterise the epitranscriptome at single-molecule resolution for multiple species and under a wide range of experimental conditions, thereby enabling the mapping of the epitranscriptome and its functions.

3.7 DNA language model interpretation reveals functional genomic elements

Julien Gagneur (TU München – Garching, DE)

Joint work of Julien Gagneur, Pedro Tomaz da Silva, Alexander Karollus, Johannes Hingerl, Gihanna Galindez, Nils Wagner, Xavier Hernandez-Alias, Danny Incarnato
Main reference Pedro Tomaz da Silva, Alexander Karollus, Johannes Hingerl, Gihanna Galindez, Nils Wagner, Xavier Hernandez-Alias, Danny Incarnato, Julien Gagneur: "Nucleotide dependency analysis of DNA language models reveals genomic functional elements", bioRxiv, Cold Spring Harbor Laboratory, 2024.
URL https://doi.org/10.1101/2024.07.27.605418

Deciphering how nucleotides in genomes encode regulatory instructions and molecular machines is a long-standing goal in biology. Supervised modeling approaches that predict molecular assays such as chromatin accessibility or RNA sequencing from genomic sequences have substantially advanced our ability to predict and design transcriptional regulatory elements, especially in mammals. However, experimental data are scarcer for other regulatory layers and species, limiting the power of supervised learning.

DNA language modeling (LM), where large models are trained on genomic sequences alone, offers a promising complementary approach for deriving effective sequence representations. First, I showed that species-aware DNA LMs, which we trained on over 800 species spanning 500 million years of evolution, capture high-order sequence and evolutionary context and yield improved sequence representations for gene expression prediction. Second, I introduced nucleotide dependencies, which quantify how substitutions at one genomic position affect DNA LM probabilities at other positions. We generated genome-wide maps of pairwise nucleotide dependencies across kilobase ranges for animal, fungal, and bacterial species. Nucleotide dependencies indicate the deleteriousness of human genetic variants more effectively than sequence alignment and DNA LM reconstruction. Regulatory elements appear as dense blocks in dependency maps, enabling the identification of transcription factor binding sites. Nucleotide dependencies also reveal bases in contact within RNA structures, including pseudoknots and tertiary contacts, with remarkable accuracy. This led to the discovery of four novel, experimentally validated RNA structures in Escherichia coli. Finally, I discussed how dependency maps reveal limitations in DNA LM architectures and training strategies. Altogether, DNA LMs and nucleotide dependency analysis open new avenues for discovering and studying functional genomic elements and their interactions.

3.8 Computational Methods for Quantification of Transcript Expression and Modifications using Long Read RNA Sequencing

Jonathan Göke (Genome Institute of Singapore, SG)

Joint work of Jonathan Göke, Chen Ying, Andre Sim, Christopher Hendra, Min Hao Ling, Yuk Kei Wan, Sui Yue, Ploy Pratanwanich

The human genome contains instructions to transcribe more than 200,000 RNAs. However, many RNA transcripts are generated from the same gene, resulting in alternative isoforms that are highly similar. Furthermore, the addition of post-transcriptional RNA modifications further impacts their function. The availability of long read RNA-Seq provides an opportunity to sequence entire RNA transcripts, enabling the analysis of individual RNA isoforms. Furhtermore the ability to sequence native RNA directly allows the computational identification of modified RNA bases. Using the information form long read and direct RNA-Seq we have developed computational methods that enable the transcript discovery and quantification for bulk, single cell and spatial long read RNA-Seq data (Bambu), the identification of m6A modifications for individual RNA molecules (m6Anet), and the quantification of differential modification rates across conditions (xPore). To evaluate and develop computational methods for long read RNA-Seq, we have further generated a systematic data resource, the Singapore Nanopore Expression Project (SG-NEx), which includes RNA-Seq data from 7 cell lines and 5 different RNA-Seq protocols. Together, these tools and resources simplify the analysis of long read RNA-Seq data and enable the discovery of new transcripts and RNA modifications with high accuracy.

3.9 SPIDR: a highly multiplexed method for mapping RNA-protein interactions

Marko Jovanovic (Columbia University, US)

Joint work of Marko Jovanovic, Erica Wolin, JK Guo, Mario Reynaldo Blanco, W Dong, D Gorhe, AA Perez, IN Goronzy, Abdurrahman Keskin, E Valenzuela, AA Abdou, Carl R. Urbinati, R Kaufhold, H. Thomas Rube, Jay Brito Querido, Mitchell Guttman

RNA-binding proteins (RBPs) play crucial roles in regulating every stage of the mRNA life cycle and mediating non-coding RNA functions. Despite their importance, the specific roles of most RBPs remain unexplored because we do not know their specific RNA binding partners. Current methods, such as crosslinking and immunoprecipitation followed by sequencing (CLIP-seq), have expanded our knowledge of RBP-RNA interactions, but are generally limited by their ability to map only one RBP at a time. To address this limitation, we developed SPIDR (Split and Pool Identification of RBP targets), a massively multiplexed method to simultaneously profile global RNA binding sites of dozens to hundreds of RBPs in a single experiment. SPIDR employs antibody-bead labeling coupled with split-pool barcoding to increase the throughput of current CLIP methods by two orders of magnitude. SPIDR identifies precise, single-nucleotide RNA binding sites for diverse classes of RBPs simultaneously. We identified several novel ribosomal RNA binders, including a novel interaction between LARP1 and 18S ribosomal RNA located within the mRNA channel on

the 40S small ribosomal subunit, and resolved this structure at 2.8 Å using single-particle cryo-electron microscopy (cryo-EM). This structure provides a potential mechanistic explanation for the role of LARP1 in translational suppression. We used SPIDR to explore changes in RBP binding upon mTOR inhibition and identified that 4EBP1 preferentially binds translationally repressed mRNAs only upon mTOR inhibition. These observations provide a potential mechanism to explain the specificity of translational regulation controlled by mTOR signaling. SPIDR enables rapid, de novo discovery of RNA-protein interactions at an unprecedented scale and has the potential to transform our understanding of RNA biology and both transcriptional and post-transcriptional gene regulation.

3.10 Transcriptional diversity and cellular plasticity in neuroblastoma

Jan Philipp Junker (Max-Delbrück-Centrum – Berlin, DE)

Transcriptional heterogeneity and phenotypic plasticity are increasingly recognized as drivers of tumor progression, metastasis and treatment evasion. While tumor heterogeneity can be measured with single cell transcriptomics, major biological questions remain unresolved since we cannot readily follow cells over time: In particular, how plastic are gene expression programs of tumor cells, and to which degree are gene expression states determined by the cell of origin or the local environment?

