Dissertation

submitted to the

Combined Faculty of Natural Sciences and Mathematics

of the Ruperto Carola University Heidelberg, Germany

for the degree of

Doctor of Natural Sciences

Presented by

M. Sc. Andrés Felipe Quintero Moreno

born in Bogotá - Colombia

Oral examination: September 21, 2021

Learning the Parts of Omics:
Inference of Molecular Signatures with
Non-negative Matrix Factorization

Referees: Prof. Dr. Benedikt Brors
Prof. Dr. Karsten Rippe

Declaration of Authorship

I hereby confirm that I have authored this dissertation independently and without the use of others sources than the ones indicated. I have not yet presented this thesis or parts thereof to a university as part of an examination or degree.

This work was carried out from June 2017 until September 2018 in the Division of Theoretical Bioinformatics at the German Cancer Research Center (DKFZ) in Heidelberg (Germany), from October 2018 until July 2021 in the Division of Neuroblastoma Genomics at the DKFZ, and in the Health Data Science Unit at the University Clinics Heidelberg under the supervision of Dr. Carl Herrmann.

Heidelberg, July, 2021

. . . . . .

Andrés Felipe Quintero Moreno

A la mamá Martha, a Caro y a toda la manada

Abstract

Background: Feature extraction and signature identification are two critical steps to understand diverse biological processes. Signatures are defined as groups of molecular features that are sufficient to identify certain genotype or phenotype. In particular, Non-negative Matrix Factorization (NMF) has been used to identify signatures in complex genomic datasets. However, running a basic NMF analysis is a challenging task with a steep learning curve and long computing time; furthermore, the usability of these algorithms is lessened by limited resources to interpret the results obtained from them. This creates a pressing need for the development of tools that mitigate such obstacles.

Results: In this study we developmented ButchR and ShinyButchR, a fast and user-friendly toolkit to decompose datasets (slicing genomics) and learn signatures using NMF. The package can be freely installed from GitHub at `https://github.com/wurst-theke/ButchRr`. We used ButchR to identify a new regulatory subtype in neuroblastoma, which showed mesenchymal charanteristics and was phenotypically associated to multipotent Schwann cell precursors. Additionally, we created a new workflow to infer regulatory relationships between genes and their *cis*-regulatory elements for individual cells, followed by inference of regulatory-signatures.

Conclusions: ButchR/ShinyButchR is an useful toolkit for analyzing multiple types of data, and inferring signatures that are able to capture relevant biological information. This toolkit is a new valuable resource to the scientific community, and it can be used to understand complex biological processes.

Zusammenfassung

Hintergrund: “Feature extraction” und “signature identification” sind zwei essenzielle Schritte zum Verständnis diverser biologischer Prozesse. Signaturen werden als Gruppen von molekularen Merkmalen definiert, die ausreichen, um einen bestimmten Genotyp oder Phänotyp zu identifizieren. Insbesondere wurde die “Non-negative Matrix Factorization” (NMF) verwendet, um Signaturen in komplexen genomischen Datensätzen zu identifizieren. Die Durchfuehrung einer einfachen NMF-Analyse ist jedoch eine anspruchsvolle Aufgabe mit einer steilen Lernkurve und langer Rechenzeit; außerdem wird die Verwendbarkeit dieser Algorithmen durch begrenzte Ressourcen zur Interpretation der daraus erhaltenen Ergebnisse verringert. Daraus entsteht ein Bedarf für die Entwicklung von Tools, welche diese Hindernisse umgehen.

Ergebnisse: In dieser Studie haben wir ButchR und ShinyButchR entwickelt, ein schnelles und nutzerfreundliches Toolkit zum Zerlegen von Datensätzen (“Slicing Genomics”) und zum Lernen von Signaturen mit NMF. Das Paket kann frei von GitHub unter `https://github.com/wurst-theke/ButchRr` installiert werden. Wir verwendeten ButchR, um einen neuen regulatorischen Subtyp im Neuroblastom zu identifizieren, der mesenchymale Eigenschaften aufwies und phänotypisch mit multipotenten Schwann-Zellvorläufern assoziiert war. Darüber hinaus haben wir einen neuen Workflow erstellt, um regulatorische Beziehungen zwischen Genen und ihren *cis*-regulatorischen Elementen für einzelne Zellen abzuleiten, gefolgt von der Inferenz von regulatorischen Signaturen.

Schlussfolgerungen: ButchR/ShinyButchR ist ein nützliches Toolkit für die Analyse verschiedener Datentypen und die Ableitung von Signaturen, die in der Lage sind, relevante biologische Informationen zu erfassen. Dieses Toolkit ist eine neue wertvolle Ressource für die Wissenschaftsgemeinde und kann zum Verständnis komplexer biologischer Prozesse verwendet werden.

Table of Contents

Acknowledgements

In the first place, I would like to thank my supervisor Dr. Carl Herrmann. Who has been a great mentor and having the opportunity of working with him guidance has been a truly enriching experience to me. He is an amazing scientist, all the time full of exciting ideas and new research topics in mind to explore. His compromise and creativity were the pillars to create truly amazing discoveries and open future research questions.

I am thankful to all my friends and colleagues from the Health Data Science Unit (HDSU) (current and alumni). In particular to Ana Luísa Costa, Anne-Claire Kröger, Daria Doncevic, Jana Dahlhoff, Dr. Ashwini Kumar Sharma, Dr. Carlos Ramirez, David Schwarzenbacher, Maximilian Kohnen, and Youcheng Zhang for their support, help, and all the shared lunches and breakfasts. Thanks to Jana, Daria, and Youcheng for beta testing ButchR and pointing out how to improve it; and to Ashwin for being there every time I needed his advice right from the beginning of my Ph.D.

I would also like to extend my deepest gratitude to my thesis advisory committee members Prof. Dr. Roland Eils, Prof. Dr. Karsten Rippe, and Prof. Dr. Christoph Dieterich, their insights and helpful advice were critical for the improvement of this project. A special thanks to my Ph.D. examiners, Dr. Sevin Turcan, Dr. Judith Zaugg, Prof. Karsten Rippe, and Prof. Benedikt Brors for reviewing my thesis and participating in the defense. I would also like to especially thank Prof. Dr. Roland Eils for allowing me to be part of the Eils Labs, it was an incredible experience to share and learn from so many scientists with diverse backgrounds.

I thank the Helmholtz International Graduate School for Cancer Research Fellowship, the Deutsches Krebsforschungszentrum (DKFZ), and all the members of the DKFZ Graduate Program Office for supporting my doctoral studies, especially to Dr. Lindsay Murrells and Heike Langlotz for their help and guidance.

My deepest gratitude to Manuela Schaefer, Corinna Sprengart, and Cathrin Hollenbach for all their help from the administrative offices of the Eils Labs and HDSU. Their support was remarkable and extremely important to organize all the bureaucratic affairs that came along during these years.

I am thankful to PD Dr. Frank Westermann and all his group in the Division of Neuroblastoma Genomics at the DKFZ, particularly Dr. Moritz Gartlgruber, Dr. Daniel Dreidax, and Selina Jansky with whom we had an incredible collaboration and many interesting discussions, to finally produce a manuscript we all are proud of.

I also want to express my gratitude to all our collaborators at the BGI-Shenzhen. Especially, to Dr. Longqi Liu and Dr. Zhouchun Shang for taking us on board of an exciting research idea such as scCAT-seq, having the opportunity of analyzing this complex dataset was truly exceptional to me.

My sincere thanks to Dr. Daniel Hübschmann, Dr. Nils Kurzawa, and Sebastian Steinhauser, for all their effort and initial work on Bratwurst, the predecessor of ButchR. It was wonderful to have such a fundamental stepping stone from which to built up ButchR.

I am extremely grateful to my mother Martha for taking care of the "Peludos" while Carolina and I decided to look for new opportunities outside of Colombia, without her help we would have not made it here, and also for the long conversations about watercolors, paints, and art in general. To Carolina, for pushing me when my mind drifts away and lending me all her strength whenever I need it. I thank her for all her support, and let's keep fighting to get "La manada" back together.

List of Publications

First author publications

- Liu, L*, Liu, C*, **Quintero, A***, Wu, L*, Yuan, Y, Wang, M, Cheng, M, Leng, L, Xu, L, Dong, G, Li, R, Liu, Y, Wei, X, Xu, J, Chen, X, Lu, H, Chen, D, Wang, Q, Zhou, Q, Lin, X, Li, G, Liu, S, Wang, Q, Wang, H, Fink, JL, Gao, Z, Liu, X, Hou, Y, Zhu, S, Yang, H, Ye, Y, Lin, G, Chen, F, Herrmann, C, Eils, R, Shang, Z. Deconvolution of single-cell multi-omics layers reveals regulatory heterogeneity. Nat. Commun. 10, 470 (2019). https://doi.org/10.1038/s41467-018-08205-7
- **Quintero, A***, Hübschmann, D*, Kurzawa, N*, Steinhauser, S*, Rentzsch, P, Krämer, S, Andresen, C, Park, J, Eils, R, Schlesner, M and Herrmann, C. ShinyButchR: interactive NMF-based decomposition workflow of genome-scale datasets. Biology Methods and Protocols, 5(1). https://doi.org/10.1093/biomethods/bpaa022
- Gartlgruber, M*, Sharma, AK*, **Quintero, A***, Dreidax, D*, Jansky, S, Park, Y, Gogolin, S, Meder, J, Doncevic D, Saary P, Toprak, UH, Ishaque, N, Afanasyeva, E, Koster, J, Versteeg R, Grünewald, TGP, Jones, DTW, Pfister, SM, Henrich, K, Nes, Jv, Herrmann, C, Westermann, F. Super enhancers define regulatory subtypes and cell identity in neuroblastoma. Nature Cancer 2, 114–128 (2021). https://doi.org/10.1038/s43018-020-00145-w

* Shared first authorship.

Other collaborative contributions

- Selina Jansky, Ashwini Kumar Sharma, Verena Körber, **Andres Quintero**, Umut H. Toprak, Elisa M. Wecht, Moritz Gartlgruber, Alessandro Greco, Elad Chomsky, Thomas G. P. Grünewald, Kai-Oliver Henrich, Amos Tanay, Carl Herrmann, Thomas Höfer, Frank Westermann. Single-cell transcriptomic analyses provide insights into the developmental origins of neuroblastoma. Nature Genetics (2021).
- Emanuel Schwarz, Dag Alnaes, Ole A. Andreassen, Han Cao, Junfang Chen, Franziska Degenhardt, Dominic Dwyer, Roland Eils, Jeanette Erdmann, Carl Herrmann, Martin Hofmann-Apitius, Tobias Kaufmann, Nikolaos Koutsouleris, Alpha T. Kodamullil, Adyasha Khuntia, Maria L. Munoz-Venegas, Markus M. Nöthen, Riya Paul, **Andres Quintero**, Heribert Schunkert, Sharma AK, Heike Tost, Lars T. Westlye, Youcheng Zhang, Andreas Meyer-Lindenberg. Identifying multimodal signatures underlying the somatic comorbidity of psychosis: the COMMITMENT roadmap. Molecular Psychiatry (2020)

List of Abbreviations

ADRN	Adrenergic
ALL	Acute lymphoblastic leukemia
AML	Acute myeloid leukemia
AUC	Area under the curve
ButchR	Slicing genomics
ChIP-seq	Chromatin Immunoprecipitation Sequencing
CI	Continuous integration
CLP	Common lymphoid progenitors
CMP	Common myeloid progenitors
CREs	Cis-regulatory elements
EMT	Epithelial-mesenchymal transition
FACS	Fluorescence-Activated Cell Sorting
GEO	Gene Expression Ommnibus
GMP	Granulocyte-monocyte progenitors
GO	Gene ontology
GRNMF_SC	Graph Regularized Non-negative Matrix Factorization with Sparse Coding
GRNs	Gene regulatory networks
GSEA	Gene set enrichment analysis
GWAS	Genome-wide association studies
H3K27ac	Acetylation of lysine 27 on histone H3
H3K27me3	Trimethylation of lysine 27 on histone H3

H3K4me1	Monomethylation of lysine 4 on histone H3
H3K4me3	Monomethylation of lysine 4 on histone H3
HSC	Hematopoietic stem cells
HMDB	Human Metabolome Database
HTS	High-throughput screenings
i2NMF	Integrative Iterative Non-negative Matrix Factorization
ICA	Independent component analysis
ICM	Inner cell mass
iNMF	integrative Non-negative Matrix Factorization
JIVE	Joint and Individual Variation Explained
jNMF	Joint NMF
LCD	Linear combination decomposition
LMPP	Lymphoid-primed multipotent
MEP	Megakaryocyte-erythrocyte progenitor
MES	Mesenchymal
MGI	Mouse Genome Informatics Web Site
MNA-HR	MYCN non-amplified high-risk
MNA-LR	MYCN non-amplified low-risk
NMF	Non-negative Matrix Factorization
MOCA	Mouse organogenesis
MOFA	Multi-Omics Factor Analysis
MPP	Multipotent progenitors
MSFA	Multi-Study Factor Analysis
MSigDB	Molecular signatures database
NB	Neuroblastoma
NRC	Neuroblastoma Research Consortium
PCA	Principal Component Analysis

PCAWG	Pan-cancer analysis of whole genomes
PCs	Principal components
PDX	Patient derived xenograft
PSNS	Peripheral sympathetic nervous system
PWM	Position Weight Matrix
ROSE	Rank Ordering of Super-Enhancers
SCPs	Schwann cell precursors
SEs	Super enhancers
SN	Substantia nigra
SNPs	Single nucleotide polymorphism
SVD	Singular value decomposition
SVM-RFE	SVM-Recursive Feature Elimination
t-SNE	t-Distributed Stochastic Neighbor Embedding
TARGET	Therapeutically Applicable Research to Generate Effective Treatments
TFs	Transcription factors
TSS	Transcription start sites
UMAP	Uniform Manifold Approximation and Projection
YAPSA	Yet another package for signature analysis

List of Figures

List of Tables

Chapter 1

Scope

1.1 Background

In the biological and clinical context, feature extraction and signature identification are two critical steps to understand diverse biological processes. Signatures are defined as groups of molecular features that are sufficient to identify certain genotype or phenotype. For instance, they have been used for the molecular diagnostic and classification of cancer, infectious diseases and genetic disorders (Fernandes and Zhang 2014), identification and characterization of pathogens (Slezak, Hart, and Jaing 2019), understanding the expression changes across tissues in autoimmune diseases (Szymczak et al. 2021), and to identify cell states in single-cell transcriptome data (Butler et al. 2018; Wolf, Angerer, and Theis 2018), among others.

In particular, the family of Non-negative Matrix Factorization (NMF) algorithms has been used in multiple opportunities to identify signatures and extract features in complex high-throughput genomic datasets (Brunet et al. 2004; Ludmil B. Alexandrov et al. 2013; Pal et al. 2014). In contrast to other methods used for these tasks (Pfeil et al. 2020; Butler et al. 2018; Stuart et al. 2019; Wolf, Angerer, and Theis 2018; Dumitrascu et al. 2019), NMF does not rely on the comparison of conditions or clusters to identify

signatures. Therefore, NMF is able to find sub-clusters or sub-types that were not initially identified during the study design. NMF works by decomposing an input matrix X into a signature matrix W and an exposure matrix H. This results in the reduction of the original data dimensionality to a small set of informative signatures. The NMF signatures can be further interrogated to extract the most relevant features associated with a biological condition. The exposure of individual samples or single cells to such signatures can be used to visualize the data structure, or as input to create an embedding with algorithms as tSNE (Van Der Maaten and Hinton 2008) or UMAP (Diaz-Papkovich et al. 2019).

Moreover, most current methods for signature identification are particularly tailored to only one data type, usually somatic mutations or gene expression. Hence, recovering signatures based on different sources of data like regulatory landscapes and interactions would be more difficult with the available methods. Therefore, more general tools and methods like NMF that can integrate and handle different types of data are currently needed. However, running a basic NMF analysis requires the installation of multiple tools and dependencies, along with a steep learning curve and computing time; furthermore, the usability of these algorithms is lessened by limited resources to interpret the results obtained from them. This creates a pressing need for the development of tools that mitigate such obstacles.

The inference of NMF-based signatures may be the key to understand complex biological processes like the regulatory differences seen in the pediatric tumor neuroblastoma, in which evidence of two regulatory states in neuroblastoma cell lines has been determined by changes in the epigenome of the cells (Van Groningen et al. 2017; Boeva et al. 2017). In addition, NMF-based signatures may be helpful to identify the regulatory links between transcription factors and *cis*-regulatory elements that drive the regulatory differences seen across different cell states.

1.2 Aims

The main objective of this thesis was to provide insights into the inference of molecular signatures from high-throughput genomic data using NMF. This was evaluated in the context of enhancer and regulatory signatures. In order to investigate these topics, the following specific aims were addressed in this study:

1. To develop a toolbox that provides a complete NMF-based analysis workflow for inferring molecular signatures.
2. To develop an application for the interactive exploration of NMF results.
3. To evaluate the regulatory variability seen in neuroblastomas by recovering enhancer signatures.
4. To model regulatory interactions and to infer regulatory signatures.

1.3 Major findings and relevance

The extraction and interpretation of signatures from high-throughput genomic data have been challenging tasks during the last decades. In this study, we addressed these problems by creating new tools and using NMF-recovered signatures. Additionally, one of the central goals of this work was to contribute to the spirit of open and reproducible research. Therefore, all the code and analyses shown here are publicly available and fully reproducible. The major products and findings of this study are:

1. Development and optimization of ButchR, a fast and user-friendly R package to decompose datasets (slicing genomics) and learn signatures using NMF. The package can be freely installed from GitHub at `https://github.com/wurst-theke/ButchR` or used from a Docker image, available at `https://hub.docker.com/r/hdsu/butchr`.
2. Development of ShinyButchR, an interactive Shiny application that uses ButchR

to execute an NMF-based analysis from start to end. ShinyButchR is publicly available at https://hdsu-bioquant.shinyapps.io/shinyButchR/ and can also be used locally from the Docker image, available at https://hub.docker.com/r/hdsu/shinybutchr.

3. Development of a new method to project any bulk or single-cell transcriptomic data onto a reference single-cell atlas, using ButchR.

4. Identification of four different regulatory subtypes in neuroblastoma (MYCN-amplified, mesenchymal, MYCN non-amplified high-risk, and MYCN non-amplified low-risk) using super enhancer-derived-signatures; which resulted in a newly described mesenchymal neuroblastoma subtype.

5. The discovered mesenchymal neuroblastoma subtype was phenotypically associated with multipotent Schwann cell precursors.

6. Creation of a new workflow to infer regulatory relationships between genes and their *cis*-regulatory elements for individual cells, followed by inference of regulatory signatures. This method works with datasets generated from technologies, in which chromatin accessibility and gene expression data are co-profiled from every single cell.

7. Creation of a new workflow to model gene regulatory networks based on the construction and quantification of cell state-specific regulons and inference of regulon-guided signatures. This method uses scRNAseq and scATAC-seq data from contextually similar datasets (i.e., same conditions, and same organism).

8. Regulatory signatures were able to capture the intra- and inter-tumor regulatory variability from two lung patient-derived xenografts, identifying groups of transcription factors that define different cell states.

9. Identification of cells from the inner cell mass in blastocyst of pre-implantation human embryos, using regulatory signatures.

10. Three interactive applications to explore the results presented in this thesis:

    - NB-SE-viz: explore the regulatory subtypes in neuroblastoma, https://nbseB087.dkfz.de.
    - NB-dev-viz: explore the developmental programs in neuroblastoma, https://adrenal.kitz-heidelberg.de/developmental_programs_NB_viz/.
    - MapMyCorona: contribution to the world effort to fight the current pandemic, https://hdsu-bioquant.shinyapps.io/mapmycorona/.

We presented in this study ButchR, a new toolkit to infer signatures and extract relevant features associated with genotypes and phenotypes using NMF. We demonstrated how ButchR is useful for analyzing multiple types of data, and how its signatures are able to capture relevant biological information. The accompanying app ShinyButchR can be effectively used to perform a complete ButchR-based analysis in an interactive fashion. This toolkit is a new valuable resource to the scientific community, and it can be used to understand complex biological processes.

Butcher and ShinyButchR were published in "ShinyButchR: interactive NMF-based decomposition workflow of genome-scale datasets. Biology Methods and Protocols" (Quintero et al. 2020). The findings related to the study of the epigenomic subtypes in neuroblastoma were published in "Super enhancers define regulatory subtypes and cell identity in neuroblastoma. Nature Cancer" (Gartlgruber et al. 2021), and the identification of regulatory signatures by integration of multi-omics data were published in "Deconvolution of single-cell multi-omics layers reveals regulatory heterogeneity. Nature Communications" (Liu et al. 2019).

1.4 Outline of the thesis

The main focus of this thesis is the inference of molecular signatures using NMF, its usage to understand regulatory differences in neuroblastoma and to create gene regulatory networks that explain different cell states. To thoroughly discuss each of these topics

this document is organized into an introductory chapter and five main parts:

- **Introduction:** in the introductory chapter, the concepts of molecular signatures, signature inference, and dimension reduction are showcased in detail. These concepts are the foundation of this thesis.
- **Part I. Tool Development:** the first part describes in detail the ButchR/ShinyButchR toolkit, showing how it can be used to infer signatures from different types of data, and the different visualization options included to understand the NMF results. This part also explains the extent of the community-building goals that motivated the development of ShinyButchR.
- **Part II. NMF to Reveal Regulatory Subtypes in Neuroblastoma:** the second part delves into the study of the regulatory subtypes in neuroblastoma. Showing how NMF-recovered signatures helped to identify the mesenchymal neuroblastoma subtype, and how ButchR allows the integration of bulk and single-cell data to determine the possible cell of origin of this subtype.
- **Part III. Tracing Identity Defined by Transcription Factor Activity:** the third part describes two new workflows used to infer regulatory interactions and signatures that reflect regulatory differences across cell states.
- **Part IV. Data Accessibility and Reproducibility:** the extent of the commitment of this project to open and reproducible research is shown in part IV. Describing how interactive applications and robust pipelines were specially made for the publications associated with this project.
- **Part V. Discussion and Conclusion:** the final part of this thesis presents a summary of the most relevant findings, and discusses the relevance of the entire work in light of the current state of the computational biology field, along with perspectives for future studies.

Chapter 2

Introduction

2.1 Molecular signatures

Molecular signatures are groups of biomolecular features (e.g., DNA sequences, genes, open chromatin sites, and CpG islands) that can be used to infer phenotypic or genotypic identity (Sung et al. 2012). The explosion of new high-throughput technologies opened the door to measure biomolecules at a scale not foreseen just 30 years ago (Lenoir and Giannella 2006). One of the consequences of such technologies was the possibility of inferring signatures by finding associations between molecular features and one particular condition (Golub et al. 1999; Fernandes and Zhang 2014; Slezak, Hart, and Jaing 2019; Sotiriou and Pusztai 2009). A signature can be represented in different ways, depending on the strategy used to infer it. For instance, when the strategy is only able to find on/off associations, then the signature would consist of a list of associated features (e.g., a list of transcription factors associated with different organs, **Figure 2.1a**), whereas if the strategy is able to distinguish the strength of the association, the signature consists of a collection of weighted features (e.g., transcription factors weighted by association to different organs, **Figure 2.1b**).

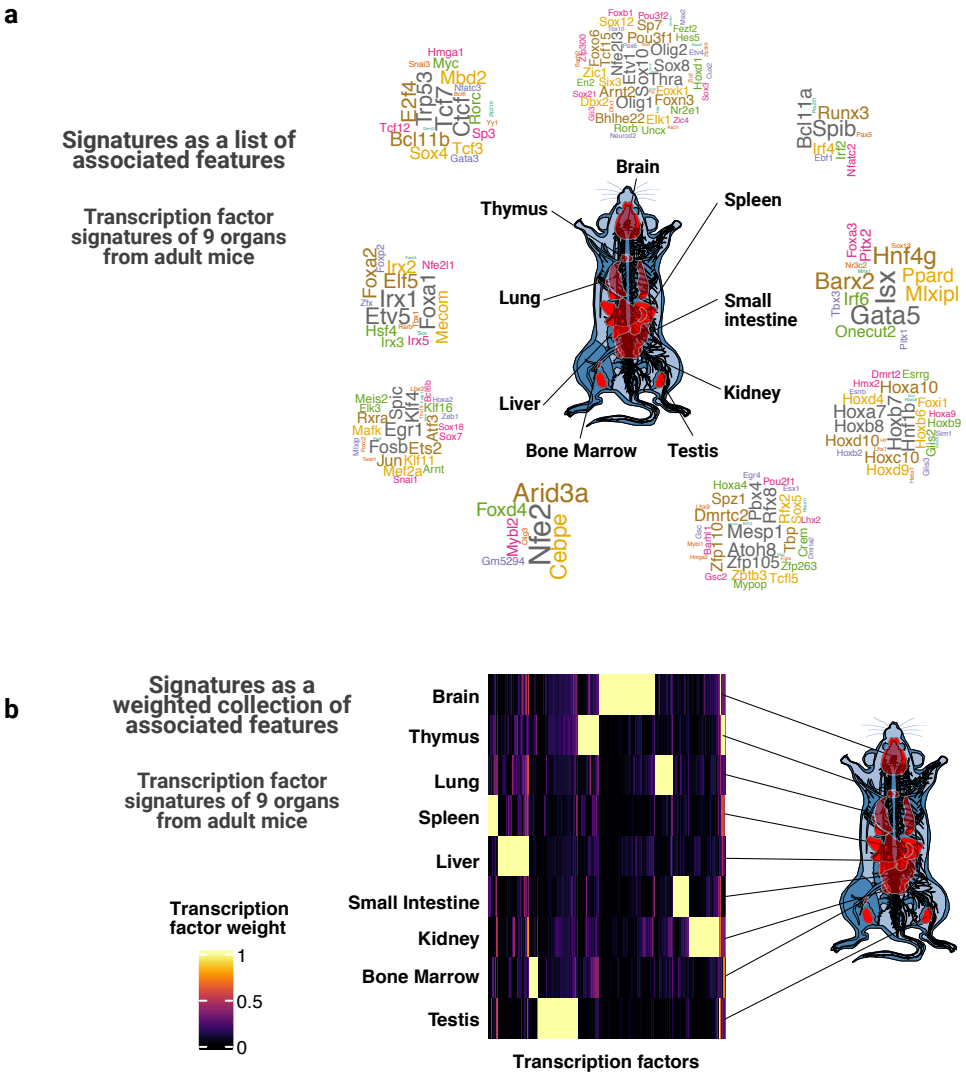


Figure 2.1: Weighted and un-weighted molecular signatures. Example of the two types of molecular signatures using transcription factor signatures of 9 organs from adult mice. **(a)** Signatures are represented as a list of associated features to every organ. **(b)** Signatures are represented as a collection of transcription factors with a weight indicating the strength of the association to every organ.

2.1.1 Usage of molecular signatures

The usage of molecular signatures extends to multiple fields. Perhaps the most known application has been the identification of tumor types in cancer and personalized medicine research (Sotiriou and Piccart 2007; Fröhlich et al. 2018; Olivier et al. 2019). In particular, the pan-cancer analysis of whole genomes (PCAWG) consortium identified point mutations from whole-genome sequencing of 4,645 and whole-exome sequencing data of 19,184 tumor samples to infer 67 mutational signatures (Ludmil B. Alexandrov et al. 2020). Expression signatures are also used in the classification of breast (Sotiriou and Pusztai 2009; Lal et al. 2017), lung (Seijo et al. 2019), and kidney cancer (Graham et al. 2018), among others. Besides cancer research, one of the most common usages of molecular signatures nowadays is the identification of cell types in single-cell transcriptomic data, in which gene signatures (also known as "marker genes") are used to evaluate if individual cells belong to a certain cell state (Kolodziejczyk et al. 2015; Stegle, Teichmann, and Marioni 2015).

Automatically testing large numbers of potential drugs or the combinatorial effect of multiple compounds for activity against biological targets (high-throughput screenings, HTS) has been a well-established strategy for drug discovery by pharmaceutical companies (Wildey et al. 2017; Mayr and Bojanic 2009). The first level of information that can be extracted from such studies are signatures of drug effects, that generally explain changes in gene expression in response to treatment. One of the goals for personalized medicine is to identify signatures that can be used to predict drug sensitivity, meaning that a predictive model will return the most appropriate treatment for every patient (Nevins and Potti 2007).

