Machine learning-assisted identification of factors affecting variability in multi-omics data

Dissertation

Zur Erlangung des Grades

Doktor der Naturwissenschaften

Am Fachbereich Biologie
Der Johannes Gutenberg-Universität Mainz

Sofya Lipnitskaya

geb. am 06.03.1994 in Wolgograd

Mainz, 2024

Dekan: Prof. Dr. Eckhard Thines

1. Berichterstatter:

2. Berichterstatter:

Tag der mündlichen Prüfung: 30.01.2025

Abstract

Recent advances in high-throughput technologies together with computational innovations have enabled the studying of biological systems at multiple levels, giving rise to integrative omics approaches. Multi-omics research refers to efforts that combine multiple omics datasets—including genes, transcripts, and proteins—obtained from the same samples to improve our understanding of biological processes. Over the past decades, omics technologies have led to new insights on complex molecular mechanisms underlying abnormal phenotypes and diseases, thus revolutionizing biomedical and biological research. This has resulted in the generation of a large volume of biological data, including that available in open-access sources. Nonetheless, comprehensive analysis of such data is not trivial and is particularly hampered by high dimensionality, noisy nature of the data, as well as the lack of standardized data analysis methods and pipelines. Therefore, it is necessary to focus on the integration of the omics data in the context of phenotypes and conditions of interest, which motivated the current research.

This thesis investigates factors affecting biological and technical variability in the context of transcriptomics studies by applying Machine Learning (ML) and Integrative Data Analysis (IDA). In particular, the thesis proposes design and implementation of: (I) a bioinformatics pipeline (FAVSeq) for identification of key effectors for variation in multimodal RNA Sequencing (RNA-Seq) profiles from matched bulk and single-cell experiments and (II) an analysis tool for ML- and IDA-based studying of alternative splicing regulome

(regulAS) comprising large-scale RNA-Seq from cancer and healthy patients from public omics data sources.

Findings and tools presented in this thesis provide a basis for further experimental investigations of identified factors, as well as subsequent improvements at the level of RNA-Seq data preparation along with downstream analysis that allow to facilitate the fundamental research and biomedical applications based on RNA sequencing technologies.

Zusammenfassung

Neueste Fortschritte im Bereich der Hochdurchsatztechnologien in Verbindung mit Innovationen im Bereich der Datenverarbeitung haben die Untersuchung biologischer Systeme auf mehreren Ebenen ermöglicht, was zu integrativen Omics-Ansätzen geführt hat. Multi-Omics-Forschung bezieht sich auf Bemühungen, die mehrere omics-Datensätze—einschließlich Genen, Transkripten und Proteinen—aus denselben Proben kombinieren, um unser Verständnis biologischer Prozesse zu verbessern. In den letzten Jahrzehnten haben omics-Technologien zu neuen Erkenntnissen über komplexe molekulare Mechanismen geführt, die anormalen Phänotypen und Krankheiten zugrunde liegen, und damit die biomedizinische und biologische Forschung revolutioniert. Dies hat zur Erzeugung einer großen Menge biologischer Daten geführt, die auch in frei zugänglichen Quellen verfügbar sind. Die umfassende Analyse solcher Daten ist jedoch nicht trivial und wird insbesondere durch die hohe Dimensionalität und die Verrauschung der Daten sowie durch das Fehlen standardisierter Datenanalysemethoden und Pipelines erschwert. Daher ist es notwendig, sich auf die Integration von Omics-Daten im Kontext

von Phänotypen und Bedingungen von Interesse zu konzentrieren, was die aktuelle Forschung motiviert.

In dieser Dissertation werden die Faktoren untersucht, die die biologische und technische Variabilität im Rahmen von Transkriptomikstudien beeinflussen, indem maschinelles Lernen (ML) und integrative Datenanalyse (IDA) angewandt werden. Insbesondere schlägt die Dissertation das Design und die Implementierung vor: (I) einer Bioinformatik-Pipeline (FAVSeq) zur Identifizierung von Schlüsselfaktoren für die Variation in multimodalen RNA-Sequenzierungsprofilen (RNA-Seq) aus aufeinander abgestimmten Bulk- und Einzelzellexperimenten und (II) eines Analysetools zur ML- und IDA-basierten Untersuchung des alternativen Spleißreguloms (regulAS), das groß angelegte RNA-Seq-Daten von Krebs- und gesunden Patienten aus öffentlichen Omics-Datenquellen umfasst.

Die in dieser Dissertation vorgestellten Ergebnisse und Werkzeuge bilden die Grundlage für weitere experimentelle Untersuchungen der ermittelten Faktoren sowie für nachfolgende Verbesserungen auf der Ebene der RNA-Seq-Datenvorbereitung und der nachgelagerten Analyse, die es ermöglichen, die Grundlagenforschung und biomedizinische Anwendungen auf der Grundlage von RNA-Sequenzierungstechnologien zu erleichtern.

Contents

1 Introduction

> *Let our advance worrying become advance thinking and planning.*
>
> — **Winston Churchill**

1.1 Multiomics data in biomedical sciences

The foundation of systems biology lies in the view of biological phenomena as systems, which are made up of a large number of complex internal and external components that interact at different levels to determine the functional properties of the systems. For instance, tumor development can be affected by a combination of genomic alterations (e.g., mutational events), which may be associated with dysregulated gene expression profiles possibly leading to changes in protein abundance and structure (e.g., distinctive protein isoforms). Thus, these are interdependent processes that cannot be viewed separately, which necessitates the use of integrative (combined) approaches in biomedical research and gave rise to omics studies.

This chapter illustrates common omics approaches and discusses underlying biological and technical challenges in data acquisition and integration. Moreover, it provides an overview of the available multi-omics repositories and toolsets, acknowledging current challenges and prospects in large-scale omics data analysis. Here, the particular focus is on transcriptomics technologies

enabling the characterization of gene expression profiles of different levels and modalities, which resulted in many discoveries in a wide range of scientific areas and experimental conditions, thus, opening up new opportunities for biomedical research.

1.1.1 Approaches for multimodal and omics research

The multitude of levels of interactions found in biological systems can be measured using a variety of omics technologies, such as genomics, transcriptomics, proteomics, and other high-throughput methods. Single-omics levels represent different data modalities and can be viewed as complementary to the central dogma of molecular biology (Figure 1.1). The latter defines a unidirectional causal flow from genome to transcriptome, proteome, metabolome and phenotype. Accordingly, the information content of an organism is recorded in the DNA of its genome (genomics) and expressed through transcription (transcriptomics), as mRNA intermediates, which, in turn, can direct the synthesis of proteins during the further translation and thus, framing proteome (proteomics) and physiological traits (phenotype). Thus, integration of multi-omics technologies in the context of system biology, aims primarily at the better understanding (e.g., through identification) of causal genes, mRNAs, proteins, and other pathways for phenotypes.

While being informative in some aspects, each of these views could lack another ones that are important for particular use cases. For instance, transcriptomics is easy and comprehensive, but what typically matters in the cell is protein expression. At the same time, proteomics complements transcriptomics to learn post-transcriptional regulation of gene expression, but often has lower coverage. As transcriptional changes often arise from mutations under pathological conditions - these necessitate the corresponding

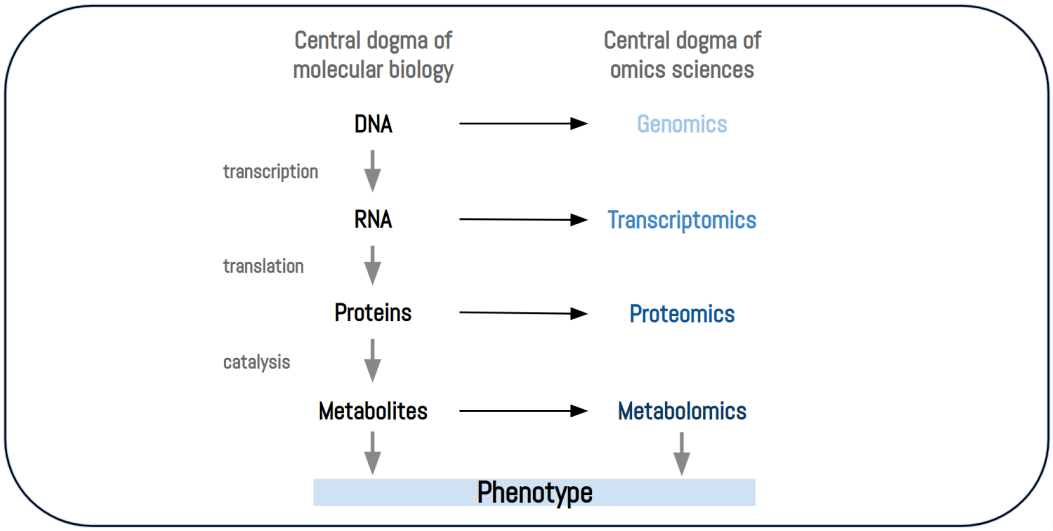

Fig. 1.1: Central dogma of molecular biology and its corresponding omics disciplines.

mutation-related genomic alteration data contained in genomics, etc. This clearly shows that simultaneous measurements at multiple levels are required to better understating the mechanisms underlying human diseases, such as cancer [7].

In recent years, rapid development of high-throughput technologies has led to their wider use and, consequently, to the exponential growth of the amount of available data of diverse nature. This has turned out into a variety of studies that provided the research community with a number of independent datasets, many of those showing different views (e.g., genomics and transcriptomics) of the same problem. Since having two or more identical studies of the same problem is not common, the published datasets usually complement each other in some aspects and have particular overlaps that make it possible to judge how consistent they are (Figure 1.2). For instance, there are two widely used techniques in the field of transcriptomics: microarray, which quantify a set of predetermined sequences, and RNA sequencing (RNA-Seq), which uses high-throughput sequencing to record all transcripts.

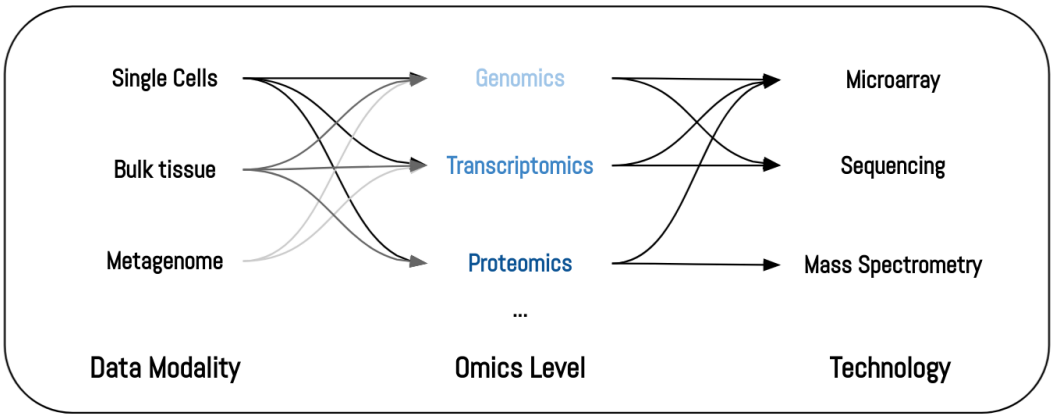

Fig. 1.2: Complementary multi-omics data of different modalities (single cells, bulk tissue, heterogeneous populations of species) generated using array-/sequencing-/mass spectrometry-based technologies. Genomics and transcriptomics are usually based on Next-generation sequencing (NGS) technologies, whereas proteomics is driven by mass-spectrometric ones. Here, the general overview of multi-omics data acquisition approaches is shown, and different modalities, levels and technologies are not necessarily overlapped. For instance, while sequencing technologies can be used for profiling single-cell genomics and transcriptomics data, the former (genomics data) can be also measured by microarray-based comparative hybridization at the single-cell resolution.

With the further development and improvement of detection technologies and analytical methods of high-throughput technologies, another branch of omics research originated, namely multimodal omics. The latter refers to the integrative study of molecular activity of data (e.g., gene expression data in transcriptomics) provided at different resolutions (e.g., bulk and single-cell modalities) and measured by same or different omics technologies (e.g., RNA sequencing) [97, 60, 136].

In this thesis, the particular focus is on multimodal transcriptomics based on Next-generation sequencing (NGS) for measuring gene expression data, which is elaborated in more detail in the following chapters.

Tab. 1.1: Comparison between bulk and single-cell RNA-Seq technologies.

	bulk RNA-Seq	scRNA-Seq
Goal	• provide the average gene expression level • analyze differences between samples and conditions	• provide the distribution of expression levels for each gene • analyze differences between cell types and states
Experimental protocols	• RNA is extracted from all cells • Reverse transcription converts RNA to cDNA to facilitates ligation of sequencing adaptors • Amplification • Mapping via identifying origin and position of each read	• RNA is extracted from isolated cells (sometimes more than 10^6) • Reverse transcription and amplification are similar to bulk, but scRNA-Seq also includes tagging of individual transcripts by UMI and barcodes to identify mRNAs for each gene in a cell • Quasi-Mapping via identifying of origin for each read

1.1.2 Transcriptomic profiling using RNA-Seq technologies

The information regarding transcriptomes has significantly contributed to discovering molecular mechanisms and identifying therapeutic targets in different research settings [112, 2, 107, 3]. Given the intermediate role of RNA between genome and proteome (Figure 1.1) the quantification of gene expression using a high-throughput sequencing assay known as RNA-Seq has since become a standard tool in the biomedical and life sciences. This subchapter presents a more detailed overview of two transcriptomics technologies based on NGS, which are commonly considered, namely bulk and single-cell RNA-Seq (Figure 1.2). In the processing pipelines of RNA sequencing experiments RNA transcripts are reverse-transcribed into complementary DNA (cDNA), the cDNA second strand is synthesized and then amplified, before being sequenced. While both protocols are similar in general, scRNA-Seq also include isolation of single cells as well as reverse transcribing and amplifying either the whole-transcript length or the prime tagged ends (e.g., 3' end) of each mRNA (Table 1.1).

Thus, the main differences between two NGS technologies is that scRNA-Seq allows determining distinct expression profiles for each cell within the specific environment, while the bulk provides the gene expression averaged across all cells.

As the technology improved, the amount of data generated by each transcriptome experiment increased. As a result, data analysis techniques were constantly being adapted to analyze ever larger amounts of data more accurately and efficiently. Transcriptome databases grew and became increasingly useful as more and more transcriptomes were collected and transmitted by researchers. It would have been nearly impossible to interpret the information contained in the transcriptome without the context of experiments performed in previous and/or other studies, which opens the potential for integrative transcriptomics studies.

1.1.3 Multiomics repositories for studying biological systems

In a multiomic workflow, RNA-Seq data resulting from joint research efforts allows the examination of thousands of gene expression profiles, while the ability of individual practitioners and laboratories is usually limited to the amount of data to generate and analyze experimentally at different omics levels and modalities. The integration of multi-omics datasets for the same samples has been used for a better understanding of the specific features of biological systems that cannot be uncovered by studying only a single data modality or level. There have been a number of attempts at creating comprehensive multi-omics profiles for healthy and cancer patients across different pathological conditions. Here, some acknowledged publicly

Tab. 1.2: Multi-omics open-source repositories and incorporated data types.

Repository	Multiomics dataset	Description	URL
TCGA	RNA-Seq, DNA-Seq, miRNA-Seq, SNV, CNV, DNA methylation, RPPA	33 cancer types	`https://www.cancer.gov/tcga`
GTEx	Gene expression and regulation data	Up to 53 non-diseased tissues	`https://gtexportal.org`
ICGC	Whole genome sequencing, genomic variations	22 cancer types (data release 28)	`https://icgc.org`
CCLE	Gene expression, copy number, and sequencing data	729 cell lines of 23 cancer types	`https://portals.broadinstitute.org/ccle`
TARGET	Gene expression, miRNA expression, copy number, and sequencing data	24 pediatric cancer types	`https://ocg.cancer.gov/programs/target`

Abbreviations: TCGA – The Cancer Genome Atlas; GTEx – The Genotype-Tissue Expression; ICGC – International Cancer Genomics Consortium; CCLE – Cancer Cell Line Encyclopedia; SNV – single-nucleotide variant; CNV – copy number variation; RPPA – reverse phase protein array.

available resources that aid multi-omics data integration initiatives are listed in Table 1.2.

The Cancer Genome Atlas (TCGA) is the one of the largest collections of multi-omics datasets including genomics, transcriptomics, proteomics, and clinical data on about 20,000 patients available for more than 33 different types of cancers. This initiative represents a landmark source for multi-omics methods benchmarking which is widely used by the research community aiming at new discoveries in cancer research and treatment. The Genotype-Tissue Expression (GTEx) is another comprehensive source for studying tissue-specific gene expression from more than 50 non-diseased tissues majorly for molecular assays including RNA-Seq data. Therapeutically Applicable Research to Generate Effective Treatments (TARGET), a program similar to TCGA, designed for functional assessment of molecular and genomic changes affecting pediatric cancers of 24 different types based on gene expression, miRNA expression, sequencing and clinical data.

Apart from these dedicated databases for multi-omics, International Cancer Genomics Consortium (ICGC), Cancer Cell Line Encyclopedia (CCLE) as well as National Center for Biotechnology Information Gene Expression Om-

nibus (NCBI GEO; https://www.ncbi.nlm.nih.gov/geo/) and European Nucleotide Archive (ENA; https://www.ebi.ac.uk/ena/) provide a comprehensive collection of multi-omics data, including genomics and transcriptomics generated from multiple platforms and arrays.

UCSC Xena (https://xena.ucsc.edu/) is a web-based data integration and visualization tool aimed for multi-omics (gene, exon and protein expression data modalities, somatic mutations, and clinical data) comparison (e.g., using statistical analysis) of more than 1500 datasets (e.g., TCGA, GTEx, CCLE, ICGC, TARGET) across 50 cancer types [92]. Moving forward with downstream analysis, the Ensembl BioMart (https://www.ensembl.org/) relates to an open source data management system and data mining tool allowing to retrieve functional annotation and map genomic identifiers (e.g., transcript length, GC-content, gene-associated genomic loci) for the integration of omics datasets.

Therefore, domain experts can benefit from such open-access sources comprising data of different assays, platforms and omics levels, that can be later integrated and analyzed using various frameworks and tools. One of such examples relates to the associative analysis of alternative splicing based on transcriptomics data related to prospective regulatory candidates and splicing patterns, which will be discussed in section 4.2 in greater detail.

1.1.4 Challenges and prospects in integrative omics analysis

Effective large-scale omics research requires reliable and robust data integration and analysis design in order to ensure flawless execution of the computational workflow and, more importantly, to gain correct insights

on biological phenomena. Integrative data analysis (IDA) of multi-omics is an emerging interdisciplinary field of study that investigates strategies, algorithms and methodologies combining multiple datasets from different sources and applies computational and systems biology approaches in solving complex biomedical problems. IDA-based approaches enable researchers not only to validate previous studies but also to identify new patterns in an intra- as well as an inter-dataset fashion, which was indicated in many omics studies covering a variety of domains, including cancer and splicing research [120, 133, 41].

In addition, from the reproducibility point of view, a variety of study-specific implementations makes it often not feasible to re-evaluate experiments by other researchers. Here, the use of a unified experimental framework defined in terms of IDA ensures comparability and thus validity of results. Therefore, the only reliable study design can lead to flawless execution of the computational multi-omics workflow. There have been several comprehensive guidelines for integrative omics analysis proposed over the past few years to address this challenge [66, 110, 4, 28]. Figure 1.3 demonstrates a generalized IDA workflow for multi-/omics data supporting the Findable, Accessible, Interoperable, and Reusable (FAIR) research [66]. In order to provide a reliable basis for manageable and reproducible ML-aided research, one has to consider using a standard approach implemented as a framework using common programming languages (e.g., Python). Such a solution provides the user with an interface that establishes a clear and unambiguous contract connecting all the steps of the pipeline. This allows reducing the amount of code to write, since many stages that do not differ from experiment to experiment are already implemented and their validity is tested beforehand. Thus, an end-user (domain-expert) could focus on the task-specific challenges and leave all boilerplate implementation details.

Handling omics datasets relates to a number of challenges that, in general, belong to several major groups: data wrangling, heterogeneity and dimensionality [66]. Specifically, when it comes to merging datasets—for instance, gene expression profiles from different studies,—one has to deal with multiple transformation and mapping steps for each of the datasets exclusively as well as for all datasets simultaneously [85]. Here, data wrangling refers to data alteration actions as a general pre-processing step, with main difficulties stemming from the lack of clearly defined mappings across data from different sources, which is quite common [97]. The aforementioned mappings cover the whole data acquisition process, starting from wet-lab experimental protocols to publishing and version management of datasets. Another source of data integration challenges comes from the diversity of the datasets, namely data heterogeneity [2]. Indeed, when solving different tasks, results are produced usually on their own (which are unique very often) formats that cannot be merged directly (i.e. without any pre-processing steps). Such multistep merging operation usually involves data normalization and aims to harmonize representation of the data provided by highly diverse sources (e.g., multi-omics integrative databases). This unified representation serves as a basis for subsequent downstream tasks through the use of feature selection and/or extraction.

Finally, irrespective of the fact that the search for overlaps between datasets might be highly non-trivial, the scale of the problem might impose challenges with respect to computational feasibility. Given that asymptotic complexity of many data processing algorithms is quadratic or even worse, one can very often face a problem of so-called curse of dimensionality [8]. Here, at some point, the amount of data (features and/or samples) to process becomes so large that it is not possible anymore to perform the analysis in a realistic amount of time [79]. In order to overcome the challenge, one relies on techniques aiming to reduce computational costs. For instance, one

can reduce the number of features through either selection (e.g., Recursive Feature Elimination) or extraction (e.g., Principal Component Analysis).

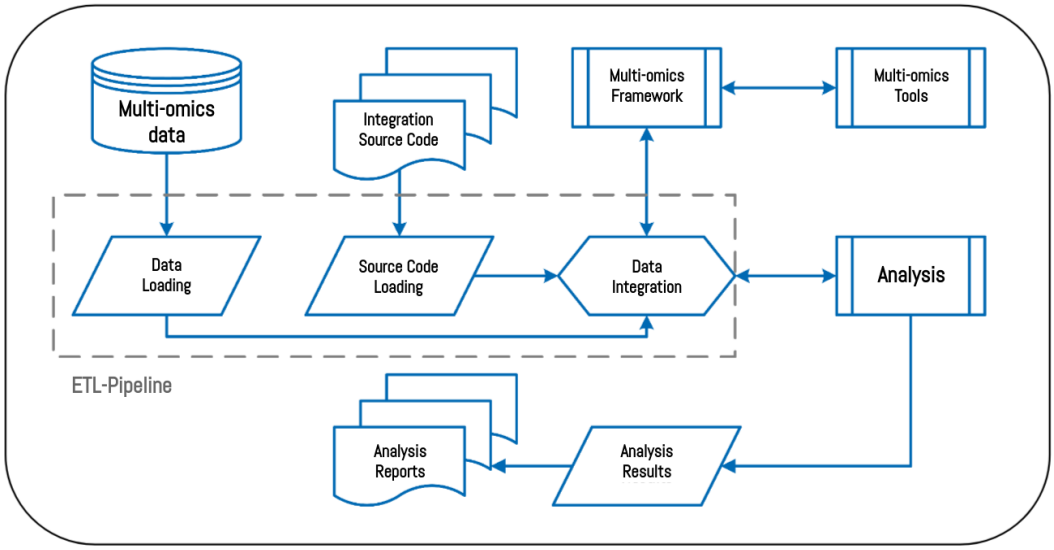

Fig. 1.3: Integrative data analysis (IDA) workflow for multi-/omics data. The joint analysis of diverse data sources consists of a sequence of standard steps. First, multi-omics data should be loaded from the data source(s) (upper left) into the analysis pipeline. Simultaneously, the pipeline requires data-specific algorithms (source code, upper left-center) that shape data integration and analysis. The data integration part relies on multi-omics tools and frameworks (upper right), which could be re-used by other studies. After the data integration is done (middle, right-center), the ETL-part is considered to be completed. The subsequent data analysis utilizes the integral data and provides the results that are to be summarized as analysis reports (bottom).

The data integration involves usually steps that are consistent from task to task. Before performing the actual data analysis (e.g., machine learning-based one), one acquires data, transforms it into a useful structure and loads it for the subsequent analysis (e.g., into a regression model) as an input. These data preparation steps could be described formally in terms of an ETL-pipeline (Extract, Transform and Load). The subsequent data analysis part (e.g., feature selection pipeline) consists of several common steps as well. The latter are easier to automate due to stronger requirements regarding the input and output data formats. The final part, namely analysis results,

provides output summaries (e.g., as tables and visual reports) that can be then used by domain experts for further interpretation and downstream studies.

The main challenge imposed by the data preparation defined in terms of the ETL-pipeline is the lack of clearly defined steps in its beginning. As noticed previously, the input format of an analysis pipeline (which is an output of the ETL-pipeline, namely Load-part) is usually defined quite strictly. Hovewer, adherence to formal format requirements, data schemes, along with data processing best practices and community standards is a subject for further improvement among open-access bioinformatics tools. This leads to particular challenges when it comes to automation of the aforementioned steps. Moreover, since every study is unique in some aspects, many researchers prefer to solve the data pre-/processing task in an ad-hoc fashion. Additionally, data preparation of big omics imposes particular challenges caused by the scale of the problem [110, 98]. All of that motivates the development of novel standardized computational tools (software and frameworks) that support domain-experts in performing effective and reproducible research.

As the later focus of this study will be on machine learning (ML)-assisted omics analysis, that itself can comprise of several common steps (e.g., model training and testing, feature selection), the programming framework should also include support for the aforementioned steps, similarly to the ETL-pipeline. Such an integration allows performing an IDA-assisted omics research in an end-to-end way, making it easier to find optimal solutions and to reveal underlying relationships in complex data.

1.2 Machine learning and its applications in biology

Machine learning approaches have been extensively used to discover biological patterns and relationships in multi-omics data across a variety of knowledge domains and research areas [102, 105, 89]. Nevertheless, integrative analysis is often challenging due to the data complexity originated from the underlying biological and technical factors encoded in the input data collected at different omics levels across different technologies and platforms. Such challenges in data representation and analysis can be also subjected to the lack of optimal end-to-end algorithms as well as due to the limited amount of standardized extendable multi-omics workflows (e.g., pipelines and software) devised for analysis of data of different modalities given a broad range of study goals.

Historically, the majority of machine learning algorithms were derived from statistical methods aiming to test hypotheses built upon data. Exponential growth of computational capacities enabled wider spread of such techniques in application to a variety of problems. The statistical nature of many ML-algorithms had turned out into the split of the methods into two major classes, namely supervised and unsupervised.

(1) Supervised task or building a predictive model based input (independent variables or the features) and desired output (dependent variable or the target) data can be further subdivided depending on the target variable to be predicted by a learning algorithm. Thus, classification and regression are aimed at prediction of discrete outcomes (e.g., disease or heathy status) or continuous quantities (e.g., survival rate), respectively. For instance, one of the classification tasks applied in biomedical research would be to utilize

gene expression profiles and clinical data of patients to help prognosing diseases, such as cancer. In this case, a non-/response outcome for different cancer-drug combinations based on pre-treatment biopsies can serve as a target variable for a binary classification algorithm in order to predict cancer prognosis and treatment response [48].

(2) Unsupervised task or building a predictive model based only on input data, on the other hand, includes dimensionality reduction (e.g., exploratory analysis of tissue dissimilarity in multidimensional data) and clustering analysis (e.g., cancer subtype identification). While the latter group of methods relied on the internal properties of the input data aiming to reveal complex intra-data relationships, here, the focus is on the supervised approaches.

The aims of this section are to provide an overview of the current state of the field, inform on difference between machine learning approaches and non- vs. linear algorithms, overview feature selection methods and discuss some important biomedical discoveries including those that were done based on model-based feature selection analysis of transcriptomics data.

1.2.1 Supervised learning algorithms for complex data analysis

Supervised learning implies the presence of additional information about the target. In case of regression problems, the target is provided in the form of continuous values of a dependent variable. Estimation of relationships between dependent variables and one or more independent variables (or predictors) using statistical methods refers to the regression analysis. Such relationships can be described by first- (linear) or higher-order (non-linear) equations. The most widely known example of the regression problem is a

linear regression task, which is formulated under the assumption that relationships between predictors and the corresponding target can be described in a linear fashion.

While the formulation of the regression problem provides the assumption on the order of relationships between inputs and outputs, there is another key component that allows to solve the problem, namely the objective (or loss) function. Through the minimization of the loss function, one aims to build a model that, on the one hand-side, best fits to the distribution of the training data and, on the other hand-side, is able to generalize well to the new data [33]. Finding the balance between two goals that are contradicting one another is also known as bias-variance tradeoff [65]. On the whole, bias and variance errors exist in a trade-off state and relate to the concepts of overfitting and underfitting, which will be described in more detail in the following paragraphs. One usually uses mean squared error (MSE) as an objective function for the linear regression task. Specifically, solving the linear regression (LR) problem using MSE objective function is known as ordinary least squares (OLS), where parameters that predict y from X are estimated – to some extent, given stochastic error vector – by a system of linear equations (Equation 1.1) those approximate solutions can be found by estimating directly (Equation 1.2).

$$y = X\beta + \epsilon \tag{1.1}$$

$$\hat{\beta} = (X^T X)^{-1} X^T y \tag{1.2}$$

However, linear relationships in the data are not always the case [38]. Wrong choice of an assumption on the order of relationships may lead to underfitting

(high bias and low variance) of the model built upon the data to analyze. In such cases, the use of higher-order (non-linear) models can be beneficial from the minimum prediction error point of view.

One widely used non-linear descendant of the linear models' family of methods is a multilayer perceptron (MLP), which is referring to the class of feed-forward artificial neural networks (ANN) [103]. From the internal structure point of view, an MLP can be considered as two or more linear models stacked sequentially, with a non-linear activation function in-between (Figure 1.4). Here, the key difference is the activation function, which allows for approximating non-linear dependencies within the data. Thus, introduction of non-linear operators turns the whole model into a non-linear one.