Neuroblastoma is a neural crest derived malignancy of the peripheral nervous system and is the most common and deadliest tumor of infancy. Neuroblastoma is characterized by high phenotypic heterogeneity but low genetic diversity. Here, we combined zebrafish models of neuroblastoma, single-cell transcriptomics, massively parallel lineage tracing, and transplantation of tumor cells to i) systematically dissect intra- and inter-tumor transcriptional heterogeneity, ii) to experimentally measure the plasticity of tumor states, and iii) clarify to which degree tumor states are determined by cell of origin or local environment. We discovered a large spectrum of distinct neuroblastoma states in zebrafish, which can be broadly classified according to physiological states or cell of origin. Importantly, the tumor states in zebrafish reflect the transcriptional diversity of patient samples, including an alk-expressing and a ribosomal gene expression program, both of which are associated with poor prognosis in patients. Analysis of CRISPR/Cas9- inserted lineage barcodes revealed a gradient of plasticity across tumor states, with lower plasticity in tumor states related to cell of origin compared to tumor states associated with physiological processes. Transplantation of zebrafish tumors into embryos showed that even the least plastic tumor states can be reprogrammed upon exposure to a different signaling environment, an observation which has important consequences for future therapeutic strategies.

In summary, I presented a comprehensive dataset integrating computational dissection of tumor expression programs, lineage tracing, and transplantation of tumor cells, to measure tumor cell plasticity and elucidate how cell of origin and local environment cooperate to shape tumor expression profiles.

3.11 Gene expression and RNA regulation in healthy and diseased cells

Claudia Kutter (Karolinska Institute – Stockholm, SE)

Main reference Jonas Nørskov Søndergaard, Christian Sommerauer, Ionut Atanasoai, Laura C Hinte, Keyi Geng, Giulia Guiducci, Lars Bräutigam, Myriam Aouadi, Lovorka Stojic, Isabel Barragan, Claudia Kutter: "CCT3-LINC00326 axis regulates hepatocarcinogenic lipid metabolism", Gut, Vol. 71(10), pp. 2081–2092, BMJ Publishing Group, 2022.
URL https://doi.org/10.1136/gutjnl-2021-325109
Main reference Christian Sommerauer, Carlos J Gallardo-Dodd, Christina Savva, Linnea Hases, Madeleine Birgersson, Rajitha Indukuri, Joanne X Shen, Pablo Carravilla, Keyi Geng, Jonas Nørskov Søndergaard, Clàudia Ferrer-Aumatell, Grégoire Mercier, Erdinc Sezgin, Marion Korach-André, Carl Petersson, Hannes Hagström, Volker M Lauschke, Amena Archer, Cecilia Williams, Claudia Kutter: "Estrogen receptor activation remodels TEAD1 gene expression to alleviate hepatic steatosis", Molecular Systems Biology, Vol. 20(4), pp. 374–402, 2024.
URL https://doi.org/10.1038/s44320-024-00024-x
Main reference Ionut Atanasoai, Sofia Papavasileiou, Natalie Preiß, Claudia Kutter: "Large-scale identification of RBP-RNA interactions by RAPseq refines essentials of post-transcriptional gene regulation", bioRxiv, Cold Spring Harbor Laboratory, 2021.
URL https://doi.org/10.1101/2021.11.08.467743
Main reference Sandhya Malla, Devi Prasad Bhattarai, Paula Groza, Dario Melguizo-Sanchis, Ionut Atanasoai, Carlos Martinez-Gamero, Ángel-Carlos Román, Dandan Zhu, Dung-Fang Lee, Claudia Kutter, Francesca Aguilo: "ZFP207 sustains pluripotency by coordinating OCT4 stability, alternative splicing and RNA export", EMBO reports, Vol. 23(3), p. e53191, 2022.
URL https://doi.org/10.15252/embr.202153191
Main reference Carlos J. Gallardo-Dodd, Christian Oertlin, Julien Record, Rômulo G. Galvani, Christian Sommerauer, Nikolai V. Kuznetsov, Evangelos Doukoumopoulos, Liaqat Ali, Mariana M. S. Oliveira, Christina Seitz, Mathias Percipalle, Tijana Nikić, Anastasia A. Sadova, Sofia M. Shulgina, Vjacheslav A. Shmarov, Olga V. Kutko, Daria D. Vlasova, Kseniya D. Orlova, Marina P. Rykova, John Andersson, Piergiorgio Percipalle, Claudia Kutter, Sergey A. Ponomarev, Lisa S. Westerberg: "Exposure of volunteers to microgravity by dry immersion bed over 21 days results in gene expression changes and adaptation of T cells", Science Advances, Vol. 9(34), p. eadg1610, 2023.
URL https://doi.org/10.1126/sciadv.adg1610

Liver cancer incidences, especially hepatocellular carcinoma (HCC), are increasing worldwide and unless detected early, patient survival remains extremely low. Several underlying conditions, such as metabolic dysfunction-associated steatotic liver disease (MASLD, formerly NAFLD), can lead to liver cancer. Identifying the series of molecular changes that drive liver diseases towards becoming malignant remains a significant challenge. To better understand the molecular phenotypes of liver cancer cells, we conducted a comprehensive characterization of molecular factors in MASLD and HCC. Through de novo transcriptome assembly, development of new binding assay (RAPseq), RBP regulatory network and gene regulatory network-based drug repurposing, we identified and validated key regulators, including transcription factors (TEAD1), canonical and non-canonical RNA-binding proteins (CCT3, AQR and PES1) and long noncoding RNAs (LINC00326). Perturbation of these factors resulted in major effects on gene regulation and the transcriptome, leading to shifts in the liver (cancer) cell phenotype. Further investigation revealed their involvement in distinct regulatory processes. The most noticeable connection affected lipid metabolism, whereby perturbation of the regulatory network led to decreased lipid accumulation and increased lipid degradation in cellulo as well as diminished tumor growth in vivo. Subsequent targeted inhibition with a small molecule reduced MASLD by suppressing lipogenic pathways, opening new avenues for future MASLD treatments and HCC preventions. Rewiring of transcriptional programs also affects immune cells of healthy volunteers exposed to microgravity through dry immersion. T cells adapt by rewiring their coding and noncoding transcriptome after 21 days of stimulated weightlessness. These remodeling cues persist even after re-exposure to normal gravity. Notably, these changes mirror those observed in crew members at the international space station (NASA's twin study). Given the growing interest in spaceflights, this finding provides the foundation for developing effective tests and countermeasures for deep-space exploration.

3.12 m6A sites in the coding region trigger translation-dependent mRNA decay

Julian König (Institut für Molekulare Biologie – Mainz, DE), Christoph Dieterich (Universitätsklinikum Heidelberg, DE), Kathi Zarnack (Universität Würzburg, DE)

Joint work of Julian König, Miona Corovic, Christoph Dieterich, Zhou You, Kathi Zarnack

N6-Methyladenosine (m6A) is the predominant internal RNA modification in eukaryotic messenger RNAs (mRNAs) and plays a crucial role in mRNA stability. Here, using human cells, we reveal that m6A sites in the coding sequence (CDS) trigger CDS–m6A decay (CMD), a pathway that is distinct from previously reported m6A-dependent degradation mechanisms. Importantly, CDS m6A sites act considerably faster and more efficiently than those in the 3' untranslated region, which to date have been considered the main effectors. Mechanistically, CMD depends on translation whereby m6A deposition in the CDS triggers ribosome pausing and transcript destabilization. The subsequent decay involves the translocation of the CMD target transcripts to processing bodies (P-bodies) and recruitment of the m6A reader protein YTHDF2. Our findings highlight CMD as a previously unknown pathway, which is particularly important for controlling the expression of developmental regulators and retrogenes.