In general, there are two classic approaches to use signatures for the evaluation of the association between a sample or individual cell to one genotype or phenotype. In the first case, known signatures are used to evaluate the association using a statistical test or deep learning model **(Figure 2.2a)**. Gene set enrichment analysis (GSEA) or similar tests are often used to estimate the association (Subramanian et al. 2005; Hänzelmann, Castelo,

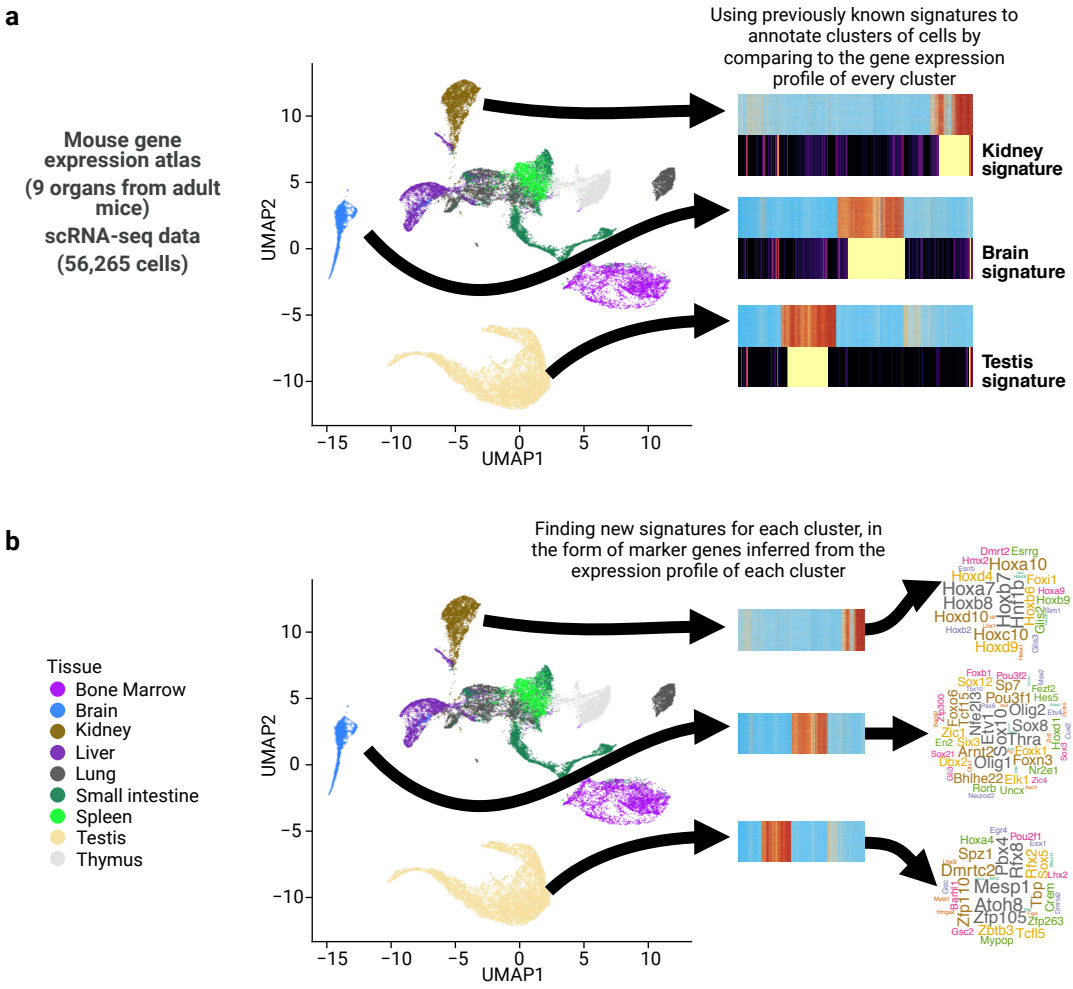


Figure 2.2: Usage of molecular signatures. Example of the usage of molecular signatures for the annotation of clusters of cells using single-cell transcriptomic data of 9 organs from adult mice (Han et al. 2018). **(a)** The expression profiles of every cluster are compared to known signatures using statistical tests like GSEA. **(b)** A gene signature (marker genes) is inferred from each cluster.

and Guinney 2013), or alternatively the collection of known signatures can be used as a training set for machine learning methods like CIBERSORT (Newman et al. 2015). In the second case, the signatures are learned *de novo* from the data, and samples or cells are linked to one signature. The identity of each signature is then determined by finding the association to known biological and clinical variables, or by extracting the most important features of the signature (e.g., marker genes) and contrasting them to other signatures or available literature **(Figure 2.2b)**. For instance, some of the widely known single-cell transcriptomics analysis packages like Seurat (Butler et al. 2018), Scanpy (Wolf, Angerer, and Theis 2018), and Monocle (X. Qiu et al. 2017), use variations of this strategy to identify cell types and annotate clusters of cells.

The programming language and statistical software R (R Core Team 2020) is the perfect analysis platform for molecular signatures, as it counts with a vast collection of packages included in the Bioconductor repository (Huber et al. 2015) for the analysis of biological data. For instance, *signatureSearch* is an R package for searching a query gene expression signature against a database of signatures, which included algorithms to work with multiple types of expression signatures allowing to perform functional enrichment analysis and the reconstruction of drug-target networks (Duan et al. 2020). Other packages like *SigsPack* (Schumann et al. 2019) and *YAPSA* (Hübschmann et al. 2020) can be used to find the exposure of individual tumor samples to a catalog of mutational signatures, these packages also allow to determine the confidence intervals of the estimated exposures providing a precise determination of genomic lesions in cancer patients. In particular, *YAPSA* has been a valuable tool for the analysis of rare tumors by identifying target regions for alternative treatments in cancer patients in which standard therapy options were not sufficient (Horak et al. 2017). Other packages like *SigCheck* (Stark and Norden 2020) are useful to validate molecular signatures by evaluating the performance of one signature against random signatures of the same length, other known signatures, or by performing permutations in the associated data or metadata.

2.1.2 Signature collections and databases

Multiple initiatives have compiled databases for different types of molecular signatures. These resources include signatures from gene expression data, metabolome, and proteome signatures, among others. Such databases are one of the main reasons why the usage of signatures has extended through all fields in genomics, as they provide collections of validated signatures that can be readily used from downstream analyses. The molecular signatures database (MSigDB) is perhaps the most complete database of gene signatures nowadays (Liberzon et al. 2011). This database contains molecular signatures obtained from different types of data and approaches, i.e., some of the signatures are inferred by manual curation and others only using computational tools, and they can be derived based on the genomic location of the genes, pathway information like Reactome (Jassal et al. 2020), regulatory links between *cis*-regulatory elements and gene promoters, expression data, and gene ontology (GO) terms (Carbon et al. 2021). The current version of the MSigDB (v7.2) database contains more than 30,000 signatures; however, in some instances, there is redundancy and inconsistency between these signatures. Addressing this problem, Liberzon et al. (2015) introduced a new set of "hallmark signatures" in the database that summarizes redundant information and is curated by experts.

There are also signature databases that compile data exclusively from one type of biomolecular feature. Databases like the GOLM metabolome database (Kopka et al. 2005) and the Human Metabolome Database (HMDB) (Wishart et al. 2007), compile signatures solely based on metabolite data that were quantified under different tissues and perturbation studies. These types of databases are helpful in drug discovery studies as they recover signatures that are related to changes in the metabolome in response to a drug. Additionally, the HMDB also contains links between metabolites and proteins associated with them. Other databases like InterPro (Blum et al. 2021) and ProTargetMiner (Saei et al. 2019) are compiling signatures based on proteome data. In particular, InterPro contains signatures that can be used by the software InterProScan (Jones et al. 2014) as predictive models for protein classification and domain identification. On the other

hand, ProTargetMiner is a database targeted to help in drug discovery studies in cancer, built from proteomes of cancer cell lines treated by different compounds.

Since the emergence of all the new technologies for profiling gene expression and chromatin accessibility from single cells, cell type identification is a recurrent problem during the analysis of such types of data. Aiming to help in this process, databases of marker genes in different tissues and cell types have been released. For instance, the PanglaoDB contains gene signatures (marker genes) for more than 170 cell types in mouse and human, encompassing 29 different tissues (Franzén, Gan, and Björkegren 2019), most of these signatures are available for both species, streamlining the process of converting between homologous genes. The CellMarker database (Z. Zhang et al. 2019) also aimed to provide marker genes for human and mouse cell types; it was created by the manual curation of more than 100,000 publications, and it contains signatures for 467 human cell types and 389 mouse cell types. There are also databases like SCDevDB (Z. Wang, Feng, and Li 2019) that focus on compiling signatures from developmental pathways identified by finding differentially expressed genes.

Taken together, all these databases provide a wide arrange of pre-computed molecular signatures that can be used in studies in which there is no need of inferring new signatures, or can also be used to validate or identify the phenotypic characteristics of newly inferred signatures.

2.2 Signature inference and feature extraction

Multiple methods have been proposed for the inference of molecular signatures, ranging from data-driven approaches (Bergstrom et al. 2019; Haradhvala et al. 2018; Pfeil et al. 2020; F. Li et al. 2013) to manual curation of scientific publications (Liberzon et al. 2015; Burge et al. 2012). While many methods have the potential to be used in different types of omics data, the majority of the packages and toolkits only focus on one type of data (Sung et al. 2012). Due to the extensive availability of transcriptomic data from bulk

samples and single cells, the identification of gene signatures is by far the most studied field in signature inference in genomics (Fröhlich et al. 2018). Therefore, most of the methods for signature inference are based on the identification of differentially expressed genes by comparing between two conditions (i.e., usually case-control studies). This strategy has been used to identify expression signatures that can be used as predictors of good or poor prognosis in breast (Vijver et al. 2002), lung (Lu et al. 2006), colon (Salazar et al. 2011), and gastric cancer (Cho et al. 2011) among many others. An inherent problem of this type of approach is that they rely on the comparison of two conditions; therefore, it will be more difficult to identify signatures that define subtypes or subclusters in the data.

Due to the current increase and accessibility of single-cell profiling techniques, one of the most relevant uses for the identification of gene expression signatures is the annotation and determination of cell-type identity. Analysis packages such as Seurat (Butler et al. 2018), Scanpy (Wolf, Angerer, and Theis 2018), and Monocle (X. Qiu et al. 2017) determine marker genes that characterize a cell state by extending the classic approach of differentially expressed genes identification. In this case, the single cells are clustered into groups of cells that share similar transcriptomic profiles, and the marker genes are identified by reducing the problem to a comparison of two conditions, i.e., comparing the expression profiles of the cells in one cluster to all the other cells. Whereas this approach has become the standard to identify marker genes in well-defined clusters of cells, it will be challenging to extract signatures that explain continuous processes as differentiation, in which the cells do not form well-defined clusters but rather move along a trajectory (Trapnell 2015; Bendall et al. 2014; Moignard et al. 2015; Buettner et al. 2015).

Besides finding signatures *de novo* (i.e., using only the dataset under consideration), semi-supervised and supervised methods can also be used to guide the identification of relevant features in different clusters of samples. Packages as *SCINA* (Z. Zhang et al. 2019) can be used to determine phenotypic identity using a training set of known signatures for previously identified classes, this package infers new sub-classes and the

corresponding genes associated with them (gene signatures) by using an alternate optimization of the probability estimation for cell type assignment (i.e., E step) and gene expression distribution (i.e., M step). *Biosigner* (Rinaudo et al. 2016) is another package that uses a training set of known signatures for signature discovery. The workflow implemented in this package consists of three main steps: 1) bootstrap resampling of the dataset restricted to the set of features included in the known signatures to build a classifier for two classes, 2) feature ranking for every resampled subset, and 3) selection of significant features. These three steps are iteratively repeated until all candidate features are significant, this final set of features will constitute the new signature. Other packages as *MarkerPen* (Y. Qiu et al. 2020) are aimed to find marker genes of different cell types by refining lists of previously published markers. This is done by comparing bulk transcriptomic profiles to the list of potential marker genes and adding or removing markers based on penalized principal component analysis.

2.3 Difficulty of finding signatures

The inference of new molecular signatures has been a challenging task since the onset of high-throughput profiling technologies. Some of the factors that have contributed to a large extent towards this difficulty are the inconsistency of the metadata associated with publicly available datasets, and the heterogeneity in the strategies for processing raw files for such data. Although databases like the NCBI Gene Expression Omnibus (GEO) (Barrett et al. 2013) or the EBI BioStudies (Sarkans et al. 2018) have become the standard to share high-throughput data produced from biological studies, there is not a global consensus in the scientific community on how to record categorical annotations derived from patient data or experimental conditions. Thus, inferring molecular signatures from previously published data still involves a long manual process of data curation, which can be more complex when datasets from multiple studies are going to be integrated. However, authors like L. Wang, Wang, and Chang (2016) carried out an innovative study to infer gene expression signatures from public data deposited on GEO by splitting the

workload in microtasks assigned to 70 participants of an online course. Every participant was in charge of manually identifying relevant datasets associated with drug, disease, and gene perturbation studies, followed by curation of the associated metadata, and identifying gene signatures by performing a differential gene expression analysis contrasting two conditions using the tool GEO2Enrichr (Gundersen et al. 2015). Studies like these open new venues for inferring molecular signatures by showing how crowdsourcing projects can help with this difficult problem. Nonetheless, in the last years, other projects like recount3 (Collado-Torres et al. 2017), UCSC Xena (Goldman et al. 2020), and UCSC Toil (Vivian et al. 2017) have focused on re-processing publicly available datasets to eliminate the computational batch effect originated by the usage of different tools across multiple studies. Therefore, using data from these resources will allow a consistent and robust integration across studies and will reduce the complexity of gene expression signature inference. Furthermore, the creation of single-cell gene expression and chromatin accessibility atlases for different species will help to remove these difficulties, as all the data and accompanying annotation will be processed and annotated in a systematic way. Examples of such atlases are the recently published human atlases of fetal gene expression (Cao et al. 2020) and chromatin accessibility (Domcke et al. 2020).

By definition, all high-throughput techniques are designed to measure large numbers of biomolecular features in one sample, and despite all the challenges created by inconsistent data processing and metadata annotation, perhaps the biggest difficulty for signature inference has always been the extraction of meaningful information from such high dimensional data (Mramor et al. 2005; Mirza et al. 2019).

Even when considering data originated from one single study, in which there are no computational batch effects or annotation inconsistencies, the challenges that arise from the so-called *curse of dimensionality* will always be present. This expression was coined by Richard Bellmann referring to the exponential increase in volume as a consequence of adding extra dimensions to a feature space (Bellman 1966; Keogh and Mueen 2017). Specifically, the *curse of dimensionality* in life-sciences originates when the number of

measured features greatly exceeds the number of samples, which cause overfitting the model; when the number of features is so large that only a few of them show significant differences across groups of samples or cells as a consequence of redundancy between features; or when the number of measured features is so large that effectively renders classical analysis methods unusable (L. Wang, Wang, and Chang 2016).

2.4 Dimensionality reduction for signature inference

In order to mitigate the curse of dimensionality, several methods (e.g., principal component analysis, factor analysis, non-negative matrix factorization, among others) have been proposed that are able to reduce the data volume to a smaller set of meta-features (Fodor 2002; S. Huang, Chaudhary, and Garmire 2017). In the context of signature inference, dimensionality reduction can be used to identify signatures that explain the essential information of the data, i.e., to extract the signal from the noise. Furthermore, the evaluation of the feature contribution towards each of the meta-features provides a measurement of its relevance to explain certain biological processes (Townes et al. 2019; Bartenhagen et al. 2010).

Most of the dimensional reduction methods used for signature identification belong to the family of unsupervised learning methods (Wong, Li, and Zhang 2016; C. Xu and Jackson 2019). The value of these methods in the life sciences, and in particular in genomics is that they do not require a pre-defined problem structure, i.e., the training set does not have a response vector (Duda, Hart, and Stork 2001; Fodor 2002). This means that in the initial stage of a research project, when no clear origin of the biological variability can be determined or when interrogating data originated from complex populations, it is possible to use these methods to transform the original measurements of biomolecules into less complex and more meaningful signatures (Libbrecht and Noble 2015). In general, the objective of using dimension reduction methods for signature inference is to search for structures and patterns in the original data X by a transformation of the original

variables, resulting in a simpler representation $\tilde{X}$ of the data ($\tilde{X} = \varphi(X)$) (Duda, Hart, and Stork 2001).

Furthermore, in many cases, the generation of omics labeled data is not possible or it is cost-prohibitive. For instance, in single-cell transcriptomics assays, it is possible to use Fluorescence-Activated Cell Sorting (FACS) to enrich one sample for specific cell types based on the presence of pre-selected cell surface markers, effectively generating pre-labeled scRNA-seq data (Baron et al. 2019; Attaf et al. 2020). But on the other hand, it would be nearly impossible to account for surface markers for all possible cell types present in one sample.

A detailed description of some of the most relevant dimension reduction methods that have been used for signature inference and feature extraction is shown in the following sections.

2.4.1 Principal component analysis

Principal Component Analysis (PCA) is the most known algorithm for dimension reduction. Originally invented by Pearson (1901) and later by Hotelling (1933), it has been rediscovered in multiple fields, highlighting its importance and intuitive derivation (Jolliffe and Cadima 2016). The objective of PCA **(Figure 2.3, equation** (2.1)**)** is to find a linear variable transformation to reduce the dimension of the original data $X \in \mathbb{R}^{m \times n}$ without losing information to produce a new matrix $z \in \mathbb{R}^{m \times d}$. Thus, the number of features in the new matrix z will be less than the number of features in the original data matrix X ($d << n$) (Duda, Hart, and Stork 2001).

$$\begin{gathered} z = \varphi(X) = X \cdot V^T \\ \text{where: } X \in \mathbb{R}^{m \times n};\ z \in \mathbb{R}^{m \times d};\ \text{and } d << n \end{gathered} \tag{2.1}$$

In order to transform the original data matrix X, PCA derives an optimal projection matrix V. Different algorithms have been found to derive the PCA projection matrix,

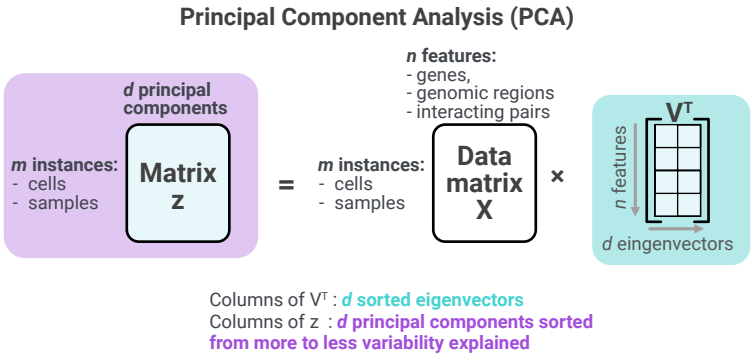


Figure 2.3: Schematic representation of the Principal Component Analysis (PCA) reduction. An optimal projection matrix V^T is found to transform the data matrix X into a matrix z of reduced dimension.

most of these algorithms are deterministic, meaning that there is an optimal unique solution for every matrix X (Hastie 2017). One of the most widely used strategies to derive V is to compute the eigenvectors and the corresponding eigenvalues of the scatter matrix $X \cdot X^T$ (Duda, Hart, and Stork 2001). This is usually done using singular value decomposition (SVD) because of its numerical stability in comparison to eigendecomposition (Nakatsukasa and Higham 2013). An overview of the PCA algorithm can be seen in **Algorithm 1**.

The result of the PCA dimension reduction selects the coordinate system where the data shows the most variance, i.e., the variance of the new features in z is maximized, and the principal components are sorted in decreasing order of the fraction of variance explained (Hastie 2017). An additional important property of PCA is that new features are pairwise uncorrelated, meaning that if different but overlapping biological states are present in the data, PCA may not be the most optimal dimension reduction strategy.

PCA is generally used as a pre-processing step in many other algorithms and workflows. For instance, a common practice to visualize clusters of cells from single-cell transcriptome or chromatin accessibility profiles is to use the UMAP (Diaz-Papkovich et al. 2019)

or tSNE (Van Der Maaten and Hinton 2008) algorithms on the transformed matrix z.

Algorithm 1: Principal component analysis

Input : X : Data matrix $X \in \mathbb{R}^{m \times n}$

d : Desired number of dimension $d < n$

Output: $\tilde{X}$: Data matrix $\tilde{X} \in \mathbb{R}^{m \times d}$

1 Center X, i.e., for each column in X substract the column mean

2 Compute scatter matrix $S = X \cdot X^T$

3 Compute eingen decomposition $S = V \cdot \Lambda \cdot V^T$

4 Sort the eigenvalues in Λ from largest to smallest such that $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_n$

5 Sort the eigenvectors in V following the order of the sorted eigenvalues

6 Compute the new features $\tilde{x}_{i,j} = X_i \cdot V_j^T$ where V_j^T = eigenvectors for $j = 1 \ldots d$

7 Return the transformed data matrix $\tilde{X}$

Regarding the inference of molecular signatures, PCA has been used in many different settings. To exemplify, GO_PCA (Wagner 2015) combines PCA with gene ontology enrichment analysis to identify sets of genes that show similar expression patterns, and that have closely related biological functions. This is done by performing PCA and testing every principal component for an association of functionally related genes. PCA was also used in combination with a deep neural network classifier to perform feature extraction for protein structure prediction (Melo, Cavalcanti, and Guimarães 2003). In a similar approximation, Kavitha et al. (2018) coupled PCA to the classification algorithm SVM-Recursive Feature Elimination (SVM-RFE) (Guyon et al. 2002) to recover gene signatures for acute myeloid leukemia (AML) and acute lymphoblastic leukemia (ALL). Furthermore, Berglund, Welsh, and Eschrich (2017) proposed a series of procedures to assess the validity of gene signature scores inferred using PCA, these procedures are based on the comparison of the principal components (PCs) of a gene signature against a set of PCs from randomized gene signatures (i.e., in this instance a signature score is obtained by performing PCA including only the signature genes and using PC1 as the signature

score). In particular, the signature coherence is measured by comparing the amount of variance explained in PC1, the robustness compares the ratio of variance explained in PC1 and PC2, and the uniqueness compares the correlation value of the PC1 of the gene signature to the PC1 obtained by including all genes.

2.4.2 Independent component analysis

Independent component analysis (ICA) was initially proposed to solve the cocktail party problem or unmixing problem, in which the idea is to recover the individual voice recordings from a group of speakers in the same room (Hastie 2017). This idea has been extended to multiple fields including genomics, in which the most natural application is to recover meaningful signatures from a high dimensional dataset (in this instance the signatures are analogous to the speakers, and the biomolecular features to the microphones in the room used to record the voice of the speakers) (Sompairac et al. 2019).

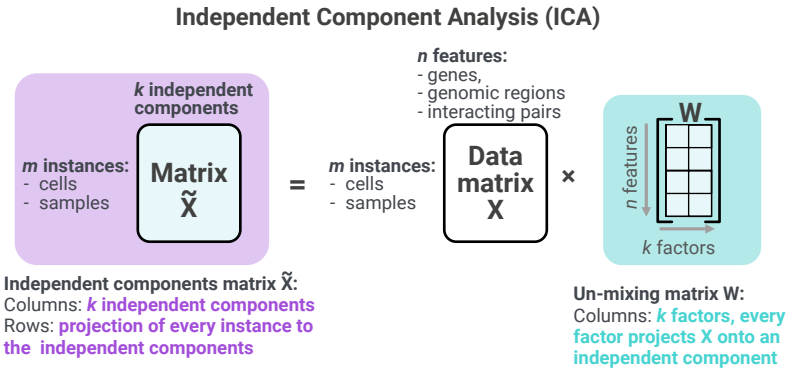


Figure 2.4: Schematic representation of the Independent Component Analysis (ICA) reduction. To apply ICA the matrix X has to be whitened beforehand. ICA decomposes X into the matrix $\tilde{X}$ with statistically independent columns and the orthogonal matrix A.

ICA is similar to PCA, in the sense that the objective is to find a linear variable transformation $\varphi(X)$ for the original data $X \in \mathbb{R}^{m \times n}$ to determine a matrix $\tilde{X} \in \mathbb{R}^{m \times k}$ of reduced dimension **(Figure 2.4)**. In the case of ICA, this transformation is done to find factors that are as mutually independent as possible, which means that the matrix A in **equation** (2.2) is orthogonal. One of the advantages of ICA over PCA is that the decomposed factors are easier to interpret as they have the same importance, and are not sorted by degree of variance explained (Hastie 2017; Duda, Hart, and Stork 2001).

$$
\begin{gathered}
X = \varphi(X) = \tilde{X} \cdot A + \epsilon \\
\text{where: } X \in \mathbb{R}^{m \times n};\ \tilde{X} \in \mathbb{R}^{m \times k};\ A \in \mathbb{R}^{k \times n};\ \text{and } k << n \\
\text{Which is equivalent to:} \\
\tilde{X} = X \cdot W + \epsilon \\
\text{where: } W \text{ is the pseudo-inverse of } A; W = (A^T A)^{-1} A^T
\end{gathered}
\tag{2.2}
$$

Different algorithms have been proposed to solve the ICA problem, and perhaps the most used implementation is FastICA from Hyvärinen and Oja (2000), depicted in **Algorithm 2**. The ICA problem is not convex, meaning that there is no global minimum, and the final results will depend on the initialization of the column vectors of the unmixing matrix W.

ICA has been used to infer molecular signatures from multiple types of genomic data and following different strategies. For instance, ICA can be used to maximize the independence of metagenes (i.e., gene signatures) (Kairov et al. 2017; S. I. Lee and Batzoglou 2003; Biton et al. 2014) or metasamples (i.e., sample signatures) (Meng et al. 2016). The first case corresponds to applying ICA as depicted in **Figure 2.4**, in which the data matrix X contains the samples in the rows, and the biomolecular features in the columns, while the second case corresponds to applying ICA to a transposed data matrix X (i.e., features in the rows, and samples in the columns). Due to its property for un-mixing factors, ICA is a popular choice for the deconvolution of bulk omics samples into individual cell types. In this context, the R package DeconICA (Czerwinska 2018) implements FastICA to estimate cell type proportions from bulk RNA-seq samples.

Algorithm 2: Independent component analysis

Input : X : Data matrix $X \in \mathbb{R}^{m \times n}$

k : Desired number of factors $k < n$

Output: $\tilde{X}$: Independent components matrix $\tilde{X} \in \mathbb{R}^{m \times k}$

W : Un-mixing matrix $W \in \mathbb{R}^{k \times n}$

Center X, i.e., for each column in X substract the column mean

Whitening of the matrix X by projecting the data onto the k principal components

for $d \leftarrow 1$ **to** k **do**

    initialize W_d randomly

    while W_d *changes* **do**

        Optimize the columns of W via Newton iterations

        $w^+ = \frac{1}{N} X_i^T g(X \cdot W_d) - \frac{1}{N} W_d^T g'(X \cdot W_d)$

        $w^+ = w^+ - \sum_{j=1}^{d-1} (w^+ \cdot W_j) W_j$ Orthogonalization

        $w^+ = \frac{w^+}{||w^+||}$ Normalization

    end

end

Compute $\tilde{X} = X \cdot W$

Return the matrix $\tilde{X}$ and W

One of the crucial steps in the analysis of single-cell RNA-seq data is to reduce the dimension of the original data matrix. By default, popular analysis toolkits as Seurat (Butler et al. 2018), and Monocle (X. Qiu et al. 2017) use PCA for such task. However, as the independent factors learned by ICA could be more informative in some cases, the option to reduce the data dimensionality using ICA is also implemented.

2.4.3 Non-negative matrix factorization

Non-negative matrix factorization (NMF) (Seung and Lee 1999) is another method of the big family of unsupervised dimensionality reduction algorithms. The goal of the NMF

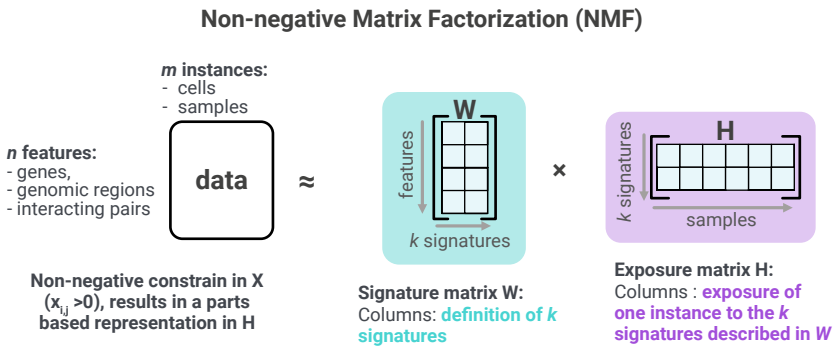


Figure 2.5: Schematic representation of the NMF decomposition. A non-negative input matrix is decomposed into a signature matrix W and an exposure matrix H. The non-negative constrain in X results in parts based representation of the data (additive factors).