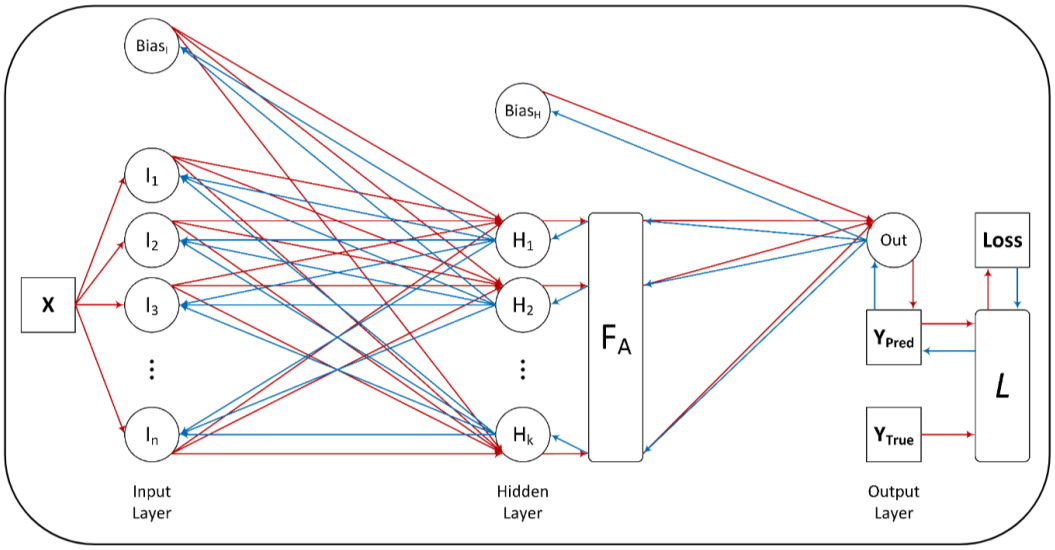

Fig. 1.4: Overview of MLP. Circles depict neurons, squares are vectors or scalars, rounded rectangles denote functions. Here, X represents an input vector, I_i – input neurons, H_j – hidden neurons, F_A – activation function, and L – loss function. Data flows are colored according to their direction: red – forward step, blue – backpropagation step. An arrow denotes multiplication, with circles being related to sum operation. Value of the loss function measures the model's prediction error, and the weight update signal is proportional to the gradient of loss with respect to the weight. Feature importance scores can be computed by accumulating gradients of the loss with respect to the input values.

More specifically, an MLP consists of an input, an output and one or more (thus, multilayer) hidden structure units called layers. Each layer consists of a number of sum-and-multiply units combined with an activation function. Those units are called artificial neurons, and the complexity of the model is proportional to their number. While the activation after the input and output layers is usually an identity function (in other words, there is no activation), a hidden layer uses a non-linear activation, such as logistic function (sigmoid), hyperbolic tangent (tanh), rectified linear unit (ReLU) shown in Figure 1.5.

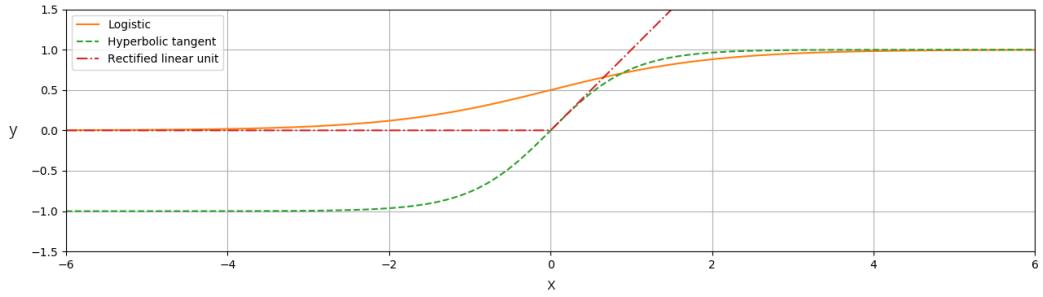

Fig. 1.5: Non-linear activation functions. Logistic function (orange solid) and hyperbolic tangent (green dashed) map input values to the bounded range ([0, 1] and [-1, 1], respectively) and have continuous first derivatives. In contrast, the output domain of the rectified linear unit (red dash-dotted) is not bounded from above ([0, +]) and its derivative has a discontinuity point at zero.

With the number of trainable parameters, the complexity of the model grows simultaneously, as well as its performance. However, if the model is way excessively complex, this can lead to so-called overfitting, i.e., when an algorithm gives accurate predictions for training data, but performs poorly for the test (unseen) one. The latter, on a high level, relates to poor generalization ability of the trained model, which is caused by the increased sensitivity to the noise presented in data (high variance) and by the reduced systematic error of the model (low bias). Thus, an overfitted model tends to "memorize" training data points instead of building an appropriate approximation of an underlying distribution (i.e., generalization ability is being affected). Such an issue can be prevented by either controlling the complexity (i.e., number

of parameters to learn) or using regularization terms. Regularization is a technique introducing auxiliary loss term(s) describing the internal state of the model and penalizing, for instance, large values of trainable parameters. Commonly, one applies L_1 or L_2 regularization (also known as Lasso or Ridge, respectively) to the vector of estimated coefficients. While the L_1 regularization relies on the $||\beta||_1$ norm, the latter accounts squared coefficients by the L_2 norm $||\beta||_2^2$.

Optimal values of a model's hyper-parameters - which include but are not restricted to the aforementioned regularization coefficients and number of parameters to learn - should be considered as unique for every new task. Thus, these values are to be tuned in order to get the best model performance, given available data. The hyper-parameter tuning is done based on the model's performance on the validation subset of the data, which provides an unbiased evaluation of a model fit on the training dataset. Unlike LR, fitting an MLP to the training data requires an iterative (numerical) approach instead of the (analytical) direct one. Specifically, it consists of forward and backward steps that repeat until the model converges, i.e. the value of its objective stops to improve. The forward step implies computation of the model's outputs based on the input. In contrast, the backward pass provides numerical estimation of true values of a multidimensional derivative (gradient) of the objective (loss) function with respect to model's parameters. The gradient values are inputs for a specific optimization algorithm—such as stochastic gradient descent (SGD), adaptive moment estimation (Adam) [63],—which adjusts the parameter values proportionally to them [99, 63]. The possibility to trace gradient estimations from the loss to the input values provides an implicit support to score importance of the model's input dimensions (features).

Applied to the feature importance assessment in a supervised setting, the following algorithms, MLP and LR, will be particularly examined in the

current work. Using both non- and linear types of ML models enables to disclose the nature of relationships between predictor and target variables, as well as to validate the choice of most important subsets of features obtained through the use of different methods.

1.2.2 Reducing data complexity through feature selection

As noted before, exponential growth of computational capacities turned into wide use of machine learning algorithms to solve various tasks. For the aforementioned models (LR and MLP) the complexity during model fitting is approximately $O(C^2N)$. Here C is the number of input features and N refers to the number of training samples. Quadratic $(O(N^2))$ and higher-order polynomial computational complexity depending on the number of input features is also known as curse of dimensionality.

The growth of an input space's number of dimensions leads to the fast growth of time required to perform computations. Given that the number of features dominates over the amount of samples, one can consider reducing the computational costs of model training through the selection of a subset of most important features. Despite the fact that there are ways to partially overcome such an issue by parallelizing calculations, one should argue that: 1) it is not always possible due to properties of a specific algorithm; 2) in general, it would be preferable to have a faster solution. Feature selection can be also considered as an optimization problem because features are qualitative variables. Moreover, from practical applications, it allows the identification of the most important input variables.

One can apply different strategies to reduce the number of input features that form three major groups: filter, wrapper and embedded methods [44, 83]. Filter methods apply an external procedure to assess the importance of features. In contrast, wrapper methods utilize an ML-model that is to be used later to measure the influence of features on its performance. Finally, embedded methods are those that perform feature scoring using the model of interest's internal mechanisms (Figure 1.6).

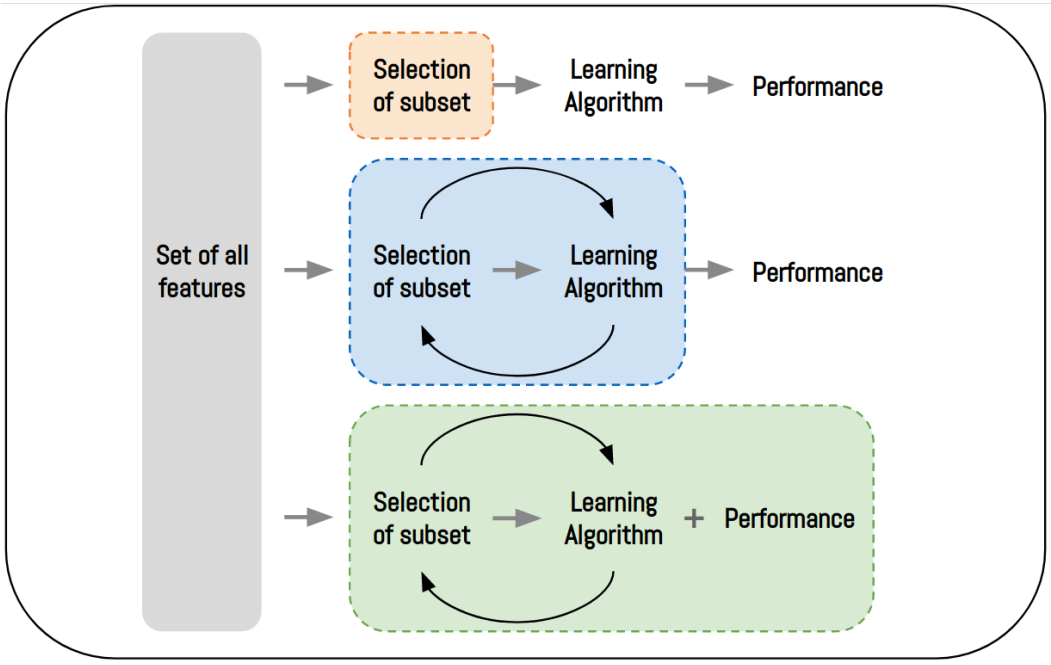

Fig. 1.6: Overview of basic approaches for feature selection. One can see that the involvement of the actual ML-algorithm into the selection process increases from filter to embedded methods. While the filter methods (orange; for instance, Pearson's correlation coefficient) ignore the model fitting step and the wrapper (blue; for instance, Recursive feature elimination) rely on specific ML-algorithms, the embedded methods (green; for instance, L1 regularization) combine fitting and selection implicitly, thus increasing computational "payload".

Although such grouping clearly distinguishes approaches to feature selection, one should highlight that, depending on the specific task, their combinations are possible, and so there could be no universal approach that suits the best in every situation. Here are more specific examples provided. For the regression task, filter methods include ones that are based on correlations (e.g., Pearson's correlation coefficient) or on mutual information. Correlation-based feature filtering implies calculation of correlation coefficients between each feature and target variable, thus measuring the strength of the linear relationship. Since Pearson's correlation coefficient (PCC) lies in the range from -1 to 1, the most important features have an absolute PCC value of 1 (high absolute PCC - high feature importance). Mutual information-based

feature filtering measures both linear and nonlinear dependence between features and is done by computing information gain for a feature and a target variable [121]. Mutual information close to 0 bits suggests nearly zero dependency between two random variables, thus making such low-ranked features first candidates for removal.

Unlike filter methods, wrapper ones apply a model of interest internally in order to perform feature importance scoring. Recursive feature elimination (RFE) belongs to the family of wrapper methods and performs feature selection sequentially, by eliminating non-informative features in several steps. Thus, at each step, features are left out one-by-one from the model, the performance is evaluated and then, the least important is eliminated [42]. In general, RFE is a model-agnostic framework, with the only requirement to provide feature scoring by the model on every step of the selection algorithm. Thus, RFE considers keeping only a top-ranked feature after every turn, which results in a subset of most important input features in the end.

While filter and wrapper methods imply the use of some external procedure in order to select the most relevant features, embedded methods rely on internal structures and algorithms of models of interest. In particular, the aforementioned LR and MLP allow identifying features that influence predictions of a target variable the most. For LR, feature ranking can be done directly (same as fitting) by using vector of coefficients that are multiplied by the input features (Equation 1.1) also known as zero-order scoring method [104]. Unlike LR, MLP-based assessment of feature importance relies on first-order method, which is based on accumulation of gradient values (Figure 1.6) during the model training [104, 71]. Specifically, for the input vector X, the importance of the i^{th} feature X_i can be calculated as a sum over N_{Epochs} epochs of absolute values of partial derivatives of the loss $\frac{dL_k}{dX_i}$

with respect to X_i, where k is the index of the current epoch, as shown in Equation 1.3.

$$Imp(X_i) = \sum_{k=1}^{N_{epochs}} \left| \frac{dL_k}{dX_i} \right| \tag{1.3}$$

As earlier discussed, quadratic complexity of many machine learning approaches is constraining their use without any limits, with respect to computational costs, even despite recent advances in the development of hardware. From that point, embedded methods (as the MLP-based one described earlier) are highly beneficial, since they perform feature ranking (and selection) implicitly during the model fitting [83]. In contrast, both filter and wrapper methods imply using additional procedures to evaluate feature importance score, which may lead to the valuable increase of computational time, thus reducing the actual computational “payload", given finite time and cost budget. While there were several methods to reduce the complexity of wrapper-based feature selection proposed [125, 95], the later focus will be on embedded methods as they are considered as more efficient and computationally less complicated than wrapper methods, while retaining a similar performance.

1.2.3 Reconstruction of missing data using imputation methods

While integrating omics data, one of the most frequent issues relates to the presence of missing and/or corrupted values. Depending on their nature, the values may be not present for various reasons. In case of scRNA-Seq data, the samples can be incomplete due to, for instance, low coverage or incorrect downstream processing. While it is rare to have fully corrupted data, the

amount of ones that lack at least one sample becomes substantial, and it is not unusual to have more than a half of such partially correct measurements [81]. Since the number of corrupted values may be substantial over the number of correct ones (e.g., as it appears to be the case for scRNA-Seq data), the exclusion of all but full data could lead to the introduction of various biases, which would likely affect the downstream analysis [109]. In order to handle such data incompleteness, one can apply a missing data replacement technique namely imputation. Imputation refers to a group of methods allowing substituting corrupted values with more plausible ones retrieved based on other available information in the data. These methods vary from very simplistic (e.g., constant substitution) to advanced (e.g., regression-based). As the former approaches provide rough approximations for incomplete samples, the quality of the data imputed using them may be severely affected.

Moving from simple to more sophisticated techniques, one can use, for instance, substitution by constant, by mean value, or apply ML-based approaches. If one decides in favor of a constant-substitution strategy (e.g., zero-based), it should be noted that such an imputation may introduce biases into the data. The reason for this is the shift of the underlying distribution towards the constant used for the actual substitution. At some degree, this kind of bias could be attenuated by adjusting the constant value to the population average. This leads to another imputation strategy namely substitution by mean value. Here, the overall distribution mean (first central moment) will not be affected; however, other moments of the distribution (e.g., variance) may change significantly.

In order to overcome such a limitation, one can apply advanced approaches such as ML-based ones. In this can, the imputation problem can be turned into the regression task, so the missing value is treated as target, while other

omics measurements serve as predictors. Simple linear regression model is able to produce quite accurate substitution values; however, its performance may suffer from the missing values among predictor variables. On the other hand, unlike the described regression models, k-Nearest Neighbors (k-NN) model imputes the missing value based on the subset of k most similar correct samples (neighbors) rather than on the whole dataset (or its statistics). In particular, the imputed value is the average of the values present in the neighbors. Such an approach allows restricting the search space for the substitution, thus reducing uncertainty and improving data quality.

As described above, there exist a variety of imputation strategies that differ in speed and accuracy. Depending on specific requirements, one can decide which method is the most suitable, given the problem at hand. Recent studies favour towards ML-based approaches among other methods [57]. The ability to follow the underlying distribution truncates possibly introduced biases, thus improving results of the subsequent analysis of biological data.

1.2.4 Biomedical discoveries based on machine learning

As we can notice, the total number of papers on data integration and ML-based analysis of multi-omics data has significantly increased over the past years. Figure 1.7 shows the number of publications indexed on the Web of Science website (Clarivate Analytics, 2020) with different key topics. As became evident, integrative and ML-based analysis of multi-omics data show the significant impact on the scientific community. In the last 5 years, there have been more publications on the topic of multi-omics integration and machine learning-based analysis, which have gained popularity in biomedical sciences and implied multi-omics analysis based on data from open repositories and platforms.

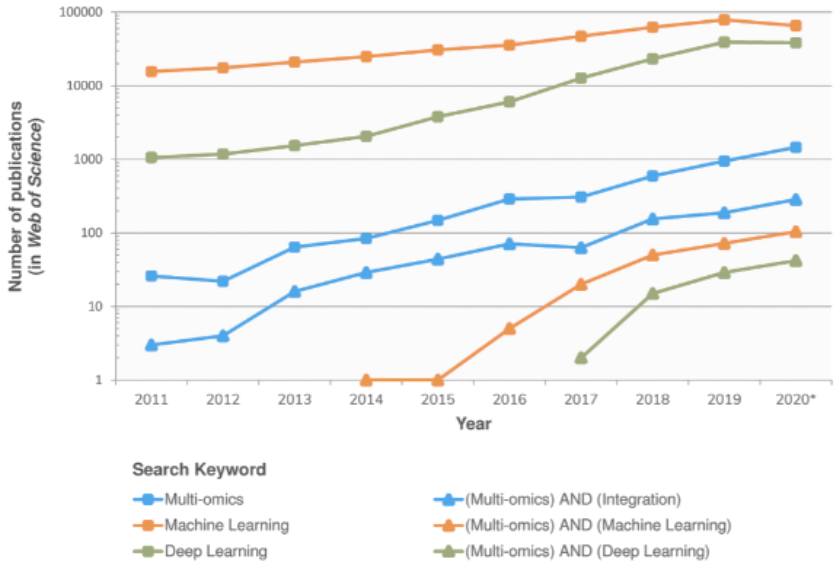

Fig. 1.7: Research contributions in multi-/omics, machine learning, and intergrative data analysis based on publications. Adapted from Reel et al. 2021 [102].

Furthermore, in practice, biological data is usually represented in tabular form (features vs. samples), which makes it useful to apply machine learning methods. While classical statistical approaches (e.g., PCC) are helpful for assessing particular properties of the data, they are unable to reveal underlying dependencies between elements. Unlike statistical methods, ML-based ones provide useful tools to disclose complex relationships in the data in addition to descriptive statistics.

While machine learning is a broad term, in recent years, a class of techniques called deep learning (DL) was set apart from other ML-methods to distinguish representation learning approaches based on artificial (usually convolutional) neural networks [69]. Here, the representation learning term describes a set of algorithms enabling an ML-model to automatically extract relevant features from the raw input data [9], thus disclosing correspondences between inputs and the matching labels. Although DL has been widely applied in other related fields, interest in multi-omics analysis is still more limited. Moreover,

the DL-based analysis of large-scale datasets is challenging to perform as it requires specialized high-throughput infrastructure (e.g., storage, GPUs). In comparison to deep learning based approaches, machine learning models are also easier to interpret by design, which is not the case for deep neural networks that are considered rather as "black boxes".

The application of machine learning- and integrative-based approaches in multi-omics provides crucial insights in understanding the mechanisms underlying biological processes and molecular functions. As evident from the increasing number of studies, many computational tools and methods have been developed over the past decade to address various problems in multi-omics analysis [102]. For instance, in the area of transcriptomics it was achieved by applying a model-based feature selection embedded approach for identification of cancer-relevant genes [83].

Similarly, data assessment using unsupervised learning methods was done to identify tumor and normal cells as well as stromal subtypes of human squamous cell carcinoma [60]. Other studies associated with application of model-based imputation methods were used for profiles and genes expression data [62]. Some other discoveries associated with ML-based analysis of omics data applied to splicing and transcriptomics research [76, 21, 127] as also discussed in following chapters.

1.3 Gene expression variability in transcriptomics data

This section presents a more detailed overview of computational challenges encountered while analyzing bulk and single-cell RNA-Seq data, including

the effect of potential confounding factors (i.e., so-called dropouts) on variability between gene expression measurements which is frequently observed between these technologies. While the earlier chapter related to omics technologies primary addresses library preparation differences between bulk and single-cell RNA-Seq experiments, the current one highlights the differences in computational pipelines applied for gene expression data, discusses computational challenges and opportunities to be considered for integrative analysis of RNA-Seq data, including multimodal transcriptomics.

1.3.1 RNA-Seq perspectives and issues to be addressed

As earlier discussed in subsection 1.1.2, scRNA-Seq technologies have greatly revolutionized transcriptomics research providing high-resolution insights into cellular heterogeneity and cell-type specific expression, especially in rare cell types (e.g., in disease states) as shown in Table 1.3. While traditional RNA-Seq methods were not primarily designed for precise characterization of cell type composition, single-cell technologies remain noisy and time-/costly compared to the bulk. The high-dimensionality and sparsity of scRNA-Seq data as well as the presence of missing values (dropouts) which may interfere with subsequent analysis and lead to data misinterpretation [21, 55]. scRNA-Seq also provides substantial amounts of low-quality data from cells that are destroyed, dead or mixed with others. Accordingly, the quality control step is crucial to identify and filter out these cells.

On the other hand, while having its own advantages and limitations, the bulk technology offers a robust and cost-effective (e.g., see Experimental aspects in Table 1.3) approach widely applied for detecting population-level associations (e.g., differential expression, gene regulation). Moreover, many bulk RNA-Seq studies performed in recent years have resulted in large-scale

Tab. 1.3: Challenges and opportunities in bulk and single-cell RNA sequencing.

	bulk RNA-Seq	scRNA-Seq
Experimental aspects		
	• Minimal initial mRNA constrain	• Low mRNA capture efficiency
	• Lower resolution	• Dropout amplification
	• Preserved and frozen samples (e.g., tumor biopsies)	• Batch effect
		• More costly
Computational challenges		
	• Insufficient for studying heterogeneous systems (early embryos, brain tissue)	• Missing data
		• High sparsity and dimensionality
	• Deconvolution required to estimate cell type composition	• QC required to identify low-quality cells
Computational analysis		
	• Differential expression analysis	• Cell subpopulation identification
	• Identification of biomarkers	• Heterogeneity analysis
	• Detection of aberrant splicing events	• Trajectory inference

datasets that are available in open-access repositories, such as TCGA and GEO (Table 1.2). For instance, given the extensive amount of available RNA-Seq data, estimation of cell-type composition in bulk data, commonly referred to as deconvolution, can be used to dissect cell-type specificity hidden at the population level with the reference to single-cell data [124]. Another possibility for the extended analysis beyond a single modality include investigations of factors affecting gene expression variability in RNA-Seq experiments, which is elaborated in the following subsection 1.3.2.

1.3.2 Advances in multimodal transcriptomics for analysis of RNA-Seq variability

There have been different protocols and frameworks published to date [20, 29, 52], which makes it difficult for new practitioners to appreciate all of the steps required to properly perform the downstream computational analysis of RNA-Seq data taking into account experimental and computational aspects of RNA-Seq technologies [24, 36]. To address these challenges considerable attention has been devoted to comparing the results obtained using different sequencing libraries and platforms, thus, opening opportunities for multimodal RNA-Seq research [74, 21, 80, 18]. In particular, multimodal RNA-Seq implies the extended analysis beyond a single data modality which can be performed based on IDA of multiple RNA-Seq datasets, such as bulk and single-cell RNA-Seq (Figure 1.8).

The joint analysis of data provided using different RNA-Seq technologies was shown to be beneficial in different applications, such as cell type decomposition [59, 30], tissue heterogeneity estimation [34], gene expression variability estimation [24, 18]. Therefore, the combined use of single-cell and bulk RNA sequencing for a specific tissue or condition contributes to the improved accuracy and precision of the final analysis (Table 1 from Ziegenhain et al. 2017).

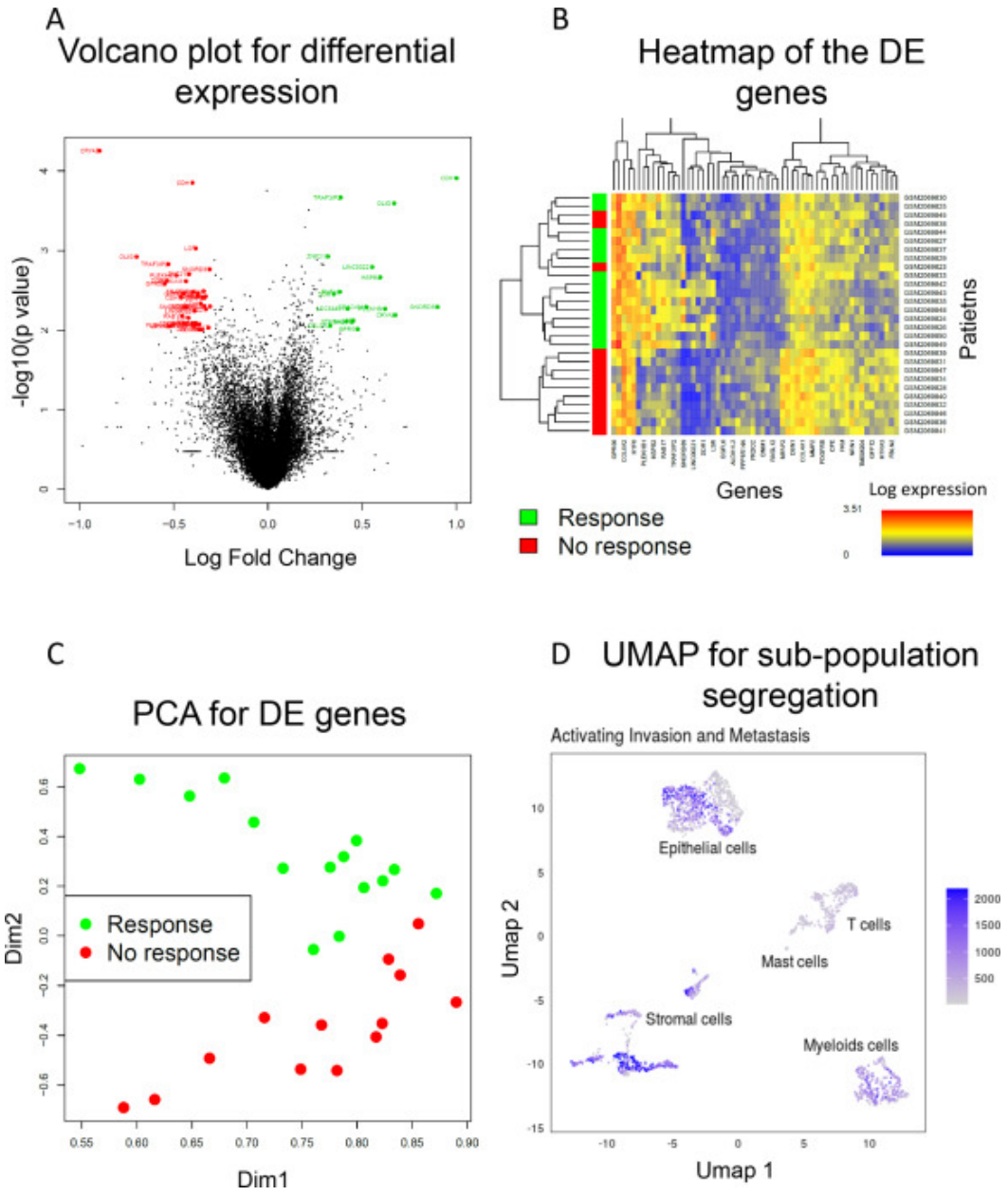

Fig. 1.8: Common computational approaches for analyzing multimodal RNA-Seq data provided by bulk and single-cell technologies, including differential expression (DE) analysis using hierarchical clustering and dimensionality reduction methods to discover changes in gene expression profiles across conditions and experimental groups. Here, PCA - Principal Component Analysis [53], UMAP - Uniform Manifold Approximation and Projection [78]. Adapted from Kuksin et al. 2021 [67].

As discussed earlier, RNA sequencing data differ appreciably. Specifically, compared to the bulk, scRNA-Seq data is markedly sparse (missing or zero-level measurements), which may be subjected to either biological reasons, where gene is not expressed at the time of isolation prior sequencing, or technical factors, where gene is expressed, but is not detected due to limitations related experimental protocols (here and later as dropouts). In addition, the ratio of such genes with missing expression significantly varies among individual cells, and it remains unclear whether these differences are due to technical rather than biological variations. There is much room for confounding factors contributing to the variability, including batch effect, capture and amplification efficiency, dilution of cell libraries or sequencing depth [50, 46]. As factors affecting technical and biological variability in RNA-Seq data have been of high interest in recent years, its in-depth analysis based on multimodal RNA-Seq towards the design of improved RNA-Seq data analysis workflow is motivated as further elaborated in section 4.1.

1.4 Integrative analysis of splicing landscapes through RNA-Seq

One particular use case of biomedical studies based on omics technologies relates to the analysis of alternative splicing events using large-scale RNA-Seq across different conditions, which became possible since the bulk technology was developed and given the availability of large-scale transcriptomics databases, such as TCGA and GTEx. The current as well as the results chapters introduce biological background related to alternative splicing in tumors, its regulatory mechanism, and computational approaches used for the identification of prospective gene candidates affecting changes in splicing events.