3.13 (RNA-based) Spatiotemporal methods to play chess with single cells

Gioele La Manno (EPFL – Lausanne, CH)

The developing brain is like a complex chess game with millions of pieces – each belonging to one of hundreds of distinct cell types interacting and differentiation. While the single-cell revolution has revealed the foundation of this "game," leading us to identify these pieces, becoming true "Masters" is still ahead. With spatial transcriptomics, we would like to interpret each snapshot of the cells in the tissue well enough, distinguish normal from pathological states, and predict "future moves". Here, we present two complementary approaches advancing toward this dream. On the temporal front, we introduce VeloCycle, an advanced RNA velocity framework that tracks cell cycle-driven expression dynamics in real time, allowing us to predict cellular trajectories with unprecedented accuracy. On the spatial front, we present PointillHist, a GNN-based mapper that reveals the precise organization of 800 distinct cell states across the embryonic brain. When applied to study folate deficiency, we uncover the differential susceptibility of radial-glial populations and a "catastrophic flipping" of patterned territories. Together, these tools reveal new rules governing normal and pathological development.

3.14 Acute depletion of UPF1 reveals the temporal dynamics of the NMD-regulated transcriptome in human cells

Markus Landthaler (Max-Delbrück-Centrum – Berlin, DE)

Joint work of Volker Boehm, Damaris Wallmeroth, Paul O Wulf, Luiz Gustavo Teixeira Alves, Oliver Popp, Michael Riedel, Emanuel Wyler, Marek Franitza, Jennifer V. Gerbracht, Katja Becker, K. Polkovnychenko K, S. Del Giudice, Nouhad Benlasfer, Philipp Mertins, Niels H. Gehring

The helicase UPF1 acts as the central essential factor in human nonsense-mediated mRNA decay (NMD) and is involved in various other mRNA degradation processes. Given its multifunctionality, distinguishing between mRNAs regulated directly and indirectly by UPF1 remains a critical challenge. We engineered two different conditional degron tags into endogenous UPF1 in human cell lines to probe the consequences of UPF1 rapid depletion. UPF1 degradation inhibits NMD within hours and strongly stabilizes endogenous NMD substrates, which can be classified into different groups based on their expression kinetics. Extended UPF1 depletion results in massive transcript and isoform alterations, partially driven by secondary effects. We define a high-confidence UPF1-regulated core set of transcripts, which consists mostly of NMD substrates. NMD-regulated genes are involved in brain development and the integrated stress response, among other biological processes. In summary, UPF1 degron systems rapidly inhibit NMD, providing valuable insights into its roles across various experimental systems.

3.15 Mapping in vivo protein-RNA binding in plants using individual-nucleotide resolution crosslinking and immunoprecipitation (plant iCLIP2)

Martin Lewinski (Universität Bielefeld, DE)

Joint work of Martin Lewinski, Mirko Brüggemann, Tino Köster, Marlene Reichel, Thorsten Bergelt, Katja Meyer, Julian König, Kathi Zarnack, Dorothee Staiger
Main reference Martin Lewinski, Mirko Brüggemann, Tino Köster, Marlene Reichel, Thorsten Bergelt, Katja Meyer, Julian König, Kathi Zarnack, Dorothee Staiger: "Mapping protein-RNA binding in plants with individual-nucleotide-resolution UV cross-linking and immunoprecipitation (plant iCLIP2). Nat Protoc. 2024 Apr;19(4):1183-1234, Epub 2024 Jan 26. PMID: 38278964.
URL https://doi.org/10.1038/s41596-023-00935-3

RNA-binding proteins (RBPs) play essential roles in plant physiology and development, yet methods to comprehensively map their transcriptome-wide binding landscapes remain underdeveloped compared to those available for other model organisms. Existing cross-linking and immunoprecipitation (CLIP) techniques, relying on UV-mediated covalent linking of RNAs and their associated RBPs in vivo, purification of the resulting RNA-protein complexes, and identification of co-purified RNAs via high-throughput sequencing, have primarily been applied to mammalian cells and translucent tissues. We established plant iCLIP2, a robust protocol for individual-nucleotide-resolution CLIP (iCLIP) specifically adapted to plants, enabling detailed exploration of RBP binding sites across plant transcriptomes. Our method addresses key experimental challenges unique to plant systems. We optimize UV dosage for effective RNA-protein cross-linking in plant tissues and use epitope-tagged RBPs expressed under native promoters in loss-of-function mutants to ensure physiologically

relevant interactions are captured. Employing immunopurification, we establish stringent conditions for isolating RNA-protein complexes, even in the presence of the high endogenous RNase activity in plants, which we mitigate by incorporating RNase inhibitors and modifying the RNA fragmentation process. Coupling these innovations with the iCLIP2 improved library preparation workflow significantly improves sequencing efficiency and reproducibility. Apart from the lab protocol, we introduce a detailed pipeline to identify Arabidopsis RBP binding sites, from sequencing read alignment to precise mapping of cross-linking events. This is facilitated by the R/Bioconductor package BindingSiteFinder, designed to pinpoint reproducible RBP binding sites across the transcriptome. The iCLIP2 protocol for plants, which can be executed in 5 days by researchers experienced in RNA handling and molecular biology, provides a powerful tool to systematically uncover the binding landscapes of RBPs in plants, paving the way for deeper insights into their regulatory roles in plant biology.

3.16 Generative machine learning to model cellular perturbations

Mo Lotfollahi (Wellcome Sanger Institute – Cambridge, GB)

The field of cellular biology has long sought to understand the intricate mechanisms that govern cellular responses to various perturbations, be they chemical, physical, or biological. Traditional experimental approaches, while invaluable, often face limitations in scalability and throughput, especially when exploring the vast combinatorial space of potential cellular states. Enter generative machine learning that has shown exceptional promise in modeling complex biological systems. This talk will highlight recent successes, address the challenges and limitations of current models, and discuss the future direction of this exciting interdisciplinary field. Through examples of practical applications, we will illustrate the transformative potential of generative ML in advancing our understanding of cellular perturbations and in shaping the future of biomedical research.

3.17 lncRNA-mediated transcription regulation: new links in human evolution

Yael Mandel-Gutfreund (Technion – Haifa, IL)

Joint work of Amir Argoetti, Dor Shalev, Galia Polyak, Noa Shima, Hadas Biran, Tamar Lahav, Tamar Hashimshony, Yael Mandel-Gutfreund

Main reference Amir Argoetti, Dor Shalev, Galia Polyak, Noa Shima, Hadas Biran, Tamar Lahav, Tamar Hashimshony, Yael Mandel-Gutfreund: "LncRNA NORAD Modulates STAT3/STAT1 Balance and Innate Immune Responses in Human Cells via Interaction with STAT3". Nature Communications 16 (1).