(Figure 2.5, equation (2.3)) is to decompose a data matrix $X \in \mathbb{R}^{n \times m}$ into a signature matrix $W \in \mathbb{R}^{n \times k}$ and an exposure matrix $H \in \mathbb{R}^{m \times k}$, such as $X \approx WH$, where k is the total number of factors (i.e., signatures). In contrast to PCA and ICA, NMF imposes a non-negative constrain in the input data matrix X, resulting in enhanced interpretability of the decomposed factors as their combination is additive.

$$
\begin{gathered}
X = W \cdot H + \epsilon \\
\text{where: } X \in \mathbb{R}^{n \times m};\ W \in \mathbb{R}^{n \times k};\ H \in \mathbb{R}^{k \times m};\ \text{and } k << m \\
\text{With objective function:} \\
W, H = \underset{W,H}{\arg\min} \, ||X - WH||_F^2
\end{gathered}
\tag{2.3}
$$

NMF is by itself a big family of algorithms. Although many different approaches have been proposed in the last decade to solve the NMF decomposition, most of these approximations are similar to the original multiplicative update rules initially proposed by Seung and Lee (1999) **(Algorithm 3)**. As well as ICA, the solution of the NMF objective function is not convex. Thus, the results will vary depending on the strategy or method used to initialize the matrices W and H.

Algorithm 3: Non-negative matrix factorization

Input : X : Data matrix $X \in \mathbb{R}^{n \times m}$

k : Factorization rank (desired number of factors) $k < n$

Output: W : Signature matrix $W \in \mathbb{R}^{n \times k}$

H : Exposure matrix $H \in \mathbb{R}^{k \times m}$

initialize W and H randomly

for $i \leftarrow 1$ **to** T *or until convergence* **do**

  Update H:

  $h_{num} = W^T \cdot X$

  $h_{den} = W^T \cdot W \cdot H$

  $H = H * \frac{h_{num}}{h_{den}}$ elementwise multiplication and division

  Update W:

  $w_{num} = X \cdot H^T$

  $w_{den} = W \cdot H \cdot H^T$

  $W = W * \frac{w_{num}}{w_{den}}$ elementwise multiplication and division

end

Return the signature matrix W and the exposure matrix H

Originally proposed for image analysis (Seung and Lee 1999), NMF has been extended to multiple fields. For instance, NMF can be used to: create a recommender system for customer preferences in online sales platforms or movie streaming services (Sheng Zhang et al. 2006; W. Song and Li 2019; T. Li et al. 2006; Shi 2020), identify email subcollections or text mining (M. W. Berry and Browne 2005), and for signal denoising (Wilson et al. 2008). Moreover, in the life sciences NMF is a particularly useful tool because many biological processes are shaped by non-negative contributions (Brunet et al. 2004), e.g., measurement of mRNA transcript levels, transcription factor activation, and protein activity, among others. NMF has been used in different settings for the analysis of genomic data, including *de novo* identification of mutational signatures (Ludmil B. Alexandrov et

al. 2013; Pal et al. 2014), cell-type classification (Shao and Höfer 2017), and metagene extraction (Brunet et al. 2004; Moffitt et al. 2015; Y. E. Li et al. 2017).

Algorithm 4: Graph regularized NMF with sparse coding

Input : X : Data matrix $X \in \mathbb{R}^{n \times m}$

G : Square matrix $G \in \mathbb{R}^{m \times m}$ representing a graph

k : Factorization rank (desired number of factors) $k < n$

Output: W : Signature matrix $W \in \mathbb{R}^{n \times k}$

H : Exposure matrix $H \in \mathbb{R}^{k \times m}$

initialize W and H randomly

if G *is empty* **then**

  Compute and adjacency graph G between the columns of X

end

$D_{i,j} = \sum_j G_{i,j}$

for $t \leftarrow 1$ **to** T *or until convergence* **do**

  Update H:

  $h_{num} = 2(W^T X + \lambda G H) - \alpha$

  $h_{den} = 2(H W^T W + \lambda D H)$

  $H = H * \frac{h_{num}}{h_{den}}$ elementwise multiplication and division

  Update W:

  $w_{num} = X \cdot H^T$

  $w_{den} = W \cdot H \cdot H^T$

  $W = W * \frac{w_{num}}{w_{den}}$ elementwise multiplication and division

  $W = \frac{W}{\sum_i W_{i.j}}$

end

Return the signature matrix W and the exposure matrix H

Among all the NMF algorithms, the Graph Regularized Non-negative Matrix Factorization with Sparse Coding (GRNMF_SC) and all its related variations are of particular interest for the analysis of biological data. The GRNMF_SC algorithm **(Algorithm 4)**

can be used to incorporate previous knowledge of the relationship between the columns of the input matrix X, which can be exploited in those cases where the association between samples is already available from the metadata. Furthermore, if previous knowledge of the interaction between the biomolecular features is already available, the GRNMF_SC decomposition can be performed in the transposed matrix X, i.e., transposing the matrix X depicted in **Figure 2.5** to obtain a matrix with samples in the rows and features in the columns. The previous knowledge is given to GRNMF_SC as a square matrix $G \in \mathbb{R}^{m \times m}$ representing a graph between columns of the input matrix X, the values in G correspond to the weight of the edges connecting nodes in the graph. GRNMF_SC can also be used as a regular unsupervised method by computing an adjacency graph between the columns of X. Variations of this algorithm have been already used to: predict the association of aberrant microRNAs with diseases (Gao et al. 2020; Xiao et al. 2018), cluster cancer samples, and extract relevant genes by combining multiple sources of gene expression information (Yu et al. 2019; C. Y. Wang et al. 2019), and to predict interactions between long non-coding RNAs and microRNAs (M. N. Wang et al. 2020).

Some R packages have native NMF implementations that can be applied to decompose multiple types of data. In particular, *NMF* (Gaujoux and Seoighe 2010) uses the multiplicative rules shown in **(Algorithm 3)**, and *NNLM* proposes an NMF algorithm using sequential coordinate-wise descent (X. Lin and Boutros 2020). Other packages are designed to work with only one type of genomic data. For instance, SigProfiler (Bergstrom et al. 2019) and SignatureAnalyzer (Haradhvala et al. 2018) use NMF to derive mutational signatures from somatic mutations; scPNMF uses NMF to extract consensus features from single-cell data (D. Song et al. 2021). However, none of these packages implements more than one NMF algorithm, and usually the biological significance of the NMF signatures has to be assessed using other statistical packages. Furthermore, despite its enhanced interpretability, the usage of NMF to infer molecular signatures requires fine-tuning the optimal factorization rank, which results in long computing times for large data matrices.

2.4.4 Uniform manifold approximation and projection

A drawback of linear dimensionality reduction methods, such as PCA, ICA, and NMF is the difficulty to visualize clusters and preserve the global structure from complex datasets by inferring two or three factors, as a consequence of non-linear interactions among the measured variables. In contrast, non-linear dimensionality methods like t-Distributed Stochastic Neighbor Embedding (t-SNE) (Van Der Maaten and Hinton 2008) and Uniform Manifold Approximation and Projection (UMAP) (Diaz-Papkovich et al. 2019) have been specifically tailored to visualize clusters and data structures in as few as possible dimensions. However, the computational cost of these algorithms is too high for big datasets (e.g., scTNA-seq and scATAC-seq), and it is not feasible to use them directly on the original data. To circumvent this limitation, a common practice is to perform one first round of linear dimensionality reduction using PCA followed by t-SNE or UMAP.

The goal of algorithms such as UMAP, t-SNE, and kernel PCA is to apply a non-linear coordinate transform on the data, by first mapping the data non-linearly into an augmented feature space. In particular, the UMAP algorithm **(Algorithm 5)** consists of two main steps, projecting the data onto a high-dimensional graph representation, and then optimizing it into a low-dimensional graph which preserves the global structure of the data. The initial high-dimensional graph also known as the "fuzzy simplicial complex" represents the likelihood that two points are connected by weighting the edges of the graph. The construction of this graph starts by extending a radius from each data point and connecting two points whenever their radii overlap. The graph is called "fuzzy" because UMAP decreases the likelihood of a connection as the radius increases.

Despite it has been said that UMAP is better than t-SNE and other dimensional reduction algorithms in preserving the global structure of the data (Diaz-Papkovich et al. 2019), the hyperparameters of the algorithm allow a trade-off between preserving local or global structure. For instance, a low number of neighbors k considered for the construction of the "fuzzy simplicial complex" will push the algorithm towards clusters with a lower

granularity, while losing the definition of the global structure. Likewise, a large number of neighbors will produce well-defined global clusters, at the expense of losing the resolution of the local structure.

Algorithm 5: UMAP

Input : X : Data matrix $X \in \mathbb{R}^{m \times n}$

k : the neighborhood size to use for local metric approximation

d : Dimension of the target reduced space (desired number of factors) $d < n$

min-dist : an algorithmic parameter controlling the layout

n-epochs :controlling the amount of optimization work to perform

Output: Y : Reduced dimension matrix $Y \in \mathbb{R}^{m \times d}$

Construct weighted graph:

for $x \in X$ **do**

  fs-set$[x]$ = LocalFuzzySimplicialSet(X, x, k)

end

top-rep$[x] = \bigcup_{x \in X}$ fs-set$[x]$

Optimize graph layout:

Y = SpectralEmbedding(top-rep, d)

Y = OptimizeEmbedding(top-rep, Y, min-dist, n-epochs)

Return the reduced dimension matrix Y

2.5 Signature inference by data integration

Gene expression data was originally thought to be sufficient to unravel the complex mechanisms that underlie the regulation of gene expression in the cell. However, it has become clear that epigenomic changes play a crucial role (Heintzman and Ren 2009). For instance, chromatin conformation changes are necessary for the recruitment of transcription factors to the gene promoter in order to start transcription (Spitz and Furlong 2012). Furthermore, it has been found that distal regulatory elements (i.e., non-coding

sites in the genome that are not in close proximity to the gene promoter) are highly specific across different cell types in comparison to the gene promoter (Corces et al. 2018; Yao, Berman, and Farnham 2015). Thus, explaining the interplay between gene expression and the epigenome is the key to understanding the distinct patterns of regulation that give rise to different cell types, and diseases (Corces et al. 2018). To exemplify, the integration of bulk gene expression (i.e., RNA-seq) and chromatin accessibility (i.e., ATAC-seq) data has been used to identify gene signatures in human α- and β-cells (A. M. Ackermann et al. 2016), identify active transcription factors and target genes in infantile hemangiomas (X. Li et al. 2020), and to explain the hematopoietic development and leukemia evolution in humans (Corces et al. 2016).

Less than one decade ago was not possible to quantify gene expression at the single-cell resolution. However, it is now becoming a common practice in many studies. Furthermore, it is also currently possible to measure more than one type of biomolecular feature in the same single cell (i.e., multi-omics data). In this regard, technologies like scCAT-seq (Liu et al. 2019) and SHARE-seq (Ma et al. 2020) allow the simultaneous profiling of gene expression and chromatin accessibility, or like scNMT-seq that provides simultaneous measurements of gene expression, DNA methylation, and chromatin accessibility (Clark et al. 2018). All these technologies expanded the limits of the potential knowledge extracted from single-cell studies to a new level. With such data, it is feasible to create models that explain the coordinated regulation between gene expression and chromatin changes (Danese et al. 2019; Stuart et al. 2019).

2.5.1 Strategies to integrate multi-omics datasets

The integration of either bulk or single-cell multi-omics data can be performed by finding a set of meta-features that are partially explained by multiple data types (S. Huang, Chaudhary, and Garmire 2017). Therefore, these meta-features can be used to identify signatures and patterns of regulation that capture more information than only using one type of data. Originally designed for the analysis of only one dataset at the same time,

many dimensionality reduction methods (e.g., PCA, ICA, factor analysis, NMF) have been adapted to integrate multiple data types, to learn a common set of meta-features across data modalities (Cantini et al. 2021; S. Huang, Chaudhary, and Garmire 2017; Meng et al. 2016). For instance, Joint and Individual Variation Explained (JIVE) (Lock et al. 2013) builds upon PCA, Multi-Omics Factor Analysis (MOFA) (Argelaguet et al. 2018, 2020) and Multi-Study Factor Analysis (MSFA) (De Vito et al. 2019) are extensions of factor analysis, and Joint NMF (Shihua Zhang et al. 2012) and Integrative NMF (Yang and Michailidis 2015) are variations of the original NMF algorithm. In general, these methods work by using different "views" (i.e., data modalities like gene expression, DNA methylation, and chromatin accessibility) of the data instances (i.e., samples or single cells), and combining them into factors that are partially explained from different features.

The meta-features shared across data modalities and identified using integrative approaches can be of two different natures. In the first place, in experiments in which two or more biomolecular features were measured for all the samples or cells under consideration, the meta-features represent groups of samples or cells that have similar biological properties. On the other hand, in studies where the same type of genomic data was measured in different experiments (e.g., scRNA-seq data from different patients), the meta-features represent groups of biomolecular features (e.g., genes, genomic regions).

2.5.2 Joint non-negative matrix factorization

Among the multi-omics dimensional reduction methods, variations of the NMF algorithms are of particular interest for signature inference because of the additive relationship between the decomposed factors. This property is helpful to understand the amount of information that every genomic layer is contributing towards one factor. The first approach to extend NMF for the analysis of multi-omics data was proposed by Shihua Zhang et al. (2012) with the Joint NMF (jNMF) algorithm **(Figure 2.6, equation** (2.4)**)**,

which has been the building stone for many other improvements and variations of multi-omics NMF algorithms, such as integrative NMF (Yang and Michailidis 2015; Chalise and Fridley 2017), coupled NMF (Duren et al. 2018), and orthogonality-regularized NMF (Strazar et al. 2016).

The goal of jNMF is to decompose n data matrices $[X_1, X_2 \dots X_n]$ (each matrix representing one genomic modality) into n signature matrices $[W_1, W_2 \dots W_n]$ and one exposure matrix H **(equation (2.4))**. The relative contribution of every feature to the decomposed signatures is encoded in the corresponding matrix W_i, while every row of the matrix H corresponds to one signature.

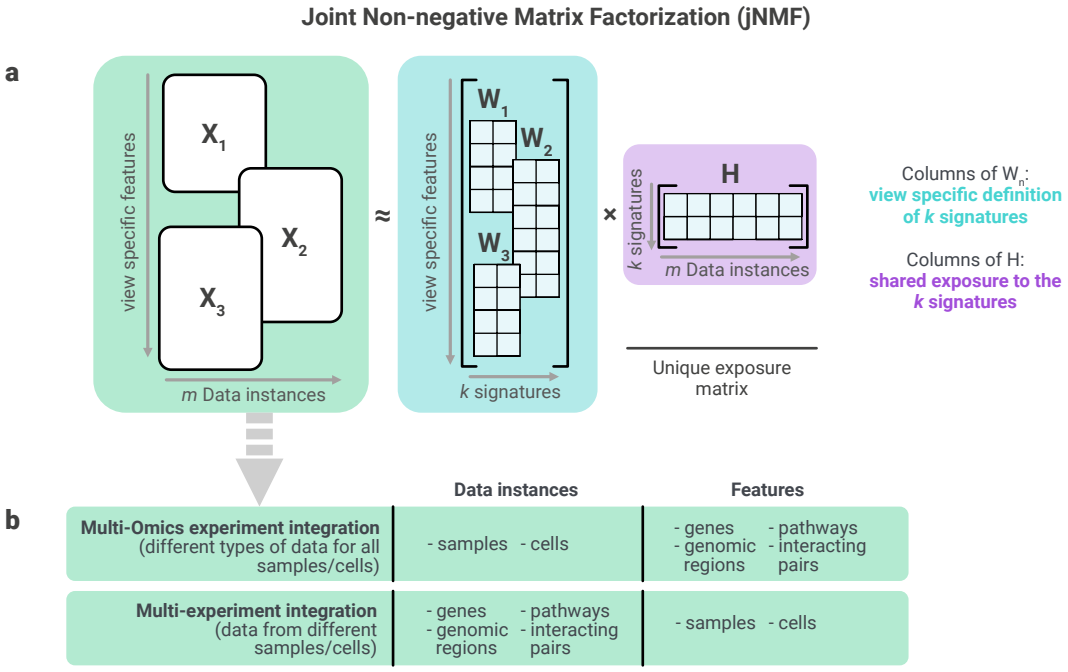


Figure 2.6: Schematic representation of the joint NMF (jNMF) algorithm. **(a)** Two or more non-negative input matrices are decomposed into view-specific signature matrices W_i, and a exposure matrix H. **(b)** Input data matrices X_n must share all the features defined in the columns.

$$\min \sum_{i}^{n} ||X_i - W_i H||_F^2$$

where:

$$\begin{aligned} [X_1, X_2 \dots X_n] &: \text{ List of } n \text{ data matrices } X_i \in \mathbb{R}^{d_i \times m} \\ d_1, d_2 \dots d_n &: \text{ number of rows of each data matrix } X_i \\ m &: \text{ number of shared columns for matrices } X_1, X_2 \dots X_n \end{aligned} \tag{2.4}$$

Algorithm 6: Join non-negative matrix factorization

Input : $X_1, X_2 \dots X_n$: List of n data matrices $X_i \in \mathbb{R}^{d_i \times m}$

$d_1, d_2 \dots d_n$: number of rows of each data matrix X_i

m : number of columns of matrices $X_1, X_2 \dots X_n$

k : Factorization rank (desired number of factors) $k < m$

Output: $W_1, W_2 \dots W_n$: List of n signature matrices $W_i \in \mathbb{R}^{d_i \times k}$

H : Exposure matrix $H \in \mathbb{R}^{k \times m}$

initialize $W_1, W_2 \dots W_n$ and H randomly
for $t \leftarrow 1$ **to** T *or until convergence* **do**
    Update H:
    $h_{num} = \sum_i W_i^T \cdot X_i$
    $h_{den} = (\sum_i W_i^T \cdot W_i) \cdot H$
    $H = H * \frac{h_{num}}{h_{den}}$ elementwise multiplication and division
    Update $W_1, W_2 \dots W_n$:
    for $i \leftarrow 1$ **to** n **do**
        $w_{num} = X_i \cdot H^T$
        $w_{den} = W_i \cdot H \cdot H^T$
        $W_i = W_i * \frac{w_{num}}{w_{den}}$ elementwise multiplication and division
    end
end
Return the signature matrices $[W_1, W_2 \dots W_n]$ and the exposure matrix H

As shown in **Figure 2.6a** the columns of the input data matrices have to be shared across all data modalities. This means that jNMF is able to integrate multi-omics experiment datasets for which different biomolecular features were measured for the same sample or cell. Thus, the number of measured features for each data type can be different across views, and only the columns have to be shared. On the other hand, jNMF can also be used for the integration of multi-experiment data, i.e., datasets comprised of a collection of matrices quantifying the same features across different experimental settings. In those cases, the input matrices have to be transposed, resulting in matrices where the data instances (i.e., samples or cells) are in the rows, and the columns are the shared biomolecular features across all experiments **(Figure 2.6b)**. Independently of the type of integrative analysis, the jNMF algorithm always takes as input a list of matrices with shared columns across them (n views), returning one signature matrix for every view and one shared matrix H **(Algorithm 6)**.

2.5.3 Integrative non-negative matrix factorization

One of the disadvantages of jNMF is that this algorithm is only able to capture the homogeneous effect (i.e., the effect that is shared across all data modalities). Thus, the signatures inferred with jNMF are not able to distinguish the effect that comes from only one data modality (i.e., specific or heterogeneous effect). The iNMF algorithm proposed by Yang and Michailidis (2015), aims to decompose n data matrices $[X_1, X_2 \dots X_n]$ into n signature matrices $[W_1, W_2 \dots W_n]$, n view specific exposure matrices $[H_{v1}, H_{v2} \dots H_{vn}]$, and one shared exposure matrix H **(Figure 2.7, equation** (2.5)**)**.

$$\min \sum_{i}^{n} ||X_i - W_i(H + H_{vi})||_F^2 + \lambda \sum_{i}^{n} ||W_i H_{vi}||_F^2$$

where:

$$[X_1, X_2 \dots X_n]: \text{ List of } n \text{ data matrices } X_i \in \mathbb{R}^{d_i \times m}$$

$$d_1, d_2 \dots d_n: \text{ number of rows of each data matrix } X_i$$

$$m: \text{ number of shared columns for matrices } X_1, X_2 \dots X_n \tag{2.5}$$

Although the iNMF algorithm **(Algorithm 7)** is capable of capturing the heterogeneous effect on the matrices H_{vi}, it comes at the price of increasing execution times. Furthermore, the hyperparameter λ has to be tuned to control the amount of heterogeneous effect that is captured in these matrices.

The R package LIGER implements iNMF to integrate multiple scRNA-seq datasets (Welch et al. 2019), using a multi-experiment data integration approach **(Figure 2.7b)**. LIGER is designed to work with single-cell transcriptomic data, and the integration is performed by using the set of common genes across multiple datasets. This algorithm has become more popular in the last years, as it has been shown to be among the top-performing methods for single-cell data integration (Cantini et al. 2021; Forcato, Romano, and Bicciato 2021).

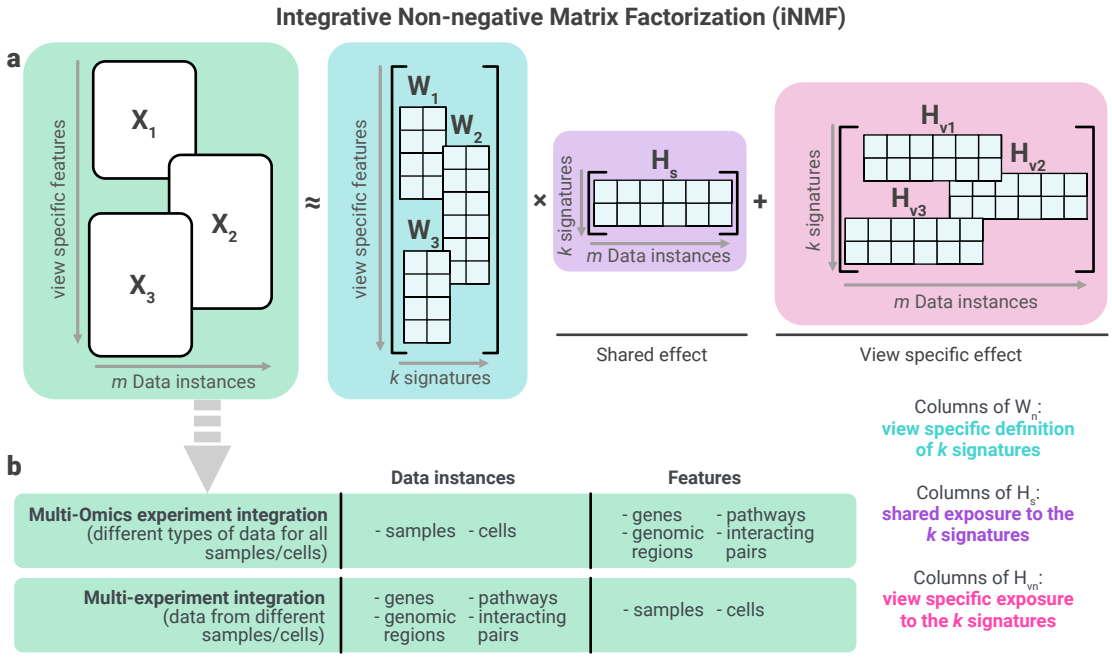


Figure 2.7: Schematic representation of the integrative NMF (iNMF) algorithm. **(a)** Two or more non-negative input matrices are decomposed into view-specific signature matrices W_i, a shared exposure matrix H_s, and view-specific exposure matrices H_{vn}. **(b)** Input data matrices X_n must share all the features defined in the columns.

Algorithm 7: Integrative non-negative matrix factorization

Input : $X_1, X_2 \dots X_n$: List of n data matrices $X_i \in \mathbb{R}^{d_i \times m}$

$d_1, d_2 \dots d_n$: number of rows of each data matrix X_i

m : number of columns of matrices $X_1, X_2 \dots X_n$

k : Factorization rank (desired number of factors) $k < m$

Output: $W_1, W_2 \dots W_n$: List of n signature matrices $W_i \in \mathbb{R}^{d_i \times k}$

H : Shared exposure matrix $H \in \mathbb{R}^{k \times m}$

$H_{v}1, H_{v}2 \dots H_{v}n$: List of n view specific exposure matrices $H_i \in \mathbb{R}^{k \times m}$

initialize $W_1, W_2 \dots W_n$ and H randomly

for $t \leftarrow 1$ **to** T *or until convergence* **do**

  Update H:

  $h_{num} = \sum_i W_i^T X_i$

  $h_{den} = \sum_i (W_i^T W_i) \cdot (H + H_{v}i)$

  $H = H * \frac{h_{num}}{h_{den}}$ elementwise multiplication and division

  Update $W_1, W_2 \dots W_n$:

  for $i \leftarrow 1$ **to** n **do**

    $H_c = H + H_{v}i$

    $w_{num} = X_i \cdot H_c^T$

    $w_{den} = W_i \cdot (H_c H_c^T + \lambda H_{v}i H_{v}i^T)$

    $W_i = W_i * \frac{w_{num}}{w_{den}}$ elementwise multiplication and division

  end

  Update $H_{v}1, H_{v}2 \dots H_{v}n$:

  for $i \leftarrow 1$ **to** n **do**

    $h_{num} = \sum_i W_i^T X_i$

    $h_{den} = (W_i^T W_i) \cdot (H + \lambda H_{v}i)$

    $H_{v}i = H_{v}i * \frac{h_{num}}{h_{den}}$ elementwise multiplication and division

  end

end

Return $[W_1, W_2 \dots W_n]$, H, $[H_{v}1, H_{v}2 \dots H_{v}n]$

2.6 Perspectives for signature inference using NMF

Among the dimensionality reduction methods, NMF is a promising candidate to derive molecular signatures. However, applying NMF algorithms to big datasets is a challenging task due to the long computing times. Similar to ICA, factor analysis, and other unsupervised methods for dimension reduction, the optimal number of factors (i.e., factorization rank) has to be manually selected, which usually leads to underestimation or overestimation of the real signature number present in the data.

As detailed in the Scope of this thesis, the main objective of this work was to provide insights into using NMF as a tool for the inference of molecular signatures.

Part I. Tool Development

Community building and open research are becoming the central pillars of modern research. Creating collaborative networks and sharing resources help the scientific community to progress and also to speed up research projects that otherwise would take decades to accomplish. From the software and method development point of view, open research has been essential to share and improve complex algorithms that build the base of nowadays bioinformatics and computational biology.

In particular, the complete family of Non-negative Matrix Factorization (NMF) algorithms has been used in multiple opportunities to better understand complex datasets in the life sciences, by decomposing matrices into signatures (i.e., the most essential parts of the data). Nevertheless, the usability of these algorithms is lessened by limited resources to interpret the results obtained from them. Therefore, here (**Part I**) we describe ButchR, a new R package implementing multiple NMF algorithms, and a collection of visualization tools to understand the most essential features of a high-throughput genomic dataset (chapter: "ButchR: NMF suit to slice genome-scale datasets"). Being community building and resource sharing a key to bring tools as ButchR to everyone's hands we also present here ShinyButchR a free to use interactive application that implements the main features from ButchR (chapter: "ShinyButchR: Interactive analysis and exploration of NMF results"). To extend the usage of ButchR to complex multi-omics datasets, we also detail i2NMF a new ButchR-based workflow that tackles the inference of common and individual signatures across omics (chapter: "i2NMF: An integrative approach to discover dataset-specific effects").

Chapter 3

ButchR: NMF suit to slice genome-scale datasets

Disclosure: The results presented in this chapter have been published in Quintero et al. (2020) and reproduced here with the permission of Oxford University Press, license number 5011370897521.

Non-negative Matrix Factorization (NMF) has been widely used for the analysis of genomic data to perform feature extraction and signature identification due to the interpretability of the decomposed signatures (Brunet et al. 2004; Ludmil B. Alexandrov et al. 2013; Pal et al. 2014). However, running a basic NMF analysis requires the installation of multiple tools and dependencies, along with a steep learning curve and computing time. To mitigate such obstacles, we developed ButchR and ShinyButchR (Quintero et al. 2020), a novel NMF suit that provides a complete NMF-based analysis workflow, allowing the user to perform matrix decomposition using NMF, feature extraction, interactive visualization, relevant signature identification, and association to biological and clinical variables.

The aim of ButchR and ShinyButchR is to provide a fast and scalable NMF framework, which enables the user to decompose an input matrix X into a signature matrix W and

an exposure matrix H (**Figure 2.5**). This results in a low-dimensional representation of the input dataset, identifying signatures/factors which help to understand the underlying biological processes and potential differences occurring between different samples.

3.1 ButchR

Several R packages (R Core Team 2020) have implemented NMF algorithms (Gaujoux and Seoighe 2010; X. Lin and Boutros 2020). Nevertheless, extracting relevant biological and clinical information from genome-scale datasets may be challenging given the size of the typical datasets. In particular, feature extraction (e.g., relevant genes, genomic regions) and signature identification (e.g., patterns of gene expression that can be associated with biological processes) are two of the most important tasks during data analysis. Thus, ready-to-use and biological-oriented software is of great importance to allow fast data exploration and analysis.