1.4.1 Relevance of alternative splicing and its regulation by RBPs

Splicing is a conserved biological process implying the transformation of primary transcripts (pre-mRNA) into its mature forms (mRNA) that can direct the synthesis of multiple protein isoforms during the further translation (Figure 1.9A). Alternative splicing (AS) of pre-mRNA is regulated by RNA-binding proteins (RBPs) defining which exons are included in the resulting transcripts. RBPs act as trans-factors controlling the splice site choice by binding to cis-acting elements (binding sites of RBPs), followed by the repressed or enhanced splice site recognition and spliceosome assembly.

Regulation of mRNA alternative splicing by RBPs is crucial for generating biological diversity in mammalian genomes and this mechanism is especially

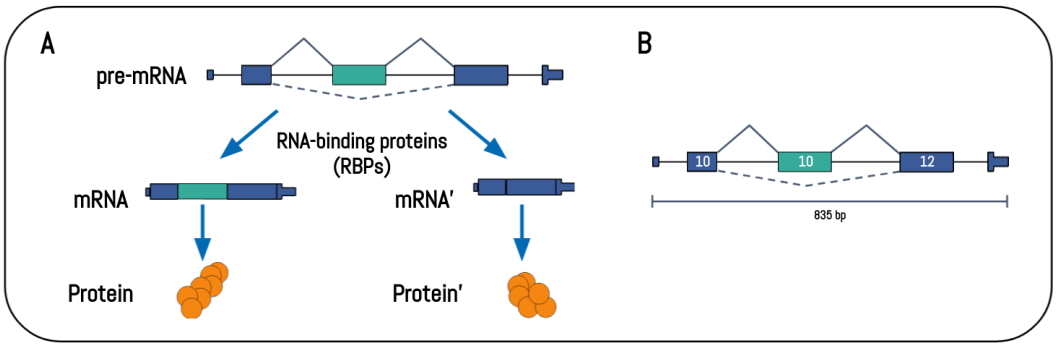

Fig. 1.9: Alternative splicing mechanism. *(A)* Alternative splicing and its regulation by RBPs. *(B)* Skipping the exon 11 in proto-oncogene RON (MST1R) results in alternatively spliced isoform RONΔ165, that promotes tumor progression.

challenging in pathological conditions [39, 125]. Splicing perturbations are common in cancers resulting in the altered expression of specific trans-regulatory factors associated with a tumorigenic state. However, mRNA-RBP interactions are highly variable and regulatory mechanisms of splicing in tumor progression that involve the combination of large numbers of RBPs remain incomplete [122, 13].

1.4.2 Understanding of splicing decisions of the proto-oncogene RON

One example for the regulation of cis-regulatory elements in alternative splicing by RBPs is exon 11 in the proto-oncogene MST1R (RON). Using deep mutagenesis, a very dense cis-regulatory landscape for the skipping exon 11 in RON event was found and almost every sequence mutation in the exon and the surrounding introns leads to a change in splicing outcomes. Accordingly, the region contains a large number of RBP binding motifs, RBP binding can be detected by iCLIP measurements, and the splicing of exon 11 is altered upon siRNA-mediated knockdown of various RBPs [93, 14]. RON

exon 11 (RONΔ165) functions as a so-called cassette exon which is either included or excluded during splicing. The neighboring exons are constitutive, so that the inclusion isoform contains exon 10-12, whereas skipping gives rise to a short isoform comprising only exon 10 and 12 Figure 1.9B).

Inappropriate pre-mRNA skipping of this exon changes protein function and results in expression of the constitutively active RONΔ165 isoform, which promotes tumor progression and is associated with metastatic cancer phenotypes. RONΔ165 is upregulated in solid tumor tissues, including breast, colon, ovarian and pancreatic cancers [132]. Previous studies have reported several RBPs to regulate RON splicing [43, 14, 84], and the further investigation of its splicing decisions in tumor cells enabling to shed light on its molecular mechanisms under pathological conditions.

Other proteins of perturbed splicing patterns are well-known apoptosis regulators are FAS/CD95 and MAP3K7, whose alternative splicing has been implicated in enhanced or reduced apoptosis-activity. FAS/CD95, a transmembrane cell surface protein binding to its ligand FASL, the inclusion of those exon 6 generates a membrane-bound receptor that promotes apoptosis in tumor patients [114]. The regulation of its splicing was shown to be modulated by such RBPs as RBM5, TIA-1 and EWS [94]. Another interesting candidate for the analysis of AS regulatory machinery is MAP3K7, belonging to mitogen-activated protein kinases and is considered as one of central regulators of cell fate decisions during developmental processes and apoptosis. Alternative splicing results in skipping of its exon 12 giving a tumorigenic isoform. Regulators RBM47 and ESRP1/2 were shown to promote the exon inclusion in MAP3K7 [131].

1.4.3 Computational approaches for analysis of alternative splicing regulome

Analysis of alternative splicing events allows to understand the requirements for correct selection of exons. There are several methods reported to identify regulators of AS using RNA-Seq data. For instance, it is common to identify RBP-mediated splicing events by knocking down or over-expressing the RBP and performing RNA-Seq [27, 72, 73, 6, 111]. These methods are, however, time-/costly and limited by the number of prospective candidates for the analysis. Regulatory mechanisms of AS are highly complex and the multidimensionality and noisy nature of transcriptomic expression data makes the analysis challenging. One of the powerful tools to simultaneously characterize alternative splicing outcomes and RBP expression patterns in a genome-wide manner using bulk RNA-Seq. Although the gene expression data for over thousands of possible regulators of AS are publicly available, there are no standardized computational approaches to dissect the most relevant one for splicing decisions and decrease the complexity of the subsequent wet lab analysis, which motivates the study described in section 4.2.

2

Aims of the Thesis

The last decade has been witnessing a rapid growth in multi-omics efforts, which enabled investigations of biological systems on a broader scale, thereby offering new insights into molecular mechanisms underlying varying physiological conditions and diseases. This resulted in the increased importance of integrative data analysis (IDA) approaches for exploring and interpreting data generated by such omics technologies, making IDA an integral component of biomedical research and life sciences.

That said, since every biological study is unique in one aspect or another, researchers may often prefer to address data collection and processing tasks on an ad hoc basis, which challenges a seamless integration of diverse data, especially when dealing with heterogeneous data sources. This necessitates the development of standardized computational tools (software packages and frameworks) that can support domain experts in conducting efficient and reproducible research given the ever-growing body of knowledge and data over time, which also motivated the current work.

Therefore, in this thesis, I combined data from laboratory experiments conducted by colleagues, along with publicly available datasets provided by the genomic community, to develop novel ML-assisted computational tools—FAVSeq and regulAS—aimed at analyzing transcriptomic data across sources and technologies towards discovering relevant signals and addressing biological questions.

More specifically, the thesis focuses on investigation of confounding factors affecting variability in multimodal omics data from bulk and scRNA-Seq (section 4.1 in the Results) and identification of splicing regulators in RNA-Seq data from human patients across pathological conditions and tissues (section 4.2 in the Results).

The following chapters provide an overview of typical omics workflow, focusing on integration and downstream analysis methods that have a potential to extend/combine transcriptomics data to/with other omics, thus, extending applications of the tools proposed in this work. The thesis also discusses machine learning algorithms and methods commonly applied for the omics analysis, including non-/linear supervised machine learning approaches, model-based feature selection, and imputation of missing values. Finally, this thesis provides an overview of the recent applications of ML and IDA approaches to omics data, with a particular focus on biomedical and alternative splicing studies.

3 Methods

3.1 Datasets

RNA-Seq data from matched experiments

To analyse the core factor affecting the difference between RNA sequencing experiments, I examined data of single-cell (scRNA-Seq) and mRNA (bulk RNA-Seq) measured on the same population of retina cells from the study of Shen et al., 2021 [106]. In short, single cell suspension from retina was originated from eight pig animals (6 pigs and 2 mini pigs), followed by library preparation done using the Single Cell 3'Reagent Kit v2 (10x Genomics). Matched sequencing experiments were performed on the same biological samples to measure gene expression patterns. In order to mitigate the difference between sequencing procedures, the same scRNA-Seq cell preparation procedure was used in both experiments. For each library pool the sequencing was performed on the Illumina HiSeq 4000 platform to generate scRNA-Seq and the bulk dataset with single-indexed paired-end and dual-indexed 2x75 bp paired-end runs.

Sequencing reads for both datasets were mapped and annotated on a gene level using the same version of Sus scrofa reference genome (version 10.2.86 primary assembly). Transcripts per million (TPM) values were computed subsequently to measure gene abundance in the RNA-Seq experiment.

Gene expression and RNA junctions for AS genes

Transcriptome profiles from tumor and healthy human donors were collected based on data generated by The Cancer Genome Atlas (TCGA; `https://www.cancer.gov/tcga`) and Genotype-Tissue Expression (GTEx; `https://www.gtexportal.org/home`) projects to investigate candidate splicing regulators across AS genes with exon-skipping events. Data of RNA junctions for tumor and healthy human samples were retrieved through Snaptron database using the Snapcount package version 1.2.0 (`https://github.com/langmead-lab/snapcount`).

In order to obtain the number of reads per sample, we used the coordinates for the upstream and downstream junctions surrounding the alternative exon as well as the skipping junction using the query j_x function. For instance, the following coordinates on chromosome 3 were queried to collect reads of the alternatively spliced exon 11 as well as the surrounding ones for MST1R data: "$49895881 - 49895960$", "$49896108 - 49896194$" and "$49895881 - 49896194$" for left, right and the skipped exons, respectively.

RBPs gene expression signatures for RON were measured as Fragments Per Kilobase of transcript per Million (FPKM). FPKM reflects the normalised estimate of gene expression in paired-end RNA-seq and is calculated as the following: 1) divide the total amount of reads in each sample by the 'per million' scaling factor of 10^6. 2) normalize read counts for sequencing depth by its division by the 'per million' scaling factor, resulting in reads per million (RPM) and 3) normalize RPM for the gene length by its division by the length of the gene, in kilobases or 10^3 base pairs of RNAs.

3.2 Experimental setup

FAVSeq

Creation of pseudobulk from scRNA-Seq

In order to enable the comparative analysis of data from different experiments, single cell counts were aggregated by summing up raw read counts across single cells from each sample, followed by its gene length and per sample normalization giving transcript counts per million.

Extraction and aggregation of gene-associated features

Data corresponding to genes were integrated from different sources, such as COMPARTMENTS [10], BioMart [108] (v.86) and Ensembl v.86 (parsed from GTF) annotation data [26]. As features representing a length (i.e., transcript length) consist of duplicates, the longest entries were chosen for the analysis. For the same reason (presence of duplicates), an averaged GC content was calculated for the corresponding feature. Chromosome feature provides the information on chromosome localization, where numerical entries (i.e., chromosome 2) are designated by their chromosome numbers.

A feature representing cellular compartments was generated as described below. First, genes with the most reliable entries were selected w.r.t. the reliability index [118], evidence codes were joined into top-level groups as suggested by [5]. As the compartment dataset consists of missing values corresponding to a single gene, duplicated entries were removed based on the most evident results, i.e., the higher priority was assigned to the entries verified by experiments and curated information. Subsequently, the subgroups

were merged into the top-level group (e.g., "Organelle" and "Organelle part" into "Organelle") in order to reduce the complexity of high-dimensional data. Most commonly-validated compartments were selected among duplicates. Finally, compartment names were updated by the corresponding GO term indices ("Membrane" localization into "GO:0016020" GO term into "16020" GO encoded) in order to preserve associations between the categories.

Imputation of missing values in generated feature matrix

Since some of the generated features consist of missing values, these values were imputed using a k-Nearest Neighbors (k-NN) approach. The exact behavior of the k-NN imputation model depends on the chosen distance metric. Here, given a sample containing a missing value indicator (NaN) for a feature, the value is being replaced by the average feature value calculated for $K = 100$ samples that are closest to the considered one, according to the Euclidean distance.

Calculation the averaged gene expression difference between experiments

In order to define a quantitative metric for comparing sequencing experiments, the aggregated per-sample gene expression difference was measured and used as a dependent variable for the regression analysis. Log-transformation of gene expression values was done to decrease variability within the sequencing data and make it conform more closely to the normal distribution. Gene expression data was represented as an $M^{|G|x|S|}$ matrix of genes vs. samples for S∈A∈B, where A and B are pseudo- and bulk RNA-Seq, respectively, and $|S_A| = |S_B| = N$. Then, the slope and intercept of the regression line can be estimated as shown in Equations 3.1-3.2, so that the sum

of squared errors (SSE) is minimized (Equation 3.3). Then, the difference is measured as the observable error from the estimated coefficients of OLS.

$$\hat{g_B} = M_B\beta + \varepsilon \tag{3.1}$$

$$\beta = (M_A^T M_A)^{-1} M_A^T g_A \tag{3.2}$$

$$SSE = \sum_{i=1}^{|G|} \sum_{j=1}^{|S_B|} (g_B - \hat{g_B}) \tag{3.3}$$

Assessment of feature importance using model-based feature selection

In order to identify the most relevant contributors to the variability in RNA-Seq experiments, we utilized embedded and wrapper feature selection methods to access the importance of a variety of factors (i.e., genomic and transcriptomic) on the quantitative (gene expression levels) and qualitative (dropout events) difference in matched bulk and single-cell RNA-Seq dataset.

To choose the most suitable model w.r.t. feature selection, we trained and evaluated different machine learning algorithms in 5-fold cross validation, as well as benchmarked them against the baseline estimator (linear and logistic regression models for regression and classification tasks, respectively) in order to access the goodness of predictions. Random forest is the one of embedded approaches that was employed to rank features w.r.t. their influence on changes in the quantitative difference and leverage performance of decision trees, while mitigating its tendency to overfitting. For the regression

task, standard deviation served as a measure of impurity, those decrease has been defining hierarchical split of the training samples [101]. Afterwards, we estimated importance of a feature proportionally to its closeness to the tree's root node, which is defined through the information gain [15]. Finally, averaging by trees in the forest produces the joint estimation of the feature importance. In order to rank features w.r.t. their relevance for dropouts (classification task), we applied an MLP model [87]. Assessment of feature importance scores using this model was done through the accumulation of absolute values of gradient of the loss w.r.t. input during the model training [104].

Optimization of hyper-parameters of regression and classification models by the 5-fold CV grid-search over all possible combinations in the parameter space as the following: 1) Split the data into 4 training and 1 test folds. 2) Perform feature pre-processing to make values follow the normality assumption. 3) Train the model on the training folds and evaluate on the test fold. 4) Choose hyper-parameters corresponding to the best model. Hyper-parameter search for the random forest model was done for the number of estimators (from 32 to 1024), their maximum depth (4-32), minimum number of samples required to split an internal node (2-16) and the minimum number of samples in a leaf node (1-8). Optimization of hyper-parameters for the MLP model was done for the number of neurons in hidden layers (32-1024), learning rate (0.001-0.1), and L2-regularization term (α; 0.0001-0.1). The MLP's output was mapped non-linearly using logistic activation function, and its weights were adjusted during the training using Adam optimizer [64].

Identification of the core relevant features using recursive feature elimination

To identify an optimal subset of the features relevant for the gene expression difference between matched RNA-Seq experiments, we applied recursive feature elimination (RFE) [45], which performs an iterative elimination of least scored features w.r.t. to an external estimator (e.g., random forest). In contrast to the full search through all possible feature combinations, the RFE-based selection solves the task in a linear time, thus providing a valuable speed-up.

The optimal number of features is determined by computing index of the first stationary point of the objective function based on the RFE-provided objective values $\mathcal{L} \in \mathcal{R}^{N_{features}}$ using first- (Δ) and second-order (Δ^2) differences (Equation 3.4).

$$N_{optimal} = \left\lfloor \frac{1}{N_{folds}} \sum_{n=1}^{N_{folds}} \underset{N_{features}}{argmin} \left(\Delta sign\left(\Delta^2 \mathcal{L}_n\right) = 0 \right) \right\rceil + 2 \tag{3.4}$$

regulAS

Preparation of RBPs gene expression data

Due to differences in sample and processing pipelines across RNA-Seq data, samples for RON were normalized and corrected for study-specific biases as suggested by Wang et al., 2018 [123]. Accordingly, for each of 7966 samples, expression levels of 2039 RBPs were available. The following pre-processing of RNA-Seq gene expression has been applied to gene expression data: RBPs with missing values in more than 70% of samples were filtered out and the only tissues with the sufficient number of samples were chosen for the analysis, resulting in 2036 RBPs for RON data.

Preparation of RNA junctions data

The junction reads data were collected for 15468 TCGA and GTEx samples in total. Due to the differences in data sources for RNA-Seq and splicing datasets, the splicing profiles for the exon 11 of RON were available in the total of 3343 tumor samples with the corresponding RBPs expression data. Pre-processing of splicing profiles for the exon 11 of RON included filtering of non-reliable splicing profiles ($k < 10$ reads), as shown in Figure 4.11. For each patient sample, alternative splicing of RON exon 11 was then quantified using percentage-of-spliced-in.

Measuring changes in splicing events as PSI

Percentage-of-spliced-in (PSI) was used as a measure of alternate splicing changes, which represents the percentage of a gene's transcripts that include

a specific exon or splice site. PSI quantification of skipped exon event was calculated as follows:

$$PSI = \frac{IR}{IR + ER} \tag{3.5}$$

where IR and ER are the total number of Inclusion and Exclusion Reads, respectively.

Designing the workflow to select relevant regulators of splicing decisions for RON

While choosing models, it was aimed to preserve the diversity of various approaches. Analysis based on both non- and linear models allows us to consider the complex relationships between splicing machinery's interplayers and possible nonlinearity of data across different conditions. The following machine learning algorithms were evaluated to predict splicing efficiency (PSI) values based on gene expression data of putative RBPs for RON exon 11: support vector regressor (SVR) with C regularization [32], neural network (MLP), Bayesian ridge regression [12] models, linear regression regularized with combined L_1 and L_2 penalties [137], as well as ensemble methods, such as ensemble trees and random forest [16].

Model training and testing were done based on a 5-fold cross-validation method. Optimization and selection of most suitable hyper-parameters of models were done based on the Grid-search method which also results in the most accurate predictions. Performance has been assessed using Mean squared error (MSE). Spearman correlation coefficient was used to evaluate the goodness of models to predict splicing events.

Knockdown verification of RBPs by siRNA

Knockdown (KD) experiments were done using siRNA method in MCF7 breast cancer cell line. Experimental protocol was used as reported in Braun et. al, 2018 [14]. Quantification of the effect of individual top-rank RBPs on isoform levels of RONΔ165 analyzed using RT-qPCR.

Comparison of RBP ranks using nDCG

Normalized Discounted Cumulative Gain (nDCG) score was used to evaluate the degree of agreement between feature ranking results obtained from "in-silico" and "in-vitro" experiments. In particular, nDCG [58] was computed to compare relevance scores derived using ML-approach based on analysis of open-access RNA-Seq from omics data sources with those from siRNA knockdown experiment (later as the reference rank) through a comparison of robust Z-score measurements for 153 RBPs [93]. This allowed to assess the relevance of candidate RBPs in the regulation of alternative splicing across different physiological conditions and experiment types.

nDCG was calculated by dividing the DCG of the ranking by the reference or ideal DCG (IDCG), normalizing the score to a value between 0 and 1, with the latter indicating a perfect match between top ranks.

$$nDCG@K = \frac{DCG@K}{IDCG@K} \tag{3.6}$$

For the cumulative gain (CG), K corresponds to the number of top-relevant elements in the ranking and rel_i is a graded relevance of the feature at position i.

$$CG@K = \sum_{i=1}^{K} rel_i \tag{3.7}$$

Then, in the discounted cumulated gain (DCG) formula, $log_2(i+1)$ is used to reflect the logarithmic reduction factor to penalize the rel_i value proportionally to the position of the result. DCG measures relevance of items in the ranking compared to the reference list for the top-K entries. It gives higher scores to relevant predictions that appear higher in the list.

$$DCG@K = \sum_{i=1}^{K} \frac{rel_i}{log_2(i+1)} \tag{3.8}$$

Finally, for the ideal DCG (IDCG), which represents the maximum possible DCG that can be achieved for a given set of relevance labels, the score is obtained for a list of top relevant items ordered by the reference rank descendingly.

Functional protein association analysis

Pathway enrichment analysis and network reconstruction of molecular associations for top relevant RBP candidates identified were done using the R interface to EnrichR and STRING databases [68, 113].

Statistical analysis

Correlation analysis was based on Pearson (Equations 3.9-3.10) and Spearman's rank (Equations 3.11-3.12) correlation coefficients. A correlation coefficient is used to measure a strength of relationship between two random variables $X, Y \in \mathcal{R}$, where a value of 1 represents a perfect positive

relationship, -1 represents a perfect negative relationship and 0 represents no association found between the variables.

$$\rho_{XY} = \frac{\text{cov}(X, Y)}{\sigma_x \sigma_y} \tag{3.9}$$

$$r = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2 \sum_{i=1}^{n}(y_i - \overline{y})^2}} \tag{3.10}$$

Spearman's rho coefficient, which can be viewed as a rank-based version of Pearson's correlation coefficient, is used to assess non-linear associations between the variables.

$$\rho = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)}, \tag{3.11}$$

$$d_i = R(X_i) - R(Y_i), \tag{3.12}$$

where d is the pairwise distances of the ranks of the raw scores (X_i, Y_i) and n is the number of observations.

After analysis of variances of PSI for tumor and normal samples, a two-sided t-test with Welch-Satterthwaite correction [126] for the case of not equal variances (Welch's t-test) was used to compare splicing efficiency across these two populations of samples obtained from TCGA and GTEx data.

Availability of data and materials

The raw data of matched bulk and single-cell RNA-Seq experiments supporting the findings of this work are available at the European Nucleotide Archive (ENA) under accession number PRJEB43819. The source code of the FAVSeq pipeline for analyzing multimodal RNA-Seq data applied in this study is available on the GitHub at `https://github.com/slipnitskaya/FAVSeq`.

RNA-Seq expression profiles for putative regulators and junctions data for AS genes of interest for tumor and healthy human samples are being retrieved from TCGA and GTEx data repositories. The regulAS software package for integrative analysis of alternative splicing regulome using RNA-Seq is available at `https://github.com/slipnitskaya/regulAS`.

4 Results

4.1 Integrative analysis of factors contributing to the technical variability between bulk and single-cell RNA-Seq experiments

4.1.1 Preamble

This work was initiated during my stay in the Global Computational Biology and Digital Sciences department at Boehringer Ingelheim Pharma GmbH & Co. KG in Biberach an der Riß, Germany. The efforts resulted in the paper, for which I designed and implemented the computational framework, performed data mining and ML-assisted analysis, and wrote the manuscript with input from all collaborators. Parts of this chapter relate to the manuscript submitted in: Machine learning-assisted identification of factors contributing to the technical variability between bulk and single-cell RNA-seq experiments (Lipnitskaya, S., Shen, Y., Legewie, S., Klein, H., Becker, K.; `https://doi.org/10.1101/2022.01.06.474932`). The source code of the FAVSeq pipeline applied in this study is available in the GitHub repository.

4.1.2 Key highlighs

Recent studies in the area of transcriptomics performed on single-cell and population levels reveal a noticeable variability in gene expression measurements provided by different sequencing technologies. Due to increased noise and complexity of single-cell RNA-Seq (scRNA-Seq) data over the bulk experiment, there is a substantial number of variably-expressed genes and so-called dropouts, challenging the subsequent computational analysis and thus, leading to false positive discoveries. In order to investigate factors affecting technical variability between RNA sequencing experiments of different technologies, we performed a systematic assessment of single-cell and bulk RNA-Seq data, which have undergone the same pre-processing and sample preparation procedures (Figure 4.1).

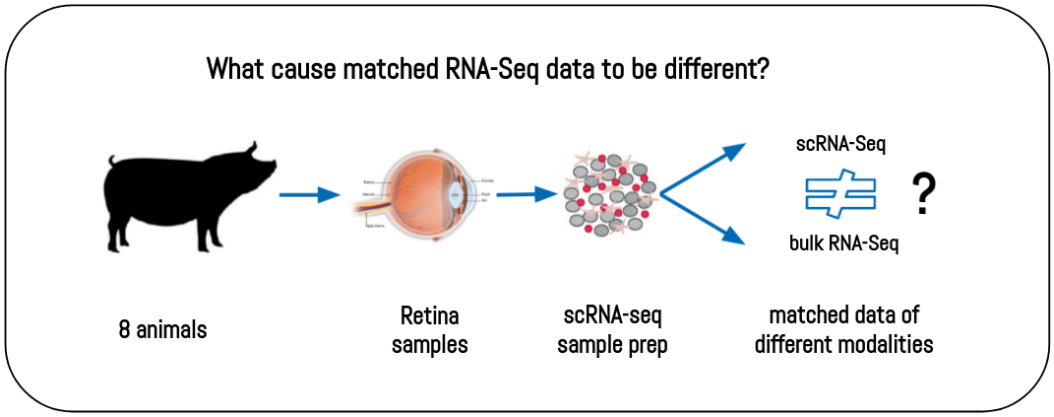

Fig. 4.1: Objectives for IDA- and ML-based analysis of factors affecting technical variability in multimodal RNA-Seq using diabetic retinopathy data.

This thesis proposes a computational pipeline for integrative analysis of multimodal RNA sequencing data of matched experiments, namely FAVSeq. This setup allowed us to investigate technical variability in data, and identify factors affecting variation in gene expression and occurrence of dropouts. Our analysis indicates that variability between gene expression measurements as well as dropout events are not exclusively caused by biological variability,

low expression levels, or random variation. Also, we found the 3'-UTR and transcript lengths as the most relevant influencers of the observed variation between matched RNA-Seq experiments, while the same factors together with cellular compartments were shown to be associated with dropout events. We further outline prospective applications and developments of the identified factors for improvements in RNA sequencing technologies.

4.1.3 Background

High-throughput single-cell RNA sequencing (scRNA-Seq) provides a powerful tool for profiling gene expression patterns at single-cell resolution that has revolutionized transcriptomic studies and advanced the knowledge of biological systems. RNA samples for bulk sequencing are typically derived from a heterogeneous population of cells and thus, cell-type specific transcriptomic changes may be lost, as the final data represents an average expression across thousands of cells from different types [90]. However, while single-cell technology allows to overcome some limitations of standard bulk RNA-Seq, data complexity and noise increases as well as the detection limit and RNA amount present a challenge to effectively identify and filter out low-quality genes for reliable and reproducible results. Therefore, in addition to the stronger variability of single-cell technology compared to traditional bulk RNA sequencing, the subsequent downstream analysis gives rise to new computational challenges in analyzing data and interpreting the findings.

The increased noise in scRNA-Seq data can be explained by both biological and technical reasons that encompass, in particular, lower amount of input material, batch effects, amplification biases, cells being in distinct phases of the cell cycle, and transcriptional bursts [86]. Another source of challenges

accompanying computational analysis relates to dropout events. Previous studies showed that single-cell technology typically suffers from dropouts, i.e., existence of transcripts which are present in a biological replicate or a cell, but not detected by the sequencing technology. In contrast to RNAs that are simply not present at the time of cell isolation, dropouts have non-zero expression but can not be identified due to limitations of experimental protocols as well as other biological and technical reasons [51].

Often these dropouts are caused by low expression, but additional factors may contribute. For instance, dropouts can be potentially caused in different ways by difficulty in isolating single cells from low starting input volumes, cell-specific capture efficiency (e.g., inefficient mRNA to cDNA capture, dilution of cell libraries, and amplification) as well as by the low amounts of mRNA in individual cells [100]. Additional challenges affecting the difference in RNA-Seq experiments can arise from differences in library preparation protocols [115, 22] and computational downstream analysis pipelines [56]. Several studies aimed at understanding the difference between matched RNA-Seq experiments performed a comparative analysis of dataset with available matched bulk and scRNA-Seq samples [61, 116, 129]. However, these studies have been usually limited by the number of sequenced cells and analysed samples [75].

All of that motivated us to perform a detailed comparison of RNA-Seq experiments in order to investigate the most relevant factors affecting the difference between them. To do so, we analysed paired data—scRNA-Seq and bulk from the same sample—to limit the biological sample-to-sample variation. Hereby, the derived knowledge can be employed further in order to identify genes that can be affected by assay-based deviations as well as to define experimental stages to be adjusted.

4.1.4 Matched bulk and single-cell experiments allow for a detailed analysis of gene expression differences

With the aim to assess sources of variation provided by different RNA-Seq technologies, we based our analysis of gene expression measurements performed in multiple biological replicates (hereinafter referred to as samples). Thus, considering distinct samples allows for improving the efficiency of statistical testing and the reliability of the findings. In order to perform a comparative analysis of bulk and scRNA-Seq experiments, both sequencing technologies were applied to eight retina Sus scrofa samples. The unified sample preparation procedures ensured high comparability of resulting measurements. Specifically, all samples were processed according to the 10x Chromium single cell RNA-Seq sample preparation protocol—including single cell dissociation—before fractions were split for the two experiments providing matched scRNA-Seq and bulk RNA-Seq measurements. Subsequently, RNA-Seq libraries for bulk sequencing were prepared using the NEBNext Ultra mRNA, followed by paired-end Illumina sequencing (2 x 75 bp). The scRNA-Seq experiment, performed using paired-end sequencing 10x Single Cell 3' v2, produced a total of 2,111,208 cells with an average number of 263,901 cells per sample (Figure 4.2A). A more detailed summary of the experimental protocols for the matching RNA sequencing experiments can be found in the Methods section and Table 4.1.

In order to compare gene expression measurements provided by different experiments, we aggregated scRNA-Seq data by summing up raw read counts across single cells from each sample, followed by its gene length and per sample normalization, which we termed "pseudobulk". The latter was shown to be highly similar ($\rho = 0.96$) to original measurements provided at the single-cell level (Figure 4.2B).