URL https://doi.org/10.1038/s41467-025-55822-0

Long non-coding RNAs (lncRNAs) are pivotal regulators of cellular processes via binding to RNA-binding proteins (RBPs). By employing the RNA interactome capture technology coupled with enhanced CLIP (eCLIP) we identified an interaction between the lncRNA NORAD and an unconventional RBPs, the immune related transcription factor STAT3, in embryonic and differentiated human cells. NORAD knockdown experiments reveal its

role in facilitating STAT3 nuclear localization and suppressing antiviral gene activation. In NORAD-deficient cells, STAT3 remains cytoplasmic, allowing the interferon stimulating protein STAT1 to enhance antiviral signaling. Analysis of RNA expression data from in-vitro experiments and clinical samples demonstrates NORAD depletion upon viral infection, supporting its role in antiviral activity. Additionally, evolutionary conservation analysis suggests this regulatory role is restricted to humans, potentially owing to the introduction of an Alu element in hominoids. These findings suggest NORAD functions as a modulator of STAT3-mediated immune suppression, adding to the understanding of lncRNAs in immune regulation and their evolutionary adaptation in host defense mechanisms.

3.18 Unlocking the Potential of AI for High-Throughput Immunotherapy Drug Discovery Through RNA Splicing

Miguel Ángel Manzanares Serrano (Envisagenics – New York, US)

Joint work of Miguel Ángel Manzanares Serrano, Devon Trahan, Gayatri Arun, Hasan Zumrut, Kendall Anderson, Martin Akerman, Nicole Williams, Sakshi Gera, Sanjana Shah, Shaleigh Stanton, Taylor Floyd

Main reference Jie Wu, Martin Akerman, Shuying Sun, W. Richard McCombie, Adrian R. Krainer, Michael Q. Zhang: "SpliceTrap: a method to quantify alternative splicing under single cellular conditions", Bioinform., Vol. 27(21), pp. 3010–3016, 2011.
URL https://doi.org/10.1093/BIOINFORMATICS/BTR508
Main reference Martin Akerman, Oliver I. Fregoso, Shipra Das, Cristian Ruse, Mads A. Jensen, Darryl J. Pappin, Michael Q. Zhang, Adrian R. Krainer: "Differential connectivity of splicing activators and repressors to the human spliceosome. Genome Biol 16:1–18, 2015
URL https://doi.org/10.1186/s13059-015-0682-5
Main reference SungHee Park, Mattia Brugiolo, Martin Akerman, Shipra Das, Laura Urbanski, Adam Geier, Anil K. Kesarwani, Martin Fan, Nathan Leclair, Kuan-Ting Lin, Leo Hu, Ian Hua, Joshy George, Senthil K. Muthuswamy, Adrian R. Krainer, Olga Anczuków: "Differential Functions of Splicing Factors in Mammary Transformation and Breast Cancer Metastasis", Cell Reports, Vol. 29(9), pp. 2672–2688.e7, 2019.
URL https://doi.org/10.1016/j.celrep.2019.10.110
Main reference Alyssa D Fronk, Miguel A Manzanares, Paulina Zheng, Adam Geier, Kendall Anderson, Vanessa Frederick, Shaleigh Smith, Sakshi Gera, Robin Munch, Mahati Are, Priyanka Dhingra, Gayatri Arun, Martin Akerman: "Development and validation of an AI/ML platform for the discovery of splice-switching oligonucleotide targets",Mol Syst Biol. 2024 Jun;20(6):676-701, Epub 2024 Apr 25. PMID: 38664594; PMCID: PMC11148135.
URL https://doi.org/10.1101/2022.10.14.512313

Alternative splicing (AS), a process that generates multiple mRNA isoforms from a single gene, plays a crucial role in cancer biology. Through RNA-seq analysis, we can identify AS-derived neoepitopes from aberrantly spliced transcripts that hold promise as targets for immuno-oncology therapies (IO). In this context, Envisagenics introduces SpliceCore, a drug discovery platform capable of analyzing thousands of RNA-seq samples with a unique reference transcriptome encompassing over 14 million splicing events, which provides an unparalleled search space for IO target discovery. Upon splicing quantification, SpliceCore uses AI algorithms to further prioritize novel neoepitopes based on antibody accessibility, tumor specificity and patient prevalence. Here we present a case study for a novel isoform expressed in Non-Small Cell Lung Cancer (NSCLC). This novel drug target emerges as a promising antibody-drug conjugate (ADC) for NSCLC. We validated its expression using mass spectrometry, western blot, and novel binders developed in-house. Membrane localization and internalization assays confirm the target isoform suitability for ADC modality. In summary, SpliceCore exemplifies the application of AI to unlocking novel therapeutic targets. By focusing on neoepitopes, we pave the way for personalized and effective cancer therapies.

3.19 Towards universal models or protein-RNA interactions

Annalisa Marsico (Helmholtz Zentrum München, DE)

RNA Binding Proteins (RBPs) are central to RNA processes. Predicting their binding sites on RNA is crucial for understanding and post-transcriptional regulation in health and disease. I introduced RBPNet, a novel deep learning model that predicts CLIP-seq cross-link count distributions directly from RNA sequence at single nucleotide resolution, eliminating the need for peak calling. Trined on up to a million regions with sufficient read coverage, RBPNet generalises effectively across diverse CLIP datasets. It corrects CLIP biases by modelling raw signals as a mixture of protein-specific and background noise, therefore detecting RBP-specific binding motifs over noise, and it uses Integrated Gradients (IG) for variant impact scoring via in silicon mutagenesis. In the second part, I presented panRBPNet, an extension of RBPNet for multi-task learning, capable of modelling multiple RBP profiles simultaneously. panRBPNet improves RBP predictions over single-task models, refines spicing mutation scores and captures meaningful RBP-RNA interactome representations from raw CLIP data. We demonstrate its foundational nature in several downstream applications, such as splicing junction prediction, prediction of translational efficiency and classification of coding and non-coding RNA species.

3.20 shRNAI: a deep neural network for the design of highly potent shRNAs

Jin-Wu Nam (Hanyang University – Seoul, KR)

Joint work of Jin-Wu Nam, Seokju Park

In this seminar, I introduced our BIGLab's research on lncRNA and bifunctional RNA, highlighting our recent advancements in RNA-based therapeutic strategies. A key focus was the development of shRNAI, a convolutional neural network model designed to predict highly potent miRNA-mimicking short hairpin RNA (shRNAmir) guide RNAs (gRNAs).

shRNAI demonstrated superior performance in identifying effective gRNAs, even when trained solely on sequence data, outperforming previous algorithms. By integrating additional features related to shRNAmir processability and target site context, we developed an improved model, shRNAI+, which exhibited further enhanced performance across both public datasets and our own experimental validations. Despite being initially trained on a CNNC motif-free pri-miR-30 backbone, shRNAI also showed improved efficacy when applied to CNNC motif-containing backbones.

During discussions with RNA and AI researchers, an important point was raised regarding the comparison of shRNAI-designed siRNAs with FDA-approved RNAi drugs that incorporate RNA modifications. It was noted that these modifications could influence the efficacy and comparability of newly designed siRNAs, posing challenges in direct benchmarking.