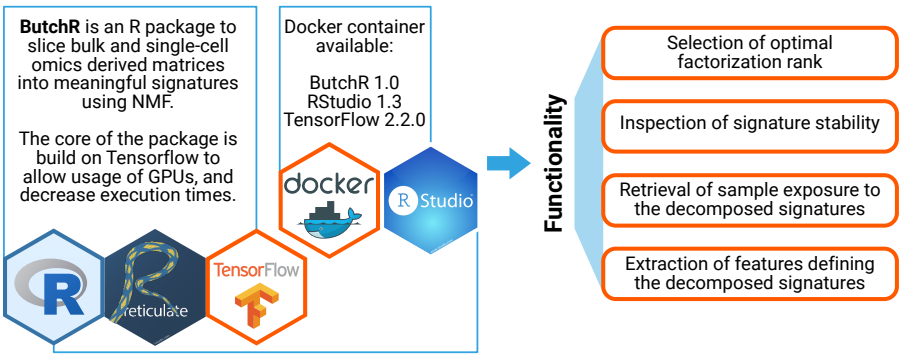


Figure 3.1: Schematic representation of the R package ButchR framework.

ButchR is implemented as an R package, providing solvers for algorithms of the NMF family, functions for downstream analysis, a rational method to determine the optimal factorization rank, and a novel feature selection strategy. All the NMF algorithms included in ButchR are implemented on TensorFlow (Abadi et al. 2016), which allows its highly efficient execution under multiple systems (e.g., CPU, GPU, and TPUs systems).

To retrieve the decomposition results, ButchR uses the Reticulate framework (Allaire et al. 2017), connecting Python and R in a seamless way (**Figure 3.1**).

In comparison to the classic implementations of the NMF algorithms, in ButchR we use a novel way to evaluate the convergence of the matrix decomposition. Instead of using the Frobenius norm of the residuals from the original matrix with the product of matrices W and H ($||X - WH||^2_{Fro}$) as the objective function, we evaluate the stability of decomposition. In this approach, at the end of every iteration (i.e., after the matrices W and H have been updated) each sample/cell is assigned to the signature to which it shows the highest exposure. If the assignment does not change for a total of n iterations, the decomposition stops and returns both matrices W and H. By evaluating the stability of the decomposition, and not the quality o the reconstruction, we ensure that the signatures learned by the NMF are indeed representing the commonalities between samples/cells because every signature will always contain a set of representative samples/cells that will show high a degree of exposure.

To decompose a single matrix, we implemented the NMF algorithm firstly described by Seung and Lee (1999) (**Figure 2.5, Algorithm 3**) as well as the Graph Regularized Non-negative Matrix Factorization with Sparse Coding (C. Lin and Pang 2015) (**Algorithm 4**). Furthermore, ButchR is also designed to decompose multiple matrices at the same time and learn a shared set of signatures. In order to achieve this, we implemented the joint NMF algorithm (Shihua Zhang et al. 2012) (**Figure 2.6, Algorithm 6**) and the integrative NMF algorithm (Yang and Michailidis 2015) (**Figure 2.7, Algorithm 7**).

Due to the stochastic nature of the NMF algorithms, they will produce different results with every execution. In order to find an optimal solution, ButchR performs the complete decomposition over a set of different random initializations of the matrices H and W and returns the best decomposition at the end.

We made ButchR freely available to install and use at `https://github.com/wurst-theke/ButchR` under the GPLv3 license. Additionally, we also created a Docker image con-

taining ButchR and all its dependencies, alongside test datasets, to help users in running and testing ButchR under any system. The Docker image is available at https://hub.docker.com/r/hdsu/butchr.

3.2 Proof of concept: Extracting signatures of the human hematopoietic system

Although NMF has been applied to high-throughput genomic data, it is of high importance to prove that the implemented algorithms work properly and that the visualizations created by the package produce meaningful and insightful results. For this purpose, here we show a proof of concept analysis, using the well-described RNA-seq dataset of different labeled cell types from the human hematopoietic system (Corces et al. 2016) (see "Appendix A: Data description" for a description of all data used in this work). We used this dataset to extract signatures and perform feature extraction. From the beginning, we expected to recover signatures for each major cell type group, namely hematopoietic stem cells (HSC) and multipotent progenitors (MPP); committed progenitors cells such as Lymphoid-primed multipotent (LMPP), common myeloid progenitors (CMP), lymphoid progenitors (CLP), granulocyte-monocyte progenitors (GMP), and megakaryocyte-erythrocyte progenitor (MEP); and differentiated cells, as shown in **Figure 3.2a**.

3.2.1 Optimal factorization rank and signature stability

As the NMF algorithm uses the factorization rank k (i.e., number of signatures) as a hyperparameter, it is in the hands of the user to select a valid or optimal factorization rank before the decomposition by default. This number is usually equivalent to the number of classes (e.g., cell types, cancer subtypes, treatments). However, it may also be challenging to determine this number in datasets where no previous information about the data stratification is known. Furthermore, selecting the factorization rank based on

previous knowledge could lead to a loss of information, as a consequence of underestimating k due to unknown or rare classes present in the data. Therefore, we addressed this challenge by allowing the user to select a range of factorization ranks before running the decomposition, and guide the selection of the optimal factorization rank k by producing a diagnostic plot of the NMF results across all selected ranks (**Figure 3.2b**). In this plot, the following six metrics are shown for every initializing condition:

1. **Frobenius error:** measures the quality of one decomposition, i.e., how close it is to the original matrix X (Wu et al. 2016).

$$\text{FrobError}(W, H) = \|X - WH\|_F^2 \tag{3.1}$$

2. **Coefficient of variation:** measures the quality and stability of several decompositions for one factorization rank, i.e., how consistent are the NMF decompositions after different initialization (Wu et al. 2016).

$$\begin{aligned} \mu_{\text{Frob}} &= \frac{1}{B}\sum_{b}^{B}\text{FrobError}(W_b, H_b) \\ \text{CoefVar}([W_1, H_1]\dots[W_B, H_B]) &= \frac{\sqrt{\frac{1}{B}\sum_{b}^{B}(\text{FrobError}(W_b, H_b) - \mu_{\text{Frob}})^2}}{\mu_{\text{Frob}}} \end{aligned} \tag{3.2}$$

3. **Mean Amari distance:** measures the instability of several decompositions for one factorization rank (Wu et al. 2016).

$$\begin{aligned} d(W_b, W_{b+1}) &= \frac{1}{2K}\left(2K - \sum_{j=1}^{K}\max_{1\le k\le K} C_{kj} - \sum_{k=1}^{K}\max_{1\le j\le K} C_{kj}\right) \\ \text{meanAmari}(W_1 \dots W_B) &= \frac{1}{B}\left(\sum_{b}^{B-1} d(W_b, W_{b+1})\right) \end{aligned} \tag{3.3}$$

where:

K : factorization rank

B : number of decomposed matrices for rank K

C : cross-correlation matrix between W_b and W_{b+1}

4. **Sum of silhouette width:** the silhouette coefficient is a measure of how similar

one data instance is to other instances of a data cluster (Rousseeuw 1987). In the NMF context, it measures the consistency of the NMF signatures over several decompositions for one factorization rank. In ButchR it is calculated by: *(i)* concatenating all matrices $W \in \mathbb{R}^{n \times k}$ decomposed for one factorization rank k into a matrix $A \in \mathbb{R}^{n \times kB}$, *(ii)* computing the cosine distance matrix D from matrix A, *(iii)* clustering the rows of matrix D (i.e., NMF signatures) into k clusters, using "around medoids" a robust version of K-means, and *(iv)* computing the silhouette scores for all signatures **(Algorithm 8)**. In particular the "sum of silhouette width" is the result of adding together all the silhouette scores.

5. **Mean silhouette width:** this metric summarizes the consistency of the NMF decomposition, by estimating the average of the silhouette scores **(Algorithm 8)** calculated for all the NMF signatures decomposed for one factorization rank.

Algorithm 8: Silhouette width calculation for NMF signatures

Input : $W_1 \dots W_B$: Decomposed matrices $W \in \mathbb{R}^{n \times k}$ for B initializations

k : Factorization rank

Output: s : Silhouette scores vector $s \in \mathbb{R}^{kB}$

Concatenate matrices $W_1 \dots W_B$

$A = W_1 \oplus W_2 \oplus \dots \oplus W_B$

Compute cosine distance matrix and cluster with partitioning around medoids:

$D = \text{CosineDistance}(A)$

$C = \text{PartitioningAroundMedoids}(D, k)$

Compute silhouette scores:

for $i \leftarrow 1$ **to** kB **do**

  $a_i = \frac{1}{|C_i|-1} \sum_{j \in C_i, i \neq j} d(i,j)$, mean intra-cluster distance

  $b_i = \min l \neq i \frac{1}{|C_l|} \sum_{j \in C_l} d(i,j)$, mean nearest-cluster distance

  $s_i = \frac{b_i - a_i}{\max(a_i, b_i)}$

end

Return silhouette scores vector $s \in \mathbb{R}^{kB}$

---

6. **Cophenetic coefficient:** this metric is an index of the dispersion of the NMF signatures for one factorization rank (Brunet et al. 2004). It is computed as the Pearson correlation of the cosine distance matrix D and the cophenetic distance matrix C. Both matrices D and C are estimated from a matrix A, which is computed by concatenating all matrices W decomposed for one factorization rank k **(Algorithm 9)**.

Algorithm 9: Cophenetic coefficient for NMF signatures

Input : $W_1 \dots W_B$: Decomposed matrices $W \in \mathbb{R}^{n \times k}$ for B initializations

k : Factorization rank

Output: c : Cophenetic coefficient

Concatenate matrices $W_1 \dots W_B$

$A = W_1 \oplus W_2 \oplus \dots \oplus W_B$

Compute cosine and cophenetic distance matrices:

$D = \text{CosineDistance}(A)$

$C = \text{CopheneticDistance}(\text{HierarchicalClustering}(D))$

Compute cophenetic coefficient:

$c = \text{PearsonCorrelation}(D, C)$

The optimal factorization rank k is determined by minimizing the Frobenius error, the coefficient of variation, and the mean Amari distance (Wu et al. 2016), while the cophenetic correlation coefficient, and sum and mean silhouette width should be maximized (Brunet et al. 2004). In the dataset from Corces et al. (2016), we found $k = 8$ to be optimal.

Another advantage of running the NMF decomposition across a wide range of factorization ranks is that it allows the inspection of the robustness and stability of the signatures. In ButchR, we implemented a riverplot or Sankey diagram (Weiner 2017) to represent the changes in a signature across multiple factorization ranks in an intuitive visualization. A

robust signature can be identified in a riverplot as a ribbon of constant width crossing multiple nodes (see "Appendix B: How to read a riverplot" for a detailed explanation of the riverplot visualization).

In our study case, the riverplot visualization revealed a separation of stem and progenitor cells from differentiated cell types, forming two clear branches of signatures that persist across all factorization ranks (**Figure 3.2c**).

3.2.2 Sample exposure and cluster analysis

For every factorization rank, a matrix H can be retrieved. Nevertheless, it is better to use the matrix H corresponding to the optimal factorization rank. The exposure values from the matrix H can be used to soft cluster samples and understand state transitions or progressive regulatory changes. In addition, if a sample or cell is involved in multiple biological processes, this can be observed as relatively high exposure to two or more signatures. One of the most informative ways for finding the meaning of the signatures contained in the matrix H is by visualizing the exposure values in a heatmap. ButchR provides helper functions to extract the matrix H and create a heatmap using packages like ComplexHeatmap (Gu, Eils, and Schlesner 2016).

The visual inspection of the matrix H for the Corces et al. (2016) data, helped to confirm that the decomposition result generated by ButchR were in line to want we originally expected (**Figure 3.3a**), i.e., the identification of a signature for each major cell group. Moreover, the continuous exposure scores from the matrix H also revealed more information about the undergoing biological processes happening in these data, by identifying one signature with high exposure for the undifferentiated populations (hematopoietic stem cells and multipotent progenitors) and a progressive decrease in the exposure for populations with increasing differentiation (**Figure 3.3a**).

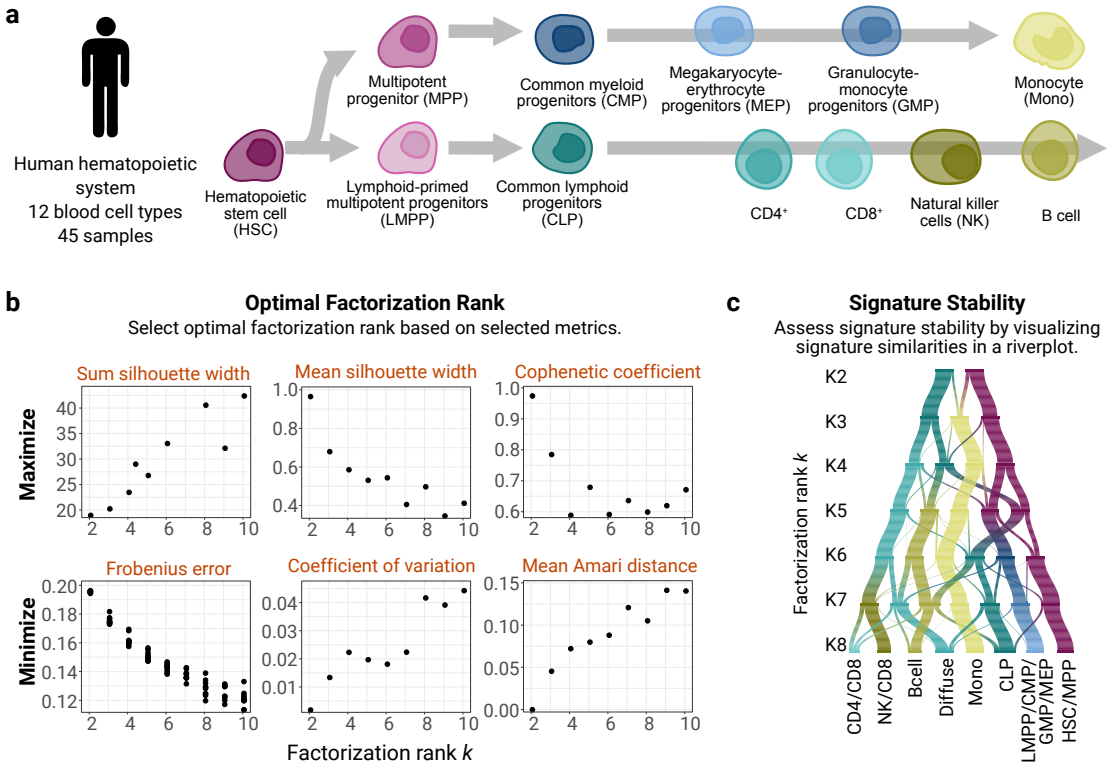


Figure 3.2: Example of a ButchR analysis **(a)** based on RNA-seq data of 12 blood cell populations and 45 samples (Corces et al., 2016). **(b)** NMF decomposition quality metrics plot. **(c)** Signature stability and hierarchy assessment by a riverplot representation of the extracted signatures at different factorization ranks. The nodes represent the signatures, the edge strength encodes cosine similarity between signatures linked by the edges. *Figure modified from Quintero et al. (2020) with permission of Oxford University Press.*

UMAP (Diaz-Papkovich et al. 2019) and tSNE (Van Der Maaten and Hinton 2008) are two algorithms extensively used to visualize high-dimensional data. However, one obstacle to using these algorithms with a large number of features is the long execution time. Thus, one of the most commonly used procedures is to perform feature selection or

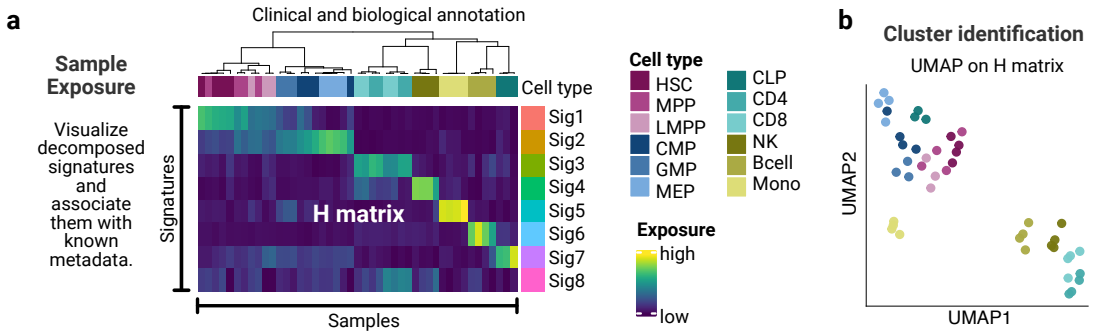


Figure 3.3: Sample exposure to the human hematopoietic system NMF signatures **(a)** Heatmap representation of the exposure matrix H showing the associated annotation features. **(b)** Cluster identification by running UMAP on the matrix H. *Figure modified from Quintero et al. (2020) with permission of Oxford University Press.*

dimensionality reduction (e.g., using PCA and selecting the top 50 principal components) before tSNE or UMAP. In ButchR, the reduced-dimensional representation of the original data (i.e., matrix H) can be used as input for UMAP or tSNE.

The UMAP visualization recreated from the matrix H, showed us an expected separation of undifferentiated populations from more differentiated cell types in the Corces et al. (2016) data, and also a distinct stratification of the differentiated cell types into sub-clusters (**Figure 3.3b**).

3.2.3 Biological annotation enrichment for NMF signatures

Despite of the visual clues provided by displaying known biological and clinical annotation on the matrix H heatmap, we developed another visualization to identify if a signature is enriched or depleted for a particular annotation variable.

In this visualization, a recovery curve is built for every category of a known categorical annotation for one signature in the matrix H. The curve is built (i) by ranking the

samples from high to low exposure score to make the x-axis of the curve, then (ii) by iterating over all ranked samples one step is increased in the y-axis if the sample is annotated for the variable under consideration. Next, the significance of the association is evaluated by computing the area under the curve (AUC) and estimating a p-value after shuffling n times the sample labels to measure the mean and standard deviation of the null distribution of AUC values.

A recovery curve follows a diagonal line for variables with no or low association, a curve with a steep increase for associated variables, and a curve with a steep drop for variables depleted in the evaluated signature. For this test case, we found that most of the cell types showed a significant association with one or two signatures (**Figure 3.4**).

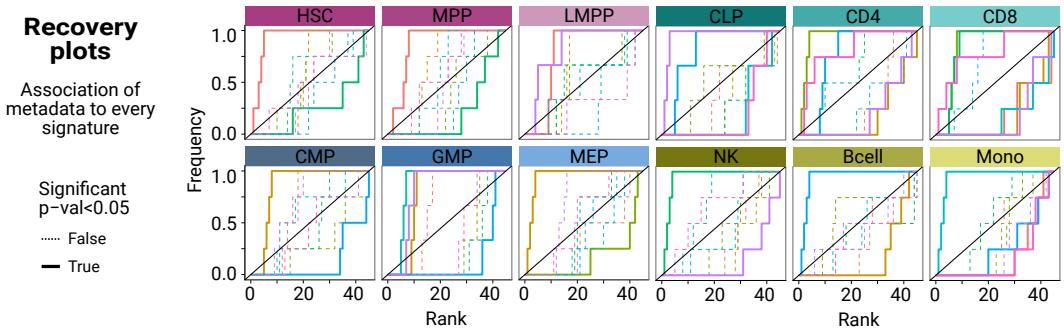


Figure 3.4: Recovery plot analysis to identify enrichment of known blood cell populations to the human hematopoietic system NMF signatures, a significant enrichment relationship is shown in a bold line. *Figure modified from Quintero et al. (2020) with permission of Oxford University Press.*

3.2.4 Feature extraction and gene set enrichment analysis

Besides soft clustering and identification of signatures with significant enrichment of know biological or clinical variables, NMF can also be used to perform feature extraction and to build groups of features that show a high contribution to the signature definition. This is a remarkable strength when applied to high-throughput genomic data because it can be

used to assign genomic features to each signature, and rank their contribution. From this perspective, we integrated a complete set of functions into ButchR to identify the variable degree of contribution of a feature to every signature, and a further classification as a signature specific feature or a multi-signature feature (**Figure 3.5**). This classification is based on performing a *k-means* clustering over every row of the matrix W with 2 clusters, effectively identifying those groups of features that show a higher contribution to the signature definition.

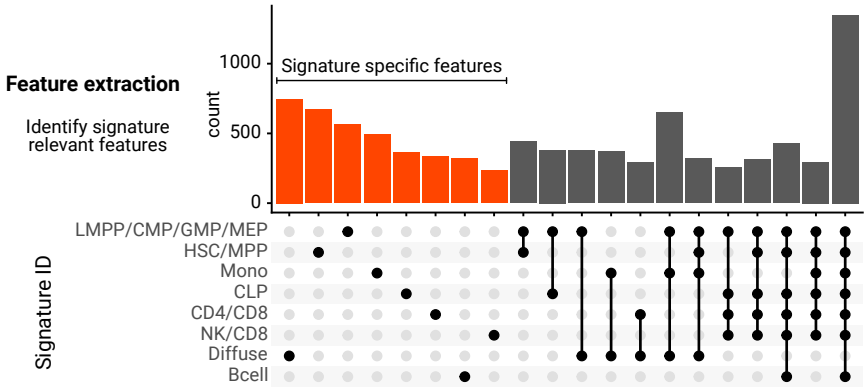


Figure 3.5: Extraction of features associated to the human hematopoietic system NMF signatures. The UpSet plot shows the number of genes that are classified as "Signature specific features" (i.e., features that mainly contribute towards only one signature) and features that are associated to more than one signature. *Figure modified from Quintero et al. (2020) with permission of Oxford University Press.*

To corroborate that the signature specific features (i.e., signature specific genes) extracted with ButchR, showed a high contribution to only one signature, we extracted and inspected the top 10% specific features from the signatures learned for the Corces et al. (2016) data, revealing groups of genes that highly support only one signature compared to the others (**Figure 3.6a**).

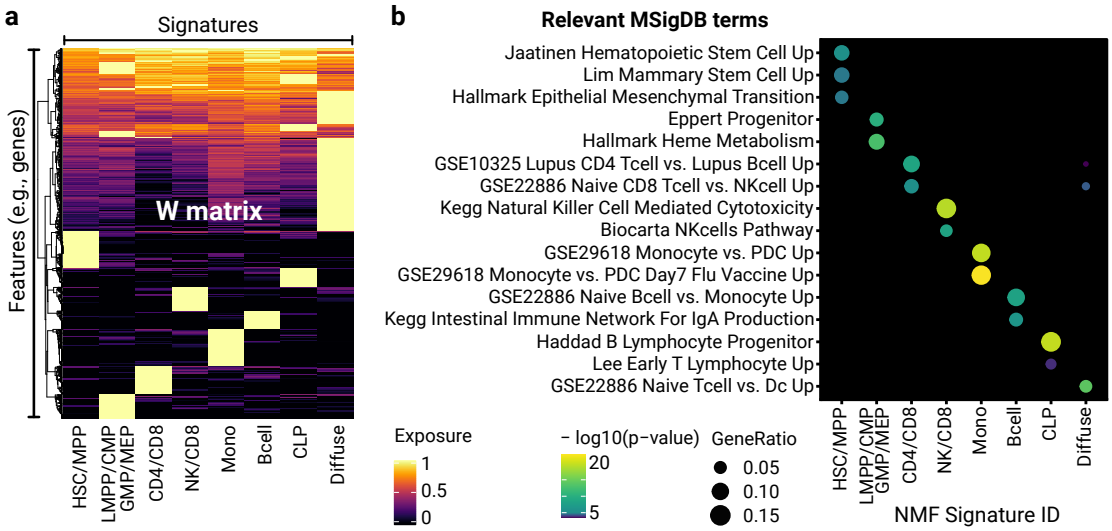


Figure 3.6: Enrichment of features associated to human hematopoietic system NMF signatures. **(a)** Feature exposure to the matrix W of the top 10% Signature specific features. The exposure values are normalized row by row. **(b)** Gene set enrichment analysis using the same set of genes displayed in (b). $-log_{10}$ of the corrected p-values are shown for representative gene set collections. *Figure modified from Quintero et al. (2020) with permission of Oxford University Press.*

The groups of signature-specific features can be further interrogated to understand the biological processes or phenotype captured by a single signature. Therefore, we performed a gene set enrichment analysis on the set of extracted features using the complete set of molecular signatures collection database (MSigDB, Subramanian et al. 2005) as a reference (**Figure 3.6b**). This provided an additional layer of validation to the array of signatures learned by ButchR, reflected as positive enrichment of gene sets that define the cell types associated with the NMF signatures. For instance:

- The HSC/MPP signature was enriched for gene sets upregulated in stem cells (Lim et al. 2010; Jaatinen et al. 2006).

- The LMPP/CMP/GMP/MEP signature captured gene sets upregulated in committed progenitor cells (Eppert et al. 2011).

- The CLP signature was enriched for gene sets up-regulated at early stages of progenitor T lymphocyte maturation (M. S. Lee et al. 2004) and in progenitor cells of B lymphocyte lineage (Haddad et al. 2004).

3.3 Chapter summary

The extraction of signatures from high-throughput data in genomics and molecular biology has been a challenging task during the last decades. We developed ButchR, a fast and user-friendly R package to decompose and learn signatures from an input non-negative matrix. ButchR also includes multiple feature extraction and visualization functions to understand and recognize the biological processes captured by the NMF signatures. The package can be installed from GitHub (`https://github.com/wurst-theke/ButchR`) and the provided Docker image (`https://hub.docker.com/r/hdsu/butchr`) allows the integration of the NMF based analysis into any existing workflow.

Chapter 4

ShinyButchR: Interactive analysis and exploration of NMF results

Disclosure: The results presented in this chapter have been published in Quintero et al. (2020) and reproduced here with the permission of Oxford University Press, license number 5011370897521.

Understanding and extracting information from the NMF results is non-trivial without appropriate representation tools. Therefore, we developed ButchR aiming to provide the community with an accessible and easy-to-use package, that represents the NMF results using a wide range of intuitive visualizations. Nevertheless, we also acknowledge the fact that not everyone in the scientific community is familiar with the R programming language, which may create a barrier to use ButchR. Therefore, we created ShinyButchR alongside with ButchR, an interactive R/Shiny application (Chang et al. 2020) to execute and explore NMF analysis in real-time. This app removes the hurdle of installing and learning to use all the software dependencies and allows any user to complete a matrix decomposition analysis from start to end inside the app itself. Additionally, aiming towards collaborative working and open research, ShinyButchR offers free computing resources available to anyone interested in using the app.

4.1 ShinyButchR

ShinyButchR is an interactive web application that can be used on any device. A fully operating version of the app is publicly hosted at https://hdsu-bioquant.shinyapps.io/shinyButchR/. Following the same approach, we took with ButchR, the complete ShinyButchR source code is freely available in GitHub at https://github.com/hdsu-bioquant/shinyButchR. Furthermore, we also made available a Docker image of the app in order to help any user interested in serving the app in a local server, which can be found on https://hub.docker.com/r/hdsu/shinybutchr.

The app was built with an intuitive user interface, consisting of two main screens (**Figure 4.1a**):

- Setup screen: consists of the interface to upload a non-negative matrix and to modify the parameters to run NMF using ButchR.

- Results screen: consists of a wide range of interactive visualizations produced from the ButchR results.

We included the RNA-seq dataset of sorted blood cell populations from Corces et al. (2016) inside the app as a demo dataset, which can be loaded by clicking on the "Demo" button on the setup screen. Thus, the NMF decomposition results shown in "ButchR: NMF suit to slice genome-scale datasets" can be fully reproduced using only ShinyButchR.

The complete NMF workflow implemented in ShinyButchR consists of the steps depicted in **Figure 4.1b**. A complete guide of how to use the app is explained in detail in the "Appendix C: ShinyButchR tutorial."

Figure 4.1: Schematic representation of a ShinyButchR NMF-based workflow. **(a)** Main screens of ShinyButchR user interface. The panel on the left shows the "Setup screen" of the app, where the user can upload a dataset and the associated annotation table, as well as tuning the parameters to run the matrix decomposition. The panel on the right shows the "Results screen," where the user is able to explore the results interactively, e.g., selection of the optimal factorization rank, clustering analysis, association to known biological and clinical factors, and signature stability assessment. **(b)** Steps performed in the ShinyButchR workflow, the setup steps (i.e., steps 1 to 6) are shown in red, the results exploration steps (i.e., steps 7 to 12) are shown in blue, and the final save results step (i.e., step 13) is shown in green. *Figure modified from Quintero et al. (2020) with permission of Oxford University Press.*

4.2 Chapter summary

Not everyone in the scientific community is familiar with the R programming language, which creates a barrier to the widespread usage of ButchR. With this in mind, we developed ShinyButchR, an interactive application to execute an NMF-based workflow from start to end. The results obtained with ShinyButchR can be imported into R to perform feature extraction and other downstream analyses.

The app is publicly available (https://hdsu-bioquant.shinyapps.io/shinyButchR/), and the provided Docker image (https://hub.docker.com/r/hdsu/shinybutchr) allows the execution and deployment in local servers.