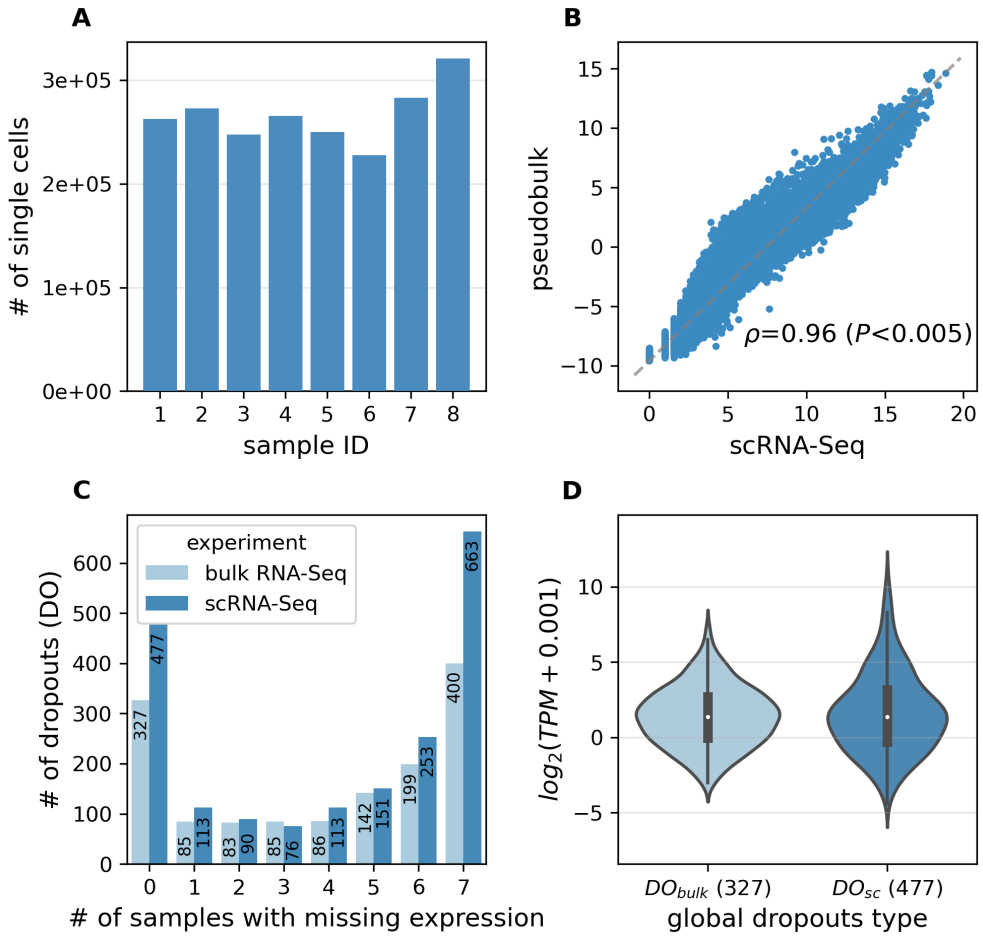

Fig. 4.2: Distributional statistics calculated for the matched single-cell and bulk RNA-Seq data. A. Number of cells measured in scRNA-Seq experiment across individual samples. B. Comparison of averaged gene expression measurements between scRNA-Seq counts and aggregated pseudobulk data. Log-transformed values are shown. C. Sparsity rate of dropouts calculated across different proportions of samples for each experiment. The middle-bar-number represents the amount of genes with missing expression observed in only one of the experiments, while being expressed in at least one sample in the matched one (sample-wise dropouts). D. Mean expression levels of genes that are global dropouts or represent missing expression in all samples in only one of the matched RNA-Seq experiments, while being detected across all samples by another technology (bars 0 in Figure 4.2C). Gene expression measurements in the matched experiment are shown.

Out of 25,322 annotated genes in the Sus Scrofa genome, 21,372 genes were detected and showed non-zero expression in at least one out of eight samples in either of the matched experiments. Among all samples, 19,965

Tab. 4.1: Comparison of experimental protocols of matched RNA sequencing experiments

Processing step	bulk RNA-Seq	scRNA-Seq
Library type	single cell library	single cell library
Library preparation	NEBNext Ultra II RNA 2x75bp	Single Cell 3' v2
Sequencing type	dual-indexed paired-end	single-indexed paired-end
Sequencing platform	Illumina HiSeq 4000	Illumina HiSeq 4000
Aligner	STAR (v2.7.3a)	STARSolo (v2.7.3a)

and 19,436 genes were detected in at least one sample of the scRNA-Seq and the bulk experiment, respectively (Figure 4.3A). Specifically, in the bulk experiment we found an average number of 17,283 genes per sample, while the scRNA-Seq provided the average number of 16,733 (Figure 4.3B), indicating a slightly lower detection rate for the single-cell protocol. Among these detected genes, 18,029 ($> 71\,\%$) were common, i.e., showed non-zero expression in at least one sample of both scRNA-Seq and bulk measurements (blue intersection in Figure 4.3A). At the same time, we observe genes with missing expression in one experimental modality (e.g., scRNA-Seq), while being detected in the matched sample provided by another experiment (e.g., bulk RNA-Seq), which we term as sample-wise dropouts (gray areas in Figure 4.3A).

Analysis of gene sets in individual experiments indicates (Figure 4.3C) that the bulk RNA-Seq provided a slightly higher proportion of the most confident genes that were detected across all samples, i.e. $97.6\,\%$ (13,100) of genes were detected by the bulk RNA-Seq in all samples against $96.4\,\%$ (12,950) detected by the scRNA-Seq experiment. The total of 12,623 genes are found to be common or expressed in all available 16 samples across matched exper-

iments (blue intersection in Figure 4.3C). Besides, we also found genes with the available expression values in 8 samples in only one of the experiments (referred later as global dropouts). Accordingly, both single-cell as well as bulk global dropouts exist with the total of 804 genes (gray areas in Figure 4.3C) that have been systematically found in the combined experiments (Figure 4.2C).

We further found a profound difference in expression profiles between bulk and pseudobulk measurements. While many common genes were detected in matched experiments, the quantitative expression levels differ between bulk and pseudobulk datasets at the global level. Comparison of mean expression levels of genes detected in both matching experiments (Figure 4.3D) indicates the higher detection rate in the bulk experiment demonstrating the consistency with reported studies [51]. Figure 4.3D shows a substantial difference in expression measurements with lower gene expression measured in the single cell experiment with an average of $log_2 TPM = 2.14$ when compared to the bulk experiment with median expression of $log_2 TPM = 4.44$.

As the ground-truth of expression in genes with missing values is unknown, we can only obtain such an information from samples of the matching experiments. Consequently, in order to ensure the validity of the derived results, we performed an ML-assisted analysis using the subset of the most confident genes (Figure 4.3C). Thus, for the analysis of the quantitative difference we used data related to common genes with available expression measurements in all samples in both RNA-Seq experiments (blue intersection area; regression task), while for the analysis of the qualitative difference (dropouts) we also include those with expression measurements detected in 8 samples in the matched experiments only (gray areas; classification task).

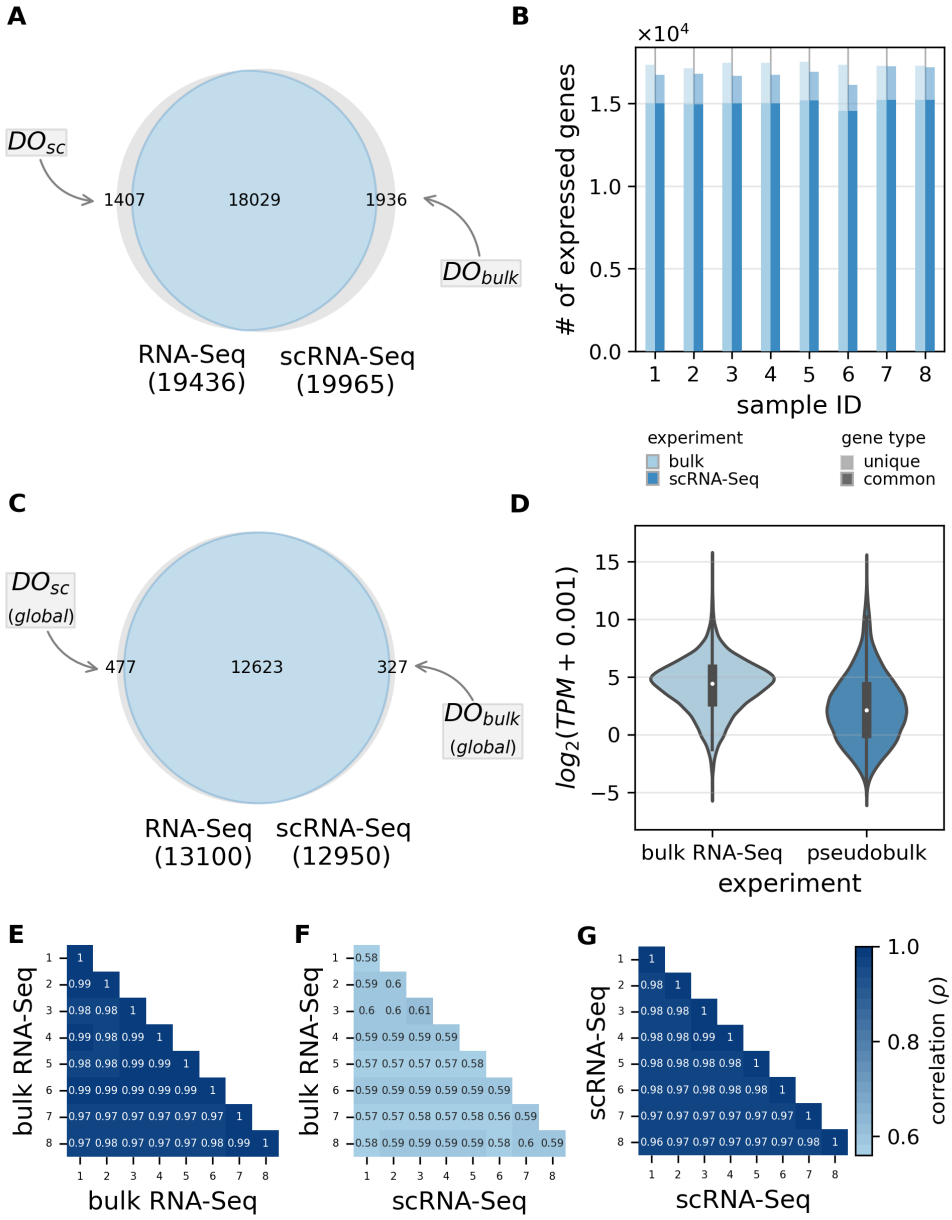

Fig. 4.3: Comparison of the bulk and single-cell gene expression profiles in matched RNA-Seq experiments. A. Amount of uniquely expressed genes detected in the bulk and scRNA-Seq, where blue color represents genes which are expressed in both experiments and gray indicates dropouts (DO) or genes with missing expression across all samples in the one of experiments. B. Quantitative analysis of common and unique (missing in one of the experiments) genes detected by different RNA-Seq technologies. C. Amount of genes detected either in all samples in both RNA-Seq experiments (intersection area, blue), or missing genes that are expressed in 8 samples (a.k.a. global dropouts) in only one of the matched experiments (disjunctive union, gray). D. Distribution of averaged expression measurements of most confident common genes detected in all samples (blue intersection in Figure 4.3C) in the bulk and aggregated scRNA-Seq experiments. E. Sample-wise correlation of transcriptomic profiles of common genes (blue intersection in Figure 4.3C) measured in the bulk experiment. F. Sample-wise correlation of transcriptomic profiles of common genes measured in the matched dataset. G. Sample-wise correlation of transcriptomic profiles of common genes measured by scRNA-Seq.

4.1.5 Technical noise is the major contributor to difference between single-cell and bulk RNA-Seq

In order to assess the contribution of technical and biological variation, we performed a sample-wise correlation analysis within and between the two sequencing experiments. Figure 4.3E-G indicates how transcriptomic profiles produced from the scRNA-Seq experiment are different to that of the bulk RNA-Seq for genes expressed across all samples and both modalities (blue intersection in Figure 4.3C). For each of the sequencing protocols, the expression measurements showed a little across-sample variation with correlation coefficients close to 1 in both the bulk RNA-Seq ($\rho > 0.97$) and aggregated into pseudobulk scRNA-Seq ($\rho > 0.96$) experiments as represented in Figure 4.3E and Figure 4.3G, respectively.

This suggests that the effect of the biological noise is negligible in the tested setup. Results also indicate a slightly increased gene expression variability level in aggregated single cells compared to bulk RNA-Seq. As droplet based scRNA-Seq systems (i.e., 10x Genomics Chromium) amplifies cDNA fragments close to polyadenylation (polyA) tails, the corresponding gene expression measurements are highly biased to the 3'-end, while the full transcript coverage was captured in the bulk data. Due to the generated single-cell data are often confounded by the quality of 3'-UTR annotation, this increase in variability of gene expression measurements was expected.

On the other hand, correlation between bulk and pseudobulk samples was comparatively low ($\rho = 0.59 \pm 0.01$), suggesting the presence of technical variation between matched RNA-Seq experiments (Figure 4.3F). Thus, the correlation analysis reveals a negligible impact of batch effects on gene expression indicating that technical noise is the major contributor to the difference between matched RNA-Seq experiments.

4.1.6 Dropouts are systematically observed in data and are only partially caused by lowly expressed genes

Dropout events relate to a common phenomenon observed in RNA-Seq data implying that specific transcripts cannot be detected by the sequencing technology [100]. As the presence of dropouts highlights possible limitations in the sequencing and/or pre-processing protocols, which, in turn, may introduce bias in downstream analysis and interpretation of the data, we investigated these genes in more detail.

Figure 4.2C provides an overview of the experiment-wise sparsity rates of dropouts identified in the matched dataset, where the sparsity rate is defined as the number of genes with missing expression ($TPM = 0$) across different proportions of samples. Results of this analysis together with Figure 4.3B indicate that sample-wise dropouts are not likely to occur randomly or by chance in matched RNA sequencing data. Here, the sparsity rate of zero indicates dropouts with the available expression values in all 8 samples in only one of the experiments.

Given that our experimental design provided gene expression measurements across several biological samples, we introduce a more specific definition of high-confidence dropouts (here and later as global dropouts), in which we consider them as genes representing missing expression in all samples in only one of the matched RNA-Seq experiments, while being detected ($TPM \neq 0$) across all samples by another technology. Based on eight matched bulk and pseudobulk samples in our dataset, we found 804 global dropouts, with a higher number in the single cell experiment (gray areas in Figure 4.3C and bar 0 in Figure 4.2C). Specifically, a quantitative analysis of global dropouts indicates 477 genes were not detected in scRNA-Seq data, while being expressed in all samples of the bulk data. Conversely, 327 genes show

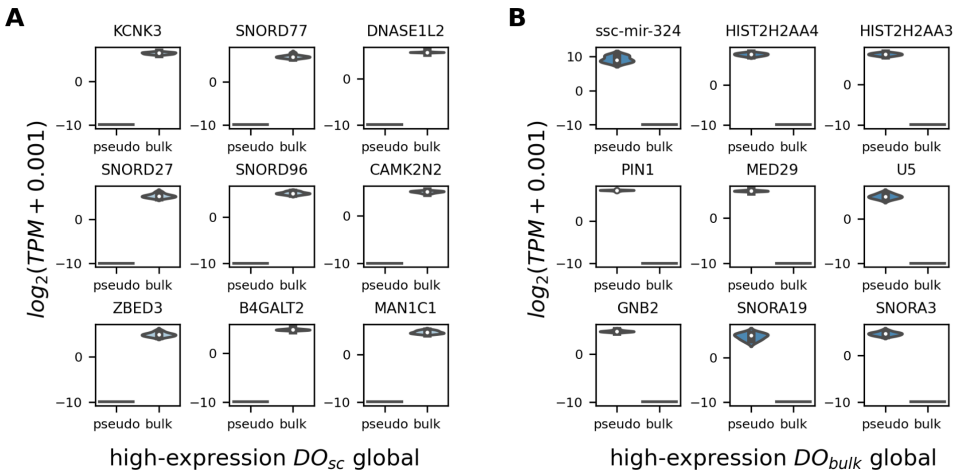

Fig. 4.4: Gene expression levels of most confident global dropouts. Measurements in the matched experiment are shown. A. Expression levels of genes that are global dropouts in scRNA-Seq and exhibit a high-expression in the bulk experiment. B. Expression levels of genes that are global dropouts in the bulk experiment and exhibit a high-expression in scRNA-Seq.

no expression in any of the bulk samples, while being strongly expressed across all samples in single cells.

Since the previous research suggests that weakly expressed genes tend to produce more differences than highly expressed genes at the single-cell level [86], we analyzed mRNA expression level of global dropouts in the matched experiment (Figure 4.2D). Subsequently, comparison of expression measurements indicates the similar average expression levels of these genes in the bulk and scRNA-Seq experiments. As expected, we also observed the majority of global dropout genes to be lowly expressed in the matched experiment. However, a substantial number of dropouts showed non-marginal (high) expression, as can also be seen for specific examples. Closer examination of global dropouts that are highly expressed in the matched experiment (Figure 4.4A and Figure 4.4B for single-cell and bulk RNA-Seq data, respectively), reveals that these genes represent variable expression patterns and therefore, cannot be explained by the low expression level only.

4.1.7 Proposing a computational approach to analyse the difference in RNA-Seq experiments

In order to identify which factors determine whether genes are differently detected in matched RNA-Seq experiments, we introduce FAVSeq (Factors Affecting Variability in Sequencing data), a machine learning-assisted pipeline, those design intends to support researchers in disclosing potential root causes of the difference—in terms of gene expression measurements and dropouts– observed between RNA-Seq technologies. FAVSeq enables to select features obtaining the strongest predictive power for estimation of technical variability between RNA sequencing modalities. The pipeline includes the following steps (Figure 4.5):

1. Create the target by calculating the ordinary least squares (OLS) residuals in gene expression levels.

2. Generate gene-associated features based on GTF annotation file and open-access databases (e.g., BioMart, Jensen Compartments).

3. Optionally, recover missing values in features using a chosen imputation approach (e.g., model-based one).

4. Optimization of hyper-parameters of models through the 5-fold cross-validated (CV) grid-search. Selection of the top-performance model.

5. Rank features w.r.t. their influence on the objective using the top-performance model and based on recursive feature elimination (RFE) in 5-fold CV.

6. Output the summary reports, including statistics for the core features that contribute to the difference between experimental modalities.

Thus, steps 1-3 provide us with the independent (features) and dependent (quantitative or dropouts-related target differences) variables. Feature selection is not a trivial task, especially in case of genomics databases that often consist of partially incomplete data, which, in turn, may affect the performance of machine learning models. To address this issue, we integrated a missing data imputation module into FAVSeq (step 3). The introduced module supports both non- and parametric imputation strategies, including k-Nearest Neighbors (kNN), that was shown to be effective to handle missing and/or corrupted values in genomic and transcriptomic data [96, 54, 117].

Then, based on the created dataset, the subset of the most relevant features is being selected using the machine learning-assisted feature selection approach (steps 4-5), which, in turn, consists of two main parts. The first part serves to select the most suitable model to predict the target difference between experiments (training, evaluation and optimization of hyper-parameters), and the latter part serves to select the subset of the most relevant features among the set of tested ones w.r.t. the objective (Mean Squared Error or Balanced Accuracy for regression and classification tasks, respectively).

While analysing features w.r.t. quantitative difference (regression task), the random forest model was used because it's agnostic to variable types (numerical or categorical). Furthermore, given that smallest and largest values of some features differ by several orders of magnitude, such a model can provide sufficient number of estimators, so individual trees cover particular ranges of the input. Also, the model preserves monotony of transformations applied to the input variables, and is not sensitive to outliers in the input samples. For the classification-based analysis of dropouts-associated features, we used a multilayer perceptron (MLP) model, those training and evaluation followed the same steps. To derive an optimal subset of features relevant for the difference between experiments, we utilized RFE, which is a model-

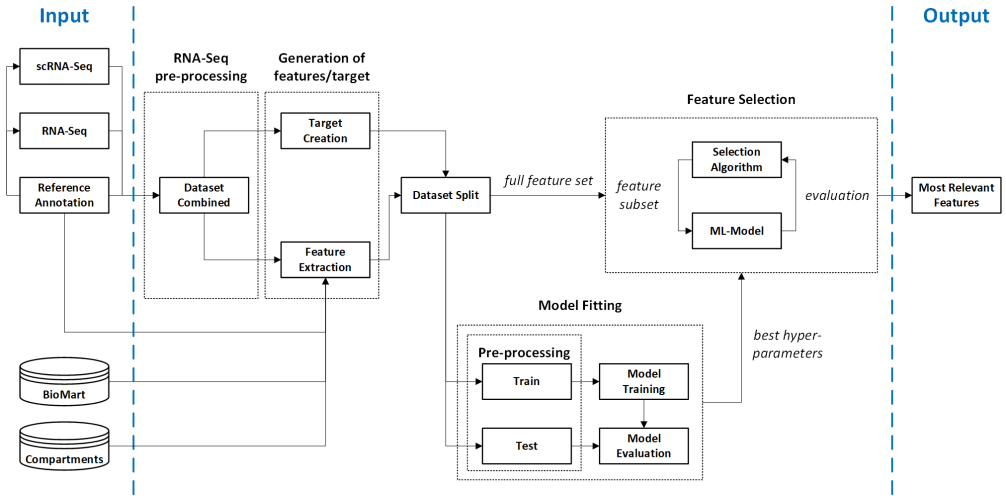

Fig. 4.5: Framework utilized in the FAVSeq pipeline for ranking and selection of features affecting the technical variability in RNA-Seq datasets of matched experiments. The raw input acquired from different sources (on the left) is being pre-processed in order to build a set of features to be fed into an ML-model (e.g., random forest). The full set of input samples is used to perform search for optimal hyper-parameters of the ML-model using 5-fold CV. The chosen hyper-parameters serve then to perform selection of most relevant features based on the underlying ML model (framework's output, on the right) using RFE feature selection technique that allows to determine most informative features, according to the ML model's scores.

agnostic meta-learning method allowing to exclude the least important ones according to the model's scores on each iteration. The algorithm works until the acquisition of the most important features w.r.t. the objective (target difference).

Finally, the pipeline provides the summary reports (step 6) in a form of tables and visuals (i.e., best hyper-parameters chosen, comparison of the performance of models), which allow interpreting the obtained results as well as guiding on further experiments to be performed.

4.1.8 Aggregating gene expression data across matched RNA-Seq experiments

In order to define a quantitative metric for comparing RNA-Seq sequencing experiments, the aggregated per-sample gene expression difference was calculated and used as a dependent variable for the regression analysis. Specifically, we calculated the averaged minimized sum of square differences between measured gene expressions in bulk and aggregated single cells using the OLS method, as described in Methods section. Log-transformation of expression values was done beforehand to make them conform more closely to the normal distribution.

Figure 4.6A represents how the quantitative difference—residuals of gene expression measurements—is estimated for the sample one for genes found in pseudo- (y-axis) and bulk (x-axis) RNA-Seq datasets, showing the higher measurements in the bulk accordingly with the analysis depicted in Figure 4.3D. Gray color indicates dropouts and blue indicates genes detected in both matched RNA-Seq experiments for the total of 13,427 genes. At a

global level (Figure 4.6B), we further notice that the quantitative difference—averaged residuals of gene expression measurements—aggregated across all samples is also higher for genes in the bulk experiments in comparison with those in the aggregated single cells. Thus, the calculated quantitative difference indicates how close measurements in matched experiments are by relating gene expression in pseudobulk to bulk RNA-Seq and serves as a basis for further regression-based feature selection analysis. At the same time, the dropouts-related difference (non-/dropout) will be used later for binary classification task while identifying features contributing to the presence of dropout events.

4.1.9 Creating the feature representation of genes by aggregating data from genomic databases

In order to identify the most relevant features for the difference between gene expression measurements, we generate prospective features representing factors for genes available in both experiments (listed in Table 4.2). These features comprise the subcellular localization of a gene product, the chromosome at which the gene is located as well as metrics describing the dimensions (e.g., transcript length, UTR length) or expression of a gene (transcript count). Collection of data from genomic databases for the further feature engineering was done for 12,623 common genes that are expressed in all samples across matched experiments (blue intersection in Figure 4.3C), also considering 804 global dropouts (gray areas in Figure 4.3C) for the further identification of features affecting the quantitative difference and dropouts events, respectively.

Tab. 4.2: Summary of the generated gene-specific features

Feature name	Description	Sources
compartment	subcellular localization	COMPARTMENTS [10]
TSS	transcription start site	BioMart [108]
GC-content	guanine-cytosine content ratio	BioMart [108]
cDNA length	length of complementary DNA	BioMart [108]
transcript count	number of transcripts	BioMart [108]
chromosome	chromosomes/scaffolds id	GTF (Ensembl v.86)
transcript length	length of transcript	GTF (Ensembl v.86)
3'UTR length	length of 3'untranslated region	GTF (Ensembl v.86)
5'UTR length	length of 5'untranslated region	GTF (Ensembl v.86)
CDS length	length of coding DNA region	GTF (Ensembl v.86)
biotype	transcript variant type (ex. coding)	GTF (Ensembl v.86)

For this purpose, a diverse data corresponding to genes were integrated from different sources, such as COMPARTMENTS [10], BioMart [108] and GTF (Ensembl v.86) annotation data [26]. Since the chosen features were of both numerical and categorical types, the latter ones were transformed into numerical representation by enumerating the categories and assigning the corresponding indices. As a result, we created a set of 10 prospective features representative for matched sequencing experiments.

Since machine learning models greatly benefit from high quality training data, it is necessary to assure such a property of the input data. One of the factors that can be potentially harmful to ML-models is multicollinearity, i.e. statistically non-independent relationships between features. In order to determine whether additional feature selection is required as a

pre-processing step before the use of the FAVSeq pipeline, we performed a cross-correlation analysis that allowed to assess the occurrence of high inter-correlations among independent variables. Its results suggest (Supplementary Figure 4.6C) that filtering of the features is not necessary, since they do not exhibit strong dependencies between each other [31]. Additionally, we performed the comparison of pairwise feature correlations according to Pearson and Spearman (upper rights and lower left in Figure 4.6C, respectively) that suggests the existence of non-linear relationships between some features (e.g., $|\rho^{GC\ content}_{Spearman}| > |\rho^{GC\ content}_{Pearson}|$), which we consider later while choosing among suitable machine learning models. With the aim to disclose a subset of the most relevant features for the aggregated target difference, we performed an ML-assisted analysis.

4.1.10 Identification of factors affecting the gene expression difference in RNA-Seq experiments

Using the proposed FAVSeq, we analysed the influence of generated features on bulk vs. scRNA-Seq gene expression differences calculated for the matched experiments. For that, we considered 5134 genes with no missing values in any tested features (data of 0% sparsity; Table 4.3).

The model performance on the gene expression difference prediction task was assessed using MSE that served as a loss metric. Additionally, we used simple linear regression as a baseline in order to show how the errors are measured on the same data in comparison with those calculated based on analysis using random forest with optimized hyper-parameters. Comparison of the performance of regression models indicate the random forest as the most suitable model to predict the quantitative difference between experiments

Tab. 4.3: Sparsity of features containing missing values for different genes subsets

Feature name	Genes with missing values (#)	
	Sparsity 17.5%	Sparsity 0%
5'-UTR length	4 266	–
3'-UTR length	3 830	–
transcript counts	3 806	–
GC content	3 806	–
TSS	3 806	–
compartment	3 631	–
CDS length	385	–
Genes total (#)	**13 427**	**5 134**
Dropouts (#)	**804**	**36**
(minority class, %)	**(6.0 %)**	**(0.7 %)**

Sparsity: ratio of missing values (NaNs) in the feature space. 0 % indicates cleaned up data (no NaNs).

based on the set of tested features (Figure 4.7A). Here, the best performance was achieved using an ensemble of 512 trees with the depth of 8. During the model training, at least 16 samples were needed before splitting tree's internal nodes, while leaf nodes contained 4 or more samples.

Subsequently, we derived an optimal subset of relevant features using recursive feature elimination based on random forest (RFE-RF) to obtain rankings within differently sized subsets of features in a 5-fold cross-validation. Figure 4.7B indicates how well the difference can be explained by the tested subset of features based on the RFE-RF approach. The cross-validation loss curve (lower is better) shows that the model error decreases with the number of features used and reaches the global minimum when using a subset of size 9. Then we determined the optimal number of features through the search for the first stationary point of the objective function by calculating second-order differences, which indicate the steepest drop of the curve. We see noticeable deceleration of the drop of the loss curve after the feature set's

size reaches value of 3, indicating these features as the most important ones w.r.t. the difference discovered between the RNA-Seq experiments (dashed vertical line in Figure 4.7B).

The same model was then used to rank features w.r.t. their influence on changes in the target variable. As random forest ranking considers both feature set completeness and non-linear interactions within it, we drawn the conclusions about the feature importance upon its scores. Closer look at the importance scores indicates the major impact of three particular features—3'-UTR, transcript length and GC content—on the quantitative difference between matched RNA-Seq experiments, as also shown in Figure 4.7C. In total, the aforementioned factors are responsible for more than $51\,\%$ of the entire relevance of the features for the regression target.

In order to provide a more in-depth understanding on how these features may vary between subsets of genes, we calculated first and second central moments of the top-identified features for the least and most different genes according to the previously calculated measure of the difference in the matched experiments, as shown in Table 4.4. Interestingly, we observe the shortened length of 3'-UTRs in the most different genes, suggesting the higher level of gene expression as well as the possible differences in their mRNA metabolisms (e.g., sub-cellular localization, stability and the rate of translation of mRNAs [82, 77]).