Overall, our study provides a robust computational framework for designing potent shRNAmir gRNAs, offering a versatile tool for RNA interference therapeutics. The insights gained from this seminar and discussions will be invaluable in refining our models and furthering our understanding of RNA-based drug development.

3.21 Unraveling the impact of long non-coding RNA processing and chromatin dissociation dynamics in gene regulation

Evgenia Ntini (FORTH – Heraklion, GR)

Joint work of Angelos Kozonakis, Stefan Budach, Aannalisa Marsico, Evgenia Ntini
Main reference Evgenia Ntini, Stefan Budach, Ulf A. Vang Ørom, Annalisa Marsico: "Genome-wide measurement of RNA dissociation from chromatin classifies transcripts by their dynamics and reveals rapid dissociation of enhancer lncRNAs", Cell Systems, Vol. 14(10), pp. 906–922.e6, 2023.
URL https://doi.org/10.1016/j.cels.2023.09.005

Long non-coding RNAs (lncRNAs) are a diverse class of molecules ($>$ 200 nt) with key roles in genome organization and gene expression regulation. While most lncRNAs undergo processing similar to mRNAs –including capping, splicing to varying degrees, and polyadenylation– they are often transcribed from enhancer-like regions or anchor points of chromosomal loops. In general, lncRNAs are enriched in the nucleus, where they exhibit variable residence times at chromatin and undergo different degrees of co- and post-transcriptional splicing, which may influence their subcellular and subnuclear localization dynamics. An open question is whether some lncRNAs remain "tethered" to chromatin post-transcriptionally through interactions with specific RNA-binding proteins (RBPs). Understanding lncRNA processing is essential for elucidating their roles in gene regulation and nuclear organization. To address this, we recently established a new method that combines pulse-chase metabolic labeling with chromatin fractionation and deep sequencing of nascent RNA from both chromatin-associated and chromatin-released RNA fractions. This approach enables us to profile chromatin dissociation dynamics of nascent RNA transcripts and measure their co- and post-transcriptional splicing (Ntini et al., 2023). We further applied machine-learning approaches to model chromatin dissociation kinetics of both lncRNAs and mRNAs, uncovering the contribution of specific mechanistic features, including splicing, and the involvement of RBPs. I then discussed unpublished data on the role of RNA processing of a chromatin-associated lncRNA, PVT1. PVT1 exhibits very slow chromatin dissociation dynamics in MCF-7 breast cancer cells, appearing chromatin-tethered at steady state. It is upregulated in several cancer types, presumably accumulating near its transcription locus. To examine how PVT1 RNA processing influences gene regulation, we extracted splicing efficiencies from all PVT1 splicing sites across hundreds of control and breast cancer samples (TCGA) and ran in silico genome-wide screens using linear regression and generalized regularized linear regression models to predict gene expression. Among genes predicted above a certain coefficient of determination (R^2) threshold, we found a significant enrichment of putative miR-200 target genes. This suggests that PVT1 splicing, which is deregulated in cancer, may be functionally linked to the regulation of miR-200 target genes. Finally, I outlined ongoing experimental approaches to validate this hypothesis.

3.22 Uncertainty quantification for deep learning based prediction of RNA binding protein targets

Uwe Ohler (Max-Delbrück-Centrum – Berlin, DE)

Joint work of Uwe Ohler, Sepideh Saran, Svetlana Lebedeva, Frederick Korbel
Main reference Sepideh Saran, Svetlana Lebedeva, Antje Hirsekorn, Uwe Ohler: "Cell-type specific prediction of RNA stability from RNA-protein interactions", bioRxiv, Cold Spring Harbor Laboratory, 2024.
URL https://doi.org/10.1101/2024.11.19.624283

RNA binding proteins (RBPs) are important regulators of RNA-based gene regulation. At our 2019 Dagstuhl meeting, I presented a multimodal, multitask convolutional neural network approach to identify target sites of RBPs. The model took RNA sequence and annotation as input and predicted binding of 60 proteins in a supervised setting.

Here, I talked about our efforts to explore approaches for uncertainty quantification (ensembles, Bayesian neural networks, Monte Carlo dropout). Our goal is to extend the use of the RBP CNN in the context of realistic scenarios where data is affected by class imbalance, mislabeling, spurious patterns. For our domain, BNNs are empirically found as superior to alternatives, and making use of uncertainty scores can indeed lead to more accurate predictions. However, BNNs are costly, and this may impact the applicability.

Additionally, I presented work to use the existing CNNs to predict molecular phenotypes via a transfer of the pretrained RBP models to downstream tasks defined by experimental measurements of RNA stability and translation. This showcased limits imposed by available data, specifically to generalize between different data sets, but also led to valuable insights for future work.

Finally, we are generating synthetic (ie real, randomized) sequences to be able to obtain training data for RNA translation on a massively parallel scale. We anticipate these data to be useful for training of generative sequence models, which are inherently limited by the number of e.g. human genes, with the aim to use ML to propose RNAs with desired properties.

3.23 Predicting Protein-RNA Binding Based on In Vivo and In Vitro Data

Yaron Orenstein (Bar-Ilan University – Ramat-Gan, IL)

Joint work of Yaron Orenstein, Ori Feldman, Hagar Chen

This talk focused on the development of computational methods to predict protein-RNA binding sites, leveraging both in vivo (eCLIP) and in vitro (HTR-SELEX) data. The research addressed the limitations of experimental methods, such as their resource-intensive nature and cell-type specificity. A novel computational approach, CellRBP, was introduced, which incorporates RNA structural features, abundance, and annotations to predict binding sites across cell types. The model demonstrated superior performance compared to state-of-the-art methods like PrismNet. Additionally, the utility of k-mer scores from in vitro data for RNA-binding modeling was discussed, highlighting their simplicity and potential for generalizability. The challenges in bridging in vitro and in vivo predictions remain an open question.

3.24 Models of maternal mRNA degradation in embryos reveal principles of its scaling by developmental pace

Michal Rabani (The Hebrew University of Jerusalem, IL)

Joint work of Michal Rabani, Mazal Tawil, Dina Alcalay, Pnina Greenberg, Shirel Har-Sheffer, Lior Fishman

Regulation of mRNA turnover is an integral part of gene expression programs, but its underlying kinetics and regulatory rules remain elusive. During early embryogenesis, thousands of pre-loaded maternal transcripts are post-transcriptionally regulated within metazoan embryos, making it an ideal system to investigate the kinetics and regulation of mRNA stability across organisms. We have recently developed QUANTA, a computational strategy to distinguish transcriptionally silent genes and analyze their regulation. QUANTA uses kinetic models to compare total and polyA+ expression patterns, and dissect quantitative rates of mRNA polyadenylation and degradation. Using QUANTA, we analyze the massive degradation of pre-loaded maternal transcripts within metazoan embryos, and compare its regulation between zebrafish, frog, mouse and human embryos. We confirm shared principles of maternal mRNA degradation in all species, and show that degradation onset and rate are proportional to the developmental pace of species. We pinpoint potential regulatory signals in 3'UTRs of species, revealing signals to accelerate maternal degradation in fast-developing species, and signals to delay degradation and enhance mRNA stability in slow-developing species. We implement a massively parallel reporter assay that is compatible with QUANTA in zebrafish embryos, and validate QUANTA's kinetic and sequence element predictions. We show that reporter degradation rates also scale by adjusting developmental pace of zebrafish embryos with external temperature, and quantify the scaling effect of 3'UTR regulatory signals. Targeted genome editing of zebrafish ARE-binding proteins, a key group of regulators of mRNA destabilization and translation, significantly reduces their expression levels in embryos and lead to developmental delays relative to wild-type embryos across temperatures. Degradation rates scales with mutants' developmental pace. Finally, we show evidence that activity of some sequence signals is affected by changes in developmental pace. These results reveal the scaling of mRNA degradation kinetics by developmental pace, and demonstrate how its sequence-based rules adjust its scaling with other biological processes.