Chapter 5

i2NMF: An integrative approach to discover dataset-specific effects

One of the main limitations of the iNMF algorithm is the assumption that exposure matrices H_S and H_{vn} share the same signature matrix W_n (**Figure 2.7**). Thus, the heterogeneous and homogeneous effects are explained by the same number of signatures. This limitation is more evident when the integration analysis is performed on a multi-view dataset containing a big amount of heterogeneous effect (view-specific effect).

In this chapter, we present Integrative Iterative Non-negative Matrix Factorization (i2NMF), a computational method to dissect cell type associated signatures from multi-omics data sets. i2NMF uses multidimensional measurements for the same sample or cell to define cell type-specific features. In this setting, i2NMF will create an integrative space using the sample/cells as common features between all matrices. In this space, the NMF signatures are explained at the same type by multiple types of data (i.e., the multidimensional measurements).

On the other hand, i2NMF can also be used in datasets from different conditions or different species, where the measured features show an overlap between datasets. In this context, the common features (e.g., genes) between datasets will be used to create the

integrative space. In this space, the NMF signatures are explained at the same time by samples/cells from all original input matrices.

The final aim of i2NMF is to explain the heterogeneous and homogeneous effect between two or more datasets using an independent set of signatures, i.e., the set of signatures explaining the homogeneous effect will be different from the signatures explaining the heterogeneous effect.

5.1 Iterative integrative NMF

i2NMF is a workflow implemented on the package ButchR (**Figure 5.1**), consisting of two different stages:

- Stage 1: Recovering the shared effect across datasets.
- Stage 2: Identification of view-specific signatures.

The input for i2NMF is two or more non-negative matrices, with a common set of features across columns, e.g., gene or sample IDs.

In the first stage, the shared effect across the set of input matrices is decomposed using iNMF (**Figure 5.1, Stage 1**), solving the equation (5.1). The shared effect is recovered in the exposure matrix H_s and is explained by the signature matrices W_{sn}. Small values in the regularization term λ will guide the decomposition to capture more shared effects (Yang and Michailidis 2015).

$$\min_{W_s \geq 0, H_s \geq 0, H_v \geq 0} \sum_{n=1}^{N} ||X_n - W_{s_n}(H_s + H_{v_n})||_F^2 + \lambda \sum ||W_{s_n} H_{v_n}||_F^2 \tag{5.1}$$

In the second stage, i2NMF decomposes the residual effect (equation (5.2)) which was not explained by the shared decomposition.

$$X_{rn} = |X_n - W_{s_n} H_s| \tag{5.2}$$

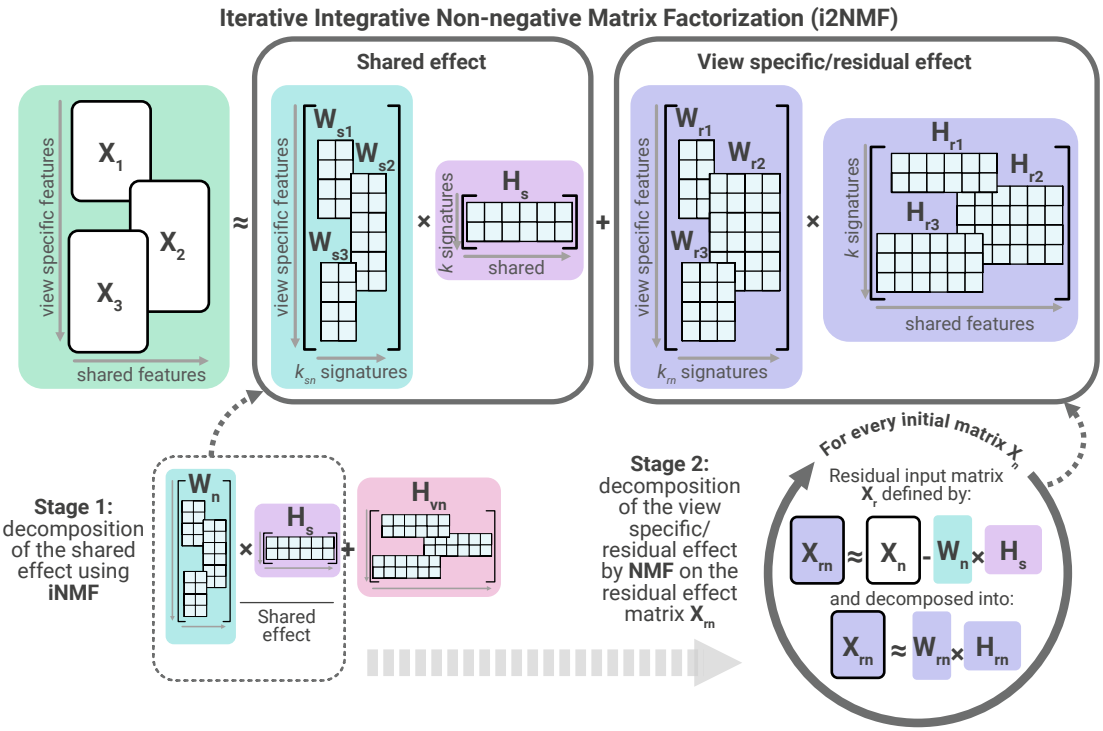


Figure 5.1: Schematic representation of the iterative integrative NMF (i2NMF) algorithm. i2NMF learns signatures that explain the shared effect between multiple matrices and also the specific effect derived from every input matrix. In order to explain the shared effect, in a first stage two or more non-negative input matrices are decomposed into signature matrices W_{sn} and shared exposure matrix H_s. On a second stage, to explain the residual effect, i.e., the view-specific effect, the residual matrix X_{rn} is decomposed into signature matrices W_{rn} and exposure matrices H_{rn}. This effectively means that the set of signatures explaining the homogeneous effect will be different from the signatures explaining the heterogeneous effect

The aim of the second stage is to recover view-specific signatures by performing a decomposition on the residual matrix X_{rn} using NMF (**Figure 5.1, Stage 2**), resulting in a matrix H_{rn} and a matrix W_{rn} for every input matrix X_n (equation (5.3)). The most important property of these matrices is that the signatures can be different in number and not shared across views, recovering in this way view-specific signatures.

$$\begin{gathered}\text{For each residual matrix } X_{rn} \\ \min_{W_{rn}\geq 0, H_{rn}\geq 0} ||X_{rn} - W_{rn}H_{rn}||_F^2\end{gathered} \tag{5.3}$$

5.2 Proof of concept: Recovering cell-specific signatures between substantia nigra of human and mouse

Using a substantia nigra (SN) single-cell RNA-seq human dataset comprised of 40,453 cells (Welch et al. 2019) and a mouse dataset comprised of 51,912 cells (Saunders et al. 2018), we show here how the i2NMF workflow implemented in ButchR can be used to integrate cross-species datasets and also retrieve view-specific signatures (**Figure 5.2a**).

5.2.1 Integration of cross-species single-cell data

After finding the set of homologous common genes between both datasets, we built the input matrices X_{mouse} and X_{human}, where the cells are in the rows and the set of homologous common genes define the columns. Then, we performed the first stage of the i2NMF workflow, integrating both input matrices X_{mouse} and X_{human} across their columns (i.e., common genes), resulting in two signature matrices W_{smouse} and W_{shuman}, and a unique shared exposure matrix H_s.

Butcher provides functions to extract, normalize, and combine the resulting signature matrices, in order to build a concatenated matrix W_S, including cells from both datasets

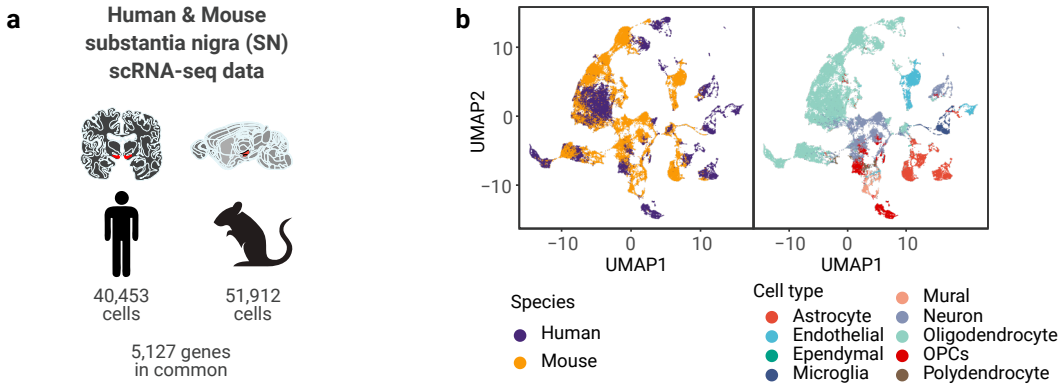


Figure 5.2: Example of an i2NMF analysis **(a)** based on substantia nigra scRNA-seq data of 40,453 human cells (Welch et al., 2019) and 51,912 mouse cells (Saunders et al., 2018). **(b)** UMAP visualization of the join W_{sn} matrices decomposed in the first stage of i2NMF.

in the rows and the common signatures in the columns.

A UMAP visualization of the matrix W_S revealed that stage 1 of i2NMF was able to integrate both datasets across species (**Figure 5.2b, left panel**) and resolve cell-type-specific clusters (**Figure 5.2b, right panel**).

5.2.2 Identification of cross-species shared signatures

To identify whether the shared signatures recovered in the matrix W_S show association to selected cell types, we performed a recovery curve analysis for every signature, and extracted the enrichment Z-score using ButchR. Inspecting the enrichment Z-scores of every signature, revealed that all the shared signatures were highly enriched for only one cell type (**Figure 5.3a**).

On the other hand, to identify if the shared signatures recovered in the matrix H_s have biological significance, we performed a gene set enrichment analysis on the common set of genes between both datasets against the complete set of molecular signatures collection database (MSigDB, Subramanian et al. 2005), ranking the genes by their corresponding

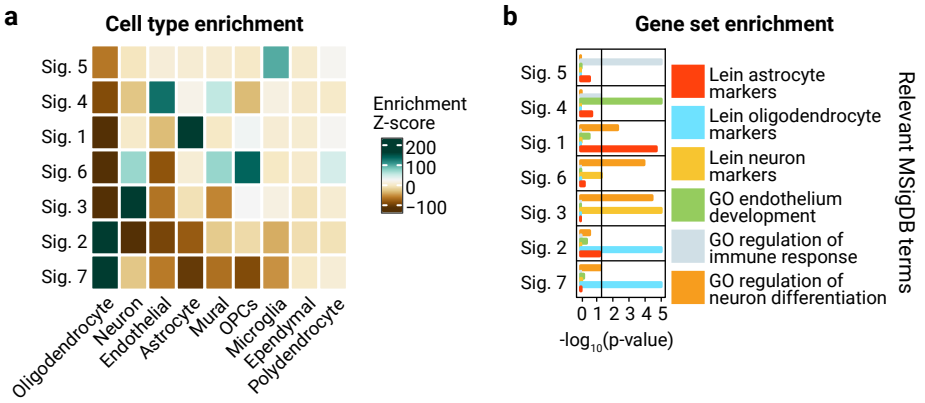


Figure 5.3: Identification of shared signatures between substantia nigra of human and mouse. **(a)** Cell type enrichment analysis (Z-score of the recovery curve analysis estimated from the join W_{sn} matrices are shown) and **(b)** gene set enrichment analysis ($-log_{10}$ of the corrected p-values are shown for representative gene set collections, GSEA analysis done on the H_s matrix) of the shared decomposed signatures in the first stage of i2NMF.

exposure scores in the signatures extracted from matrix H_s. After inspection of the enriched terms for every signature, we identified that the most enriched gene sets showed a clear correspondence to the cell types enriched in the matrix W_S (**Figure 5.3b**).

5.2.3 Recovering species-specific signatures

Finally, we performed the second stage of the i2NMF workflow to identify species-specific signatures. In order to this, we calculated the residual effect matrices X_{rmouse} and X_{rhuman} and run an NMF decomposition in them, resulting in two sets of matrices explaining the species-specific effect, $[W_{rmouse}, H_{rmouse}]$ with 41 identified signatures, and $[W_{rhuman}, H_{rhuman}]$ with 21 identified signatures.

As the mouse dataset contained several polydendrocyte subtypes that were not described in the human dataset, we focused only on the signatures enriched for this cell type to identify mouse substantia nigra-specific signatures. After performing UMAP on the

matrix $W_{r mouse}$, we found that the mouse-specific signatures were able to recover clusters corresponding to polydendrocyte subtypes (**Figure 5.4a**) and that the marker genes associated with the polydendrocyte subtypes showed high exposure scores to only one of the signatures enriched for polydendrocytes (**Figure 5.4b**).

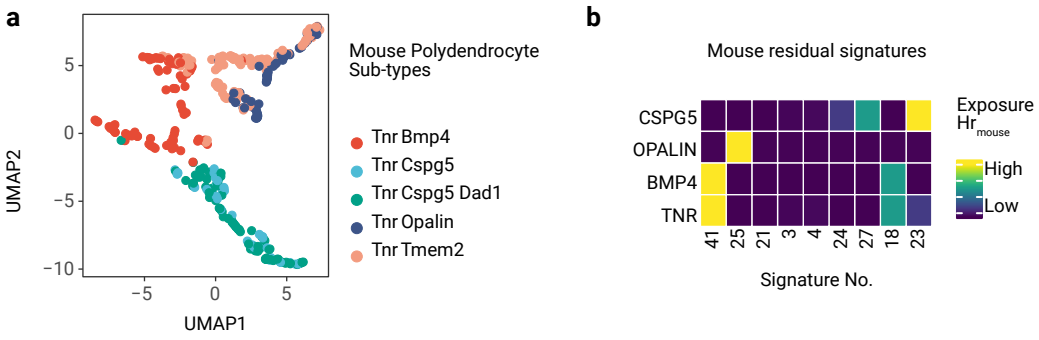


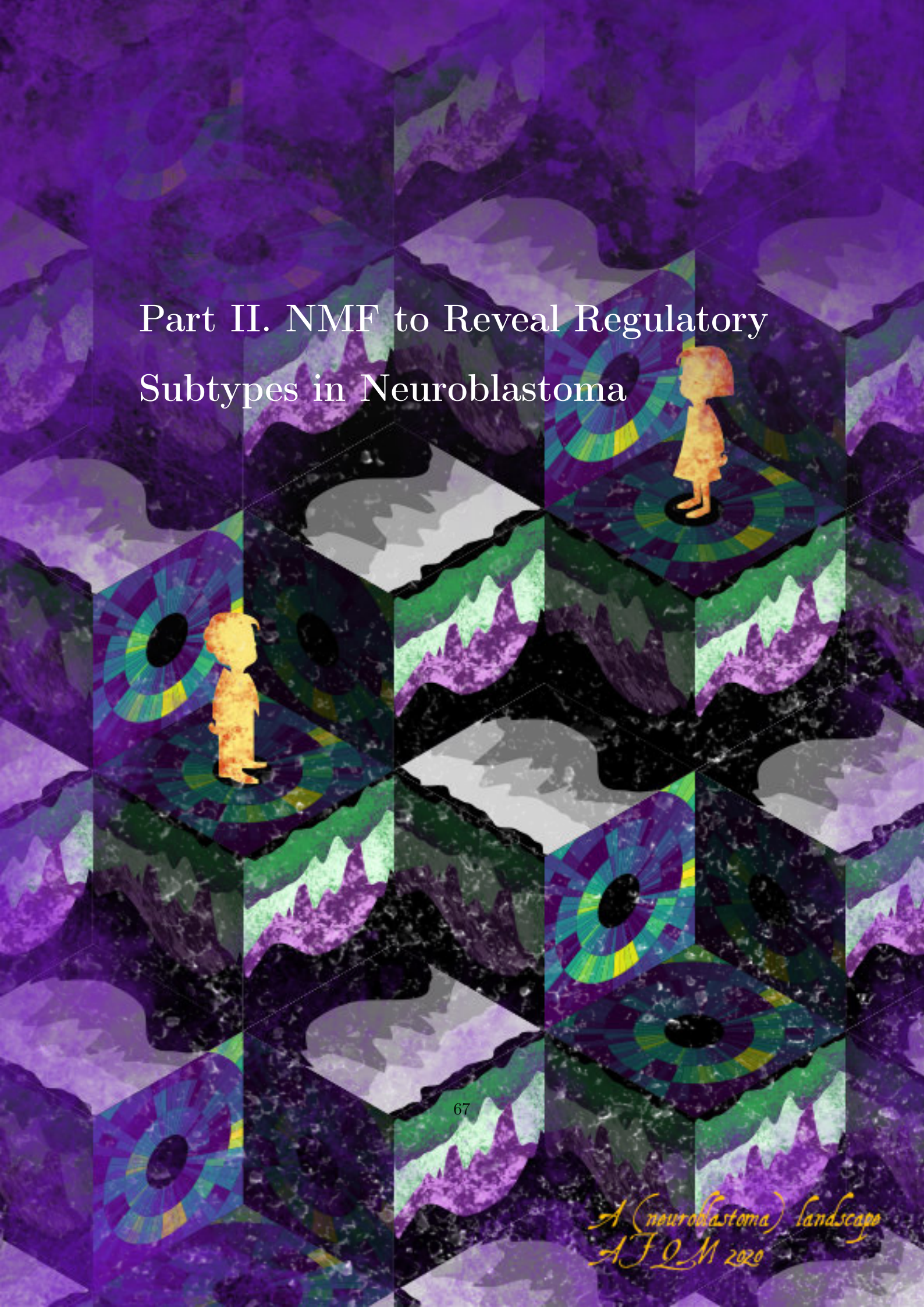
Figure 5.4: Identification of substantia nigra mouse-specific signatures in the second stage of i2NMF. **(a)** UMAP visualization of the $W_{r.mouse}$ matrix exposure scores in mouse polydendrocytes. **(b)** $H_{r.mouse}$ matrix exposure of marker genes associated with the subtypes in (a).

These results confirmed that the view-specific signatures recovered from i2NMF are meaningful, and they can be used to understand the dataset-specific granularity. In particular, the matrices W_r found by decomposing the residual effect matrices X_n can be used as input for tSNE and UMAP to resolve sub-clusters of samples/cells that were clustered together in the first stage of i2NMF. Additionally, the matrices H_r proved to be a reliable source to mine features explaining the variability associated with only one dataset.

5.3 Chapter summary

Disentangling the heterogeneous (view-specific) effect from the homogeneous (shared) effect in a multi-omics dataset is a challenging and time-consuming task. The i2NMF workflow implemented in ButchR tries to overcome some of the iNMF pitfalls by recovering true view-specific signatures. The i2NMF can be useful in cases where it is important to understand the differences between datasets and not only the commonalities.

Part II. NMF to Reveal Regulatory Subtypes in Neuroblastoma



After building and testing ButchR as a robust package to decompose datasets into the signatures that define them, we dived deep into using all the implemented features to solve a relevant scientific question, and also to determine and enable new functions that would make ButchR more flexible to interpret diverse types of questions.

The results presented here (**Part II**) are the product of a collaborative effort to understand how super-enhancers help to define regulatory subtypes in neuroblastoma (Gartlgruber et al. 2021). We show how ButchR was used to define signatures related to tumor identity and that these signatures were preserved across multiple tumor cohorts (chapter: "Neuroblastoma regulatory subtypes defined by super-enhancers"). We also explain the possible origin and genomic characteristics of a newly described neuroblastoma subtype, and at the same time how we increased the functionality of ButchR by adding a new NMF-based workflow that allows the projection of transcriptomic data of any nature (i.e., microarrays, bulk RNA-seq, scRNA-seq) onto a single cell reference atlas. (chapter: "Projection of transcriptomic neuroblastoma data onto a single-cell reference atlas").

Chapter 6

Neuroblastoma regulatory subtypes defined by super-enhancers

Disclosure: The results presented in this chapter have been published in Gartlgruber et al. (2021) and reproduced here with the permission of Springer Nature.

6.1 The molecular basis of neuroblastoma

Neuroblastoma (NB) is a pediatric tumor of the peripheral sympathetic nervous system (PSNS) derived from the neural crest. Multiple neural crest-derived precursors (e.g., neuroblasts, chromaffin cells, and Schwann cell precursors [SCPs]) are involved during the normal PSNS development to form the adrenal medulla and the sympathetic trunk (Furlan et al. 2017). It has been described that NB can originate from these sites after malignant expansion of a homogeneous population of undifferentiated neuroblasts and a few normal Schwann cells (Shimada et al. 1984; Matthay et al. 2016). However, the exact cellular origin remains to be determined.

6.1.1 Genetic predisposition

It was originally conceived that NB could be explained by a two-hit model, i.e., the first hit represents a hereditary mutation (germline mutation) and the second hit a mutation acquired after conception (somatic mutation) (Knudson and Strong 1972). Nevertheless, only 1-2% of neuroblastomas have been related to a hereditary component. The main genetic predisposition causes of "familial NB" are germline gain of function mutations in *ALK* (Mossé et al. 2008) and loss of function mutations in *PHOX2B* (Trochet et al. 2004).

6.1.2 Genetic alterations

Despite its unclear cellular origins, diverse genetic alterations have been described and characterized in NB. For instance:

- **MYCN amplification:** *MYCN* is a member of the transcription factor family MYC encoding the master regulator N-MYC, which controls the expression of several target genes, including genes that promote cell cycle progression such as *CDK4*, *CHK1*, *ID2*, and *SKP2*, as well as genes that promote cell differentiation such as *CDKL5* (M. Huang and Weiss 2013). Nearly half of all high-risk NB tumors are associated with amplification of *MYCN* (at the 2p24 amplicon), and poor prognosis (Bosse and Maris 2016). Additionally, transgenic mice overexpressing *MYCN*, developed NB 3-6 months after birth, supporting the key role of this gene in NB tumor progression (Weiss et al. 1997).

- **ALK mutation:** despite its association to familial NB, somatic mutations of *ALK* have also been found in 14% of all high-risk neuroblastomas (Bresler et al. 2014). *ALK* is located on 2p as well as *MYCN*, which can lead to co-amplification of both genes. A mouse model overexpressing *ALK* alone, led to the development of NB; while the simultaneous overexpression of *ALK* and *MYCN* resulted in an earlier onset and increased lethality (T. Berry et al. 2012).

- **1p deletion and 17q gain:** the loss of the short arm of chromosome 1 (1p) (Attiyeh et al. 2005) and the gain of up to five copies of the long arm of chromosome 17 (17q) (Bown et al. 1999) have been linked to high-risk neuroblastomas. Moreover, both events are also associated with MYCN amplification and poor prognosis (Matthay et al. 2016).

- **TERT enhancer hijacking:** the structural rearrangements that lead to placing an ectopic enhancer in proximity to a proto-oncogene are known as "enhancer hijacking." In NB cases, 25% of the patients have *TERT* promoter rearrangements, which leads to enhancer hijacking (Matthay et al. 2016). These rearrangements target the downstream and upstream regions of *TERT* and cause the positioning of groups of transcriptional active enhancers in these regions, inducing an increase in *TERT* expression. Besides, TERT is also a target of N-MYC and is involved in chromatin remodeling (Valentijn et al. 2015; Peifer et al. 2015).

6.1.3 Regulatory programs in neuroblastoma cell lines driven by super-enhancers

The concept of large groups of transcriptional enhancers, associated with genes that define cell identity was originally proposed by Whyte et al. (2013), calling them super-enhancers (SEs). In comparison to conventional enhancers, these groups of enhancers have an increased size, density of binding regions for transcription factors, and ability to activate gene expression. After their initial conception, Hnisz et al. (2013) studied the role of SEs in cancer and how key cancer driver genes generate formations of SEs at their genomic location.

The identification of active SEs starts by defining regions associated with active normal enhancers, followed by stitching together enhancers that are within 12.5 Kb of each other, and finally selecting stitched regions that show high regulatory activity (Hnisz et al. 2013; Whyte et al. 2013).

Different histone modifications are helpful to recognize the state of an enhancer or chromatin region (Bannister and Kouzarides 2011). For instance, enrichment of monomethylation of lysine 4 on histone H3 (H3K4me3) is an indicator of transcription start sites (TSS) (Schneider et al. 2004). Enrichment of monomethylation of lysine 4 on histone H3 (H3K4me1) is typically associated with gene enhancers (Hon, Hawkins, and Ren 2009), and the combination of enrichment of H3K4me1 with trimethylation of lysine 27 on histone H3 (H3K27me3) is a mark of closed or poised enhancers (Heinz et al. 2015). On the other hand, a combination of enrichment of H3K4me1 and acetylation of lysine 27 on histone H3 (H3K27ac) is a typical indicator of an active enhancer (Heinz et al. 2015). Therefore, a common strategy to find active enhancer regions is to use H3K27ac Chromatin Immunoprecipitation Sequencing (ChIP-seq) profiles and identify genomic regions enriched with this histone mark (Hnisz et al. 2013). Tools such as macs2 are helpful to accomplish this task, producing a list of "peaks" that represent regions with differential enrichment of H3K27ac signal compared to a background signal (Y. Zhang et al. 2008). Once active enhancers are identified, these are stitched together and classified as SEs using dedicated tools, such as Rank Ordering of Super-Enhancers (ROSE) (Hnisz et al. 2013).

SEs have been reported as key regulators in NB. For instance, genome-wide association studies (GWAS) revealed the association between NB susceptibility and the *LMO1* gene locus; in particular, the single nucleotide polymorphism (SNPs) rs2168101 G>T located at the first intron of *LMO1* resides within an SE region. The G-allele enables binding and positive regulation of *LMO1* by GATA3, leading to an oncogenic addiction of NB tumor cells to LMO1 (Oldridge et al. 2015). Furthermore, several of the structural rearrangements that lead to *TERT* enhancer hijacking are related to the translocation of active SEs regions (Peifer et al. 2015).

More recently, Van Groningen et al. (2017) and Boeva et al. (2017) described the existence of two predominant cell identities in NB cell lines regulated by SEs. Van Groningen et al. (2017) found that 8 out of 33 NB cell lines showed high gene expression scores

for a signature inferred from mesenchymal (MES) cells in an undifferentiated state, comprised of known MES markers such as FN1, VIM, and SNAI2. On the other hand, the remaining 25 NB cell lines had high gene expression scores for a signature of adrenergic (ADRN) markers such as GATA2, GATA3, DBH, and PHOX2A. Therefore, these findings showed that there are at least two different cell identities in NB, namely MES-type and ADRN-type. Furthermore, it was also found that the SE landscapes were different between the cell lines associated with the MES-type and the ADRN-type, with a clear association to the MES and ADRN gene signatures, including a tight regulation of key MES and ADRN transcription factors. Interestingly, inducing the expression of the MES transcription factor PRRX1 on the ADRN cell line SK-N-BE(2)-C produced repression of PHOX2B and DBH (two ADRN markers) while inducing the expression of the MES marker SNAI2. Furthermore, after 12 days of induced PRRX1 expression, the SE landscape of the SK-N-BE(2)-C cells shifted from an ADRN-type pattern towards a MES-type pattern, indicating that the MES-type and ADRN-type cells can interconvert by alterations in their regulatory landscape (Van Groningen et al. 2017; Boeva et al. 2017). Despite all these findings, it remains unclear how the ADRN-type and MES-type cell identities play a role in the development, progression, and relapse of NB tumors.

After understanding that the signatures learned by ButchR can be used to get insights into the biological variability and processes captured in high-throughput genomic datasets (see "ButchR: NMF suit to slice genome-scale datasets"), we hypothesized that the NMF decomposition would be a useful tool to explore the regulatory differences observed in NB. Therefore, in order to reconstruct the heterogeneous SE landscape observed in NB. We used ChIP-seq profiles for the histone mark H3K27ac from a diverse cohort of 60 neuroblastomas. A third of the cohort was comprised of *MYCN* amplified tumors, 25 out of the 60 samples were low-risk samples (stages 1-3, and 4s) and 35 were classified as high-risk samples (stage 4), and 11 samples were taken from relapse neuroblastomas (**Figure 6.1a**). Chromatin interaction data from bulk HiChIP profiles of the MES cell line SK-N-AS and the ADRN cell line CLB-GA were also used to elucidate the target

genes of SEs loci (**Figure 6.1b**). Additionally, H3K27ac ChIP-seq profiles for 25 NB cell lines and 579 bulk transcriptomic profiles were used to further understand the regulatory diversity in NB (all the experimental data was generated by Dr. Moritz Gartlgruber and Dr. Daniel Dreidax under the supervision of PD Dr. Frank Westermann in the Neuroblastoma Genomics department at the DKFZ).