Tab. 4.4: Characteristics associated with top-relevant identified features in least and most different genes according to the analysis of matched RNA-Seq experiments

Statistics (Mean±SD; N=100)	**3'-UTR length**	**Transcript length**	**GC content ratio**
Least different genes	944.8±1 059.5	21 051.2±22 175.2	47.4±7.9
Most different genes	506.7±787.6	48 851.2±62 090.0	49.6±7.2

Identification of factors affecting the presence of dropouts in RNA-Seq experiments

In the previous Sections, we focused on quantitative expression differences between single cells and bulk RNA-sequencing. Here, we applied the similar approach introduced in the FAVSeq pipeline for the classification task in order to identify the most relevant factors associated with the occurrence of dropouts. We asked whether the generated features in Table 4.2 affect the presence of dropouts (qualitative difference) in RNA-Seq experiments and therefore we adapted our machine learning approach to this group of genes. For this analysis we chose 13,427 genes, including 804 global dropouts, representing expression measurements across all samples in the opposite experiment (Figure 4.3C, gray areas). As the features generated in the previous step consist of missing values (Table 4.3), these values were imputed using a k-NN model as it was shown to be more accurate on dropouts prediction task (Figure 4.6D).

Unlike regression, the classification task implies estimation of discrete (class labels) rather than continuous values. In a such setup, one has to take into account class imbalance, since strong domination of one class may introduce severe bias into the model's predictions. As the presence of a dropout—and the corresponding class in the data—is a quite rare event ($< 6\,\%$), the complementary class, namely the absence of a dropout, is highly over-represented, since the corresponding samples occur roughly 15.6 times more frequently (Table 4.3). To address this issue, one can utilize the oversampling approach for the training data or choose a model, those training procedure allows to re-weight classes assigning higher importance to the under-represented class (dropouts). The latter is possible using an MLP model trained using weighted cross-entropy loss as the optimization objective. In order to account the class imbalance during the performance evaluation,

we also applied class-wise averaged recall (a.k.a. balanced accuracy) to assess the model's prediction accuracy. Results of the feature ranking w.r.t. global dropouts found in both experiments indicate that cellular compartments, 3'-UTR and transcript lengths affect the presence of dropouts in the combined bulk and scRNA-Seq dataset (Figure 4.7D).

The best model performance was achieved for the MLP containing 512 hidden neurons trained using learning rate $\eta = 0.01$ and regularization term $\alpha = 0.0001$. Exact values of the aforementioned hyper-parameters were obtained through the grid search over a predefined hyper-parameter space. Furthermore, we utilized this approach to analyse features w.r.t. their influence on experiment-wise global dropouts (including 477 and 327 in scRNA-Seq and bulk, respectively). Latter results indicate the similar factors to be relevant for dropouts in individual single-cell (Figure 4.7E) and bulk (Figure 4.7F) RNA-Seq, while the 3'-UTR exhibits a particular importance for dropouts in the single cell experiment.

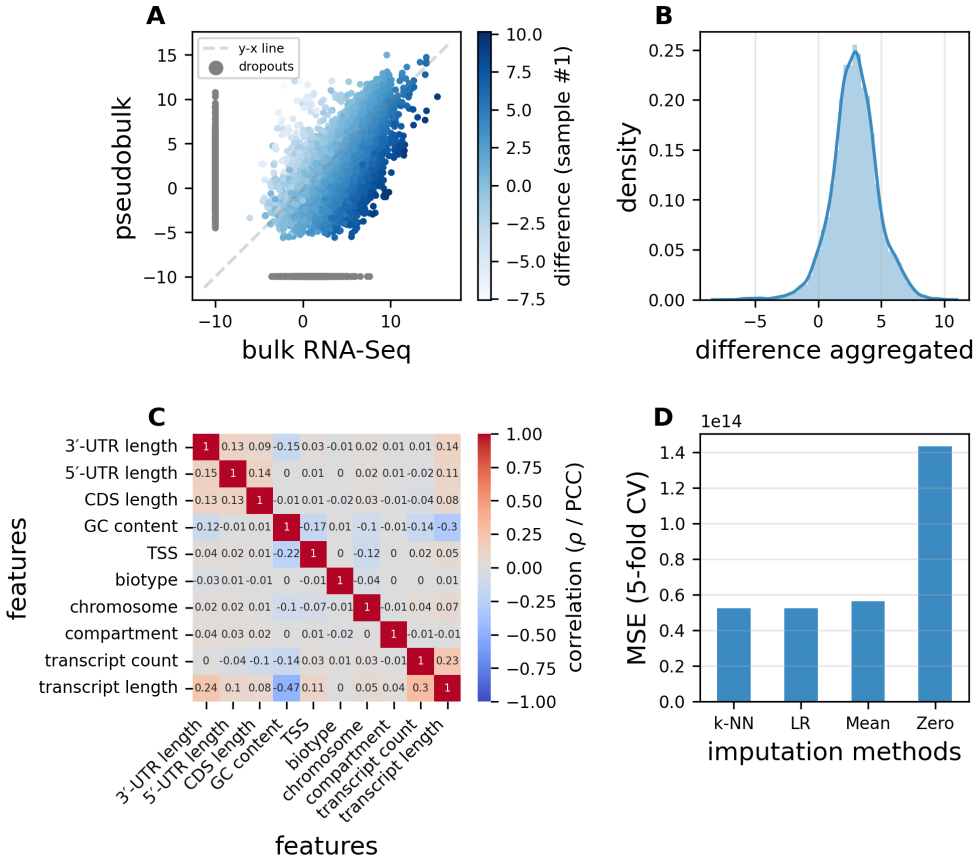

Fig. 4.6: Exploratory analysis of generated features and the target difference used in the FAVSeq pipeline. A. Gene expression difference—quantitative and dropouts-related—between the bulk and aggregated scRNA-Seq datasets measured in an individual sample for the most confident 13,427 genes detected in either of the matched experiments. Blue color indicates the most confident common genes detected in both matched experiments and gray indicates global dropouts (12,623 and 804 genes, respectively). Log-transformed values are shown. B. Distribution of the quantitative difference aggregated across 8 samples and calculated for the most confident common genes expressed in both RNA-Seq experiments. C. Cross-correlation analysis of gene-specific features filtered w.r.t. missing values (sparsity 0% in Table 4.3) and generated for the most confident 5,134 genes detected in either of RNA-Seq experiments. Color bar represents the distribution of correlation coefficients (Spearman rank and Pearson in lower left and upper right triangles, respectively), with grey color indicating values close to 0 (weak correlation). D. Performance of imputation approaches (k-Nearest Neighbors as k-NN, linear regression as LR, mean- and zero-based methods) done based on 5-fold CV in the feature reconstruction task and performed for all features with 20% of NaNs introduced manually in the training subset. Y-axis represents MSE that serves as an average error measure (lower is better) for the evaluated imputation methods.

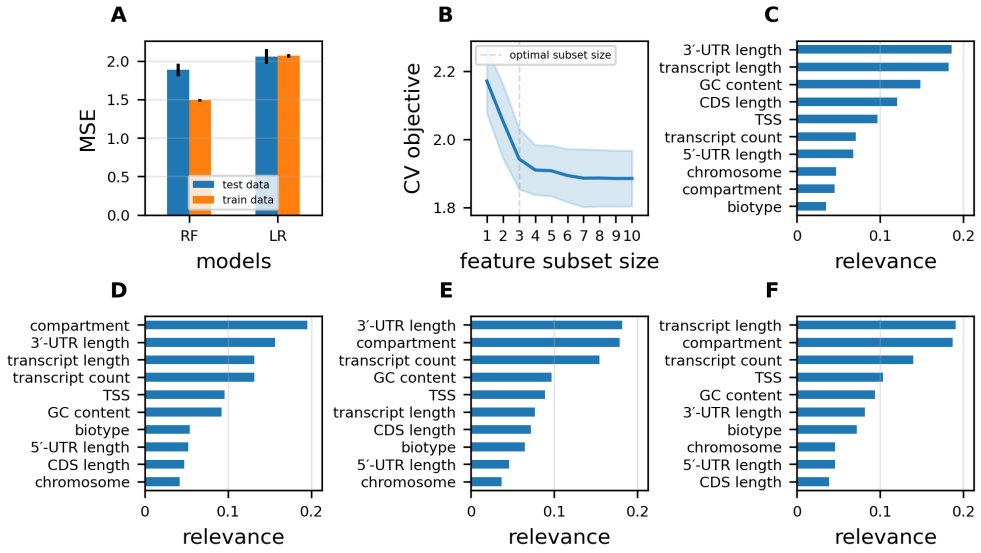

Fig. 4.7: Identification of features contributing to the quantitative and qualitative (dropouts) differences in the matched RNA-Seq experiments. A. Comparison of the performance of machine learning models—random forest (RF) and linear regression (LR)—in prediction of the quantitative difference based on tested features in terms of MSE (lower is better) in 5-fold CV. B. Solution on the optimal number of features based on RFE-RF used to examine feature importance across different subset sizes (x-axis). Cross-validation loss (y-axis) indicates how the model performs based on differently-sized subsets of features in 5-fold CV, showing the steep drop in mean loss values around the subset of three features. Dashed vertical line indicates the optimal number of features. C. Ranking results for features affecting the quantitative difference between experiments. X-axis indicates feature relevance values derived using the RFE-based-on-RF model and y-axis indicates feature names. D. Ranking results for features affecting dropouts in both experiments and derived using RFE-MLP. E. Ranking results for features affecting dropouts in scRNA-Seq and derived using RFE-MLP. F. Ranking results for features affecting dropouts in the bulk RNA-Seq and derived using RFE-MLP.

4.1.11 Concluding remarks

In this thesis we investigated sources of variation in RNA-Seq profiles obtained from the same population of biological replicates based on different RNA sequencing technologies. Based on the analysis of matched single-cell and bulk RNA-Seq data, we found high similarity in gene expression measurements for the majority of genes and also discovered that this difference is primarily subjected to a technical variation given the negligible effect of biological variation in the tested setup. In addition, we performed an in-depth analysis of dropouts which were found to be systematically present in both experiments and to be not explained by low-expression genes only, as it was generally accepted in the preceding studies.

Further analysis of multimodal RNA-Seq data using the proposed computational framework (FAVSeq) allowed to identify key factors (i.e., genomic and transcriptomic) affecting quantitative (gene expression levels) and qualitative (dropouts) difference across experiments. In particular, the performed ML-based analysis revealed that 3'-UTR and transcript lengths affect gene expression difference the most. Subsequently, we identified features associated with the occurrence of dropouts and found out the same features together with cellular compartments to be relevant for presence of global dropouts as well. We also demonstrated how to reconstruct missing values in data generated from metadata and genomic databases based on the k-NN approach.

To conclude, using the proposed computational framework, we investigated sources of variability across RNA-Seq experiments, which, in turn, builds a basis for improving interpretability and reducing complexity of downstream in-depth analysis of gene expression data provided by different RNA-Seq technologies.

4.2 Machine learning-assisted identification of alternative splicing regulators using RNA-Seq from cancer patients

4.2.1 Preamble

This work is the part of my doctoral research in the computational biology unit at the Institute of Molecular Biology (IMB) in Mainz. My contribution to the work is the following: I have designed and implemented the computational framework, performed statistical and ML-based analysis, summarized results and findings in the manuscript. The developed bioinfomatics software package regulAS used in this work can be installed from PyPI, with the open-source implementation being provided in the GitHub repository. Technical report on regulAS can be found at `https://arxiv.org/abs/2307.08800`.

4.2.2 Key highlighs

The regulation of alternative splicing (AS) is a highly complex process, which is frequently perturbed in diseases such as cancer. At the molecular level, splicing is regulated by RNA-binding proteins (RBP) that bind to pre-mRNA and control the recruitment of the splicing machinery. In order to investigate regulatory mechanisms of splicing decisions, a computational framework for identifying candidate modulators was designed to dissect complex relationships between alternative splicing events and thousands of potential RBPs using large-scale RNA-Seq from public omics data sources (TCGA and GTEx).

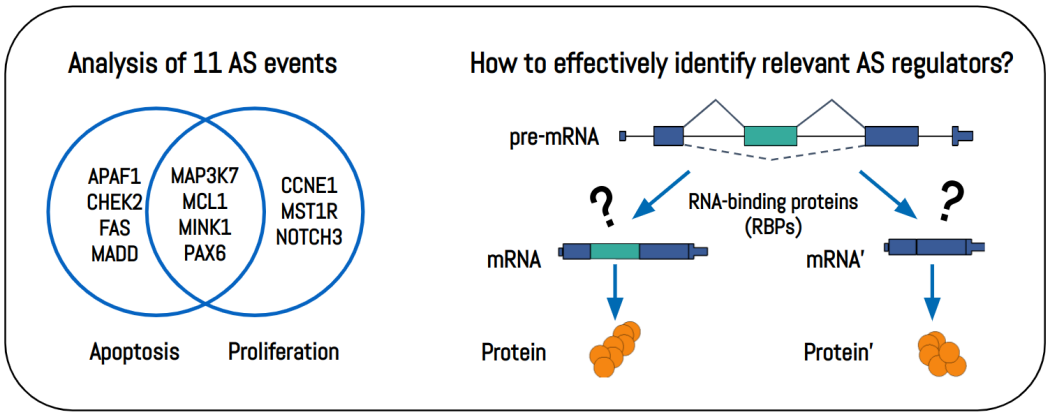

Fig. 4.8: Objectives for IDA and ML-based analysis of alternative splicing events across different conditions using RNA-Seq data from TCGA and GTEx.

This work presents an analysis of eleven apoptotic- and proliferation-related genes with exon-skipping events, whose splicing is perturbed in a variety of physiological conditions and tumors (Figure 4.8). Specifically, the focus is on the skipping event of exon 11 in the proto-oncogene MST1R (RON) which leads to constitutive activation of the resulting protein in multiple cancers. Based on the developed approach, we identified and experimentally validated RBPs controlling generation of the tumorigenic isoform RONΔ165, including SON, HNRNPH, and RBM3, that affect its splicing changes. Further analysis of additional datasets for AS genes across cancer and normal tissues allowed to identify novel RBP candidates associated with splicing machinery regulation, suggesting directions of in-depth studies towards a better understanding of RBPs-mediated regulatory mechanisms of alternative splicing changes across tumor/normal conditions and tissue types.

4.2.3 Background

Splicing is a conserved biological process implying transformation of primary transcripts (pre-mRNA) into its mature forms (mRNA) that can direct synthesis of multiple protein isoforms during the subsequent translation. Alternative splicing of pre-mRNA is regulated by RNA-binding proteins defining which exons are included in the resulting transcripts. RBPs act as trans-factors controling the splice site choice by binding to cis-acting elements (binding sites of RBPs), followed by a repressed or enchanced splice site recognition and a spliceosome assembly. Regulation of mRNA alternative splicing by RBPs is crucial for generating biological diversity in mammalian genomes, and this mechanism is especially complicated in pathological conditions [39, 125]. Splicing perturbations are common in cancers, resulting in an altered expression of specific trans-regulatory factors associated with a tumorigenic state. However, mRNA-RBP interactions are highly variable, and regulatory mechanisms of splicing in tumor progression that involve combinations of a large number of RBPs remain incomplete [122, 13].

One example of the regulation of cis-regulatory elements in alternative splicing by RBPs is exon 11 in the proto-oncogene MST1R (RON). RON exon 11 (RONΔ165) functions as a so-called cassette exon which is either included or excluded during splicing. The neighboring exons are constitutive so that the inclusion isoform contains exon 10-12, whereas skipping gives rise to a short isoform comprising only exon 10 and 12. Inappropriate pre-mRNA skipping of this exon changes protein function and results in an expression of the constitutively active RONΔ165 isoform, which promotes tumor progression and is associated with metastatic cancer phenotypes. RONΔ165 is up-regulated in solid tumor tissues, including ovarian, pancreatic, breast and colon cancers [132]. Previous studies have reported several RBPs to regulate RON splicing [43, 14, 84], and the further investigation of its splicing decisions in tumor

cells allows to shed light on its molecular mechanisms under pathological conditions.

Analysis of alternative splicing events allows to understand requirements for correct selection of exons. There are several methods reported to identify regulators of AS using RNA-Seq data. For instance, it is common to identify RBP-mediated splicing events by knocking down or over-expressing the RBP and performing RNA-Seq [27, 72, 73, 6, 111]. These methods are, however, time-/resource-costly and limited by the number of prospective candidates for the analysis. In addition, regulatory mechanisms of AS are highly complex, as the multi-dimensional and noisy nature of transcriptomic expression data makes the analysis very challenging [19, 111]. One of the powerful tools to simultaneously characterize alternative splicing outcomes and RBP expression patterns in a genome-wide manner is next-generation sequencing that made available gene expression data for thousands of potential regulators of AS. However, there are no standardized computational approaches to dissect ones that are the most relevant for splicing decisions, thus decreasing complexity of the subsequent wet lab analysis.

All of that motivated two streams of this research: (1) identification of relevant regulators of alternative splicing changes using machine learning and transcriptome sequencing data and (2) analysis of regulatory mechanisms of splicing decisions in alternatively-spliced genes (including RON exon 11) across tumors. Furthermore, this thesis proposes a design and implementation of an ML-assisted pipeline for integrative omics analysis to dissect complex relationships between RNA expression profiles of putative regulators from cancer and healthy human tissues derived from publicly available TCGA and GTEx genome projects data.

4.2.4 Proposing ML-based toolset for integrative analysis of alternative splicing regulome using RNA-Seq

Theoretical and empirical knowledge about alternative splicing mechanisms is crucial for understanding processes underlying biological diversity. Regulation of alternative splicing implies that numerous splicing regulators (e.g., RBPs) interact with corresponding cis-acting regulatory sites on the target pre-mRNA. In order to analyze these complex relationships between RBP expression and splicing outcomes, regulAS, a machine learning (ML) assisted tool was developed to dissect the most important effectors of changes in the alternative splicing machinery (Figure 4.9). Given the information on splicing event efficiency for an exon of interest (e.g. RON exon 11 skipping) calculated as PSI for each tissue sample (e.g., non-/tumor) and the RBP gene expression data measured as RNA-Seq (in the same samples), it allows the analysis of prospective splicing regulators using feature selection approach by deriving the relevance of RBPs as part of the learning algorithm [44]. In order to identify key AS regulatory factors from thousands of candidates, different regression models were evaluated to predict PSI and to find the most impactful RBP candidates.

The proposed regulAS tool incorporates model selection and assessment of feature importances providing a flexible workflow for performing ML experiments on RNA-Seq data (Figure 4.9). The working process is structured into subsequent steps that include: (1) Extraction, transformation and loading of RNA-Seq of perspective regulatory RBPs and RNA splicing profiles data (incl., filtering RNA-Seq data by read coverage, selecting tumor/normal/combined samples, selecting tissue types); (2) Training and evaluation of machine learning models to predict splicing efficiency (incl.,hyperparameter optimization, selection of top-performance setups); (3) Ranking of putative RBPs using ML

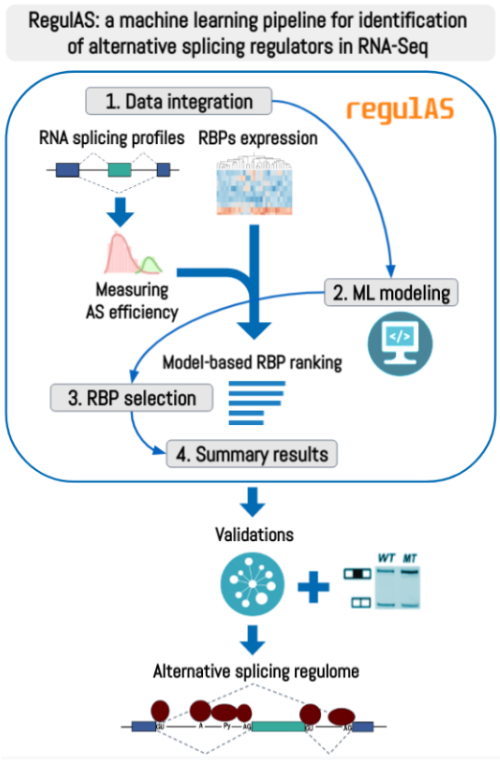

Fig. 4.9: Overview of the experimental workflow utilizing regulAS for the identification of relevant regulators of splicing events using RNA-Seq expression profiles of putative RBPs data across conditions and tissue types. Computational workflow consists of four major steps: Data extraction, transformation, and loading; Model training, evaluation, and selection; RBP (feature) ranking and selection; Summary report generation. During the validation phase, the output reports can be utilized for downstream tasks, such as functional enrichment analysis, protein-protein interaction networks, further laboratory validation of top regulatory candidates identified.

models from (2) with respect to their influence on splicing changes; (4) Generation of summary reports reflecting performed experimental analysis (incl., model training and validation results as images and/or in tabular forms; conversion of Ensemble gene ID to its symbolic aliases for visuals). The flow of data- and ML-pipelines is orchestrated by an YAML-based configuration file that defines implementation details for each aforementioned step, thus unifying those individual tasks into end-to-end processes (Figure 4.10).

Tuning of hyper-parameters of ML-models is being performed using a Grid search approach, those elements are marked in the configuration file explicitly. Here, each combination of hyper-parameters spawns a task that is submitted into a pool of parallel workers, thus increasing the speed of the tuning process linearly, with the number of CPU-cores. The information on the exact transformations and models is being assembled by the core of regulAS, then structured and stored into a database. After a task successfully finishes, its results—including model predictions and raw feature ranks—are being added to the database as well. Such careful tracking of all performed steps ensures completeness of experimental data, while the use of a relational database provides guarantees on the data integrity. The stored data can be later retrieved in order to generate the resulting reports in tabular and graphical forms. We provide implementation of the proposed experimental pipeline bundled with sample report generators, which outputs were used in this work.

To identify splicing-associated factors involved in skipping of RON exon 11, we selected a candidate panel of gene expression data of regulatory RNA binding proteins from open-access genome project databases. Resulting RNA-Seq data contain expression values of thousands of RBPs across cancer patients alongside with alternative splicing decisions in the same samples.

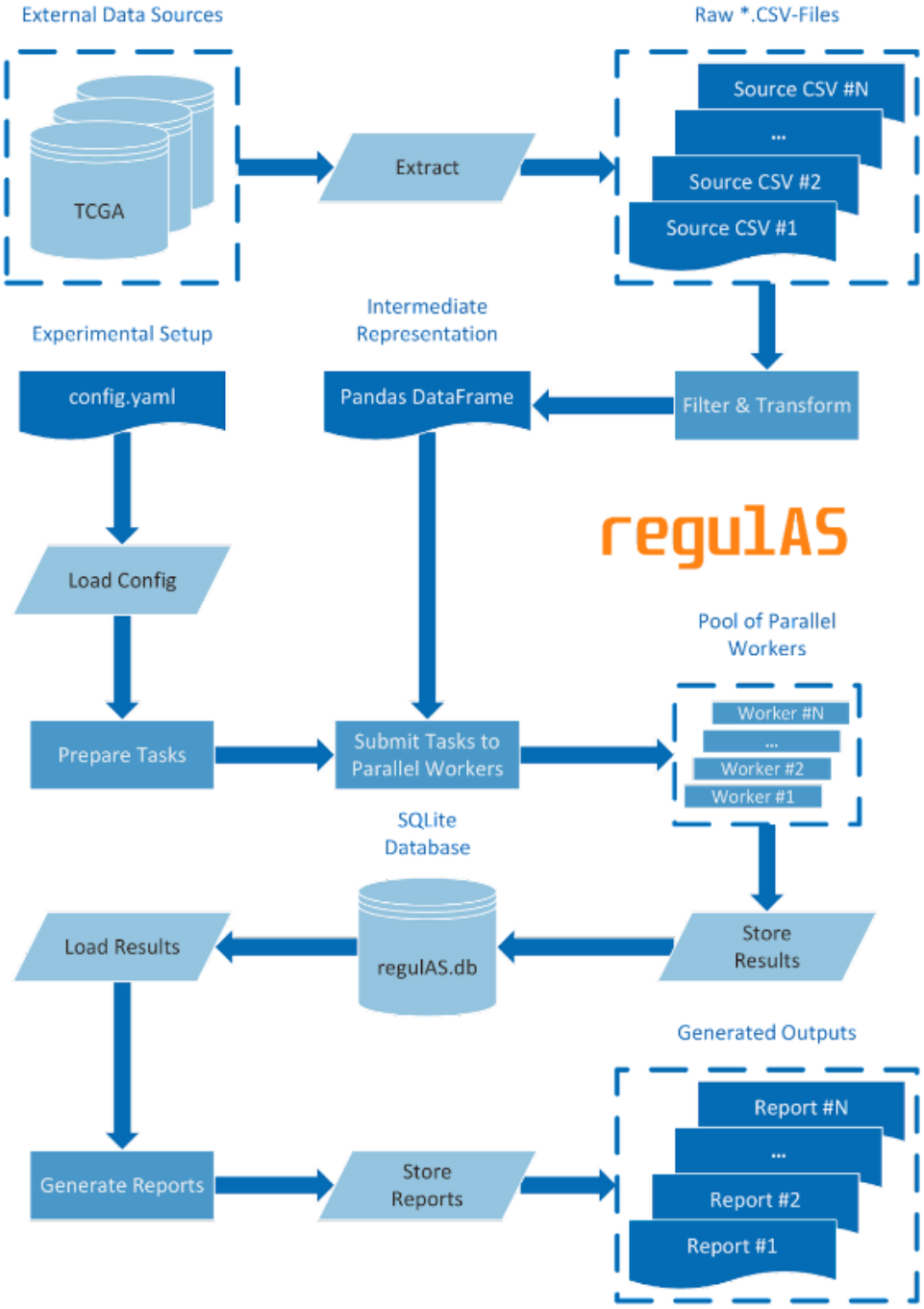

Fig. 4.10: Workflow of regulAS, a machine learning-assisted tool for integrative analysis of alternative splicing regulome using large-scale RNA-Seq from multiomics public data sources. To identify relevant regulators for splicing changes across conditions and tissues, gene expression profiles for genes of interest are being retrieved from TCGA and GTEx repositories.

4.2.5 Exploring RBP expression and RON splicing profiles across a panel of tumors

To investigate the regulatory effect of tissue-specific gene expression of bulk RBPs on splicing outcomes in RON, we analyzed transcriptome profiles derived from thousands of human donors using data from multi-omics open-source data repositories. RNA-Seq data of prospective regulators and junction data (used to calculate splicing profiles later on) in tumor patients were retrieved from the TCGA multi-omics data source and using the Snaptron searching engine [128], correspondingly. Following data collection, RNA-Seq data have been further pre-processed to choose the most quality data (e.g., by handling missing values in RNA-Seq as described below) and improving the efficiency of machine learning algorithms.

In order to filter out non-informative data, features (RBP expression) with missing values in the majority of the data samples (patient cohorts)—more than 70% are NaNs—were removed (RP11-257K9.8, RNASE9, RBMXL3), resulting in 2036 remaining RBPs considered for further analysis. Samples of tissues with the minimum size of 50 [47] have been selected to ensure that the limited sample size won't affect the performance of machine learning models (Figure 4.11A). Data thresholding with respect to the minimum sample size per tissue allows to introduce a sufficient amount of information needed to deduce properties of an underlying distribution from each tissue during the machine learning analysis step and also to be able to further investigate cancer type-specific regulators. Then, log-transformation of RBP expression values was done to decrease the range of values of sequencing data and make it conform closely to the normal distribution. To improve reliability of splicing estimates, only patients harboring at least 5 read counts for all splice junctions were considered for further analysis (Figure 4.11B).

Therefore, filtering of the unreliable data improved quality of the data and enabled further comparative analysis across different sources. Finally, the percentage of spliced in (PSI) was calculated based on junctions data in order to measure alternative pre-mRNA splicing changes in exon 11, which represented the fraction of inclusion reads relative to the total amount of exon inclusion plus skipping (Figure 4.11C).

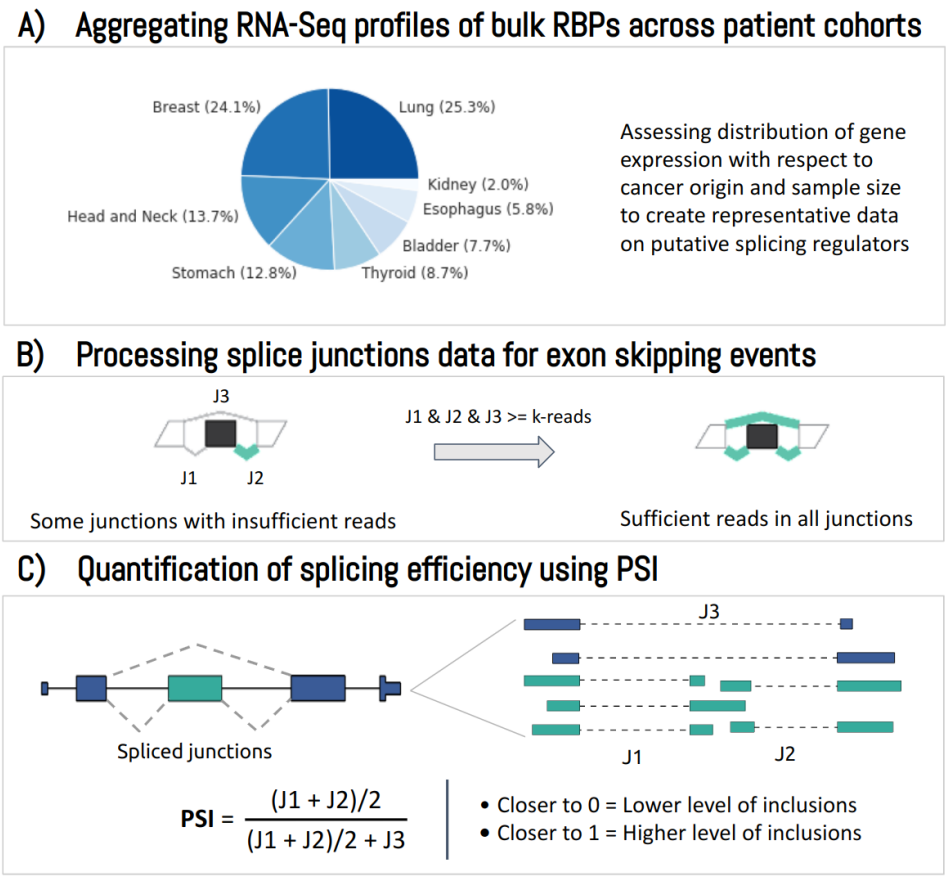

Fig. 4.11: RNA-Seq processing steps for RBP expression and splice junctions data applied in the study of RON AS. *(A)* Processing of RNA-Seq profiles of putative regulators across cancer tissues to create representative data. *(B)* Processing of RNA junctions data to subset samples with respect to the read coverage. *(C)* Quantification of splicing efficiency using PSI. J1-3 correspond to junction reads, while dividend and quotient represent proportions of inclusion and skipping to the total number of splicing events, respectively.