3.25 Comprehensive RNA binding protein analyses and deep learning uncover genetic constraints and disease associations in protein-RNA interfaces

Katie Rothamel (University of California – San Diego, US)

Joint work of Hsuan-lin Her, Brian A Yee, Xintao Wei, Evan A Boyle, Katherine L Rothamel, Steven M Blue, Sara Olson, Lijun Zhan, Jasmine R Mueller, Samuel Park, Grady G Nguyen, Jack T Naritomi, Adam Klie, Stefan Aigner, Brenton R. Graveley, Gene W Yeo

RNA-binding proteins (RBPs) orchestrate post-transcriptional processes, including splicing, cleavage and polyadenylation, and translation. We present an updated RBP resource, integrating data from 338 RBPs profiled by 414 eCLIP experiments and complementary knockdown (KD) RNA-seq datasets, comprehensively characterizing RNA elements within

human K562 and HepG2 cells. To decipher the "syntax" of RBP binding, we trained deep learning models using eCLIP data. These models were employed in a linear mixed-effects framework to link RBP binding and expression levels to splicing events, extracting position-dependent activity of splicing regulation to identify new splice regulators such as FAM120A and PPP1R10. The models scored genetic variants and quantified constraints on RBP binding sites, revealing that splicing enhancers and silencers exhibit opposing selective constraints in their binding sites and that strengthening mutations in ELAVL1 and HNRNPC sites are observed at a higher frequency than expected. Finally, we used the models to prioritize disease variants, revealing both known and novel RBP-related mechanisms of pathogenesis.

3.26 Improving sequence-to-function modeling to understand the gene-regulatory code

Alexander Sasse (Universität Heidelberg, DE)

Joint work of Alexander Sasse, Nuria A Chandra, Anna Spiro, Xinming Tu, Sara Mostafavi
Main reference Alexander Sasse, Bernard Ng, Anna Spiro, Shinya Tasaki, David Bennett, Christopher Gaiteri, Philip De Jager, Maria Chikina, Sara Mostafavi: "Benchmarking of deep neural networks for predicting personal gene expression from DNA sequence highlights shortcomings", Nature Genetics, Vol. 55, pp. 1–5, 2023.
URL https://doi.org/10.1038/s41588-023-01524-6
Main reference Nuria Alina Chandra, Yan Hu, Jason D. Buenrostro, Sara Mostafavi, Alexander Sasse: "Refining the cis-regulatory grammar learned by sequence-to-activity models by increasing model resolution", bioRxiv, Cold Spring Harbor Laboratory, 2025.
URL https://doi.org/10.1101/2025.01.24.634804

Most disease associated genetic variants are located within regulatory regions of the genome that affect gene expression and its regulation. Therefore, knowledge about the effect of variants on gene expression and its regulation is essential to understand disease susceptibility and develop personalized treatments. On the other hand, regulatory mechanisms are impossible to study experimentally on a personalized level due to their cell type specificity. Nevertheless, in the last two decades, massive amounts of gene expression and regulatory phenotypes have been measured on a genome-wide scale to explore all layers of gene expression regulation across species, individuals, cell types, and conditions. Multi-task Convolutional Neural Networks (CNN) have been applied to analyze how DNA sequence influences these measured molecular phenotypes, thereby successfully learning the sequence patterns that are recognized by regulatory factors to control these processes. This general understanding of the underlying processes theoretically enables them to determine the effect of any individual genomic variant in any cell type that the model was trained on. However, in our recent study, we found that while these models perform well on a variety of different benchmarks for predicting variant effects (Avsec et al. 2021), they fail to correctly determine the direction of the variant effect on gene expression across a set of different individuals (Sasse et al. 2023), the essential task to associate individual variants with phenotypes or disease. Attribution analysis revealed that these models tend to focus on strong eQTLs close to the TSS outside any notable motif grammar, suggesting a lack of knowledge. To address these shortcomings and increase the information that these models can learn from the available data, I am presenting four strategies: 1) Training on data with sequence variation. We present a modeling approach that directly contrasts sequence differences to predict allele-specific functional measurements from RNA-seq, ATAC-seq, and CHIP-seq 2) Training models on higher resolution. We present a modeling approach that models ATAC-seq at base-pair resolution. In addition to modeling

total accessibility to chromatin, it also models the distribution of Tn5 insertions, allowing it learn a more sensitive motif representation that distinguishes between different closely related cell types 3) Building more mechanistic models for RNA-seq. We present a model architecture that uses intronic and exonic reads to estimate RNA processing rates and predicts these from DNA and RNA sequences using a bio-physically motivated objective function. 4) Building multi-modal models that learn from multiple modalities. We present a model that uses a shared sequence representation for all data modalities but applies individual modality heads that account for the non-linear relationships between different modalities. We anticipate that future model architectures will integrate these strategies to enhance their capacity to decipher gene regulatory mechanisms from vast data collections.

3.27 Towards a comprehensive single-cell picture of RNA isoforms in mouse and human brain and their diseases – or – single-cell isoforms in time and space

Hagen Tilgner (Weill Cornell Medicine – New York, US)

Most mammalian genes encode multiple distinct RNA isoforms and the brain harbors especially diverse isoforms. Complex tissue includes diverse cell types, which employ distinct isoforms. To untangle cell-type specific brain isoform profiles, we developed the first single-cell long-read technology for »1,000 cells and fresh tissues (Single-cell isoform RNA sequencing – ScISOr-Seq [1]) as well as for frozen tissues (Single-nuclei isoform RNA sequencing – SnISOr-Seq [2]). To add spatial resolution, we developed Slide-isoform sequencing (Sl-ISO-Seq) [3]. Collectively, these long-read approaches reveal a striking difference between coordinated pairs of exons with in-between exons ("Distant coordinated exons") and without in-between exons ("Adjacent coordinated exons"): The former show strong enrichment for cell-type specific usage of exons, whereas the latter do not in mouse [1] and human brain [2]. Of note, coordinated TSS-exon pairs and exon-polyA-site pairs follow the same trend as distant coordinated exon pairs [2]. Simultaneously, autism-associated exons are among the most highly variably used exons across cell types [2]. Spatially barcoded isoform sequencing revealed that often region-specific isoform differences correlate with precise boundaries of brain structures (e.g., from the choroid plexus to the hippocampus). However, genes including Snap25 go against this trend, using a steady gradient of exon inclusion as one traverses the brain [3]. Moreover, choroid plexus epithelial cells show a dramatically distinct isoform profile, which originates from distinct exon and poly(A) site usage, but most strongly from distinct TSS usage [3]. For the NIH Brain Initiative, we have mapped single-cell isoform expression across development, brain regions and species. Neurotransmitter release and reuptake as well as synapse turnover genes harbor variability in the same cell type across anatomical regions but the same cell type traced across development shows more isoform variability than across adult anatomical regions. Moreover, most cell-type specific exons in adult mouse hippocampus behave similarly in human hippocampi. However, human brains have evolved additional cell-type specificity in splicing, suggesting gain-of-function isoforms [4]. Most recently, we have made advances in understanding the error sources of Pacific Biosciences and Oxford Nanopore long-read sequencing technologies [5] and have implemented highly accurate long-read interpretation software [6]. Finally, the concurrent measurement of