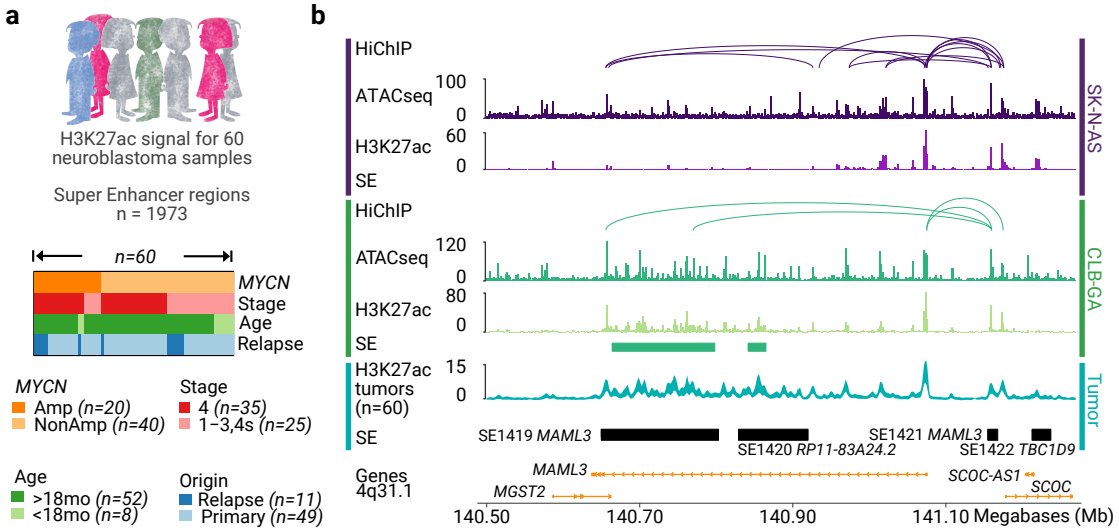


Figure 6.1: A neuroblastoma H3K27ac ChIP-seq cohort. **(a)** Characteristics of the NB tumor ChIP-seq cohort (n=60), MYCN status (MYCN; Amp = amplified), INSS stage (Stage), age at diagnosis (Age), and relapsed tumor. **(b)** Multiple layers of regulatory information like ATAC-seq, H3K27ac ChIP-seq profiles, super-enhancers, and chromatin interactions (HiChIP) integrated into this study are shown exemplarily for the MAML3 locus, which is regulated by one of the top NB SEs. This regulation of MAML3 is shown in SK-N-AS (top, purple), CLB-GA (middle, green), and the entire NB tumor ChIP-seq cohort (bottom, turquoise). The consensus SEs (black horizontal bars) from the whole cohort are depicted at the bottom. Predicted SE target genes are given beside the SE bars. Orange arrows indicate genes and orientation. *Figure taken from Gartlgruber et al. (2021) with permission of Springer Nature.*

6.2 Neuroblastoma super-enhancer signatures define epigenetic subtypes

As stated by Van Groningen et al. (2017) and Boeva et al. (2017), the alteration of the transcriptional state and lineage identity in NB is partially explained by underlying networks of SE associated with transcription factors. Thus, we identified SE regions in NB using genome-wide profiles of the enhancer mark H3K27ac across 60 neuroblastomas (NB ChIP-seq cohort), covering the different clinical and molecular subtypes (**Figure 6.1a**, Gartlgruber et al. 2021).

A total of 1,973 SE consensus regions and their target genes (**Figure 6.1b**) were found in the cohort of 60 NB tumors using the following strategy:

1. **Call H3K27ac peaks for every NB sample:** using a custom pipeline implemented in Snakemake, reads were trimmed using the Trimgalore tool (`https://github.com/FelixKrueger/TrimGalore`) and aligned using Bowtie2 (Langmead and Salzberg 2012) with standard parameters. Peaks were called using the *-callpeak* mode in MACS2 (Y. Zhang et al. 2008).
2. **Filter peaks associated with TSS regions:** to mitigate the effect of SEs associated with TSS regions, we filtered the H3K27ac peaks that were closer than 5Kb to a set of 40,512 consensus H3K4me3 peaks (found from a subset of tumor samples for which this mark was available).
3. **Call SEs for every NB sample:** SEs were identified by passing the filtered list of H3K27ac peaks to the ROSE (Hnisz et al. 2013) pipeline with standard parameters.
4. **Find SE consensus regions:** considering the union of SEs from all tumor samples we removed those that showed no overlap with another SE of the set. Then, we merged the shared SEs regions, resulting in a list of 1,973 consensus SEs.
5. **Assigning target genes to SEs:** for reliable assignments of target genes to SEs, a hierarchical strategy was followed:

- The highest-ranking criterion was the presence of physical interactions (FDR < 0.05) between the SE and a gene promoter in HiChIP chromatin interaction data from NB cells CLB-GA and SK-N-AS (**Figure 6.1b**).
- In the absence of HiChIP interaction evidence, publicly available Hi-C interaction data (GEO accession GSE63525) were used to make the SE-gene assignment.
- In the absence of Hi-C interaction evidence, a window of 1 MB around every SE was screened for genes with a significant correlation between RNA expression and SE H3K27ac signal, the gene with the strongest correlation was assigned as a target gene. A significant correlation was set to Spearman correlation greater than 0.1.
- In the absence of significant expression-H3K27ac correlation, the closest gene was assigned to the SE.

We extracted the SE H3K27ac signal for every tumor in the ChIP-seq cohort and used ButchR to extract signatures associated with epigenetic differences in NB, deriving a matrix H_{SE}. A total of four distinct signatures were identified, in which three of them corresponded to known NB subtypes (i.e., MYCN-amplified [MYCN], MYCN non-amplified high-risk [MNA-HR], and MYCN non-amplified low-risk [MNA-LR]) as can be seen in **Figure 6.2**.

To understand the nature of the fourth signature (**Figure 6.2 second row**), we hypothesized that mesenchymal features could be driving the behavior of a fraction of NB tumors similarly to what was seen in NB cell lines (Boeva et al. 2017; Van Groningen et al. 2017). We extracted features from the matrix W, and performed a gene set enrichment analysis of the signature-associated SE target genes against gene sets related to cell migration and epithelial-mesenchymal transition, computing a one-sided Fisher's exact test (**Figure 6.2 left panel**). A strong enrichment was found in this signature compared to the MYCN, MNA-HR, and MNA-LR signatures, suggesting that it defines a new NB subtype that exhibits mesenchymal characteristics.

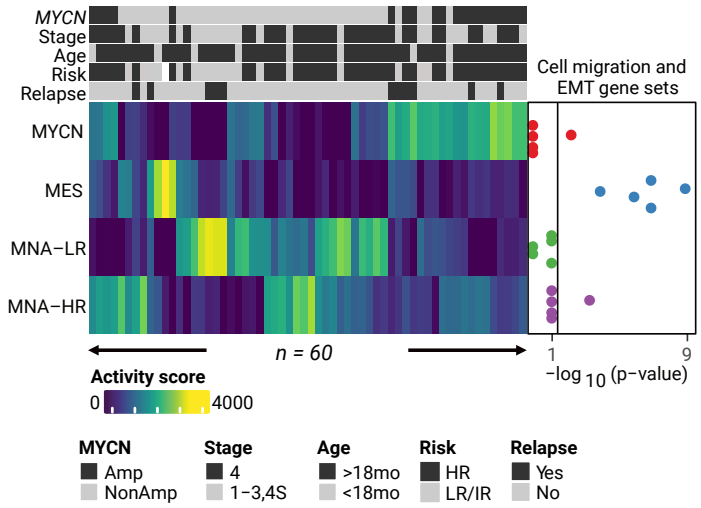


Figure 6.2: NMF analysis of the super-enhancer H3K27ac signal in NB tumors. Heatmap representation of the exposure matrix H_{SE}. The signatures (rows) are annotated as MYCN: MYCN-amplified; MNA-LR/-HR: MYCN-non-amplified low-risk/high-risk and MES: mesenchymal. Enrichment analyses of the signature-specific SE target genes among representative cell migration and epithelial-mesenchymal transition (EMT) terms are given on the right as jitterplots. P-values are computed using a one-sided Fisher's exact test. *Figure taken from Gartlgruber et al. (2021) with permission of Springer Nature.*

6.3 The mesenchymal subtype is also found in neuroblastoma cell lines

We also applied NMF on the total H3K27ac signal over the consensus SE regions from a complementary dataset comprised of 23 NB and two neural crest-derived cell lines (Gartlgruber et al. 2021). Three signatures were found: MYCN, MNA-HR, and remarkably, a third signature associated with mesenchymal processes and EMT (**Figure 6.3a**).

In order to evaluate the robustness of the MES signature, we used ButchR to decompose

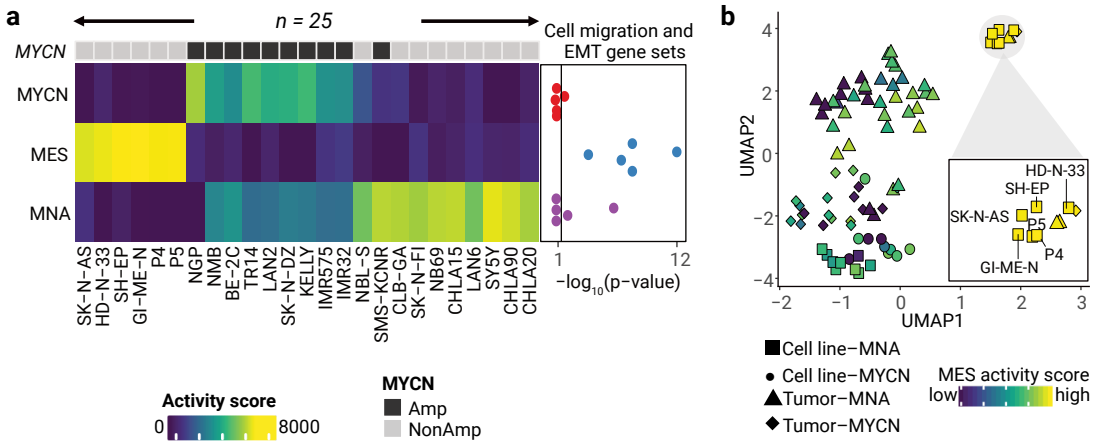


Figure 6.3: NMF analysis of the SE H3K27ac signal in cell lines (n=25). **(a)** Heatmap representation of the exposure matrix $H_{SE-Cells}$. The signatures (rows) are annotated as MYCN: MYCN-amplified; MNA: MYCN-non-amplified and MES: mesenchymal. Enrichment analyses of the signature-specific SE target genes among representative cell migration and EMT terms were done as in **Figure 6.1b**. **(b)** UMAP-based clustering of the NB tumor and cell line samples based on the H3K27ac signal over NB SEs regions. Samples are colored according to the MES activity. Samples with high MES activity (inside the grey circle) are labeled in the inset plot. *Figure taken from Gartlgruber et al. (2021) with permission of Springer Nature.*

a matrix $H_{combined}$ from tumors and cell lines at the same time. From the resulting matrix, we found a clear MES signature and subsequently assigned a combined MES activity score for tumors and cell lines. A UMAP visualization generated from the matrix $H_{combined}$ showed that those tumors and cell lines with high exposure to the MES signature formed a common cluster, demonstrating the distinctiveness of the mesenchymal phenotype (**Figure 6.3b**). In addition, the signature stability was assessed using the riverplot visualization (**Figure 6.4**). We found that the MES signature had high stability across multiple factorization ranks in both tumors ($k > 2$) and cell lines ($k \geq 2$).

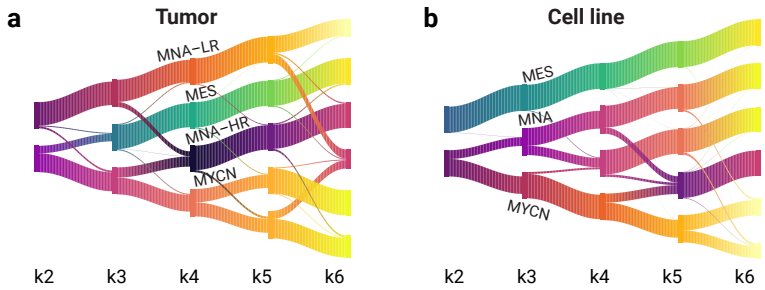


Figure 6.4: River-plot representation of the stability of the signatures extracted from the H3K27ac SE signal in **(a)** tumors and **(b)** cell lines. The vertical axis represents the different factorization ranks (which is equal to the number of signatures extracted, $k = [2..6]$) and the ribbons indicate the similarity of the signatures defined for different factorization ranks. For instance, in cell lines, the MES signature (top-most ribbon) is stable for all factorization ranks. In tumors, the MES signature appears for factorization ranks $k > 2$. *Figure taken from Gartlgruber et al. (2021) with permission of Springer Nature.*

6.4 Neuroblastoma transcriptomic signatures

We also defined SE-directed transcriptomic signatures using the expression of the SE-target genes from 579 NB tumors (RNA-seq cohort, **Figure 6.5a**, Gartlgruber et al. 2021). The decomposed exposure matrix H_{SE-Exp} resulted in four signatures, corresponding to those derived from the matrix H_{SE}, and the corresponding MES signature showed a strong association with mesenchymal features (**Figure 6.5a, right panel**).

We further tested the association between the epigenetic signatures extracted from the matrix H_{SE-Exp} and known clinical parameters, using the following strategy for every signature (e.g., MES signature from matrix H_{SE-Exp}) and every clinical annotation (e.g., relapse patients):

1. Extract signature (activity) scores of all samples.

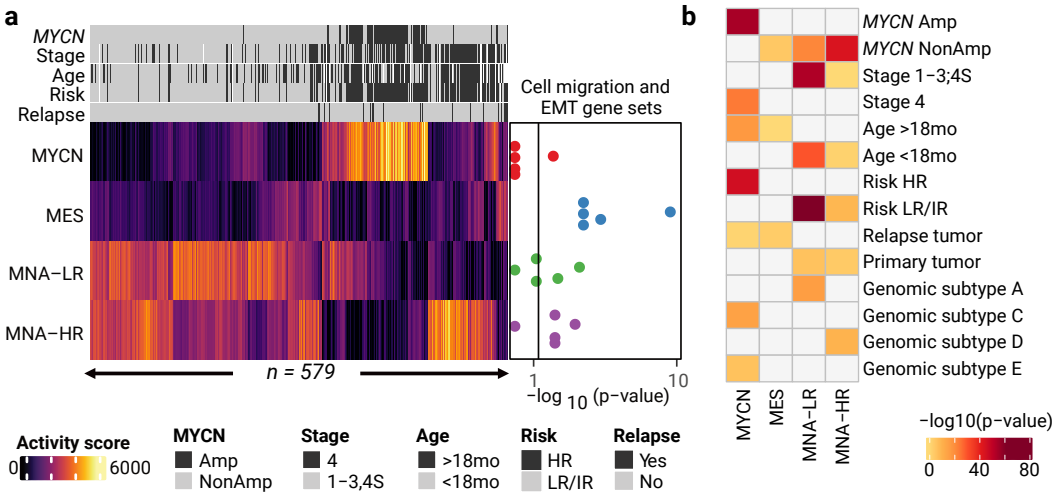


Figure 6.5: NMF analysis of NB tumors (n=579) based on the expression of the SE target genes. **(a)** Heatmap representation of the exposure matrix H_{SE-Exp}. Enrichment analyses of signature specific genes among representative cell migration and EMT terms were done as in **Figure 6.1b**. **(b)** Association of known clinical and molecular variables in neuroblastoma to signature activity taken from (a). *Figure taken from Gartlgruber et al. (2021) with permission of Springer Nature.*

2. Compare the activity scores of the set of patients annotated with the clinical annotation under evaluation against the activity scores of the rest of the patients. This was done using a one-sided Wilcoxon signed-rank test.
3. A significant association was assigned for p-values < 0.05.

«Tumors with high exposure to the MYCN signature were significantly associated with unfavorable clinical features and high-risk disease, while the MNA-LR signature was significantly associated with low-risk disease and favorable clinical features. Samples with high MYCN or MES signature scores were strongly associated with relapsed disease in the RNA-seq cohort (**Figure 6.5b**), suggesting that a substantial number of relapsed NBs exhibit MES properties» *Fragment modified from Gartlgruber et al. (2021) with the permission of Springer Nature.*

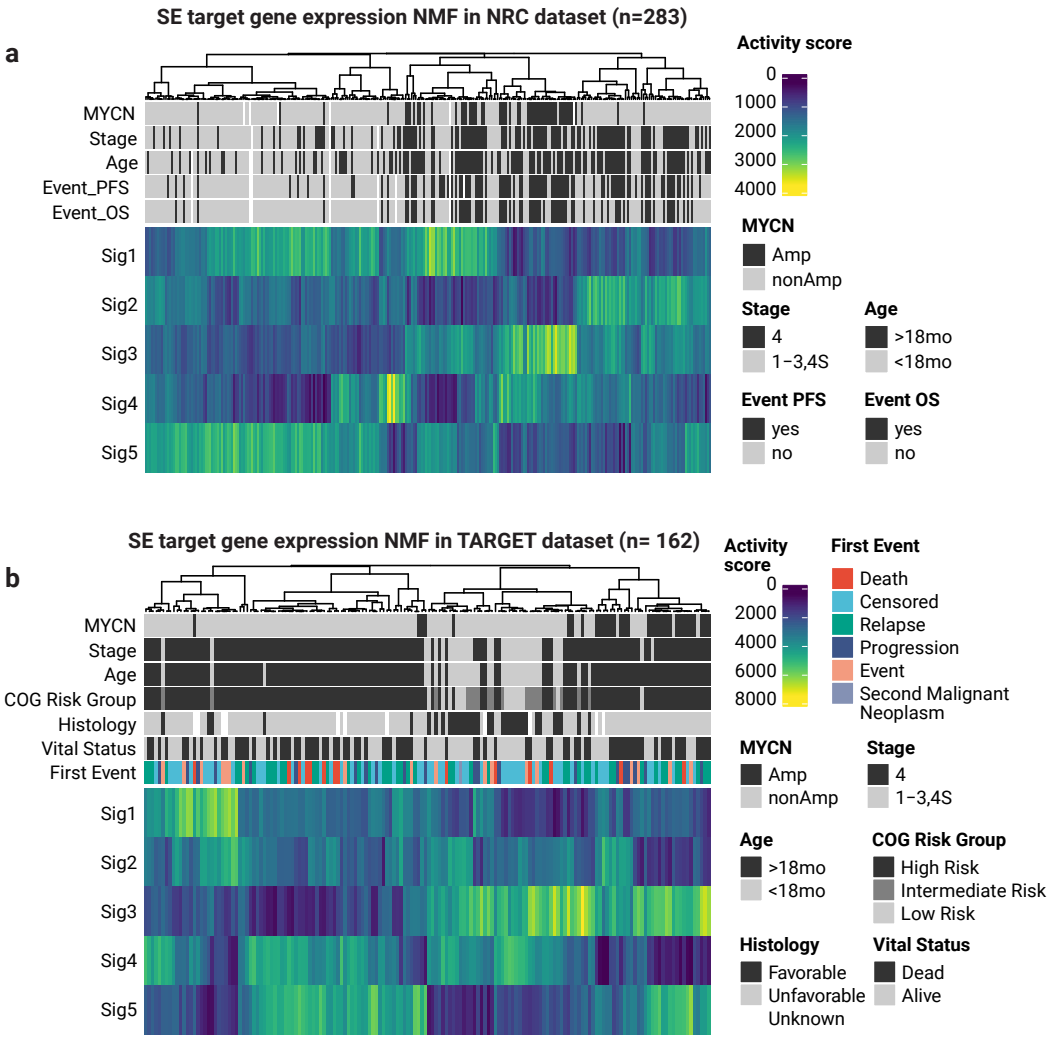


Figure 6.6: **(a)** NMF analysis based on the expression of the SE target genes in NB tumors from the NRC dataset (n=283 tumor samples, n=972 target genes) and **(b)** from the TARGET dataset (n=162 tumor samples, n=1428 target genes). *Figure taken from Gartlgruber et al. (2021) with permission of Springer Nature.*

6.5 The neuroblastoma transcriptomic signatures are found over multiple cohorts

Aiming to validate our findings in different cohorts, we used two independent NB gene expression datasets:

- NRC: Neuroblastoma Research Consortium (NRC), 162 samples (Rajbhandari et al. 2018).
- TARGET: Therapeutically Applicable Research to Generate Effective Treatments, 283 samples (Pugh et al. 2013).

We extracted the expression of the SE-target genes (i.e., target-genes found in our NB ChIP-seq cohort) from the TARGET and NRC cohorts, and used ButchR to identify SE-directed transcriptomic signatures (**Figure 6.6**).

To understand if the NRC and TARGET signatures showed correspondence to the signatures found in our RNA-seq cohort (**Figure 6.5**), we compared the gene weights (exposures) from the matrix W_{SE-Exp} (extracted from the NMF of our NB RNA-seq cohort) to the gene weights from the matrix W_{NRC} (**Figure 6.7**) and matrix W_{TARGET} (**Figure 6.8**). We were able to resolve the same subtypes identified with our SE-directed transcriptomic signatures and found high similarity among the signatures between the different datasets.

Remarkably, the evidence of a MES signature in both NRC and TARGET cohorts supports the existence of neuroblastoma subtype with mesenchymal characteristics.

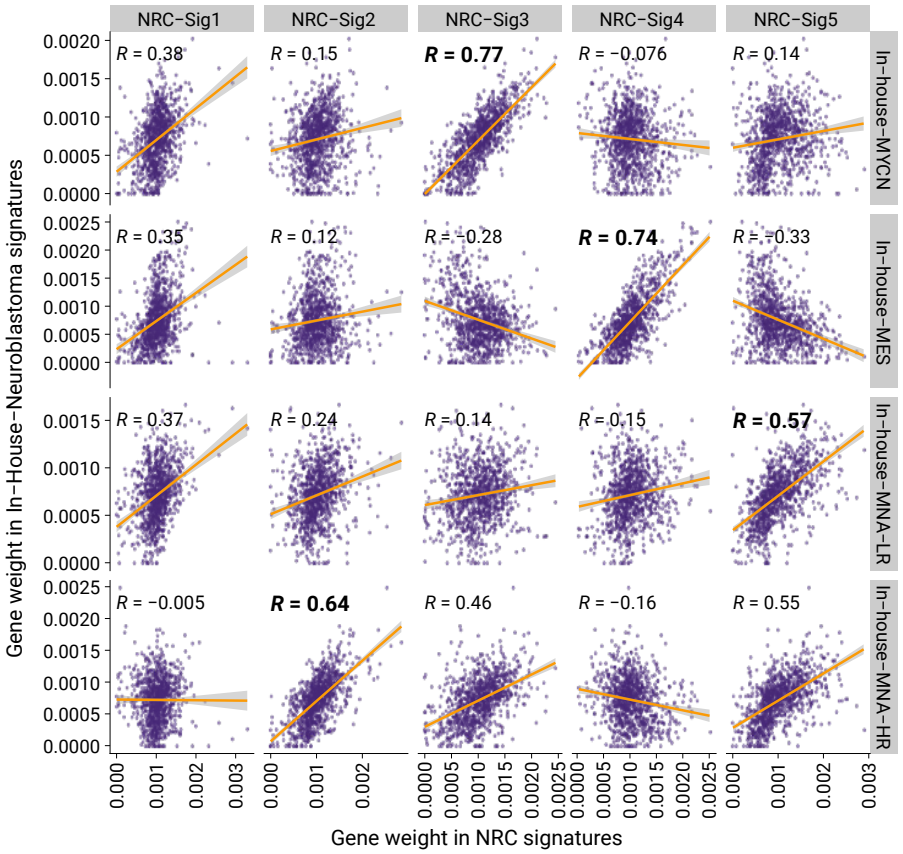


Figure 6.7: Scatter plots indicating the correlation of the signature specific gene activity between NB tumor and NRC dataset (n=972 target genes). Each dot represents a gene and the x/y coordinates represent the contribution (or weight) of the gene to the corresponding signature. Pearson's correlation values are indicated. *Figure taken from Gartlgruber et al. (2021) with permission of Springer Nature.*

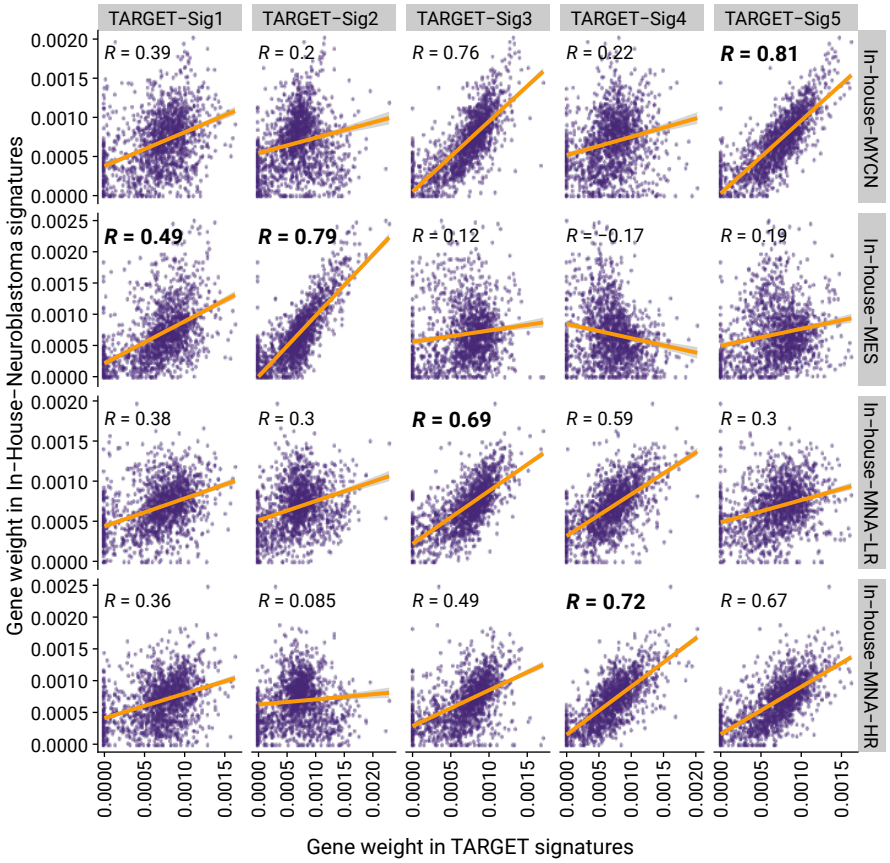


Figure 6.8: Scatter plots indicating the correlation of the signature specific gene activity between NB tumor and TARGET dataset (n=1428 target genes). Each dot represents a gene and the x/y coordinates represent the contribution (or weight) of the gene to the corresponding signature. Pearson's correlation values are indicated. *Figure taken from Gartlgruber et al. (2021) with permission of Springer Nature.*

6.6 Chapter summary

Neuroblastoma is a neuroendocrine tumor derived from the neural crest. Using transcriptional and epigenetic profiles of cell lines derived from neuroblastoma, Van Groningen et al. (2017) and Boeva et al. (2017) described the existence of two predominant cell identities (mesenchymal-type and adrenergic-type). However, the effect of the cell identity on the development, progression, and relapse of neuroblastoma tumors was unclear.

Here, using genome-wide H3K27ac profiles across 60 neuroblastomas, covering the different clinical and molecular subtypes. We used ButchR to identify four major super-enhancer-driven neuroblastoma epigenetic signatures. Three of these signatures recapitulated known clinical groups, namely MYCN-amplified, MYCN non-amplified high-risk, and MYCN non-amplified low-risk. The fourth signature showed a clear association to cell migration and epithelial-mesenchymal transition, suggesting that it defines a new neuroblastoma subtype exhibiting mesenchymal characteristics.

Chapter 7

Projection of transcriptomic neuroblastoma data onto a single-cell reference atlas

Disclosure: The results presented in this chapter have been partially published in Gartlgruber et al. (2021) and reproduced here with the permission of Springer Nature.

Neuroblastoma like other embryonal tumors originate from deregulated developmental programs at prenatal stages (Janoueix-Lerosey et al. 2008; Mossé et al. 2008; Schwab et al. 2003; Pugh et al. 2013). Identification of the specific cell(s) of origin for these tumors and their subtypes is fundamental towards understanding their development, progression, and relapse mechanisms. Neuroblastoma is thought to originate from the sympathoadrenal compartment during development. However, the exact cellular origins are unknown (Ross, Spengler, and Biedler 1983).

In order to determine the possible cell of origin of the new neuroblastoma MES subtype (see "Neuroblastoma regulatory subtypes defined by super-enhancers"), we developed a new ButchR-NMF-based workflow to map any contextually similar bulk or single-cell

transcriptomic data onto a reference single-cell data. In this chapter, we show how this new workflow was used to map neuroblastoma transcriptomic data onto two reference atlases (i.e., scRNA-seq data from mouse and human developing adrenal gland).