The dataset for the RON analysis comprised information on 1621 patients across various cancer cohorts considering gene expression of 2036 putative RBPs and PSI values of RONΔ165 (Figure 4.12).

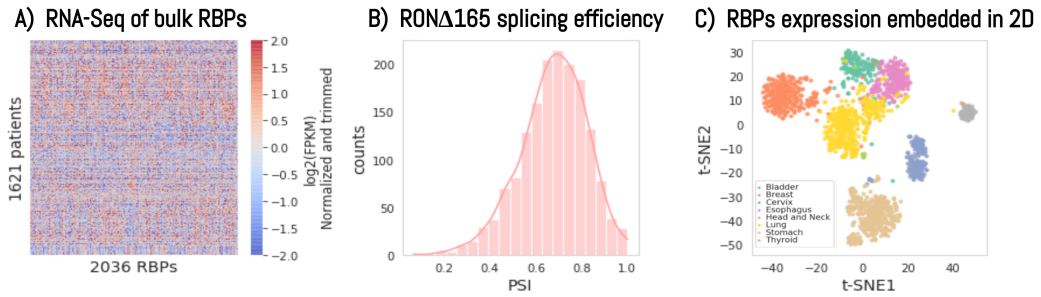

Fig. 4.12: Gene expression of putative regulators and splicing profiles from thousands of cancer patients for the analysis of RONΔ165 alternative splicing based on TCGA data. *(A)* RBP expression of putative regulators of RONΔ165 splicing. Expression level of each RBP is scaled to the same minimum and maximum range across all samples. *(B)* Distribution of RONΔ165 splicing efficiency measured as PSI across thousands of tumor patients. *(C)* Expression of RBPs across tumors embedded in 2D using t-SNE. Data is grouped by tumor origin, incl., breast and lung tissues.

As shown in Figure 4.12A, RBPs are heterogeneously expressed across different tumor patients, showing variable expression in tissues (Figure 4.12C). t-SNE embedding of gene expression data by tissues indicates distinguishable groups of RBPs corresponding to different tumors, as shown in Figure 4.12C. PSI values of the exon 11 in RON vary between 0.06 and 1 across samples, i.e., the dataset covers high and low inclusion values (Figure 4.12B). The median of PSI values is 0.68, indicating the averaged high inclusion rate of the exon 11 in the analyzed patients. Based on these findings, we reasoned to identify splicing-associated factors involved in skipping of RON exon 11 at the most. The resulting dataset was further used to perform transcriptome-wide study of regulatory mechanisms of splicing alterations for RON in tumors using ML-based modelling approach.

4.2.6 Prediction of RON splicing efficiency based on RBP expression data

Model training and evaluation on the PSI regression task were done using regulAS As a part of the data transformation, subsets of the training and test data were obtained by randomly sampling and splitting the data with a test size of 20%. The tuning of hyper-parameters of ML-models was performed using 5-fold cross validation and Grid search. We sought to select the most suitable estimator to predict splicing efficiency of the exon of interest considering different model families. The best model minimizes mean squared errors (MSE) and the goodness of models can be compared with respect to the loss function (lower is better). Comparison of performance metrics of eight models with optimized hyper-parameters to predict RONΔ165 splicing efficiency based on gene expression of prospective regulatory RBPs is shown in Figure 4.13. For RON data, the following models were evaluated - Bayesian regression, extra trees, gradient boosting, linear regression, linear regression with L1 and L2 regularization (ElasticNet), neural network, random forest, Support Vector Machine (SVM). To choose the most suitable model, these algorithms were trained and evaluated in 5-fold cross validation, with the data split into 4 training and 1 test folds.

As shown in Figure 4.13A, neural network minimizes mean squared errors on the unseen (test) data at the best. Figure 4.13B shows how well the model was able to predict the splicing efficiency of RON exon 11 based on analyzed transcriptome profiles of bulk regulators. Results indicate three models—neural network, ElasticNet, and SVM regression (SVR)—as the most suitable ones to predict PSI based on the analyzed data, showing the lowest errors on test (previously unseen) data. For the remaining seven models we observe negligible errors while fitting the training data, however, the errors

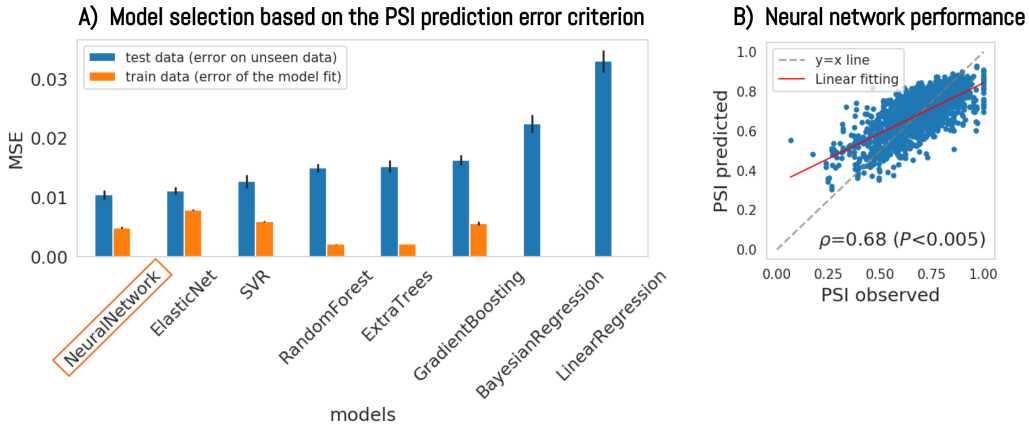

Fig. 4.13: Performance of ML models in predicting splicing efficiency of RONΔ165. *(A)* Comparison of different algorithms (x-axis) based on the PSI prediction error criterion (y-axis). Models are ordered ascendingly with respect to the value the loss function (MSE) on the test data (blue); smaller test MSE corresponds to a better model. Results indicate a neural network as the top-1 performance model to predict PSI for RON data analyzed. *(B)* Prediction ability of the trained neural network to predict PSI on the unseen test dataset. Spearman's rank correlation (ρ) and linear fitting line (red) are indicated for assessment of observed and predicted PSIs.

are high for test data. So, these models exhibit proneness to overfitting and they were not considered for further analysis. Thus, through comparing the performance of ML models on PSI prediction error criterion, the neural network model was selected as the most optimal model for predicting PSI for RON data, and therefore we applied it for feature ranking as discussed in the next section.

4.2.7 Identification of key RBPs controlling the exon 11 skipping in RON

The selection of most relevant RBPs of the alternative splicing event was done using models generalizing to the test data the best. At this point, the assumption is that RBPs with a higher rank with respect to splicing alterations,

as defined through these models, are more likely to be significant regulatory factors. To identify RBPs that influence RONΔ165 splicing changes based on the neural network's output the most, the absolute values of gradients of loss function with respect to individual RBPs were accumulated and sorted descendingly giving the importance ranking (relative importance), as shown in Figure 4.14A for the top-10 relevant RBPs found.

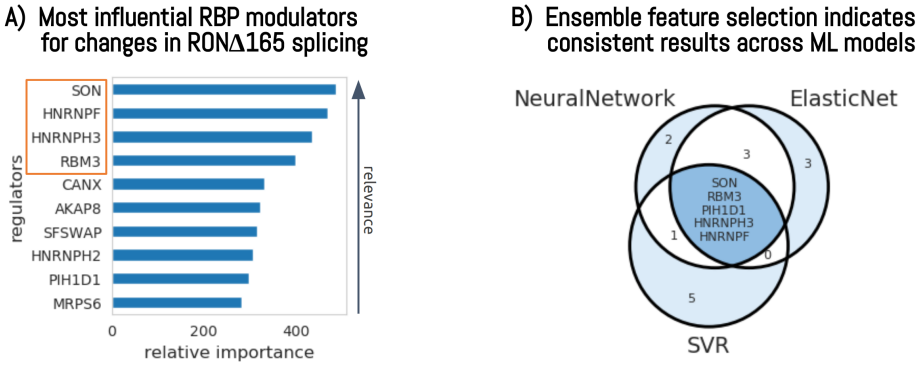

Fig. 4.14: Ranking results derived using ML-approach indicate most influentual candidate regulators for changes in alternative splicing for RONΔ165 data. *(A)* Neural network indicates SON, RBM3 and hnRNPs as the most relevant RBPs for changes in RON exon 11 AS in tumors. Relative importance (x-axis) represents the absolute gradient values or relevance weights. *(B)* Ensemble feature selection assessment based on three top-performance ML models for RON confirms these findings. Set analysis is based on top ranking results derived using a Neural Network, an Elastic Net and a Support Vector machine Regression (SVR) algorithms. Most relevant RBPs—according to consensus ranking results from ML models of different families—are indicated in the intersection and include SON, RBM3 and hnRNPs for tumor samples.

Then, ranking results from three top-performance models defined in the previous model selection step have been aggregated in order to exploit independence between different learning algorithms, to decrease biases for these models and to produce an optimal generalized ranking decision by averaging. RBP ranking using SVR was done as described in [45]. Here, an SVR model aimed to find a set of coefficients of a multidimensional linear function (weights) that best fited the training data points. Subsequently, the importance of the features was determined based on the learned weights,

or support vectors. As shown in Figure 4.14B, the ensemble ranking model based on top-15 results produces consistent results indicating SON, RBM3, hnRNPs (H3/F) as the most significant factors influencing alternative splicing changes in RONΔ165 based on tumor data.

4.2.8 Confirming candidate RBPs to be significantly associated with mRNA splicing machinery

The following analysis was done to assess what are the associations of identified RBPs with RNA splicing and mRNA processing. In order to infer the role of top identified candidates to regulate alternative splicing of RONΔ165, gene set enrichment analysis (GSEA) was performed based on data queries from the Enrichr search engine, summary results of those statatistical analysis are indicated in Figure 4.15A.

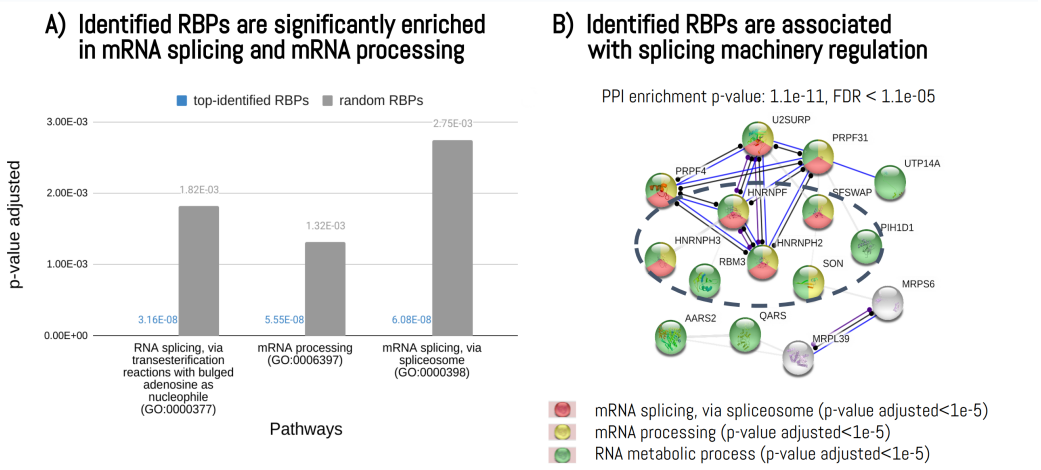

Fig. 4.15: Associations of identified RBPs with RNA splicing and mRNA processing. *(A)* GSEA results for the subset of the most informative vs. randomly sampled RBPs. *(B)* Network analysis of molecular associations between top-20 selected regulators derived using the neural network model demonstrates identified RBPs to be significantly associated with mRNA splicing regulation and mRNA processing. Molecular associations are represented by edges, while nodes are individual RBPs.

Network analysis of protein-protein interactions (PPI) of the subset of top-identified RBPs was then performed based on open-access STRING database to further validate the above findings, demonstrating that these regulators to be associated at the molecular level, indicaing the presence of interactions (shown as edges) between individual RBPs (shown as nodes), as depicted in Figure 4.15B. For instance, HNRNPH2 is known to be associated in regulation of RNA splicing as well as in pre-mRNA processing [40]. Additionally, structurally similar HNRNP H2 and F (showing stronger PPI associations as indicated in Figure 4.15B) have been implicated to regulate gene expression binding to similar sequences, confirming these RBPs to be associated at the molecular level [1]. Overall, the obtained results indicate the derived RBPs to be associated with mRNA splicing regulation and mRNA processing.

The identified here RBPs were confirmed to interact with known splicing regulators frequently altered in cancers (PRP, U2) or are themselves described as such (hnRNPH) [14, 35]. Moreover, top-rank RBP candidates form a PPI network, suggesting their role in RON splicing regulation. Therefore, identified RBP candidates to regulate RON exon 11 AS under tumor progression are proven to be involved in splicing machinery regulation and also may act together at the molecular level to modulate this process.

4.2.9 Experimental validation of SON, RBPM3, hnRNPH as modulators of RON△165 splicing

To determine whether alternative splicing changes in RON are actually driven by the derived RBPs, we performed siRNA knockdown (KD). KD experiment is usually performed to estimate the relevance of target regulators for the splicing efficiency by depletion of gene expression of a gene of interest. In

order to assess whether the depletion of identified RBPs has the impact on the alternative splicing of RON exon 11, measuring mRNA of isoform levels of RONΔ165 under KD-RBPs and performing RT-qPCR (Figure 4.16). Four splice isoforms (skipping, inclusion, full intron retention, and partial intron retention) were measured by capillary electrophoresis in MCF7 cells in control conditions and upon transfection with siRNAs targeting RBPs identified to be relevant based on our analysis.

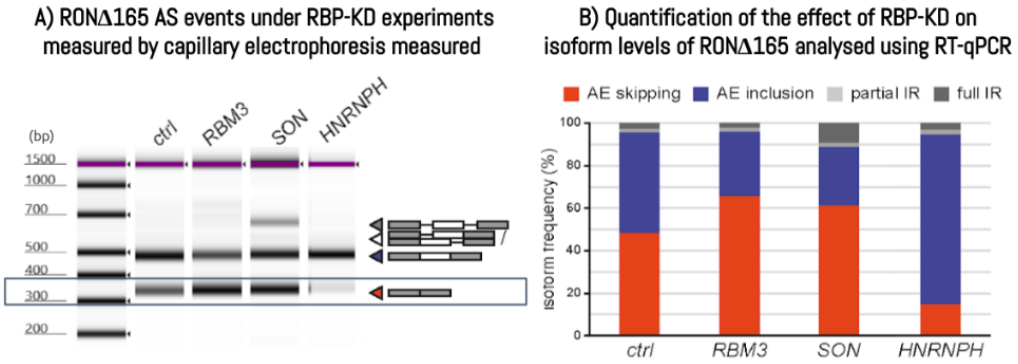

Fig. 4.16: Knockdown experiments confirm a role of identified RBPs in RON alternative splicing. *(A)* RON mRNA expression profiles in regard to exon 11 skipping under RBP-KD experiments measured by capillary electrophoresis measured in MCF7 cells in non-KD control conditions (ctrl) and upon transfection with siRNAs targeting RBPs (SON, RBM3, hnRNPH). Scheme of the splicing isoform structures is indicated at the right. Intensity represents the relative mRNA levels of different isoforms (darker is higher). Skipping of exon 11 generates its tumorigenic isoform RONΔ165 of 300 bp. *(B)* Quantification of the effect of RBP-KD on isoform levels of RONΔ165 analyzed using RT-qPCR. Colors represent ratios of differently-spliced isoforms of the exon 11 (exon skipping, exon inclusion, full intron retention, and partial intron retention). Isoform level of RON under non-KD conditions was used as a control (ctrl). Isoform changes of its skipped exon 11 affected by RBP-KD are indicated by a red color.

Figures above indicate the effect of individual top-relevant RBPs (SON, RBM3 and hnRNPH) on changes in the proportion of different alternatively spliced isoforms of RON, while the isoform level under non-KD conditions is used as a control (ctrl). Figure 4.16A indicates pronounced changes at the level of RON isoform with the skipped exon 11 (300-bp-isoform highlighted by a red

arrow) w.r.t. the control condition, meaning that the identified RBPs are likely to have an effect on RON splicing. Similarly, results of the RT-PCR experiment shown in Figure 4.16B indicate changes in the proportion of different RON isoforms under the depletion of individual RBPs (isoforms with skipped exons are indicated by a red color). Thus, these experimental results validate SON, RBM3 and hnRNPH as important regulators of RONΔ165 alternative splicing affecting its isoforms.

4.2.10 Deciphering alternative splicing events in apoptosis and proliferation related genes

To demonstrate how to leverage the proposed computational framework for the analysis of context-specific alternative splicing regulome, the study of proto-oncogene RON was extended to additional AS genes whose splicing is disrupted in various conditions and diseases. Here, our primary objective was to explore the potential of machine learning-assisted approach in capturing biological and functional associations based on large-scale RNA-Seq for thousands of RBP candidates in connection with alternative splicing to identify those most relevant for splicing decisions. Also, we wanted to investigate how AS mechanisms can be altered in tumors compared to normal tissues for a range of apoptotic- and proliferation-related genes. Among others, the following questions were covered: (1) Can we reveal transcriptomic variations that are associated with tumor states in the context of splicing alterations using public data from different omics sources? (2) Which RBP candidates for splicing changes can be identified as most relevant using machine learning approach based on RNA-Seq expression data from human tissues across cancer cohorts? (3) How do tumor-associated tissue-specific AS regulatory factors differ across methods of analysis? For the latter, to

estimate the impact of a specific RBP on splicing efficiency, we then performed a qualitative and quantitative assessment of "in silico" and "in vitro" based relevance scores using set analysis, normalized Discounted Cumulative Gain (nDCG) [17], functional enrichment and PPI characterization of most influential RBPs identified across different experimental settings.

In order to identify relevant regulatory RBPs affecting splicing events in tumors, 11 genes related to apoptosis and proliferation with exon skipping events were used for in-depth analysis of transcriptome profiling data from human tissues. Among these, for instance, Apaf1, CHEK2, Fas, and MADD are well-known apoptosis regulators, whose alternative splicing has been implicated in enhanced or reduced apoptosis-activity. One of these genes is Fas/CD95, also known Fas (later as FAS), a transmembrane cell surface protein binding to its ligand FASL, the inclusion of those exon 6 generates a membrane-bound receptor that promotes apoptosis in tumor patients [114]. The regulation of its splicing was shown to be modulated by RBPs such as RBM5, TIA-1 and EWS [94]. Another interesting candidate for the analysis of AS regulatory machinery is MAP3K7, belonging to mitogen-activated protein kinases and is considered as one of central regulators of cell fate decisions during developmental processes and apoptosis. Alternative splicing results in skipping of its exon 12 giving a tumorigenic isoform. Regulators RBM47 and ESRP1/2 were shown to promote the exon inclusion in MAP3K7 [131].

To study relevant factors for splicing alterations considering distinctions between tumor and healthy states, combined datasets of RBP expression and RNA junctions data for the genes of interest were created by aggregating tumor and normal samples by tissues across patient cohorts from TCGA and GTEx repositories. Here, we focus on the analysis of more than 150 protein components of the spliceosome with gene expression measurements from cervix adjacent tissues, given the supported siRNA-RBP knockout (KO)

screens based on cervical adenocarcinoma HeLa cells [93] for the AS genes analyzed, thus, allowing reliably compare experimental setups later on.

4.2.11 Exploring splicing patterns in transcriptome profiling data for tumor and normal samples

Upon examination of the constructed datasets from TCGA and GTEx data, differential patterns reflecting splicing changes in non- and cancerous conditions were identified. Splicing efficiency was measured as percent-spliced-in ratio based on RNA splice junctions (Equation 3.5). As indicated in Table 4.5, there is a high degree of statistically significant difference ($\alpha = 10^{-3}$) in changes in exon skipping events between tumor and normal samples for the majority of genes analyzed (8 out of 11).

Notably, for some genes, PSI values represent bimodal distribution across different sample types. Specifically, APAF1, CHEK2, MAP3K7, and MINK1 demonstrate a lower mean PSI in tumor conditions, in comparison to normal samples. PSI changes for FAS, MADD, MST1R, and NOTCH3 are in the opposite direction. According to the significance reported by the two-sided Welch's t-test, for CCNE1 and MCL1 there is no significant difference between PSI values across both conditions. For PAX6, the t-test could not be performed, as RNA-Seq profiles corresponding to this gene were limited to cervix adjacent tumor tissues based on combined data dataset explored, thus, causing an insufficient number of samples for tissues considered in this case.

We can also see that—as the PSI values averaged across healthy and tumor-associated samples, respectively—the existence of at least two modes in the distribution that are clearly identifiable for a subset of genes of interest (namely, APAF1, CHEK2, FAS, MADD, MAP3K7, MINK1, MST1R, and

Tab. 4.5: Assessment of splicing efficiency patterns across conditions for cervix adjacent human tissues based on TCGA and GTEx data. Results of a two-sided Welch's t-test are given for equality of mean PSI values between tumor and normal samples at the significance level $\alpha = 10^{-3}$ (number of degrees of freedom within the data is indicated in the column df). The obtained test statistics t indicate that the PSI values change for the majority of genes of interest (8 out of 11). Within this group, the change in PSI is positive ($\Delta\mu > 0$) for APAF1, CHEK2, MAP3K7, and MINK1, while for FAS, MADD, MST1R, and NOTCH3, PSI tends to decrease when going from tumor to normal conditions. For CCNE1 and MCL1, test does not indicate a significant difference between PSI values for both conditions. For PAX6, there were no test results reported, as the corresponding gene expression profiles data do not supply information on healthy samples.

Gene	t	df	Significance (2-tailed)	Mean Diff. ($\Delta\mu$)	Std. Error Diff. ($\sigma_{\Delta\mu}$)	95% CI of $\Delta\mu$	
						Lower	Upper
APAF1	4.9371	167.1030	$1.91{\cdot}10^{-6}$	0.1045	0.0212	0.0627	0.1462
CCNE1	-0.0525	8.1192	$9.59{\cdot}10^{-1}$	-0.0016	0.0308	-0.0724	0.0692
CHEK2	14.1268	237.7319	$2.62{\cdot}10^{-33}$	0.1341	0.0095	0.1154	0.1528
FAS	-7.7186	500.8742	$6.43{\cdot}10^{-14}$	-0.0583	0.0076	-0.0731	-0.0435
MADD	-21.9325	332.2647	$1.47{\cdot}10^{-66}$	-0.2259	0.0103	-0.2462	-0.2056
MAP3K7	21.8038	242.9562	$4.02{\cdot}10^{-59}$	0.0198	0.0198	0.3935	0.4716
MCL1	2.9375	342.4902	$3.53{\cdot}10^{-3}$	0.0055	0.0019	0.0018	0.0093
MINK1	25.8755	280.4404	$2.84{\cdot}10^{-76}$	0.3973	0.0154	0.3671	0.4276
MST1R	-4.7112	66.2374	$1.31{\cdot}10^{-5}$	-0.0850	0.0181	-0.1211	-0.0490
NOTCH3	-8.6835	535.0853	$4.66{\cdot}10^{-17}$	-0.0725	0.0084	-0.0889	-0.0561
PAX6	–	–	–	–	–	–	–

Tab. 4.6: Summary data analysis results for 11 genes with apoptotic and proliferative activity altered through alternative splicing and used for discovering prospective regulators across paired tumor and normal human samples. Pearson correlation coefficient (PCC) for test and train subsets is calculated for observed vs. predicted PSI values averaged across triplicates based on the ML-assisted analysis of combined tumor and normal samples. The value of 1 means a positive linear relationship between "in vitro"- and "in silico"-based PSI values measured for tumor and normal cervix adjacent tissues. Here, a functional effect of alternative splicing (AS) reflects changes in exon skipping across the genes of interest.

Gene	Exon	AS Effect	Samples		PCC	
			tumor	normal	test	train
APAF1	18	Apoptosis disruption	213	101	0.81	0.87
CCNE1	9	Cell cycle progression vs. Protein inactivity	740	9	0.85	0.89
CHEK2	8	Apoptosis disruption	736	176	0.87	0.91
FAS	6	Apoptosis disruption	476	253	0.77	0.84
MADD	24a	Apoptosis vs. Cell viability	426	251	0.90	0.93
MAP3K7	11a	Apoptosis vs. Cell viability	732	170	0.82	0.87
MCL1	2	Apoptosis inhibition	775	265	0.92	0.95
MINK1	18	Apoptosis vs. Cell viability	754	265	0.93	0.95
MST1R	11	Constitutively active form vs. Normal signaling	402	48	0.15	0.45
NOTCH3	16	Cell proliferation	670	238	0.93	0.95
PAX6	5a	Apoptosis vs. Proliferation	134	0	0.92	0.95

NOTCH3), so we can hypothesize that the underlying difference can be captured using machine learning methods. Thus, the observed difference in splicing patterns between tumor and healthy states for AS genes suggests exploring splicing alterations that may be associated with specific cancer-related factors, which we will investigate using regulAS.

Table 4.6 summarizes details on analyzed data, including descriptive statistics related to the corresponding RNA-Seq datasets generated (e.g., number of tumor samples), and the output statistics derived using ML-modelling experiments (e.g., performance metrics). functional roles of alternatively-spliced exons of analyzed genes are indicated as AS Effect. For instance, through alternative splicing of the exon 6, Fas can generate isoforms with different functions that mediate or inhibit apoptosis in cancer [114].

In the computational experiments, Pearson correlation coefficient (PCC) was used to indicate goodness-of-fit of the model to the data. Table 4.6 provides measurements quantified for training and test datasets used to assess how the predicted PSI values are close to the ones used for model fitting and evaluation, respectively. Performance evaluation results indicate ability of NN models to learn underlying dependencies between bulk RBP expression and target PSI data during the training phase, as well as to generalize to previously unseen samples for data with sufficient amount samples of paired tumor-normal samples available.

4.2.12 Discovering tissue-specific splicing modulators across “in silico” and “in vitro” experiments

To identify tissue-specific regulatory factors affecting splicing decisions for 11 genes with apoptosis and proliferation related genes, different approaches were utilized for RBP selection. For ML-based capture of relevant RBPs for splicing changes, relevance scores were derived using neural network models that were trained on data containing information on gene expression levels of putative regulators measured in tumor and healthy human tissues. The top candidate regulators derived from the ML-based approach were then compared with siRNA-induced knockdown results by [93]. For RBP ranking, a robust Z-score [11] was computed to estimate isoform ratios upon siRNA, thus, indicating the impact of a given knockdown in the regulation of AS [11]. As the authors have provided RBP relevance scores upon normalized difference between control and altered tumor samples stemming from the same tissue, we used them as reference ranks and constructed “in silico” study design in a similar way.

RBP selection based on ranking results from computational experiments was done by computing normalized differences between RBP relevance scores obtained for healthy and tumor-associated splicing efficiency for exon skipping events. In particular, a neural-network-derived RBP relevance score is described by Equation 4.1 and Equation 4.2.

$$\nabla^{i}_{L_j} = \frac{\partial L_j}{\partial X_i},\ \nabla_L \in \mathcal{R}^{N_{RBP} \times N_{iter}}, \tag{4.1}$$

$$Rel_i = \frac{avg_j\left(\overset{T}{\nabla}{}^{i}_{L_j}\right) - avg_j\left(\overset{H}{\nabla}{}^{i}_{L_j}\right)}{\hat{\sigma}_j\left(\overset{H}{\nabla}{}^{i}_{L_j}\right)}, \tag{4.2}$$

where $\nabla^{i}_{L_j}$ is a gradient of loss (L_j) with respect to input (X_i).

Here, during the training phase, raw gradients of input (RBP expression profiles) with respect to the loss value (Equation 4.1) were accumulated separately, for healthy (H) and tumor-associated (T) samples. Then a robust Z-score was calculated using measurements from healthy samples as a baseline for comparison, which yielded the relative importance of individual RBPs for predicting PSI in tumor-associated samples compared to healthy ones (Equation 4.2).

Next, to compare ranking results obtained from different experiments, the following metrics were considered: nDCG, Spearman's rank correlation, and the number of top-rated RBPs those ranks were available for both approaches. nDCG served as a benchmarking metric to assess how similar the results obtained through ML-based approach were to those produced by RBP-KO experiments, for which the most relevant items were positioned at the forefront of the list (Equation 3.6). The choice for the nDCG score was conditioned by

Tab. 4.7: Qualitative and quantitative assessment of the most influential factors affecting alternative splicing changes in cervical-associated tumors derived through different approaches for RBP selection. The normalized Discounted Cumulative Gain (nDCG) and Spearman's rank correlation (ρ) serve as metrics to measure similarity between "in silico" (neural networks based on tumor TCGA and normal GTEx samples from human patients) and "in vitro" (RBP-KO screen based on HeLA cells [93]) rankings for the top-25 relevant RBPs obtained via robust Z-score. nDCG values closer to 1 indicates consistency between orders of items across top ranks, ρ closer to 1 indicates similarity between two sequences. For the top results from each ranking, lists of common RBPs are provided.