chromatin and splicing patterns in post-mortem human brain reveals that distinct chromatin-transcriptome coupling states can yield different splicing patterns and points to strong convergent dysregulation of both modalities in similar cell types in Alzheimer's disease [7].

References

1 Gupta*, Collier* et al, Nature Biotechnology, 2018
2 Hardwick*, Hu*, Joglekar* et al, Nature Biotechnology, 2022
3 Joglekar et al, Nature Communications, 2021
4 Joglekar et al, Nature Neuroscience, 2024
5 Mikheenko*, Prjibelski* et al, Genome Research, 2022
6 Prjibelski*, Mikheenko* et al, Nature Biotechnology, 2023
7 Hu*, Foord*, Hsu* et al, biorxiv, 2024

3.28 mRNA multivalency enables homeostatic co-regulation of condensation-prone proteins

Jernej Ule (UK Dementia Research Institute at King's – London, GB)

Joint work of Rupert Faraway, Neve Costello Heaven, Holly Digby, Klara Kuret Hodnik, Jure Rebselj, Oscar G. Wilkins, Anob M. Chakrabarti, Ira A. Iosub, Lea Knez, Stefan L. Ameres, Clemens Plaschka, Jernej Ule
Main reference Rupert Faraway, Neve Costello Heaven, Holly Digby, Oscar G. Wilkins, Anob M. Chakrabarti, Ira A. Iosub, Lea Knez, Stefan L. Ameres, Clemens Plaschka, Jernej Ule: "Mutual homeostasis of charged proteins", bioRxiv, Cold Spring Harbor Laboratory, 2023.
URL https://doi.org/10.1101/2023.08.21.554177

The concentration of proteins containing intrinsically disordered regions (IDRs) must be tightly controlled to maintain cellular homeostasis. However, mechanisms for collective control of these proteins, which tend to localise to membraneless condensates, are less understood compared to proteins at membrane-bound organelles5. Here we report "interstasis", a homeostatic mechanism that senses the concentration of co-condensing proteins and controls their gene expression through mRNA-mediated negative feedback. Interstasis relies on multivalent mRNA regions that encode IDRs, which are reinforced by conserved codon biases and recognised by specific RNA-binding proteins. TRA2 proteins are strongest binders of multivalent purine-rich regions that encode charged IDRs, including arginine-enriched mixed charge domains (R-MCDs). Accumulation of R-MCD proteins increases the cohesion of nuclear speckles, a protein-RNA condensate, which recruits TRA2 proteins that selectively retain the purine-rich mRNAs in the speckles. This decreases further synthesis of charged proteins that are most highly prone to phase separation and are encoded by bidirectionally dosage-sensitive genes. Cdc2-like kinase (CLK) activity controls the localisation of TRA2 proteins to speckles, thereby modulating the setpoint of interstasis. Thus, interstasis is a collective feedback loop that senses the accumulation of condensation-prone speckle proteins, and then sequesters mRNAs that encode these proteins to promote their mutual homeostasis.

3.29 Cis-acting and targetable gene regulation by long noncoding RNAs

Igor Ulitsky (Weizmann Institute – Rehovot, IL)

Joint work of Igor Ulitsky, Aviv Rom, Liliya Melamed, Caroline Jane Ross, Yoav Lubelsky, Rotem Ben-Tov Perry
Main reference Aviv Rom, Liliya Melamed, Noa Gil, Micah Goldrich, Rotem Kadir, Matan Golan, Inbal Biton, Rotem Perry, Igor Ulitsky: "Regulation of CHD2 expression by the Chaserr long noncoding RNA gene is essential for viability", Nature Communications, Vol. 10, p. 5092, 2019.
URL https://doi.org/10.1038/s41467-019-13075-8

It is now clear that many intergenic regions in eukaryotic genomes give rise to a range of processed and regulated transcripts that do not appear to code for functional proteins. A subset of these are long (>200 nt), capped, and polyadenylated RNAs transcribed by RNA polymerase II and collectively called long noncoding RNAs (lncRNAs). The recent estimates are that the human genome may have >50,000 distinct lncRNA-producing loci, many of which show tissue-specific activity and dysregulation in human disease, including cancer and neurodegeneration. Given the growing number of lncRNAs implicated in human disease or required for proper development, fundamental questions that need to be addressed are: Which lncRNAs are functional? How is functional information encoded in the lncRNA sequence? Is this information interpreted in the context of the mature or the nascent RNA? What are the identities and functional roles of specific sequence domains within lncRNA genes? These are challenging questions, primarily because of the substantial heterogeneity in mechanisms utilized by lncRNAs and the current paucity of lncRNAs with well-understood mechanisms. We are tackling these questions by combination of experimental methods with a focus on lncRNA functions in early cell fate decisions and computational methods focused on lncRNA evolution. I described our efforts to decode conserved combinations of short functional sequence elements in lncRNAs, with a particular focus on the Chaserr/Chd2 axis.

3.30 Capture me if you can: identification of long noncoding RNAs in vertebrate genomes

Barbara Uszczynska-Ratajczak (Polish Academy of Sciences – Poznan, PL)

Accurate and comprehensive gene annotations are essential for understanding how genome sequences encode biological functions. For the past two decades, the GENCODE consortium has provided reference annotations for the human and mouse genomes, serving as a cornerstone for the biomedical and genomics communities. However, key gene classes such as long non-coding RNAs (lncRNAs) remain incomplete and dispersed across multiple uncoordinated catalogs, hindering progress in the field. To address this challenge, GENCODE has launched its most extensive lncRNA annotation effort to date. This initiative is based on the manual annotation of full-length targeted long-read sequencing from matched embryonic and adult tissues in both human and mouse orthologous regions. As a result, 17,931 novel human genes (140,268 novel transcripts) and 22,784 novel mouse genes (136,169 novel transcripts) have been added to the GENCODE catalog–marking a twofold and sixfold increase in transcripts, respectively, and the largest expansion since the human genome was first sequenced. These newly annotated genes exhibit evolutionary constraints, possess well-formed promoter regions, and are linked to phenotype-associated genetic variants. Their inclusion significantly enhances

the functional interpretability of the human genome by helping to explain millions of previously unaccounted "orphan" omics measurements, including transcription start sites, chromatin modifications, and transcription factor binding sites. Furthermore, this targeted approach has dramatically improved human-mouse-zebrafish ortholog assignments, tripling the number of disease-associated human lncRNAs with mouse and zebrafish counterparts.