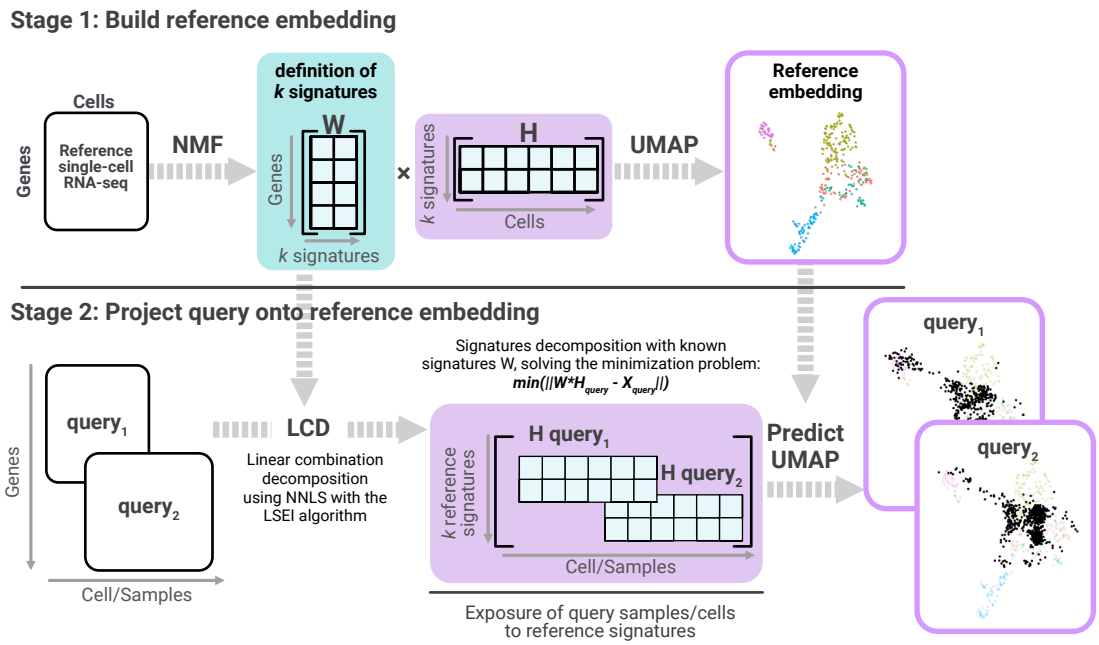


Figure 7.1: Schematic representation of the strategy to project new high-throughput transcriptomic data onto a single-cell reference atlas. **Stage 1.** The reference atlas is decomposed into the matrix W_{atlas} and the matrix H_{atlas} using NMF. An embedding of the reference atlas is build using the matrix H_{atlas}. **Stage 2.** The query data X_{query} is transformed to the lower-dimensional space H_{query} using LCD and matrix W_{atlas}. Then, the matrix H_{query} is projected onto the atlas embedding. These steps result in overlapping reference and query samples in the same latent space.

7.1 Projecting data onto an existing embedding using NMF

Single-cell transcriptomic analysis has shown outstanding potential towards understanding cell fate and diversity (Trapnell 2015; Gulati et al. 2020; Sagar and Grün 2020; Tam and Ho 2020). These rich datasets have been widely exploited for cross-inference analysis and integration between different species (Shafer 2019; Ding et al. 2019), conditions (Butler et al. 2018; Stuart and Satija 2019) or data modalities (Ma et al. 2020; Chappell, Russell, and Voet 2018; Colomé-Tatché and Theis 2018; Liu et al. 2019). However, integrative analysis between bulk and single-cell transcriptomic data remains a challenge.

We developed a workflow implemented using ButchR to project any type of high-throughput transcriptomic data (i.e., bulk and single-cell transcriptomic data) onto a single-cell reference atlas (**Figure 7.1**). This workflow is divided into two different stages, construction of a reference embedding and projection of a query onto the reference:

7.1.1 Stage 1. Construction of a reference embedding

For the first stage, ButchR is used to decompose an expression matrix from a single-cell reference atlas into a matrix W_{atlas} and a matrix H_{atlas}. In this case, the rows of the matrix correspond to genes, and the columns are the single-cells of the atlas.

The reduced-dimensional representation H_{atlas} of the original expression matrix is used to fit a UMAP model UMAP_{atlas} and plot the embedding. The matrix W_{atlas} is saved to be used in the second stage (**Figure 7.1** top panel).

7.1.2 Stage 2. Projection of a query onto the reference embedding

On the second stage, a query transcriptomic data (either bulk or single-cell transcriptomic data) is projected onto the reference embedding, this is broken down in the following steps (**Figure 7.1** lower panel):

1. **Transform the query data X_{query} to the same lower-dimensional space H_{atlas} of the reference:** using the function linear combination decomposition (LCD) from the R package YAPSA (Hübschmann et al. 2020), the exposure values of each query sample/cell H_{jquery} are found by solving the non-negative least-squares minimization problem $\min||W_{atlas}H_{jquery} - X_{jquery}||$ with the LSEI algorithm (Lawson and Hanson 1995) implemented in the R package lsei. This effectively transforms the query data into the space H_{query} which is shared with H_{atlas}.
2. **Project H_{query} onto the atlas embedding:** using the UMAP model UMAP_{atlas} the position of the query samples/cells on the reference embedding are predicted from the matrix H_{query}.
3. **Filter low-quality projections:** correlate every query data point to its nearest reference neighbors (i.e, Spearman correlation of the transcriptomic profiles). The projection quality is refined by removing lowly correlating query data points.
4. **Transfer labels from reference atlas to query:** labels from the reference atlas are transferred to every query data point by a majority vote rule of its nearest reference neighbors.

7.2 Projection of neuroblastoma transcriptomic data onto a mouse adrenal medulla reference atlas

The publicly available single-cell transcriptomic data of 384 cells, where Furlan et al. (2017) described the cellular composition and dynamics of the developing adrenal medulla in mice (E12.5), has become a staple to understand the development of the adrenal medulla, suprarenal ganglia sympathoblasts, and the onset of neuroblastoma. According to Furlan et al. (2017), the developing adrenal medulla in mice is composed of Schwann cell precursors (SCPs), early and late chromaffin cells, and a connecting Bridge population (**Figure 7.2a**).

In particular, the origin of the neuroblastoma MES subtype described in "Neuroblastoma regulatory subtypes defined by super-enhancers" (**Figure 6.5a**) could be explained by identifying similar regulatory landscapes with its cellular origin. Therefore, we explored this shared identity by projecting transcriptomic neuroblastoma data onto the reference atlas made from the developing adrenal medulla in mice.

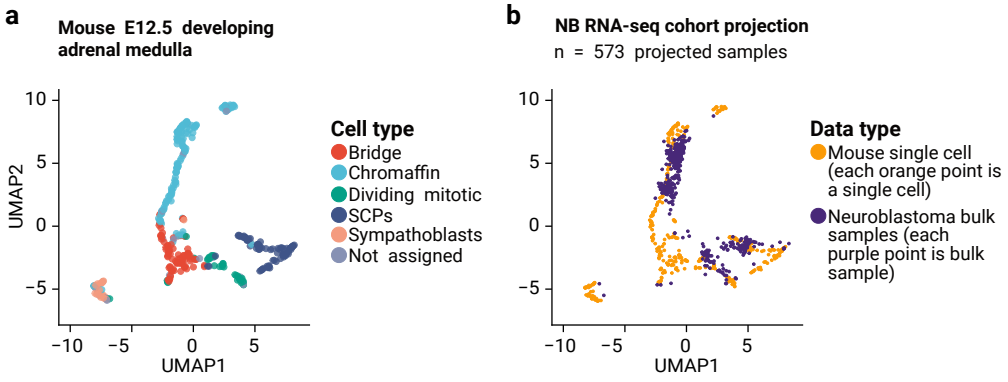


Figure 7.2: Projection of the NB RNA-seq cohort onto a mouse adrenal medulla atlas. **(a)** UMAP visualization of mouse adrenal medulla cells at E12.5. Colors indicate the inferred cell type based on marker genes. **(b)** Projection of NB tumor transcriptomic profiles (n=573) (purple dots) onto the landscape shown in (a) defined by the mouse adrenal medulla cells (orange dots). *Figure modified from Gartlgruber et al. (2021) with permission of Springer Nature.*

7.2.1 The mesenchymal subtype resembles Schwann cell precursors

We projected our cohort of bulk transcriptomic data from 579 NB tumors onto the developing mouse adrenal medulla atlas embedding (**Figure 7.2b**). In order to map the human genes to their corresponding mouse ortholog, we used the homologous master list from the Mouse Genome Informatics Web Site (MGI, Bult et al. 2019).

The labels from the mouse atlas were transferred to the NB tumors, seeking to determine

the identity of the bulk samples based on the nearest mouse-cell neighbors (**Figure 7.3a**). The majority of the samples from the NB bulk RNA-seq cohort mapped to early chromaffin cells with smaller proportions mapping to cycling cells and SCPs (**Figure 7.3a** inset plot). Remarkably, the MES subtype showed a distinct overlap with SCPs, suggesting a high degree of phenotypic similarity (**Figure 7.3b**).

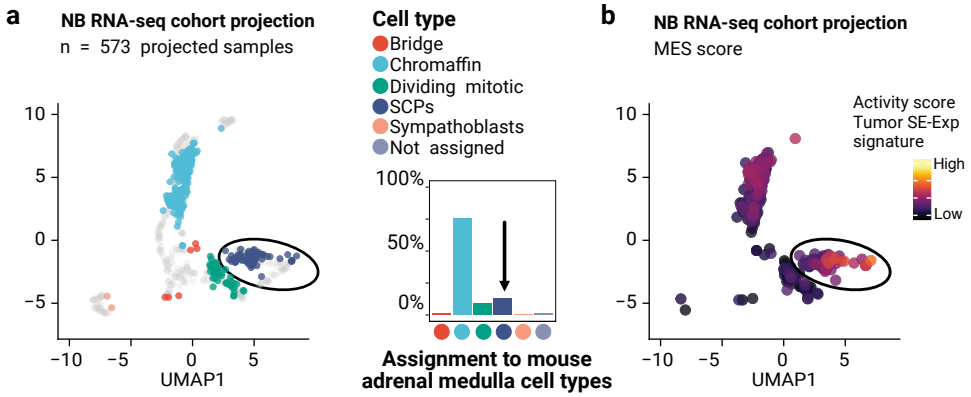


Figure 7.3: Tracing the cellular origin of the MES subtype - mouse atlas. **(a)** Projection of NB tumor transcriptomic profiles (n=573) (colored dots) onto the landscape shown in **Figure 7.2a** defined by the mouse adrenal medulla cells (grey dots). The tumor samples are colored according to **Figure 7.2a** and the quantification of the most frequent neighboring single cells is shown in the inset barplot. **(b)** The same tumor samples are colored by the MES signature activity score shown in **Figure 6.5a**. *Figure taken from Gartlgruber et al. (2021) with permission of Springer Nature.*

7.2.2 Expression of MES signature genes in cells of the developing adrenal gland

We further tested the association between the MES subtype with SCPs using the comprehensive single-cell atlas of mouse organogenesis (Cao et al. 2019) covering multiple developmental stages E9.5–E13.5. We quantified the mean expression of genes from

the MES signature in single-cells of the potential progenitor cell types for the adrenal gland, and sympathetic ganglia development (**Figure 7.4**). Interestingly, we observed that among the groups of potential progenitors, only SCPs showed a high expression of genes from the MES signature across multiple developmental times. Confirming a clear association between SCPs and the MES subtype

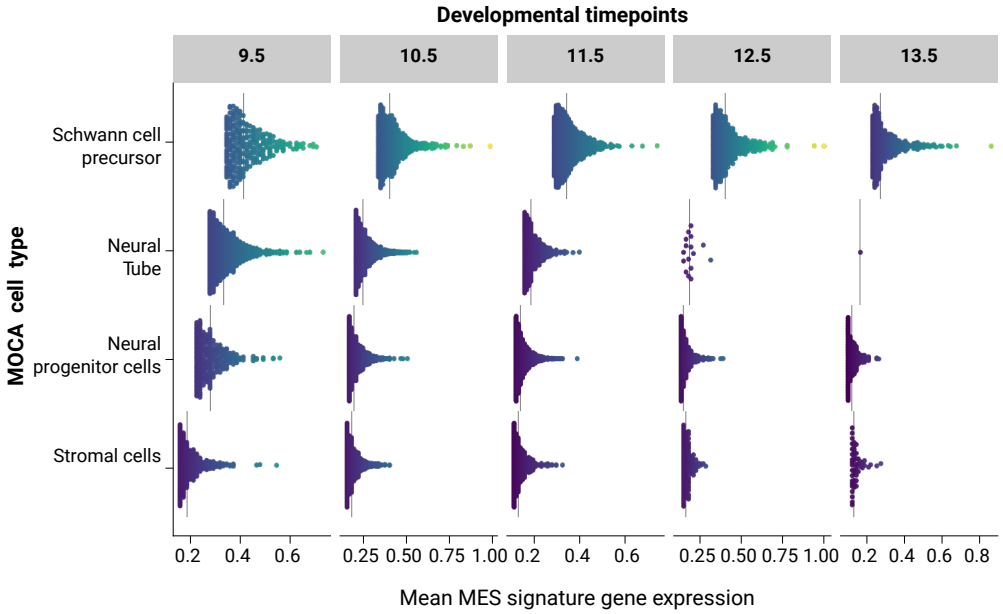


Figure 7.4: Mean expression of the mesenchymal gene signature in four selected single-cell populations from the mouse organogenesis (MOCA) dataset, indicating that the Schwann cell precursors have the highest mean mesenchymal expression across all developmental time points. *Figure taken from Gartlgruber et al. (2021) with permission of Springer Nature.*

7.2.3 Neuroblastoma mesenchymal cell lines resemble Schwann cell precursors

To deeply understand the relationship of the neuroblastoma mesenchymal phenotype to the Schwann cell precursors population, we projected data from three neuroblastoma

cell lines (i.e., KELLY, SK-N-AS, and SK-N-SH) at single-cell resolution onto the developing adrenal medulla mouse atlas. KELLY is a homogeneous cell line derived from neuroblastoma that possesses a genomic amplification of the N-myc gene, resulting in an elevated expression of the mRNA and protein products (Schwab et al. 1983). On the other hand, SK-N-AS was originally described as a mesenchymal-like cell line (S-type, expressing VIM); however, Boeva et al. (2017) reported heterogeneity of the cell identity for this cell line. SK-N-SH is also a heterogeneous cell line model containing noradrenergic (N-type, expressing neuronal markers), mesenchymal-like (S-type, expressing VIM), and intermediate cell types (I type) (Biedler, Helson, and Spengler 1973; Ciccarone et al. 1989).

Upon projecting the KELLY cells (n=636) onto the developing adrenal medulla mouse atlas, we observed that almost all cells mapped to cycling cells (**Figure 7.5a,b** top panel). On the other hand, the majority of SK-N-AS (n=603) and SK-N-SH (1,176) cells mapped to either SCPs or cycling cells (**Figure 7.5a,b** middle and lower panel).

Furthermore, to confirm that the similarities between neuroblastoma cells mapping to SCPs are related to the MES signature defined from NB tumors (**Figure 6.5a**), we quantified the mean expression of genes from the MES signature in the single-cells of neuroblastoma cell lines (**Figure 7.5c**). We found that the fraction of SK-N-AS and SK-N-SH cells mapping to SCPs exhibited the highest mean expression of MES signature genes. Taking together, these results confirmed that neuroblastoma cells with mesenchymal characteristics share similar identities with SCPs.

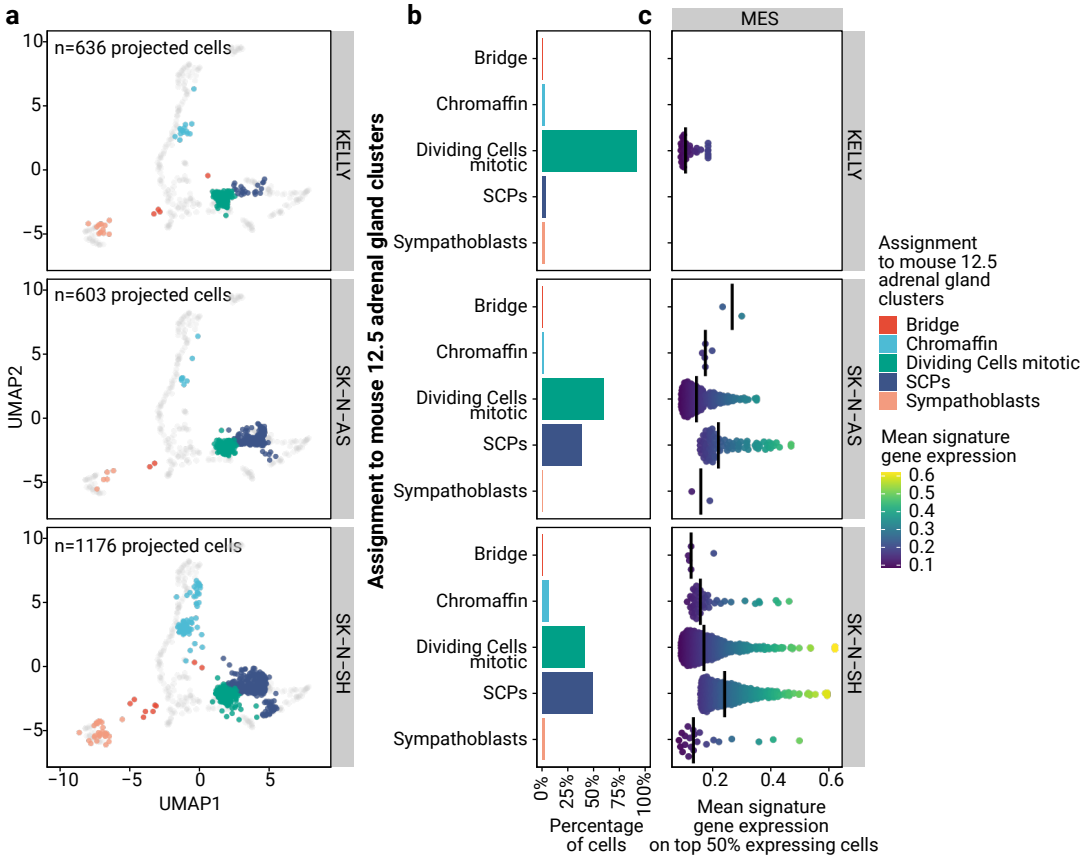


Figure 7.5: **(a)** Projection of single-cells from the KELLY, SK-N-AS, and SK-N-SH cell lines (colored dots) onto the landscape shown in **Figure 7.2a** defined by the mouse adrenal medulla cells (grey dots). **(b)** Quantification of the associated KELLY, SK-N-AS, and SK-N-SH cells based on nearest neighbors from the mouse adrenal cells. **(c)** Mean expression of the MES gene signature in the single cells of the indicated cell line by cell type, showing a higher mesenchymal activity for the cells associated with SCPs. *Figure modified from Gartlgruber et al. (2021) with permission of Springer Nature.*

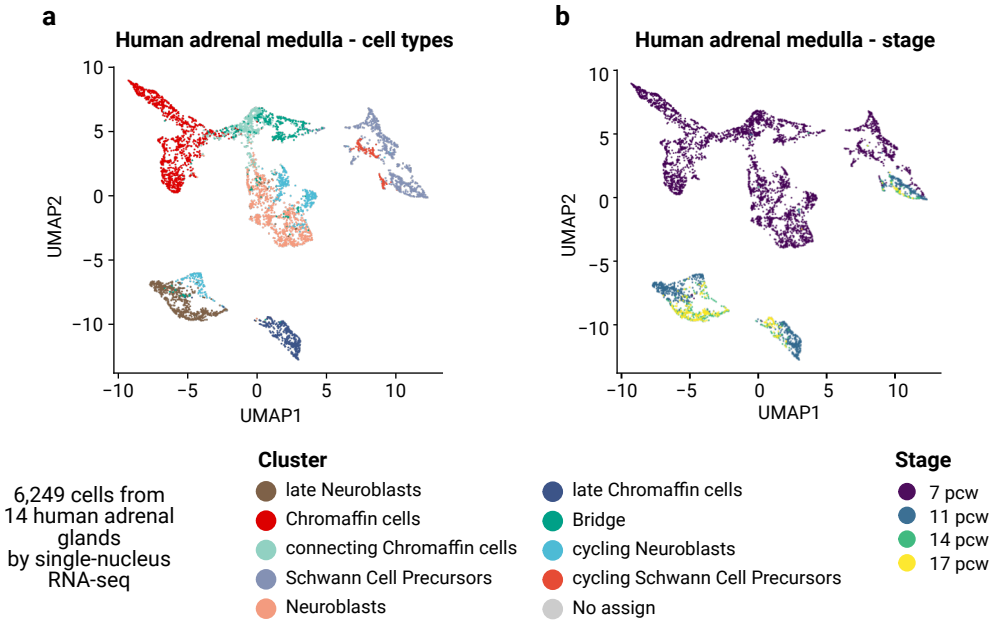


Figure 7.6: A single-cell human adrenal medulla atlas. UMAP visualization of human adrenal medulla cells at 7, 11, 14, and 17 pcw **(a)** colors indicate the inferred cell type based on marker genes. **(b)** colors indicate the developmental time points.

7.3 Projection of neuroblastoma transcriptomic data onto a human adrenal medulla reference atlas

Although the projections to the reference atlas made from the developing adrenal medulla in mice helped us to understand that the MES subtype shares similar regulatory landscapes with the SCPs, we sought to confirm that our findings were not an artifact of the intrinsic differences between human and mouse development. Therefore, we used the recently published single-cell human adrenal medulla dataset by Jansky et al. (2021), to build an atlas and to project neuroblastoma transcriptomic data onto it. For this, we

used a total of 6,249 single-cells from 14 human adrenal glands spanning the 7, 11, 14, and 17 post-conception weeks (pcw) (**Figure 7.6**).

7.3.1 The mesenchymal subtype resembles human Schwann cell precursors

We projected our cohort of bulk transcriptomic data from 579 NB tumors onto the human adrenal medulla atlas embedding (**Figure 7.7a**) and transferred the labels to determine the identity of the bulk samples (**Figure 7.7b**). Similar to what we found with the projections to the mouse adrenal medulla atlas, the MES subtype showed a distinct overlap with SCPs, confirming a high degree of phenotypic similarity in humans as well (**Figure 7.7c**).

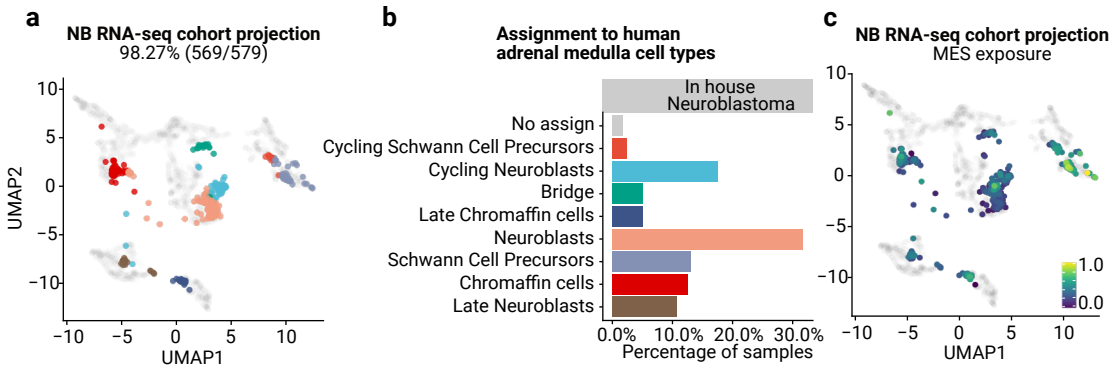


Figure 7.7: Tracing the cellular origin of the MES subtype - human atlas. **(a)** Projection of NB tumor transcriptomic profiles (n=573) (colored dots) onto the landscape shown in **Figure 7.6a** defined by the human adrenal medulla cells (grey dots). The tumor samples are colored according to **Figure 7.7a**. **(b)** Quantification of the most frequent neighboring single cells. **(c)** The same tumor samples are colored by the MES signature activity score shown in **Figure 6.5a**.

7.3.2 Single cells from tumors with mesenchymal characteristics map to Schwann cell precursors

We further hypothesized that single cells from neuroblastoma tumors with high exposure to the MES signature may be projected close to the SCPs population. As the tumor composition could be affected by the tumor identity, we used data of two tumors (i.e., NSP050-PT01 and NSP052-PT01) from the original bulk RNA-seq cohort that showed different exposures to the NMF signatures corresponding to the neuroblastoma epigenetic subtypes (**Figure 7.8a**).

We projected single-cell transcriptomic data for the tumors NSP050-PT01 and NSP052-PT01 onto the human developing adrenal medulla atlas and observed that the proportion of cells mapping to the SCPs population was higher for the tumor NSP050-PT01 than for NSP052-PT01 (**Figure 7.8b**). Considering that tumor NSP050-PT01 showed higher exposure to the MES signature than tumor NSP052-PT01, suggests that the larger proportion of cells mapping to the SCPs in neuroblastoma tumors could be related to the mesenchymal properties of the tumor.

7.4 Chapter summary

Identification of the specific cell(s) of origin for embryonal tumors like neuroblastoma and their subtypes is fundamental towards understanding their development, progression, and relapse mechanisms. We developed a new NMF-based workflow (implemented in ButchR) to map any contextually similar bulk or single-cell transcriptomic data onto a reference single-cell data, aiming to determine the possible cell of origin of the neuroblastoma MES subtype identified with ButchR (see "Neuroblastoma regulatory subtypes defined by super-enhancers").

We applied our method to map transcriptomes of neuroblastoma data onto single-cell transcriptomic profiles from developing mouse and human adrenal glands. The phe-

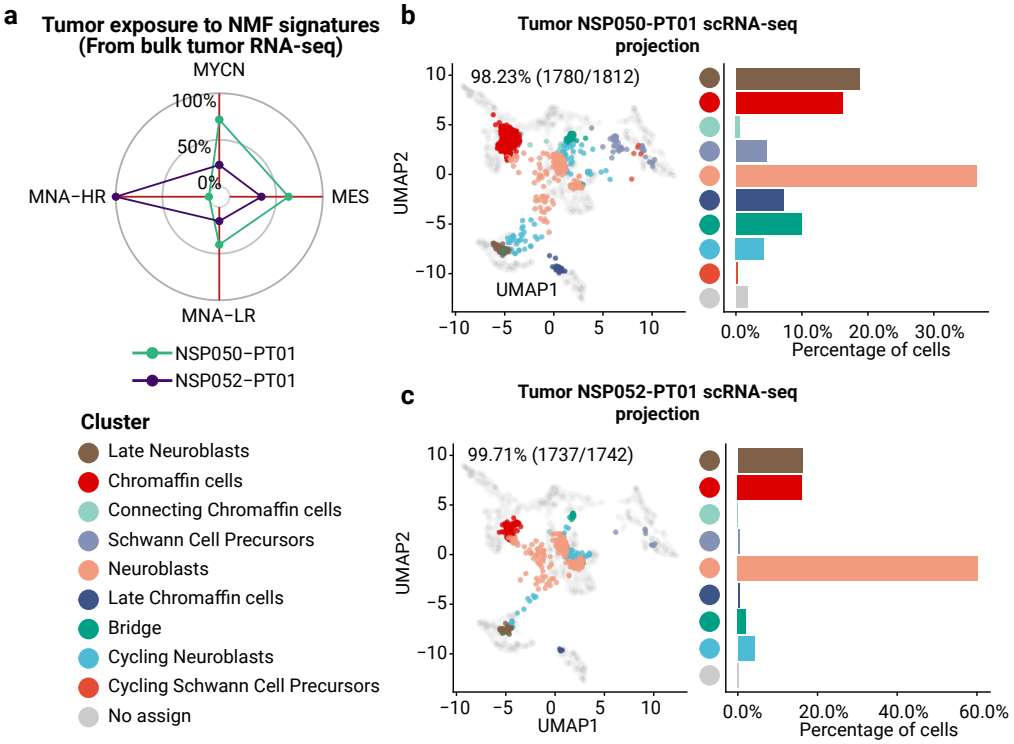


Figure 7.8: Projection of single cells from neuroblastoma tumors onto a human adrenal medulla atlas. **(a)** Relative exposure to the epigenetic signatures defined in **Figure 6.5a** of the tumors NSP050-PT01 and NSP052-PT01. **(b)** Projection of single cells from tumor NSP050-PT01 **(c)** and tumor NSP052-PT01 onto the landscape shown in **Figure 7.6a** defined by the human adrenal medulla cells (grey dots). Every cell is colored according to **Figure 7.7a**, quantification of the most frequent neighboring is shown in the inset plot.

notypic similarities of neuroblastomas entities with distinct cellular types from normal differentiation stages, including SCPs to early neuroblast/chromaffin cells, revealed neuroblastoma cellular heterogeneity. In particular, the MES subtype shared similar regulatory landscapes with multipotent Schwann cell precursors, pointing towards a strong phenotypic and molecular similarity between these two cell types.