Gene	ρ	nDCG	Matches	Influential RBPs Common
APAF1	-0.35	0.44	4	SLU7, U2AF1, SF3B1, DDX5
CCNE1	-0.10	0.33	5	SNRPC, DHX38, SF3A3, LSM2, LSM3
CHEK2	-0.14	0.35	6	U2AF1, U2AF2, SNRPB2, SF3A3, SF3B1, SF3B3
FAS	-0.03	0.47	6	SF3B3, SNRPD3, SF3B1, LSM6, PRPF19, RBM25
MADD	-0.24	0.55	6	SF3B3, MFAP1, SART1, SRRM2, TDRD9, DDX24
MAP3K7	0.43	0.41	3	MFAP1, DHX35, SART1
MCL1	0.01	0.38	3	CHERP, SNRPD2, SF3B4
MINK1	-0.25	0.45	5	U2AF2, RNPS1, U2AF1, PQBP1, SNRPB2
MST1R	-0.01	0.64	5	PRPF19, RNPS1, SNRPD3, RBM17, CRNKL1
NOTCH3	-0.04	0.45	7	SNRPB2, DDX5, RNPS1, SF3B3, LSM6, DDX41, LSM4
PAX6	0.02	0.49	4	EWSR1, U2AF2, SRSF1, PRPF4B

the fact that the RBP relevance scores form an ordered set, where one of the sets serves as a reference, while another one represents predictions derived from neural network models. Moreover, the exact relationship between the two rankings does not follow a monotonicity assumption, and the nDCG score takes into account this factor, unlike correlation-based metrics [91].

Table 4.7 shows the comparison of "in silico" (model system: neural network based on RNA-Seq gene expression data from TCGA tumor and GTEx normal samples from human cervix adjacent tissues; relevance units: robust Z-score) and "in vitro" (model system: siRNA RBP-KO screen based on cervical cancer HeLa cells [93]; relevance units: robust Z-score) RBP relevance ranking results. Here, the number of common regulators allows to measure the strength of the association between each of the ranking results and the actual size of the overlapping subset of regulators. As the nDCG score is associated stronger with the number of correctly predicted key regulators,

this metric was used to assess consistency across ranking results for more in-depth analysis of prospective RBPs. Higher nDCG values indicate more similar ranks in terms of relevance and position of features compared to the reference scores.

The obtained results indicate congruence of top-25 identified RBPs with respect to Z-scores for the majority of results. Figure 4.17 shows that the number of matches for some AS events analyzed—NOTCH3, CHEK2, FAS, MADD—varies between 6 and 7 for top-25 rankings, with around 5 matches across all the events. Although the sample size doesn't allow for comparing and assessing the degree of a relationship between in silico and in vitro ranking methods, the PPI analysis revealed for the "in silico" rank suggest a non-random nature of the discovered key RBPs. Specifically, PPI network analysis results of top-10 identified RBPs for these 4 genes suggest presense of biological relationships among these groups of RBPs, with PPI enrichment p-value ranging from $<1.0 \cdot 10^{-16}$ to to $7.31 \cdot 10^{-10}$.

Next, to identify novel relevant regulators for alternative splicing decisions in the context of ML experiments and not to be constrained by availability of other scores, feature ranking results were obtained based on NN-based apporach, as summarized in Table 4.8. PPI analysis was then performed to examine the degree of involvement of a set of top relevant RBPs identified in AS-associated pathways and reveal whether they interact in the context of alternative splicing regulome (Table 4.8). Gene set enrichment analysis indicated biological coherence among top RBP candidates for all the AS genes of interest analyzed. Evaluation of network interaction further confirmed the identified regulators to be significantly associated with a major mRNA splicing pathway across the genes analyzed in the setting of alternative splicing alterations considering tumor samples.

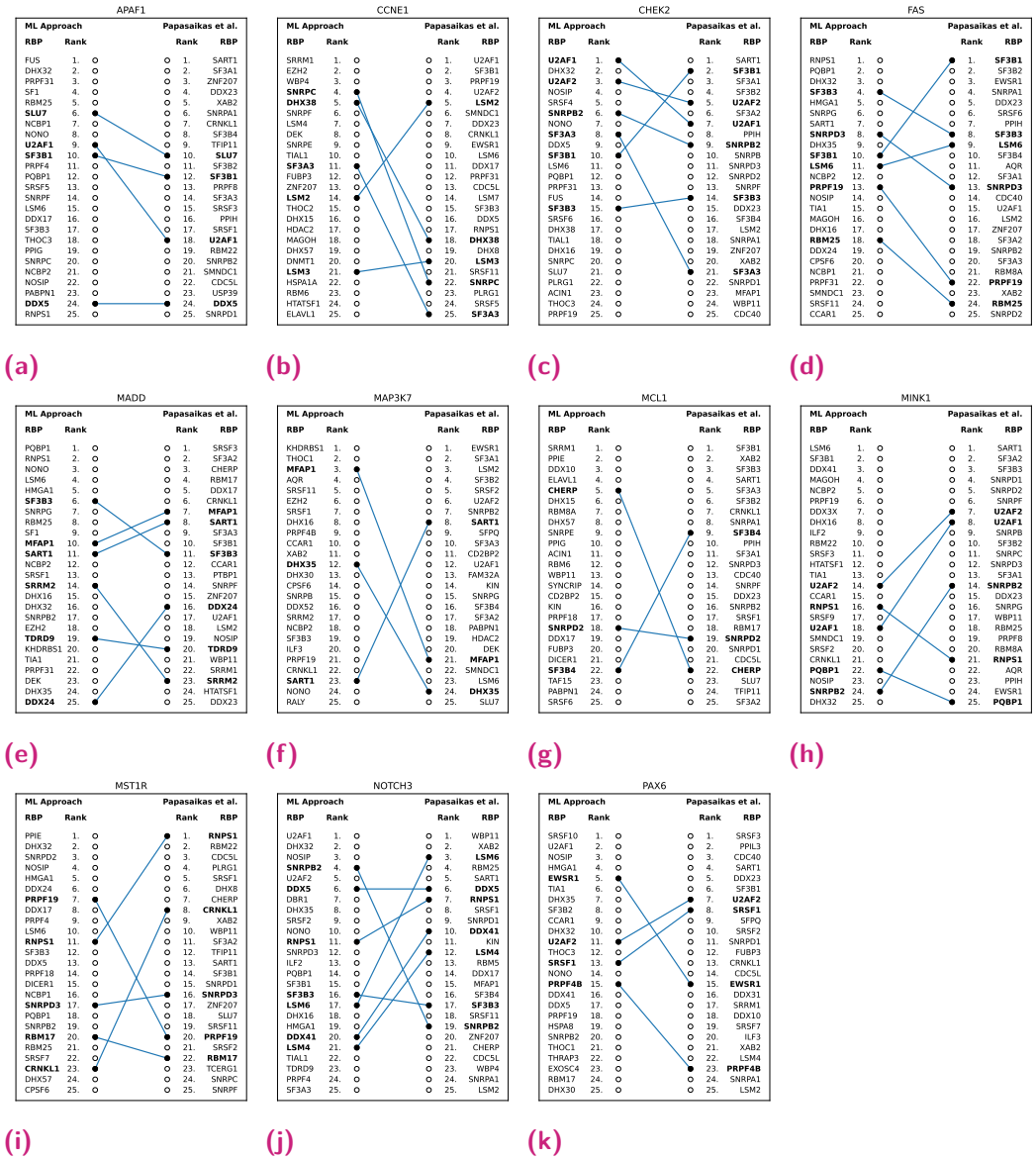

Fig. 4.17: Comparative analysis of top relevant RBP modulators for splicing changes obtained using "in silico" and "in vitro" experiments (Table 4.7). ML-based approach utilizes neural network models for deriving relevance scores based on transcriptome profiles from thousands of putative RBPs from TGCA and GTEx human patient samples (Equation 4.2). Another approach for RBP selection is based on RBP-KO screen based HeLa cells, with relevance scores being measured for median PSI indexes for each KD-AS pair [11]. Left panel presents RBP ranking results derived using ML-based approach, right panel contains ranks based on siRNA screen from [93]. Bold font indicates influential RBPs that simultaneously appear in top-25 results across experimental settings.

Summarizing the analysis results, we can conclude that the ML-assisted identification of relevant RBPs for changes in splicing events can assist downstream studies and lead to a better understanding of alternative splicing regulome across conditions and tissue types. Further study of selected candidates in the context of splicing changes provides more insights into the mechanisms of its regulation.

Tab. 4.8: Discovering tissue-specific splicing modulators for 11 genes with apoptotic and proliferative activity altered in alternative splicing for cervical tumor adjacent tissues. RBP relevance scores Equation 4.2 are obtained using neural network models based on TCGA and GTEx data. Protein protein interation (PPI) assessment and functional enrichment analysis performed for top-10 relevant RBPs suggest presense of biological connections between identified candidates forming networks, as well as their association with mRNA splicing as a whole. The enrichment p-value for PPI network and false discovery rate (FDR) for pathway enrichment analysis ($\alpha = 0.05$) are provided to indicate confidence levels for hypothesis testing in confirming physical and functional associations based on individual ranking results derived using ML-based approach via regulAS.

Gene	PPI Network (p-value)	mRNA Splicing Pathway (FDR)	STRING Network Cluster (FDR)	Top-10 relevant RBPs
APAF1	$<1.0{\cdot}10^{-16}$	$1.52{\cdot}10^{-11}$	$U2_C$ and $mRNA_S$ ($4.30{\cdot}10^{-4}$)	FUS, DHX32, PRPF31, SF1, RBM25, SLU7, NCBP1, NONO, U2AF1, SF3B1
CCNE1	$2.22{\cdot}10^{-16}$	$3.76{\cdot}10^{-9}$	$U2_C$ and Sm ($3.15{\cdot}10^{-5}$)	SRRM1, EZH2, WBP4, SNRPC, DHX38, SNRPF, LSM4, DEK, SNRPE, TIAL1
CHEK2	$7.77{\cdot}10^{-16}$	$3.76{\cdot}10^{-9}$	$U2_A$ ($1.30{\cdot}10^{-4}$), snRNP ($1.30{\cdot}10^{-4}$)	U2AF1, DHX32, U2AF2, NOSIP, SRSF4, SNRPB2, NONO, SF3A3, DDX5, SF3B1
FAS	$<1.0{\cdot}10^{-16}$	$1.52{\cdot}10^{-11}$	U4/U6 x U5 ($3.24{\cdot}10^{-5}$)	RNPS1, PQBP1, DHX32, SF3B3, HMGA1, SNRPG, SART1, SNRPD3, DHX35, SF3B1
MADD	$<1.0{\cdot}10^{-16}$	$1.52{\cdot}10^{-11}$	U4/U6 x U5 ($3.24{\cdot}10^{-5}$), snRNP ($1.67{\cdot}10^{-9}$)	PQBP1, RNPS1, NONO, LSM6, HMGA1, SF3B3, SNRPG, RBM25, SF1, MFAP1
MAP3K7	$9.15{\cdot}10^{-10}$	$1.88{\cdot}10^{-9}$	$U2_C$ and $mRNA_S$ ($1.40{\cdot}10^{-5}$)	KHDRBS1, THOC1, MFAP1, AQR, SRSF11, EZH2, SRSF1, DHX16, PRPF4B, CCAR1
MCL1	$4.67{\cdot}10^{-11}$	$3.76{\cdot}10^{-9}$	$U2_C$ and Sm ($3.30{\cdot}10^{-3}$)	SRRM1, PPIE, DDX10, ELAVL1, CHERP, DHX15, RBM8A, DHX57, SNRPE, PPIG
MINK1	$<1.0{\cdot}10^{-16}$	$1.52{\cdot}10^{-11}$	$U2_C$ and $mRNA_S$ ($8.30{\cdot}10^{-6}$)	LSM6, SF3B1, DDX41, MAGOH, NCBP2, PRPF19, DDX3X, DHX16, ILF2, RBM22
MST1R	$3.20{\cdot}10^{-10}$	$6.88{\cdot}10^{-5}$	$U2_C$ and $mRNA_S$ ($5.83{\cdot}10^{-6}$), U4/U6 x U5 ($5.83{\cdot}10^{-5}$)	PPIE, DHX32, SNRPD2, NOSIP, HMGA1, DDX24, PRPF19, DDX17, PRPF4, LSM6
NOTCH3	$2.62{\cdot}10^{-12}$	$6.12{\cdot}10^{-7}$	CRD and $mRNA_P$ ($1.31{\cdot}10^{-2}$)	U2AF1, DHX32, NOSIP, SNRPB2, U2AF2, DDX5, DBR1, DHX35, SRSF2, NONO
PAX6	$4.74{\cdot}10^{-8}$	$6.88{\cdot}10^{-5}$	–	SRSF10, U2AF1, NOSIP, HMGA1, EWSR1, TIA1, DHX35, SF3B2, CCAR1, DHX32

Abbreviations: $U2_C$ and $mRNA_S$ – U2-type spliceosomal complex, and mRNA Splicing - Major Pathway; $U2_C$ and Sm – U2-type spliceosomal complex, and Sm-like protein family complex; $U2_A$ – U2-type prespliceosome assembly; snRNP – Spliceosomal snRNP complex, and mRNA cis splicing, via spliceosome; U4/U6 x U5 – U4/U6 x U5 tri-snRNP complex; CRD and $mRNA_P$ – mRNA processing, and CRD-mediated mRNA stability complex.

4.2.13 Concluding remarks

In this thesis, we have demonstrated an application of the proposed computational framework for identifying relevant factors affecting alternative splicing changes using large-scale RNA-Seq from human samples. By analysing relationships between gene expression of prospective regulators and splicing profiles for a range of apoptosis and proliferation related genes, those functioning is altered in alternative splicing, the proposed approach allows to identify RBPs affecting alterations of exon-skipping events across cancer cohorts, thus leading to a better understanding of RBP-mediated AS regulatory mechanisms across conditions and tissues.

Our results revealed that SON, HNRNPH and RBM3 affect the skipping exon 11 events at the most for proto-oncogene RONΔ165 across cancer cohorts. The assessment of network-based protein interactions among the top RBP candidates suggest that they can be involved in a common RBP-RNA regulatory network to coordinate alternative splicing of RON in cancers. Further laboratory study of the identified RBPs in the context of alternative splicing will allow to better understand the molecular mechanisms of splicing regulation under different pathologic conditions and tissue types. For instance, as RONΔ165 is aberrantly expressed in breast cancer [130], conducting an extended analysis—such as siRNA-mediated knockdown in MCF7 cells—for the most relevant regulatory candidates can shed light on tissue-specific regulation of its splicing decisions.

The following are the main contributions of the chapter to the thesis: (1) This work applied advanced data analysis methods to learn complex relationships between gene expression of putative regulators and splicing efficiency patterns for exon skipping events using transcriptomics data from public TCGA and GTEx sources. (2) We performed a comprehensive assessment of

predictive modelling approaches on PSI prediction task using RBPs exression data, allowing to find optimal setups for downstream analyses, such as RBP selection in a supervised way, beyond simple statistical methods. (3) We applied ML-based approach utilizing neural networks to identify relevant factors affecting splicing decisions for tumor data across patient cohorts. (4) We performed an ensemble-based feature selection applied to RBP ranking results derived using ML algorithms of different families—neural network, support vector machine, regularized linear regression—that indicated consistency in relevance scores among top candidates for RON AS. (5) We then validated top-rank RBPs experimentaly through siRNA knockdown and pathway enrichment analysis to confirm identified candidates to be associated with splicing machinery regulation, suggesting their functional role in AS mechanisms for RONΔ165 under tumor conditions.

In this chapter, we also presented an in-depth analysis of the most influential RBPs for 11 AS genes between "in silico" and "in vitro" experiments. To assess impact of individual RBPs for splicing changes across different settings, relevance of the top candidates on splicing changes was measured based on the robust Z-score [11] upon results derived using both the ML-based (neural network-derived weighing based on TCGA tumor and GTEx normal samples) and gene knockdown (siRNA screen based on HeLa cancer cell line [93]) approaches. We next used the RBP selection results obtained from NN-based scoring for cancer patients data to reconstruct a network of functional interactions and perform pathway enrichment analysis [113]. The analysis revealed regulatory potential of top candidates among core spliceosomal complex components and mRNA splicing pathways, as shown in more detail in Table 4.8.

To summarize, the proposed approach can be further applied to identify relevant factors affecting splicing alterations, thus leading to a better un-

derstanding of RBP-mediated regulatory mechanisms of alternative splicing under pathological conditions and tumor origin. Moving forward, such a computational approach leveraging public datasets from large-scale omics studies offers a wide variety of opportunities for splicing and cancer research, allowing discovery of relevant affectors of AS changes from a broad range of possible candidates, thus facilitating downstream analysis tasks in a less time- and resource-demanding manner. Hence, these results can support decision-making when establishing laboratory experiments to further validate regulatory mechanisms of splicing alterations in “in vitro” settings.

5

Discussions

This thesis presents results on integrative analysis of transcriptomics data with a focus on: (I) in-depth examination of factors contributing to variability in multimodal omics of bulk and scRNA-Seq and (II) studying RNA expression profiles from human samples for investigating RBP-mediated regulatory mechanisms of alternative splicing across tumor and normal conditions. For each type of analysis, the thesis proposes design and implementation of a dedicated tool encapsulating functionality related to study factors relevant for either (I) variability in RNA-Seq data measured by different technologies (FAVSeq) or (II) alternative splicing events across cancer cohorts using open-source multi-omics data repositories (regulAS).

While “in vitro” studies can be stringent to examining a limited number of putative gene candidates under a highly controlled setting provided by, for instance, a model system and laboratory conditions that do not reflect diversity of regulatory elements and environmental contexts manifested in vivo [37], the use of computational approaches leveraging public omics and clinical data has a great potential to address these challenges. Specifically, machine learning-assisted identification of relevant gene candidates for splicing decisions using large-scale RNA-Seq data from human tissues provides a more inclusive approach for assessing factors affecting variability under different conditions, thus overcoming some of these limitations in characterizing the regulatory landscape in the genome.

The computational pipelines that were developed as a part of this thesis are intended to address specific biological challenges. Nevertheless, these approaches can be applied more broadly and be further extended to accommodate complementary omics modalities and data sources adjusted for the specifics of different pathological conditions.

All-together, the results presented in this thesis contribute to more efficient and robust multi-omics computational studies to enhance our understanding of complex molecular mechanisms underlying biological processes, such as alternative splicing of proto-oncogenes in tumor patients. Below are further elaborations on methodological improvements and approaches applied in the studies included in the thesis.

5.1 Integrative analysis of gene expression variability using FAVSeq

This thesis studied factors affecting gene expression measurements based on different RNA-Seq technologies. In the matched experiments setup, the impact of a variety of factors (i.e., genomic and transcriptomic) on the variability in RNA-Seq data was evaluated using the proposed ML-based data analysis pipeline (FAVSeq). Our results suggest that 3'-UTR and transcript lengths, as well as GC content influence the gene expression difference between the bulk and scRNA-Seq profiles. Similarly, 3'-UTR and cellular compartments were found to be relevant for dropouts at the most. Some of these identified features have been already reported to affect gene expression variability in RNA-Seq data. In particular, 3'-UTR was shown in determining expression differences in RNA-Seq.

From the technological standpoint, the potential reason for 3'-UTR to be appeared as one of the core factors affecting gene expression variability is that reads can be biased towards the 3'-UTR in single cell sequencing. Droplet-based single-cell RNA-Seq methods (e.g., 10x Genomics Chromium), have the majority of the reads mapped to UTR (mainly 3'-UTR) [135], while in the bulk both ends are typically used for tagging of mRNA fragments (cell-barcode + UMI). This could lead to the result we see, that 3'-UTR length contributes to a difference between the expression levels. Thus, the 3'-UTR in such a case should be more relevant in measuring gene abundance. Moreover, in the case of the pig data the 3'-UTR can be identified as the most influential feature because this non-model organism has less well annotated 3'-UTR regions [106], while bulk doesn't have 3'-bias, so it wasn't affected so much. In order to explore deeper the differences between scRNA-Seq and bulk, one can further utilize FAVSeq for the matched data from non-model organism (e.g., mouse as provided by Chen et al., 2020 [23]).

Additionally, the following approaches can be employed to extend the current analysis at different levels (data preparation and analysis steps), such as: A) test how additional features can affect the gene expression variability; B) further optimize the hyper-parameters of the model; C) try out other algorithms. With regards to the data preparation step, additional features extracted from the unprocessed sequencing data (e.g., BAM, FASTQ) can be considered. Thus, the feature selection analysis based on data with the least of downstream processing steps will allow to investigate in more detail the experiment-related factors affecting the technical variability between sequencing protocols. Another possibility addressing individual model's performance in the data analysis step is to further explore hyper-parameters of the model during the optimization phase. Thus, the performance of the model can be further improved by searching for more optimal values in the increased hyper-parameters' space. Accordingly, the number of hyper-

parameters can be expanded and/or the model can be calibrated by the analysis of out-of-bag (OOB) errors that serves as cross-validation loss to approximate a suitable number of trees at which the error stabilizes.

Furthermore, the ML pipeline can be extended by testing other machine learning algorithms and meta-learning approaches (e.g., Support Vector Machine, Boosting, k-Best) to achieve higher predictive power. For instance, boosting of the regression trees can be used in order to improve further the model performance on the gene expression difference prediction task. Thus, the FAVSeq-based analysis can be extended by, for instance, introduction of additional features (e.g., cell specificity)—that are dependent on the scientific question formulated by a researcher—in order to verify their relevance for the difference between RNA-Seq technologies of interest.

5.2 Extended analysis of alternative splicing regulome using regulAS

This thesis presents an ML-assisted workflow for identifying relevant factors affecting alternative splicing variation across conditions and tissues in order to alleviate searching for regulatory candidates for further in-depth studies. The computational pipeline was implemented as a software package (regulAS) utilizing open-access omics data to assist computational biology researchers and domain experts, allowing accelerate laboratory validation of identified factors in the context of alternative splicing and streamline experimental workflow as a whole. The proposed approach allowed to confirm existing and identify novel relevant factors affecting splicing alterations in proto-oncogene RONΔ165 in different tumors, including SON, hnRNP H2/3

and RBM3. Thus, the derived findings lead to a better understanding of RBPs-mediated regulatory mechanisms of its alternative splicing.

While several regulators of RON AS have been reported to modulate RONΔ165 AS [13, 70, 84], the analysis results demonstrate that top identified candidates to regulate RON AS were also shown to be involved in splicing machinery regulation in different genes expressing under tumor progression [93]. Among them, SON is an RNA splicing co-factor ensuring efficient intron removal from the transcripts containing suboptimal splice sites. SON targets genes which are involved in cell proliferation, genome stability, chromatin modifications, etc., and is also playing a role in gene regulation during cancer development and progression [49]. Also, RBM3 has been also implicated as a regulator of alternative splicing of RNAs from the genes involved in apoptosis and cell differentiation [134]. HnRNP H/F preferentially bind to poly(G)-rich sequences (G-runs) in the target exon and/or adjacent introns, and regulate alternative splicing of numerous genes [88]. Moreover, hnRNPH2—the one of the top-10 identified regulators—was shown to affect splicing of RON exon 11 [14].

The results indicate that a machine learning-based analysis is able to identify dependencies between gene expression profiles of putative regulators and splicing efficiency data (measured as PSI) from tumor and healthy human samples when applied to open-access data from TCGA and GTEx sources. Indeed, Pearson's correlation coefficient computed for the observed and predicted PSI values is consistently above 0.8 for almost all genes of interest, as shown in Table 4.6. Such strong correlation suggests that a multilayer neural network was successful in identifying dependencies between RBP expression data and PSI, which is a significant cue suggesting to perform the subsequent assessment of the RBP relevance ranks.

For the following apoptosis related AS genes—FAS, MADD, MINK1, NOTCH3, and PAX6,—the obtained results indicate substantial correspondence between top-rank RBPs across "in silico" and "in vitro" experimental setups. In particular, the value of the nDCG score for these candidate AS modulators fluctuates around 50%, with the intersection of the top-25 results by both methods ranging from 4 and 7 matches. For APAF1, MAP3K7 and MCL1, the obtained nDCG is slighly lower in comparison to the previous group, and the number of matches among top relevant RBPs identified is three to four.

Interestingly, modelling results for MST1R/RON demonstrate a low predictive power for the analyzed data on RBPs expression and splice junctions in the context of cervix adjacent tissues (PCC $\sim$0.15, Table 4.6). At the same time, when comparing RBP selection results derived in different experimental settings, we observe a relatively high nDCG for the top relevance ranks (64%) along with a moderate number of common RBPs (5 matches). Such mismatch can indicate that, although the expression levels of prospective regulators influence on the PSI value, in case of MST1R, the correspondence between those tends to be weak, thus suggesting that the MST1R-related signals can contain a significant amount of noise, which makes analysis of these data quite challenging.

By comparison with MST1R, the nDCG scores for CCNE1 and CHEK2 are rather low (33% and 35%, respectively), whereas the number of common RBPs (5-6 matches) being close to the highest result within the target gene group, with a maximum of 7 matches. This suggests that beyond the top-scored relevant RBPs (those identified by both—"in silico" and "in vitro"—experiments), the remaining regulatory candidates may have limited impact on splicing changes for these genes. Thus, the specific ordering of such low-relevance RBPs may vary across approaches and measurements.

Finally, it is worth noting that while the analysis of PAX6 was limited to study of tumor samples, the modelling results reveal a rather high RBPs-PSI predictive performance for unseen test data (PCC=92%, subsection 4.2.11), together with the presence of PPI associations for the top ranked RBPs and their links to mRNA splicing pathways (p-value=$4.74 \cdot 10^{-8}$ and $6.88 \cdot 10^{-5}$, respectively), as shown in Table 4.8. This permits to extend the interpretation of ML-based RBP selection ranking for PAX6, namely to attribute it tumor-related cross-RBP relevance, even considering the limited availability of corresponding RBP measurements from healthy samples.

Additionally, the proposed approach can be extensively applied to the analysis of other classes of potential regulatory factors for splicing alterations across tissues (e.g., TF) and over a broad range of transcriptome profiling and splice junctions data from various omics sources. It can also be extended to identify important effectors of any biological process of interest given data with continuous values for features and the target variable. Furthermore, the analysis using regulAS can be extended to further studies of co-regulated alternative splicing events (e.g., [119]) and to additional data from open-access sources, such as ENCODE [25].

.

Bibliography

[1] Serkan A Alkan, Kathleen Martincic, and Christine Milcarek. “The hnRNPs F and H2 bind to similar sequences to influence gene expression”. In: *Biochemical Journal* 393.1 (2006), pp. 361–371 (cit. on p. 96).

[2] Guillermo de Anda-Jáuregui and Enrique Hernández-Lemus. “Computational oncology in the multi-omics era: state of the art”. In: *Frontiers in oncology* 10 (2020), p. 423 (cit. on pp. 7, 12).

[3] Tallulah S Andrews and Martin Hemberg. “Identifying cell populations with scRNASeq”. In: *Molecular aspects of medicine* 59 (2018), pp. 114–122 (cit. on p. 7).

[4] Ricard Argelaguet, Britta Velten, Damien Arnol, et al. “Multi-Omics Factor Analysis—a framework for unsupervised integration of multi-omics data sets”. In: *Molecular systems biology* 14.6 (2018), e8124 (cit. on p. 11).

[5] Michael Ashburner, Catherine A Ball, Judith A Blake, et al. “Gene ontology: tool for the unification of biology”. In: *Nature genetics* 25.1 (2000), pp. 25–29 (cit. on p. 43).

[6] Francisco E Baralle and Jimena Giudice. “Alternative splicing as a regulator of development and tissue identity”. In: *Nature reviews Molecular cell biology* 18.7 (2017), pp. 437–451 (cit. on pp. 38, 84).

[7] Alev Baysoy, Zhiliang Bai, Rahul Satija, and Rong Fan. “The technological landscape and applications of single-cell multi-omics”. In: *Nature Reviews Molecular Cell Biology* 24.10 (2023), pp. 695–713 (cit. on p. 5).

[8] Richard Bellman and Robert Kalaba. “A mathematical theory of adaptive control processes”. In: *Proceedings of the National Academy of Sciences* 45.8 (1959), pp. 1288–1290 (cit. on p. 12).

[9] Yoshua Bengio, Aaron Courville, and Pascal Vincent. “Representation learning: A review and new perspectives”. In: *IEEE transactions on pattern analysis and machine intelligence* 35.8 (2013), pp. 1798–1828 (cit. on p. 28).

[10]Janos X Binder, Sune Pletscher-Frankild, Kalliopi Tsafou, et al. “COMPARTMENTS: unification and visualization of protein subcellular localization evidence”. In: *Database* 2014 (2014) (cit. on pp. 43, 72).

[11]Amanda Birmingham, Laura M Selfors, Thorsten Forster, et al. “Statistical methods for analysis of high-throughput RNA interference screens”. In: *Nature methods* 6.8 (2009), pp. 569–575 (cit. on pp. 103, 107, 111).

[12]Christopher M Bishop and Nasser M Nasrabadi. *Pattern recognition and machine learning*. 2006 (cit. on p. 49).

[13]Sophie C Bonnal, Irene López-Oreja, and Juan Valcárcel. “Roles and mechanisms of alternative splicing in cancer—implications for care”. In: *Nature reviews Clinical oncology* 17.8 (2020), pp. 457–474 (cit. on pp. 36, 83, 117).

[14]Simon Braun, Mihaela Enculescu, Samarth T Setty, et al. “Decoding a cancer-relevant splicing decision in the RON proto-oncogene using high-throughput mutagenesis”. In: *Nature communications* 9.1 (2018), p. 3315 (cit. on pp. 36, 37, 50, 83, 96, 117).

[15]Leo Breiman. “Bagging predictors”. In: *Machine learning* 24.2 (1996), pp. 123–140 (cit. on p. 46).