3.31 Deep learning for deep transcriptome data mining

Li Yang (Fudan University – Shanghai, CN)

Joint work of Zhi-Can Fu, Bao-Qing Gao, Guo-Wei Li, Fang Nan
Main reference Zhi-Can Fu, Bao-Qing Gao, Fang Nan, Xu-Kai Ma, Li Yang: "DEMINING: A deep learning model embedded framework to distinguish RNA editing from DNA mutations in RNA sequencing data", Genome Biology, Vol. 25, 2024.
URL https://doi.org/10.1186/s13059-024-03397-2
Main reference Guo-Wei Li, Fang Nan, Guo-Hua Yuan, Bin Tian, Li Yang: "SCAPTURE: a deep learning-embedded pipeline that captures polyadenylation information from 3' tag-based RNA-seq of single cells", Genome Biol, Vol. 22, 2021.
URL https://doi.org/10.1101/2021.03.17.435782

It is seemingly simple to decode the genetic information formed by the four nucleotide bases embedded in the human genome. However, the completion of the Human Genome Project (HGP) and the advancement of affordable high-throughput (or deep) sequencing technologies have revealed the intricate nature of gene expression at the whole-transcriptomic level, which makes functionally decoding the human transcriptome anything but simple. Over the last decade, my laboratory has dedicated to developing and applying new computational strategies together with novel deep sequencing technologies to study the complex transcriptomes. Recently, we further apply machine learning and deep learning models to identify hidden information embedded in the complex transcriptomic datasets. We develop a stepwise computational pipeline SCAPTURE to identify, evaluate, and quantify cleavage and polyadenylation sites (PASs) from popular 3' tag-based single cell RNA-seq datasets. We also report a DEMINING framework to directly detect expressed DNA mutations and RNA editing sites in canonical RNA-seq datasets. In this talk, I will present these home-brewed bioinformatic toolkits embedded with artificial intelligence models towards deciphering the complex human transcriptomes.

3.32 How RNA G4s and cooperative HNRNPH binding mediate switch-like splicing

Kathi Zarnack (Universität Würzburg, DE)

Joint work of Kerstin Tretow, Mario Keller, Mikhail Mesitov, Miona Ćorović, Mirko Brüggemann, Nadine Körtel, Nicolas Melchior, Anke Busch, Simon Braun, Nadja Hellmann, Heike Hänel, Susanne Strand, Stefan Legewie, Friederike Schmid, Kathi Zarnack, Julian König

We show the molecular mechanism how rG4 secondary RNA structures facilitate cooperativity in splicing regulation by the RNA-binding protein HNRNPH. By combining high-throughput in vivo and in vitro studies with theoretical modelling, we demonstrate how rG4s mediate cooperative HNRNPH binding to RNA, triggering the regulation of hundreds of exons. We

propose that rG4 unfolding by HNRNPH exposes multiple G-rich binding sites, thereby establishing indirect cooperativity, which is further amplified to achieve switch-like splicing. Additionally, we report on a predominant artifact in RNA sequencing data, resulting in the erroneous detection of unconventional splicing events, termed falsitrons.

3.33 Machine Learning for Modeling Gene Regulatory Networks from Single-Cell Sequencing Data

Jianyang Zeng (Westlake University – Hangzhou, CN)

Single-cell technologies enable the dynamic analyses of cell fate mapping. However, capturing the gene regulatory relationships and identifying the driver factors that control cell fate decisions are still challenging. In this talk, I will present our two recently proposed methods for inferring gene regulatory networks (GRNs) and identifying driver genes from high-throughput single-cell RNA sequencing (scRNA-seq) data. First, we propose DeepSEM, a deep generative model that can jointly infer GRNs and biologically meaningful representation of scRNA-seq data. In DeepSEM, we develop a neural network version of the structural equation model (SEM) to explicitly model the regulatory relationships among genes. Next, we present CEFCON, a network-based framework that first uses a graph neural network with attention mechanism to infer a cell-lineage-specific gene regulatory network from single-cell RNA-sequencing data, and then models cell fate dynamics through network control theory to identify driver regulators and the associated gene modules, revealing their critical biological processes related to cell states. Benchmark tests show that our methods achieved the state-of-the-art results and outperformed the baselines. In addition, we also demonstrated some application potentials of our methods using current available biological data. Overall, our proposed methods may provide useful tools to model gene regulatory networks from large-scale scRNA-seq data.

Participants

- Rolf Backofen
Universität Freiburg, DE
- Pavel Baranov
University College Cork, IE
- Mathieu Blanchette
McGill University – Montréal, CA
- Charlotte Capitanchik
The Francis Crick Institute – London, GB
- Christoph Dieterich
Universitätsklinikum Heidelberg, DE
- Florian Erhard
Universität Regensburg, DE
- Eduardo Eyras
Australian National University – Canberra, AU
- Julien Gagneur
TU München – Garching, DE
- Jonathan Göke
Genome Institute of Singapore, SG
- Marko Jovanovic
Columbia University, US
- Jan Philipp Junker
Max-Delbrück-Centrum – Berlin, DE
- Julian König
Institut für Molekulare Biologie – Mainz, DE
- Frederick Korbel
Max-Delbrück-Centrum – Berlin, DE
- Claudia Kutter
Karolinska Institute – Stockholm, SE
- Gioele La Manno
EPFL – Lausanne, CH
- Markus Landthaler
Max-Delbrück-Centrum – Berlin, DE
- Liana Lareau
University of California – Berkeley, US
- Martin Lewinski
Universität Bielefeld, DE
- Mo Lotfollahi
Wellcome Sanger Institute – Cambridge, GB
- Yael Mandel-Gutfreund
Technion – Haifa, IL
- Miguel Ángel Manzanares Serrano
Envisagenics – New York, US
- Annalisa Marsico
Helmholtz Zentrum München, DE
- Jin-Wu Nam
Hanyang University – Seoul, KR
- Evgenia Ntini
FORTH – Heraklion, GR
- Uwe Ohler
Max-Delbrück-Centrum – Berlin, DE
- Yaron Orenstein
Bar-Ilan University – Ramat-Gan, IL
- Michal Rabani
The Hebrew University of Jerusalem, IL
- Katie Rothamel
University of California – San Diego, US
- Alexander Sasse
Universität Heidelberg, DE
- Hagen Tilgner
Weill Cornell Medicine – New York, US
- Jernej Ule
UK Dementia Research Institute at King's – London, GB
- Igor Ulitsky
Weizmann Institute – Rehovot, IL
- Barbara Uszczynska-Ratajczak
Polish Academy of Sciences – Poznan, PL
- Li Yang
Fudan University – Shanghai, CN
- Kathi Zarnack
Universität Würzburg, DE
- Jianyang Zeng
Westlake University – Hangzhou, CN