[16]Leo Breiman. “Random forests”. In: *Machine learning* 45 (2001), pp. 5–32 (cit. on p. 49).

[17]Róbert Busa-Fekete, György Szarvas, Tamás Elteto, and Balázs Kégl. “An apple-to-apple comparison of learning-to-rank algorithms in terms of normalized discounted cumulative gain”. In: *ECAI 2012-20th European Conference on Artificial Intelligence: Preference Learning: Problems and Applications in AI Workshop*. Vol. 242. Ios Press. 2012 (cit. on p. 99).

[18]Matteo Calgaro, Chiara Romualdi, Levi Waldron, Davide Risso, and Nicola Vitulo. “Assessment of statistical methods from single cell, bulk RNA-seq, and metagenomics applied to microbiome data”. In: *Genome biology* 21 (2020), pp. 1–31 (cit. on p. 32).

[19]Enrico Capobianco. “RNA-Seq data: a complexity journey”. In: *Computational and structural biotechnology journal* 11.19 (2014), pp. 123–130 (cit. on p. 84).

[20]Hsueh-Ping Chao, Yueping Chen, Yoko Takata, et al. “Systematic evaluation of RNA-Seq preparation protocol performance”. In: *BMC genomics* 20.1 (2019), pp. 1–20 (cit. on p. 32).

[21]Geng Chen, Baitang Ning, and Tieliu Shi. “Single-cell RNA-seq technologies and related computational data analysis”. In: *Frontiers in genetics* (2019), p. 317 (cit. on pp. 29, 30, 32).

[22] Geng Chen, Baitang Ning, and Tieliu Shi. “Single-cell RNA-seq technologies and related computational data analysis”. In: *Frontiers in genetics* 10 (2019), p. 317 (cit. on p. 58).

[23] Tianyi Chen, Sehhoon Oh, Simon Gregory, Xiling Shen, and Anna Mae Diehl. “Single-cell omics analysis reveals functional diversification of hepatocytes during liver regeneration”. In: *JCI insight* 5.22 (2020) (cit. on p. 115).

[24] Ana Conesa, Pedro Madrigal, Sonia Tarazona, et al. “A survey of best practices for RNA-seq data analysis”. In: *Genome biology* 17.1 (2016), pp. 1–19 (cit. on p. 32).

[25] ENCODE Project Consortium et al. “An integrated encyclopedia of DNA elements in the human genome”. In: *Nature* 489.7414 (2012), p. 57 (cit. on p. 119).

[26] Fiona Cunningham, Premanand Achuthan, Wasiu Akanni, et al. “Ensembl 2019”. In: *Nucleic acids research* 47.D1 (2019), pp. D745–D751 (cit. on pp. 43, 72).

[27] Miri Danan-Gotthold, Regina Golan-Gerstl, Eli Eisenberg, et al. “Identification of recurrent regulated alternative splicing events across human solid tumors”. In: *Nucleic acids research* 43.10 (2015), pp. 5130–5144 (cit. on pp. 38, 84).

[28] Jeremy Davis-Turak, Sean M Courtney, E Starr Hazard, et al. “Genomics pipelines and data integration: challenges and opportunities in the research setting”. In: *Expert review of molecular diagnostics* 17.3 (2017), pp. 225–237 (cit. on p. 11).

[29] Jiarui Ding, Xian Adiconis, Sean K Simmons, et al. “Systematic comparison of single-cell and single-nucleus RNA-sequencing methods”. In: *Nature biotechnology* 38.6 (2020), pp. 737–746 (cit. on p. 32).

[30] Meichen Dong, Aatish Thennavan, Eugene Urrutia, et al. “SCDC: bulk gene expression deconvolution by multiple single-cell RNA sequencing references”. In: *Briefings in bioinformatics* 22.1 (2021), pp. 416–427 (cit. on p. 32).

[31] Carsten F Dormann, Jane Elith, Sven Bacher, et al. “Collinearity: a review of methods to deal with it and a simulation study evaluating their performance”. In: *Ecography* 36.1 (2013), pp. 27–46 (cit. on p. 73).

[32] Harris Drucker, Christopher J Burges, Linda Kaufman, Alex Smola, and Vladimir Vapnik. “Support vector regression machines”. In: *Advances in neural information processing systems* 9 (1996) (cit. on p. 49).

[33] Richard O Duda, Peter E Hart, et al. *Pattern classification*. John Wiley & Sons, 2006 (cit. on p. 17).

[34] Jean Fan, Kamil Slowikowski, and Fan Zhang. "Single-cell transcriptomics in cancer: computational challenges and opportunities". In: *Experimental & Molecular Medicine* 52.9 (2020), pp. 1452–1465 (cit. on p. 32).

[35] Clarisse van der Feltz and Aaron A Hoskins. "Structural and functional modularity of the U2 snRNP in pre-mRNA splicing". In: *Critical reviews in biochemistry and molecular biology* 54.5 (2019), pp. 443–465 (cit. on p. 96).

[36] Mattia Forcato, Oriana Romano, and Silvio Bicciato. "Computational methods for the integrative analysis of single-cell data". In: *Briefings in bioinformatics* 22.3 (2021), bbaa042 (cit. on p. 32).

[37] Xiang-Dong Fu and Manuel Ares Jr. "Context-dependent control of alternative splicing by RNA-binding proteins". In: *Nature Reviews Genetics* 15.10 (2014), pp. 689–701 (cit. on p. 113).

[38] Jean Gaudart, Bernard Giusiano, and Laetitia Huiart. "Comparison of the performance of multi-layer perceptron and linear regression for epidemiological data". In: *Computational statistics & data analysis* 44.4 (2004), pp. 547–570 (cit. on p. 17).

[39] Fátima Gebauer, Thomas Schwarzl, Juan Valcárcel, and Matthias W Hentze. "RNA-binding proteins in human genetic disease". In: *Nature Reviews Genetics* 22.3 (2021), pp. 185–198 (cit. on pp. 36, 83).

[40] Thomas Geuens, Delphine Bouhy, and Vincent Timmerman. "The hnRNP family: insights into their role in health and disease". In: *Human genetics* 135 (2016), pp. 851–867 (cit. on p. 96).

[41] Anchel González-Barriga, Louison Lallemant, Diana M Dincã, et al. "Integrative cell type-specific multi-omics approaches reveal impaired programs of glial cell differentiation in mouse culture models of DM1". In: *Frontiers in Cellular Neuroscience* 15 (2021), p. 662035 (cit. on p. 11).

[42] Pablo M Granitto, Cesare Furlanello, Franco Biasioli, and Flavia Gasperi. "Recursive feature elimination with random forest for PTR-MS analysis of agroindustrial products". In: *Chemometrics and intelligent laboratory systems* 83.2 (2006), pp. 83–90 (cit. on p. 24).

[43] Amit Gupta, Sandhya Yadav, Archana Pt, et al. "The HNRNPA2B1–MST1R–Akt axis contributes to epithelial-to-mesenchymal transition in head and neck cancer". In: *Laboratory Investigation* 100.12 (2020), pp. 1589–1601 (cit. on pp. 37, 83).

[44] Isabelle Guyon and André Elisseeff. "An introduction to variable and feature selection". In: *Journal of machine learning research* 3.Mar (2003), pp. 1157–1182 (cit. on pp. 22, 85).

[45] Isabelle Guyon, Jason Weston, Stephen Barnhill, and Vladimir Vapnik. "Gene selection for cancer classification using support vector machines". In: *Machine learning* 46.1 (2002), pp. 389–422 (cit. on pp. 47, 94).

[46] Christoph Hafemeister and Rahul Satija. "Normalization and variance stabilization of single-cell RNA-seq data using regularized negative binomial regression". In: *Genome biology* 20.1 (2019), p. 296 (cit. on p. 34).

[47] P Harris. "Testing for variance homogeneity of correlated variables". In: *Biometrika* 72.1 (1985), pp. 103–107 (cit. on p. 89).

[48] Leandro C Hermida, E Michael Gertz, and Eytan Ruppin. "Predicting cancer prognosis and drug response from the tumor microbiome". In: *Nature communications* 13.1 (2022), p. 2896 (cit. on p. 16).

[49] Christopher J Hickey, Jung-Hyun Kim, and Eun-Young Erin Ahn. "New discoveries of old SON: a link between RNA splicing and cancer". In: *Journal of cellular biochemistry* 115.2 (2014), pp. 224–231 (cit. on p. 117).

[50] Stephanie C Hicks, F William Townes, Mingxiang Teng, and Rafael A Irizarry. "Missing data and technical variability in single-cell RNA-sequencing experiments". In: *Biostatistics* 19.4 (2018), pp. 562–578 (cit. on p. 34).

[51] Stephanie C Hicks, F William Townes, Mingxiang Teng, and Rafael A Irizarry. "Missing data and technical variability in single-cell RNA-sequencing experiments". In: *Biostatistics* 19.4 (2018), pp. 562–578 (cit. on pp. 58, 62).

[52] Mingye Hong, Shuang Tao, Ling Zhang, et al. "RNA sequencing: new technologies and applications in cancer research". In: *Journal of hematology & oncology* 13.1 (2020), pp. 1–16 (cit. on p. 32).

[53] Harold Hotelling. "Analysis of a complex of statistical variables into principal components." In: *Journal of educational psychology* 24.6 (1933), p. 417 (cit. on p. 33).

[54] Wenpin Hou, Zhicheng Ji, Hongkai Ji, and Stephanie C Hicks. "A systematic evaluation of single-cell RNA-sequencing imputation methods". In: *Genome biology* 21.1 (2020), pp. 1–30 (cit. on p. 68).

[55] Byungjin Hwang, Ji Hyun Lee, and Duhee Bang. "Single-cell RNA sequencing technologies and bioinformatics pipelines". In: *Experimental & molecular medicine* 50.8 (2018), pp. 1–14 (cit. on p. 30).

[56] Byungjin Hwang, Ji Hyun Lee, and Duhee Bang. "Single-cell RNA sequencing technologies and bioinformatics pipelines". In: *Experimental & molecular medicine* 50.8 (2018), pp. 1–14 (cit. on p. 58).

[57] Sebastian Jäger, Arndt Allhorn, and Felix Bießmann. "A benchmark for data imputation methods". In: *Frontiers in big Data* 4 (2021), p. 693674 (cit. on p. 27).

[58] Kalervo Järvelin and Jaana Kekäläinen. "Cumulated gain-based evaluation of IR techniques". In: *ACM Transactions on Information Systems (TOIS)* 20.4 (2002), pp. 422–446 (cit. on p. 50).

[59] Brandon Jew, Marcus Alvarez, Elior Rahmani, et al. "Accurate estimation of cell composition in bulk expression through robust integration of single-cell information". In: *Nature communications* 11.1 (2020), p. 1971 (cit. on p. 32).

[60] Andrew L Ji, Adam J Rubin, Kim Thrane, et al. "Multimodal analysis of composition and spatial architecture in human squamous cell carcinoma". In: *Cell* 182.2 (2020), pp. 497–514 (cit. on pp. 6, 29).

[61] Luciane T Kagohara, Fernando Zamuner, Emily F Davis-Marcisak, et al. "Integrated single-cell and bulk gene expression and ATAC-seq reveals heterogeneity and early changes in pathways associated with resistance to cetuximab in HNSCC-sensitive cell lines". In: *British journal of cancer* 123.1 (2020), pp. 101–113 (cit. on p. 58).

[62] Minseung Kim, Navneet Rai, Violeta Zorraquino, and Ilias Tagkopoulos. "Multi-omics integration accurately predicts cellular state in unexplored conditions for Escherichia coli". In: *Nature communications* 7.1 (2016), p. 13090 (cit. on p. 29).

[63] Diederik P Kingma and Jimmy Ba. "Adam: A method for stochastic optimization". In: *arXiv preprint arXiv:1412.6980* (2014) (cit. on p. 20).

[64] Diederik P Kingma and Jimmy Ba. "Adam: A method for stochastic optimization". In: *arXiv preprint arXiv:1412.6980* (2014) (cit. on p. 46).

[65] Ron Kohavi, David H Wolpert, et al. "Bias plus variance decomposition for zero-one loss functions". In: *ICML*. Vol. 96. Citeseer. 1996, pp. 275–83 (cit. on p. 17).

[66] Michal Krassowski, Vivek Das, Sangram K Sahu, and Biswapriya B Misra. "State of the field in multi-omics research: from computational needs to data mining and sharing". In: *Frontiers in Genetics* 11 (2020), p. 610798 (cit. on pp. 11, 12).

[67] Maria Kuksin, Daphné Morel, Marine Aglave, et al. "Applications of single-cell and bulk RNA sequencing in onco-immunology". In: *European Journal of Cancer* 149 (2021), pp. 193–210 (cit. on p. 33).

[68] Maxim V Kuleshov, Matthew R Jones, Andrew D Rouillard, et al. “Enrichr: a comprehensive gene set enrichment analysis web server 2016 update”. In: *Nucleic acids research* 44.W1 (2016), W90–W97 (cit. on p. 51).

[69] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. “Deep learning”. In: *nature* 521.7553 (2015), pp. 436–444 (cit. on p. 28).

[70] Clare V LeFave, Massimo Squatrito, Sandra Vorlova, et al. “Splicing factor hnRNPH drives an oncogenic splicing switch in gliomas”. In: *The EMBO journal* 30.19 (2011), pp. 4084–4097 (cit. on p. 117).

[71] Philippe Leray and Patrick Gallinari. “Feature selection with neural networks”. In: *Behaviormetrika* 26 (1999), pp. 145–166 (cit. on p. 24).

[72] Ji Li, Peter S Choi, Christine L Chaffer, et al. “An alternative splicing switch in FLNB promotes the mesenchymal cell state in human breast cancer”. In: *Elife* 7 (2018), e37184 (cit. on pp. 38, 84).

[73] Jin Li, Yang Wang, Xi Rao, et al. “Roles of alternative splicing in modulating transcriptional regulation”. In: *BMC systems biology* 11.5 (2017), pp. 1–12 (cit. on pp. 38, 84).

[74] Xinmin Li and Cun-Yu Wang. “From bulk, single-cell to spatial RNA sequencing”. In: *International Journal of Oral Science* 13.1 (2021), p. 36 (cit. on p. 32).

[75] Malte D Luecken and Fabian J Theis. “Current best practices in single-cell RNA-seq analysis: a tutorial”. In: *Molecular systems biology* 15.6 (2019), e8746 (cit. on p. 58).

[76] Guillermo Marco-Puche, Sergio Lois, Javier Benitez, and Juan Carlos Trivino. “RNA-Seq perspectives to improve clinical diagnosis”. In: *Frontiers in genetics* 10 (2019), p. 1152 (cit. on p. 29).

[77] Christine Mayr. “Evolution and biological roles of alternative 3’UTRs”. In: *Trends in cell biology* 26.3 (2016), pp. 227–237 (cit. on p. 75).

[78] Leland McInnes, John Healy, and James Melville. “Umap: Uniform manifold approximation and projection for dimension reduction”. In: *arXiv preprint arXiv:1802.03426* (2018) (cit. on p. 33).

[79] Chen Meng, Oana A Zeleznik, Gerhard G Thallinger, et al. “Dimension reduction techniques for the integrative analysis of multi-omics data”. In: *Briefings in bioinformatics* 17.4 (2016), pp. 628–641 (cit. on p. 12).

[80] Vilas Menon. “Extracting new insights from bulk transcriptomics”. In: *Nature Neuroscience* 21.9 (2018), pp. 1142–1144 (cit. on p. 32).

[81] Bilal Mirza, Wei Wang, Jie Wang, et al. “Machine learning and integrative analysis of biomedical big data”. In: *Genes* 10.2 (2019), p. 87 (cit. on p. 26).

[82] Pedro Miura, Sol Shenker, Celia Andreu-Agullo, Jakub O Westholm, and Eric C Lai. “Widespread and extensive lengthening of 3’UTRs in the mammalian brain”. In: *Genome research* 23.5 (2013), pp. 812–825 (cit. on p. 75).

[83] Zahra Momeni, Esmail Hassanzadeh, Mohammad Saniee Abadeh, and Riccardo Bellazzi. “A survey on single and multi omics data mining methods in cancer data classification”. In: *Journal of Biomedical Informatics* 107 (2020), p. 103466 (cit. on pp. 22, 25, 29).

[84] Heegyum Moon, Xuexiu Zheng, Tiing Jen Loh, et al. “Identification of regulatory-RNAs for alternative splicing of Ron proto-oncogene”. In: *Journal of Cancer* 6.12 (2015), p. 1346 (cit. on pp. 37, 83, 117).

[85] Marco Moretto, Paolo Sonego, Ana B Villaseñor-Altamirano, and Kristof Engelen. “First step toward gene expression data integration: transcriptomic data acquisition with COMMAND> _”. In: *BMC bioinformatics* 20 (2019), pp. 1–9 (cit. on p. 12).

[86] Tian Mou, Wenjiang Deng, Fengyun Gu, Yudi Pawitan, and Trung Nghia Vu. “Reproducibility of methods to detect differentially expressed genes from single-cell RNA sequencing”. In: *Frontiers in genetics* 10 (2020), p. 1331 (cit. on pp. 57, 66).

[87] Fionn Murtagh. “Multilayer perceptrons for classification and regression”. In: *Neurocomputing* 2.5-6 (1991), pp. 183–197 (cit. on p. 46).

[88] Mohammad Nazim, Akio Masuda, Mohammad Alinoor Rahman, et al. “Competitive regulation of alternative splicing and alternative polyadenylation by hnRNP H and CstF64 determines acetylcholinesterase isoforms”. In: *Nucleic acids research* 45.3 (2017), pp. 1455–1468 (cit. on p. 117).

[89] Giovanna Nicora, Francesca Vitali, Arianna Dagliati, Nophar Geifman, and Riccardo Bellazzi. “Integrated multi-omics analyses in oncology: a review of machine learning methods and tools”. In: *Frontiers in oncology* 10 (2020), p. 1030 (cit. on p. 15).

[90] Dimitry Ofengeim, Nikolaos Giagtzoglou, Dann Huh, Chengyu Zou, and Junying Yuan. “Single-cell RNA sequencing: unraveling the brain one cell at a time”. In: *Trends in molecular medicine* 23.6 (2017), pp. 563–576 (cit. on p. 57).

[91] A Emin Orhan and Xaq Pitkow. “Skip connections eliminate singularities”. In: *arXiv preprint arXiv:1701.09175* (2017) (cit. on p. 105).

[92]"Pan-cancer analysis of whole genomes". In: *Nature* 578.7793 (2020), pp. 82–93 (cit. on p. 10).

[93]Panagiotis Papasaikas, J Ramón Tejedor, Luisa Vigevani, and Juan Valcarcel. "Functional splicing network reveals extensive regulatory potential of the core spliceosomal machinery". In: *Molecular cell* 57.1 (2015), pp. 7–22 (cit. on pp. 36, 50, 100, 103, 105, 107, 111, 117).

[94]Maria Paola Paronetto, Ilaria Passacantilli, and Claudio Sette. "Alternative splicing and cell survival: from tissue homeostasis to disease". In: *Cell Death & Differentiation* 23.12 (2016), pp. 1919–1929 (cit. on pp. 37, 99).

[95]Soumen Kumar Pati, Saptarshi Sengupta, and Asit K Das. "Improved Genetic Algorithm for Selecting Significant Genes in Cancer Diagnosis". In: *Progress in Advanced Computing and Intelligent Engineering: Proceedings of ICACIE 2016, Volume 2*. Springer. 2018, pp. 395–405 (cit. on p. 25).

[96]Ben Omega Petrazzini, Hugo Naya, Fernando Lopez-Bello, Gustavo Vazquez, and Lucia Spangenberg. "Evaluation of different approaches for missing data imputation on features associated to genomic data". In: *BioData mining* 14.1 (2021), pp. 1–13 (cit. on p. 68).

[97]Francesco Pettini, Anna Visibelli, Vittoria Cicaloni, Daniele Iovinelli, and Ottavia Spiga. "Multi-omics model applied to cancer genetics". In: *International Journal of Molecular Sciences* 22.11 (2021), p. 5751 (cit. on pp. 6, 12).

[98]Milan Picard, Marie-Pier Scott-Boyer, Antoine Bodein, Olivier Périn, and Arnaud Droit. "Integration strategies of multi-omics data for machine learning analysis". In: *Computational and Structural Biotechnology Journal* 19 (2021), pp. 3735–3746 (cit. on p. 14).

[99]Boris T Polyak and Anatoli B Juditsky. "Acceleration of stochastic approximation by averaging". In: *SIAM journal on control and optimization* 30.4 (1992), pp. 838–855 (cit. on p. 20).

[100]Peng Qiu. "Embracing the dropouts in single-cell RNA-seq analysis". In: *Nature communications* 11.1 (2020), pp. 1–9 (cit. on pp. 58, 65).

[101]J. Ross Quinlan. "Induction of decision trees". In: *Machine learning* 1.1 (1986), pp. 81–106 (cit. on p. 46).

[102]Parminder S Reel, Smarti Reel, Ewan Pearson, Emanuele Trucco, and Emily Jefferson. "Using machine learning approaches for multi-omics data analysis: A review". In: *Biotechnology Advances* 49 (2021), p. 107739 (cit. on pp. 15, 28, 29).

[103]Frank Rosenblatt. “The perceptron: a probabilistic model for information storage and organization in the brain.” In: *Psychological review* 65.6 (1958), p. 386 (cit. on p. 18).

[104]Dennis W Ruck, Steven K Rogers, and Matthew Kabrisky. “Feature selection using a multilayer perceptron”. In: *Journal of Neural Network Computing* 2.2 (1990), pp. 40–48 (cit. on pp. 24, 46).

[105]Stephen-John Sammut, Mireia Crispin-Ortuzar, Suet-Feung Chin, et al. “Multi-omic machine learning predictor of breast cancer therapy response”. In: *Nature* 601.7894 (2022), pp. 623–629 (cit. on p. 15).

[106]Yang Shen, Carina Bruckmaier, Miao Sun, et al. “ScRNAX: cross-species transfer of high quality 3’UTR annotation for single cell RNA-Seq”. In: *GigaScience* (in review) (cit. on pp. 41, 115).

[107]Yu Shi, Zuhua Chen, Juanjuan Gao, et al. “Transcriptome-wide analysis of alternative mRNA splicing signature in the diagnosis and prognosis of stomach adenocarcinoma”. In: *Oncology reports* 40.4 (2018), pp. 2014–2022 (cit. on p. 7).

[108]Damian Smedley, Syed Haider, Benoit Ballester, et al. “BioMart–biological queries made easy”. In: *BMC genomics* 10.1 (2009), pp. 1–12 (cit. on pp. 43, 72).

[109]Meng Song, Jonathan Greenbaum, Joseph Luttrell IV, et al. “A review of integrative imputation for multi-omics datasets”. In: *Frontiers in genetics* 11 (2020), p. 570255 (cit. on p. 26).

[110]Indhupriya Subramanian, Srikant Verma, Shiva Kumar, Abhay Jere, and Krishanpal Anamika. “Multi-omics data integration, interpretation, and its application”. In: *Bioinformatics and biology insights* 14 (2020), p. 1177932219899051 (cit. on pp. 11, 14).

[111]Yijun Sun, Jin Yao, Le Yang, et al. “Computational approach for deriving cancer progression roadmaps from static sample data”. In: *Nucleic acids research* 45.9 (2017), e69–e69 (cit. on pp. 38, 84).

[112]Stanislaw Supplitt, Pawel Karpinski, Maria Sasiadek, and Izabela Laczmanska. “Current achievements and applications of transcriptomics in personalized cancer medicine”. In: *International Journal of Molecular Sciences* 22.3 (2021), p. 1422 (cit. on p. 7).

[113]Damian Szklarczyk, Annika L Gable, Katerina C Nastou, et al. “The STRING database in 2021: customizable protein–protein networks, and functional characterization of user-uploaded gene/measurement sets”. In: *Nucleic acids research* 49.D1 (2021), pp. D605–D612 (cit. on pp. 51, 111).

[114]J Ramón Tejedor, Panagiotis Papasaikas, and Juan Valcárcel. “Genome-wide identification of Fas/CD95 alternative splicing regulators reveals links with iron homeostasis”. In: *Molecular cell* 57.1 (2015), pp. 23–38 (cit. on pp. 37, 99, 102).

[115]Sam Tracy, Guo-Cheng Yuan, and Ruben Dries. “RESCUE: imputing dropout events in single-cell RNA-sequencing data”. In: *BMC bioinformatics* 20.1 (2019), pp. 1–11 (cit. on p. 58).

[116]Cole Trapnell, Davide Cacchiarelli, Jonna Grimsby, et al. “The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells”. In: *Nature biotechnology* 32.4 (2014), pp. 381–386 (cit. on p. 58).

[117]Olga Troyanskaya, Michael Cantor, Gavin Sherlock, et al. “Missing value estimation methods for DNA microarrays”. In: *Bioinformatics* 17.6 (2001), pp. 520–525 (cit. on p. 68).

[118]Mathias Uhlen, Per Oksvold, Linn Fagerberg, et al. “Towards a knowledge-based human protein atlas”. In: *Nature biotechnology* 28.12 (2010), pp. 1248–1250 (cit. on p. 43).

[119]Jernej Ule and Benjamin J Blencowe. “Alternative splicing regulatory networks: functions, mechanisms, and evolution”. In: *Molecular cell* 76.2 (2019), pp. 329–345 (cit. on p. 119).

[120]Lorea Valcarcel-Jimenez, Alice Macchia, Natalia Martin-Martin, et al. “Integrative analysis of transcriptomics and clinical data uncovers the tumor-suppressive activity of MITF in prostate cancer”. In: *Cell Death & Disease* 9.10 (2018), p. 1041 (cit. on p. 11).

[121]Jorge R Vergara and Pablo A Estévez. “A review of feature selection methods based on mutual information”. In: *Neural computing and applications* 24 (2014), pp. 175–186 (cit. on p. 24).

[122]Eric Wang and Iannis Aifantis. “RNA splicing and cancer”. In: *Trends in cancer* 6.8 (2020), pp. 631–644 (cit. on pp. 36, 83).

[123]Qingguo Wang, Joshua Armenia, Chao Zhang, et al. “Unifying cancer and normal RNA sequencing data from different sources”. In: *Scientific data* 5.1 (2018), pp. 1–8 (cit. on p. 48).

[124]Xuran Wang, Jihwan Park, Katalin Susztak, Nancy R Zhang, and Mingyao Li. “Bulk tissue cell type deconvolution with multi-subject single-cell expression reference”. In: *Nature communications* 10.1 (2019), p. 380 (cit. on p. 31).

[125] Yan Wang, Jing Liu, BO Huang, et al. “Mechanism of alternative splicing and its regulation”. In: *Biomedical reports* 3.2 (2015), pp. 152–158 (cit. on pp. 25, 36, 83).

[126] Bernard L Welch. “The generalization of ‘STUDENT’S’problem when several different population varlances are involved”. In: *Biometrika* 34.1-2 (1947), pp. 28–35 (cit. on p. 52).

[127] Jennifer Westoby, Pavel Artemov, Martin Hemberg, and Anne Ferguson-Smith. “Obstacles to detecting isoforms using full-length scRNA-seq data”. In: *Genome Biology* 21 (2020), pp. 1–19 (cit. on p. 29).

[128] Christopher Wilks, Phani Gaddipati, Abhinav Nellore, and Ben Langmead. “Snaptron: querying splicing patterns across tens of thousands of RNA-seq samples”. In: *Bioinformatics* 34.1 (2018), pp. 114–116 (cit. on p. 89).

[129] Angela R Wu, Norma F Neff, Tomer Kalisky, et al. “Quantitative assessment of single-cell RNA-sequencing methods”. In: *Nature methods* 11.1 (2014), pp. 41–46 (cit. on p. 58).

[130] Quan Yang, Jinyao Zhao, Wenjing Zhang, Dan Chen, and Yang Wang. “Aberrant alternative splicing in breast cancer”. In: *Journal of molecular cell biology* 11.10 (2019), pp. 920–929 (cit. on p. 110).

[131] Yueqin Yang, Juw Won Park, Thomas W Bebee, et al. “Determination of a comprehensive alternative splicing regulatory network and combinatorial regulation by key factors during the epithelial-to-mesenchymal transition”. In: *Molecular and cellular biology* 36.11 (2016), pp. 1704–1719 (cit. on pp. 37, 99).

[132] Hang-Ping Yao, Yong-Qing Zhou, Ruiwen Zhang, and Ming-Hai Wang. “MSP–RON signalling in cancer: pathogenesis and therapeutic potential”. In: *Nature reviews Cancer* 13.7 (2013), pp. 466–481 (cit. on pp. 37, 83).

[133] Peng Yu, Jin Li, Su-Ping Deng, et al. “Integrated analysis of a compendium of RNA-Seq datasets for splicing factors”. In: *Scientific data* 7.1 (2020), p. 178 (cit. on p. 11).

[134] Yu Zeng, Dana Wodzenski, Dong Gao, et al. “Stress-Response Protein RBM3 Attenuates the Stem-like Properties of Prostate Cancer Cells by Interfering with CD44 Variant SplicingRBM3 and Prostate Cancer”. In: *Cancer research* 73.13 (2013), pp. 4123–4133 (cit. on p. 117).

[135] Xiannian Zhang, Tianqi Li, Feng Liu, et al. “Comparative analysis of droplet-based ultra-high-throughput single-cell RNA-seq systems”. In: *Molecular cell* 73.1 (2019), pp. 130–142 (cit. on p. 115).

[136] Chenxu Zhu, Sebastian Preissl, and Bing Ren. “Single-cell multimodal omics: the power of many”. In: *Nature methods* 17.1 (2020), pp. 11–14 (cit. on p. 6).

[137] Hui Zou and Trevor Hastie. “Regression shrinkage and selection via the elastic net, with applications to microarrays”. In: *JR Stat Soc Ser B* 67 (2003), pp. 301–20 (cit. on p. 49).