Part III. Tracing Identity Defined by Transcription Factor Activity

So far, we have used ButchR to decompose matrices that were built in a traditional way. For instance, we used gene expression matrices resulting from the measurement of mRNA levels for all genes (chapter: "ButchR: NMF suit to slice genome-scale datasets"), matrices containing epigenomic measurements like the signal of histone marks, or gene expression matrices from curated set genes (chapter: "Neuroblastoma regulatory subtypes defined by super-enhancers").

While these approaches proved to be successful for understanding mechanistic processes for groups of biological entities (i.e., cells or bulk tumor samples), explaining more complex processes would require the construction of initial matrices using alternative strategies. For instance, explaining the regulation of gene expression by the simultaneous action of *cis*-regulatory elements.

Here (**Part III**), we devised two new methods for building matrices that can be decomposed in regulatory signatures using ButchR. In the first case, binary matrices explaining the on/off status of regulatory links between genome elements were built by using data generated from state of the art methods, in which it is possible to co-profile chromatin accessibility and gene expression from the same single cell (chapter: "Understanding gene expression regulation with scCAT-seq"). In the second case, we generated matrices that served as a proxy of TF activity by using cell-state specific regulons (chapter: "Deconvolution of regulon-guided signatures"). Some of the results presented here are product of an international collaboration to reveal regulatory heterogeneity by combining single-cell multi-omics layers (Liu et al. 2019).

Chapter 8

Understanding gene expression regulation with scCAT-seq

Disclosure: The results presented in this chapter have been partially published in Liu et al. (2019) and reproduced here with the permission of Springer Nature.

The developmental destination of a cell (cell fate) and the transcriptional output of the gene regulatory networks (cell state), defined by modules formed of transcription factors and their target genes (regulon) are tightly coupled to the interplay between the epigenetic and transcriptomic landscapes of the cell (Spitz and Furlong 2012; Shema, Bernstein, and Buenrostro 2019; Moris, Pina, and Arias 2016). Therefore, measuring the epigenomic and transcriptomic characteristics of single cells is a key element towards understanding the patterns of regulatory relationships between these two elements.

Until recently, it was not possible to study the direct relationship between the transcriptome and the chromatin accessibility in the same single cell. However, this limitation has been overcome by recent advances in single-cell technologies; in particular, techniques like scCAT-seq (Liu et al. 2019) and SHARE-seq (Ma et al. 2020) provide measures of mRNA expression (scRNA-seq) and chromatin accessibility (scATAC-seq) from the same single cell.

In this chapter, we present a new methodology to infer regulatory relationships between genes and their *cis*-regulatory elements (CREs), followed by the identification of regulatory signatures that can help to define the cell state using ButchR. To this end, we used three scCAT-seq datasets generated by Liu et al. (2019), in which simultaneous profiling of scRNA-seq and scATAC-seq was done for cells of three cell lines, two patient-derived xenografts (PDX) tissues, and pre-implantation human embryos (**Figure 8.1**). Every dataset consists of a gene expression matrix, a set of peaks found from the scATAC-seq data, and a matrix of peak counts.

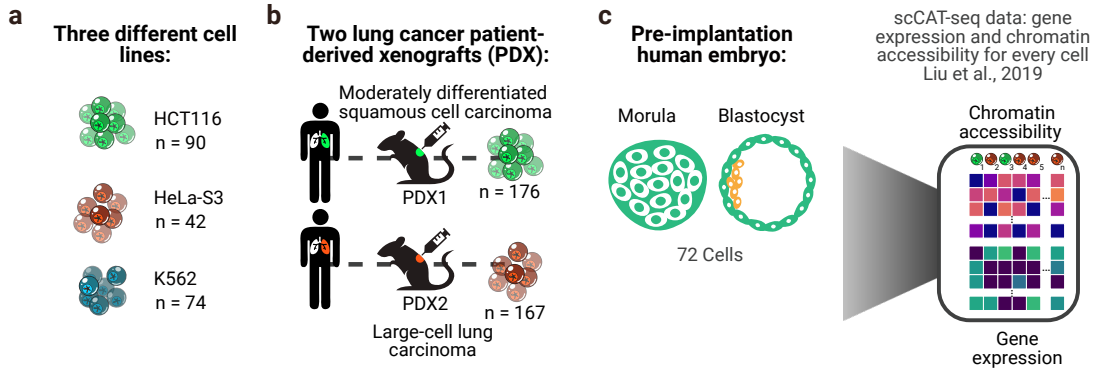


Figure 8.1: Schematic representation of used scCAT-seq datasets. scCAT-seq provides an accurate genome-wide measure of both chromatin accessibility and gene expression for every single cell. Here, we used three datasets generated by Liu et al. (2019); **(a)** three cell lines; **(b)** two lung patient-derived xenografts (PDX) tissues; **(c)** and pre-implantation human embryos. *Figure modified from Liu et al. (2019) with permission of Springer Nature.*

8.1 Prediction of regulatory relationships using scCAT-seq

Gene expression regulation is mediated by the simultaneous action of multiple CREs (Lenhard, Sandelin, and Carninci 2012; Spitz and Furlong 2012), making them a key intermediate component in Gene regulatory networks (GRNs) (Davidson 2010). Therefore,

to thoroughly describe and predict cell states, it is necessary to understand the complex interplay between TFs and the CREs that act on their target genes, as well as how these regulatory relationships affect gene expression. Therefore, we used scCAT-seq data from three cell lines (K562, HeLa-S3, and HCT116) (Liu et al. 2019) (**Figure 8.1a**) to predict regulatory relationships between TF and CREs, taking advantage of the availability of the two omics layers (i.e., gene expression and chromatin accessibility) across single cells.

8.1.1 Three strategies to predict regulatory relationships

It has been shown that the correlation of the accessibility between *cis*-regulatory elements provides an effective approach to define regulatory links (Buenrostro et al. 2015). However, with the additional layer of information provided from scCAT-seq (i.e., gene expression), it is also possible to devise further strategies that effectively integrate the two omics layers to predict regulatory relationships.

We used the following three strategies to define regulatory relationships between TFs *cis*-regulatory elements and a target gene in the cell lines dataset:

Strategy 1:

This strategy is based on correlation of promoter and distal peaks signal, it uses only scATAC-seq data (**Figure 8.2** Strategy 1):

1. For every gene promoter, find the list of nearby accessible regions (i.e., peaks located in the window 1 Mb upstream-downstream of the promoter).
2. Estimate the Spearman correlation between the signal of peaks located at gene promoters and every nearby peak.
3. Assign a regulatory link when the Spearman correlation was above 0.25.

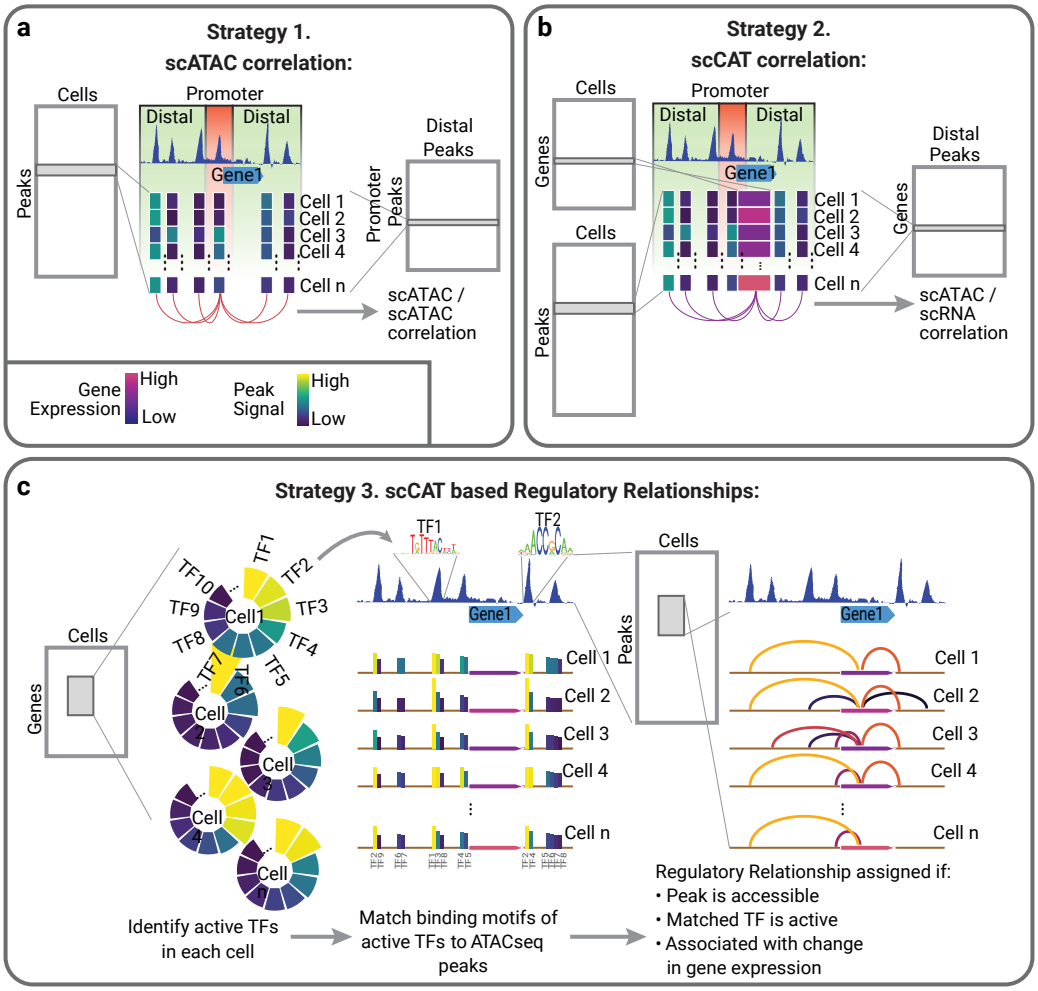


Figure 8.2: Inferring regulatory relationships between CREs and genes by scCAT-seq. Overview of three strategies for inferring regulatory relationships. Strategy 1: based on correlation of promoter and distal peaks signal. Strategy 2: based on correlation of gene expression and distal peaks signal. Strategy 3: based on TF activity, accessibility of CREs, and effect on gene expression. See the text for a detailed explanation. *Figure modified from Liu et al. (2019) with permission of Springer Nature.*

Strategy 2:

The second strategy is based on the correlation of gene expression and distal peaks signal, it uses scRNA-seq and scATAC-seq data (**Figure 8.2** Strategy 2):

1. For every gene, find the list of nearby accessible regions (i.e., peaks located in the window 1 Mb upstream-downstream of the gene promoter).
2. Estimate the Spearman correlation between the gene expression and the signal of every nearby peak.
3. Assign a regulatory link when the Spearman correlation was above 0.25.

Strategy 3:

The third strategy is based on TF activity, accessibility of CREs, and effect on gene expression, it uses scRNA-seq and scATAC-seq data (**Figure 8.2** Strategy 3):

1. **Identification of active TFs for every cell by pySCENIC (Van de Sande et al. 2020):**
    - Starting from the gene expression matrix, define regulons based on the co-expression of TFs and their target genes across cells.
    - Quantify the regulon enrichment in each cell by measuring the area under the recovery curve (AUC) of the genes that define each regulon.
    - Classify individual TFs as active or inactive in each cell based on the bimodal distribution of the AUC scores of the corresponding regulon.
2. **Identification of gene-associated active and accessible regions in every cell:**
    - For every gene find the list of nearby accessible regions (i.e., peaks located in the window 1 Mb upstream-downstream of the gene promoter).
    - Match the binding motifs of active TFs to the list of nearby accessible regions using the R package Biostrings.
    - For every cell, classify the accessible regions as active, when at least one motif matched with at least 95% of the highest possible score for the given motif

Position Weight Matrix (PWM).

3. **Assignment of regulatory relationships:**
    - For every active and accessible region that was found in at least 10% of the cells, classify the cells in two groups (i.e., cells with and cells without the active and accessible region).
    - Perform a Wilcoxon rank-sum test between the two groups, comparing the expression of the gene associated with the active and accessible region.
    - Assign a regulatory relationship between the gene and the active and accessible region if the presence of the region was associated with a significant change in the gene expression (Wilcoxon test p-value < 0.05).
    - *Note:* by using this strategy, it is possible to recover more than one regulatory relationship for every gene. This is a reflection of the underlying complexity of GRNs.

One of the disadvantages of using strategies based on correlations like *Strategy 1* and *Strategy 2* is that the dictionary of inferred regulatory relationship will be defined across all cells and not in a cell-wise specific manner. On the other hand, our proposed *Strategy 3* will be able to infer regulatory relationships for every cell under consideration, allowing the comparison of the regulatory landscape across individual cells.

After applying these three strategies to the cell lines dataset, we found that the largest number of regulatory relationships was identified using *Strategy 3* (62,769), compared to *Strategy 1* (46,813) and *Strategy 2* (21,219) (**Figure 8.3a**). We observed that only about a third of the regulatory links predicted using exclusively chromatin accessibility data (*Strategy 1*), were shared with the regulatory links that effectively combined both omic layers available from the scCAT-seq cell lines dataset. Suggesting that the dynamics of the epigenomic and transcriptomic regulation in GRNs can not be fully explained using only chromatin accessibility data (**Figure 8.3a**).

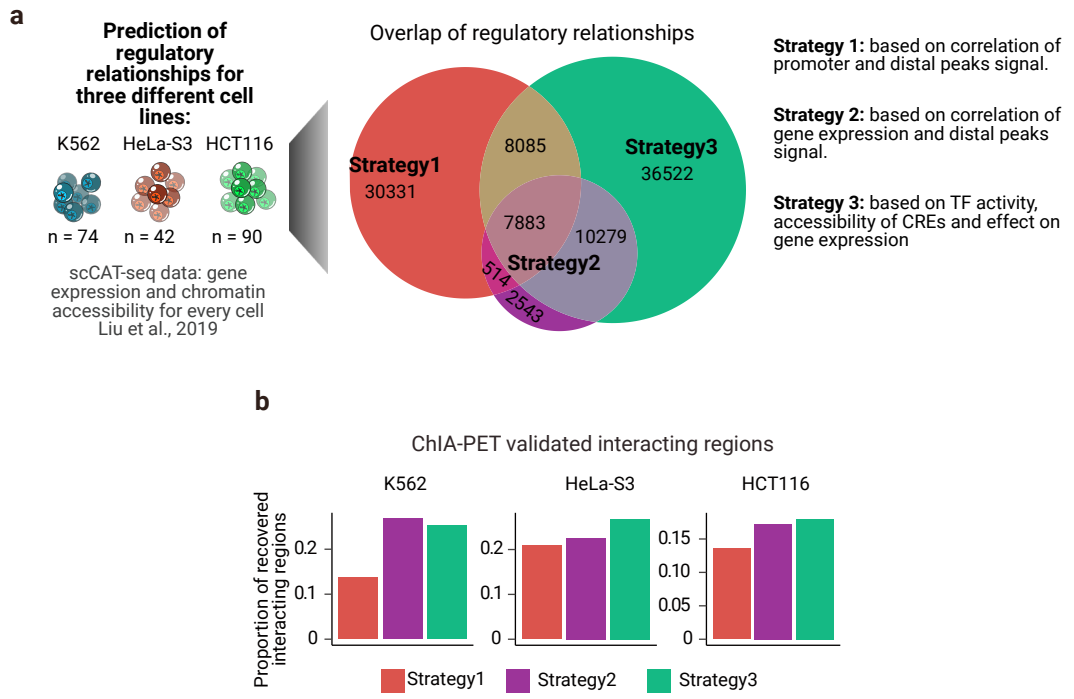


Figure 8.3: Validation of the prediction of regulatory relationships. **(a)** Venn diagram showing the number of overlapping regulatory relationships identified by the three strategies. **(b)** Proportion of ChIA-PET validated regulatory relationships identified by the three strategies in K562 (left), HeLa-S3 (middle), and HCT116 (right) single cells. *Figure modified from Liu et al. (2019) with permission of Springer Nature.*

8.1.2 Validation of the prediction of regulatory relationships

In order to validate the accuracy of the regulatory relationships predicted using the proposed strategies, we counted the number of regulatory relationships that were also recovered using chromatin interaction analysis by paired-end tag sequencing (ChIA-PET) (G. Li et al. 2010). We overlapped publicly available interaction profiles for the evaluated cell lines (i.e., K562, HeLa-S3, and HCT116) (Teng et al. 2015) to our predicted

regulatory links (**Figure 8.3b**), and found that for all three cell lines, the proportion of links validated by ChIA-PET interactions were higher when using the multi-omics-based methods (*Strategy 2* and *Strategy 3*) in comparison to using only scATAC-seq (*Strategy 1*). Suggesting that the interaction between CREs and gene expression could be a better model to understand GRNs than models based only on the interaction between CREs at different genomic locations.

Regarding *Strategy 2* and *Strategy 3*, we selected the latter to perform the rest of our analyses based on the largest number of validated regulatory relationships, and that with this strategy it is possible to predict regulatory links for every single cell.

8.1.3 Using ButchR to find regulatory signatures

With the potential to define regulatory relationships for every single cell (i.e., using *Strategy 3*), we further hypothesized that the hidden patterns of common regulatory pathways used by groups of cells that share similar cell states could be revealed by finding regulatory signatures using ButchR.

We generated a binary matrix X_{reg} for the cell lines dataset, where columns represent single cells and every row is one regulatory relationship identified between an accessible site and one gene. Values of 1 indicate the presence of the regulatory link in the indicated cell. Using ButchR, we then decomposed the matrix X_{reg} into the matrices W_{reg} and H_{reg} with an optimal factorization rank $k = 3$. A closer inspection of the matrix H_{reg} (**Figure 8.4a**), revealed that each of the recovered signatures showed a high correspondence to only one of the cell lines under consideration.

To understand the regulatory links explaining every signature and whether they showed any relation to the biology of the associated cell line, we extracted signature associated features (i.e., regulatory relationships that show high contribution towards the definition of the signature under consideration) from the matrix W_{reg} using ButchR, and visualized the genomic distribution of a few representative examples across the three cell lines

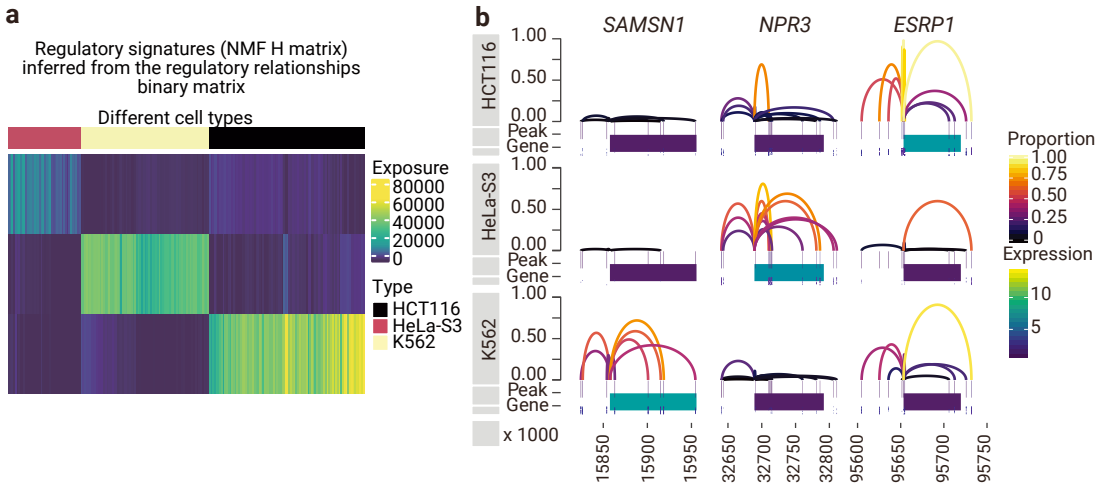


Figure 8.4: NMF analysis of the regulatory relationships inferred for the scCAT-seq cell lines dataset. **(a)** Heatmap representation of the exposure matrix H_{reg} decomposed for the K562, HeLa-S3, and HCT116 cell lines. Values represent the exposure of every cell to the regulatory signatures. **(b)** Regulatory relationships for the indicated genes in each cell line. Each panel contains three tracks: the top track shows the regulatory relationship between one peak and the gene (linking them with an arch), where the height and color of the arch show the proportion of cells that share the regulatory relationships; the middle track shows the genomic location of the gene and the associated peaks, where the color of the gene shows the mean expression in each cell line; the bottom track shows the accessible states (on and off) for each peak in every single cell. *Figure taken from Liu et al. (2019) with permission of Springer Nature.*

(**Figure 8.4b**). For instance, we found a higher number of regulatory relationships to the known oncogene SAMSN1 in K562 cells compared to the other cell lines. As *SAMSN1* is preferentially expressed in multiple myeloma, and K562 is a myelogenous leukemia cell line, this suggests that our workflow is able to recover regulatory signatures associated with the cell phenotype.

We found similar results for the HeLa-S3 and HCT116 cell lines. To exemplify, a large number of regulatory links were found to be associated with the gene *NPR3* in HeLa-S3 cells, which correlates to a high expression of this gene in HeLa cells (https://www.proteinatlas.org/ENSG00000113389-NPR3/cell, Uhlen et al. 2017); similarly, a large number of links were identified to the *ESRP1* gene in HCT116 (https://maayanlab.cloud/Harmonizome/gene_set/HCT116/BioGPS+Cell+Line+Gene+Expression+Profiles, Rouillard et al. 2016)

Taken together, the results of the prediction of regulatory relationships using *Strategy 3* and its subsequent decomposition in regulatory signatures for the K562, HeLa-S3, and HCT116 cell lines showed us that this workflow is an effective alternative to understand the regulation of CREs and gene expression, and how regulatory relationships cell states.

8.2 Unveiling tumor regulatory heterogeneity

We further tested whether the proposed workflow was able to recover regulatory signatures related to tumor identity. To this end, we inferred regulatory relationships for the single cells of the scCAT-seq PDX tissues dataset (**Figure 8.1b**). This dataset includes 176 cells from a moderately differentiated squamous cell carcinoma (PDX1) and 167 cells from a large-cell lung carcinoma (PDX2).

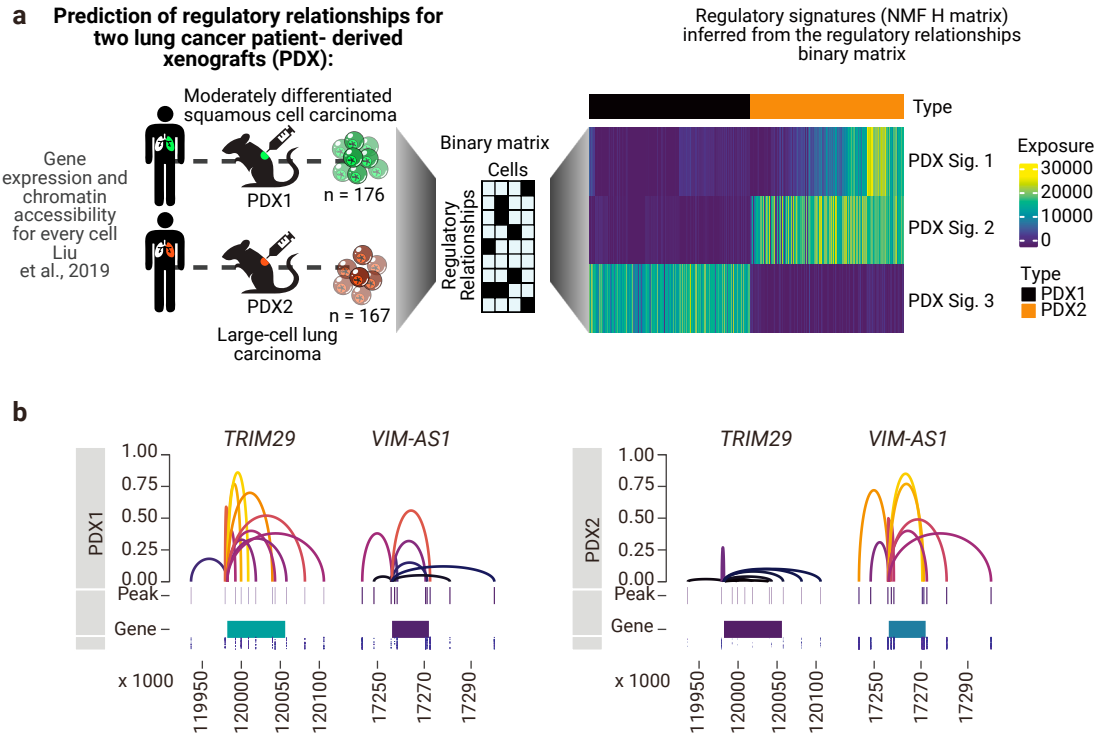


Figure 8.5: NMF analysis of the regulatory relationships of PDX tissues. **(a)** Heatmap representation of the exposure matrix H_{reg} decomposed from single cells of two PDX tissues. Values represent the exposure of every cell to the regulatory signatures. **(b)** Regulatory relationships for *TRIM29* and *VIM-AS1* in PDX1 and PDX2, tracks are built as in **Figure 8.4b**. *Figure modified from Liu et al. (2019) with permission of Springer Nature.*

8.2.1 Tumor regulatory signatures

We used ButchR to decompose the regulatory relationships binary matrix into regulatory signatures. In this case, the optimal factorization rank was $k = 3$. The resulting matrix H_{reg} was sufficient to separate PDX1 from PDX2 (**Figure 8.5a**). Interestingly, *PDX Signature 3* was highly correlated only to PDX1 cells, whereas *PDX Signature 2* only to PDX2 cells. On the other hand, only a fraction of the PDX2 cells showed a varying

degree of exposure to *PDX Signature 1*, hinting at the presence of different cell states in this tumor.

Given the clear regulatory differences between PDX1 and PDX2 revealed by the decomposed regulatory signatures, we extracted regulatory relationships that showed high contribution towards the definition of *PDX Signature 2* and *PDX Signature 3*, to find regulatory modules that drive these differences. We found *TRIM29* to be among the top regulated genes in PDX1. This gene product has been reported to mediate metastasis in lung squamous cell carcinoma via regulation of the autophagic degradation of E-cadherin (W. Xu et al. 2020), which corresponds to the original PDX1 tumor type (i.e., moderately differentiated squamous cell carcinoma). On the other hand, *VIM-AS1* was found to be one of the top regulated genes in PDX2. VIM-AS1 is a long noncoding RNA that promotes colorectal (Rezanejad bardaji, Asadi, and Yaghoobi 2018) and prostate cancer (Z. Zhang et al. 2019) progression inducing EMT. Therefore, the clear upregulation in PDX2 for the *VIM-AS1* regulatory module makes it a candidate to evaluate if it may also be involved in lung cancer progression.

8.2.2 Intra-tumor variability

We further investigated the possible intra-tumor heterogeneity present in PDX2, evidenced by a varying degree of exposure to the *PDX Signature 1*. To understand if the variability in the regulatory landscape was an artifact of the selection of factorization rank in the NMF (i.e., $k = 3$), we generated a riverplot visualization for factorization ranks 2 to 6 (**Figure 8.6a**). We found that *PDX Signature 1* was originated only from a PDX2-specific signature at $k = 2$ and that the signature was stable even until $k = 6$. Thus, this proves that *PDX Signature 1* is not an artifact and it is indeed capturing intra-tumor differences from PDX2.

We reasoned that the group of regulatory relationships specific for *PDX Signature 1* and *PDX Signature 2* could help to identify genes that were driving the regulatory variability in PDX2 cells. ButchR was used to extract signature-specific features from these

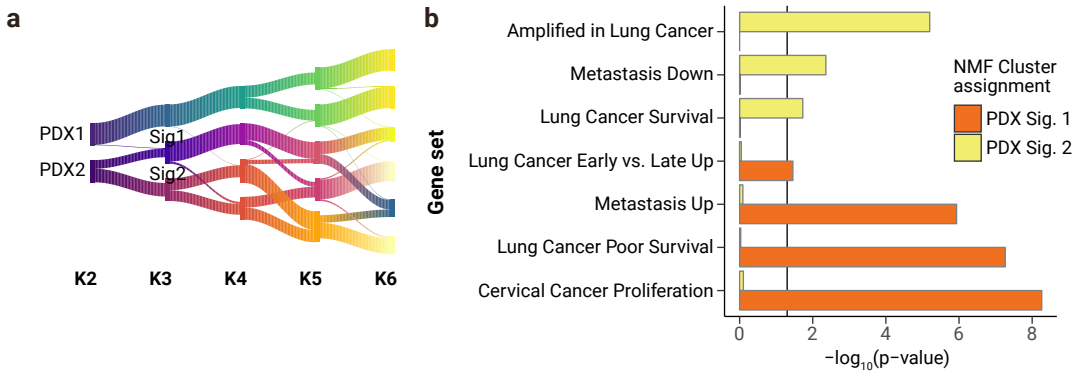


Figure 8.6: Intra-tumor variability in PDX2. **(a)** Riverplot from the NMF decomposition of the regulatory relationships binary matrix for PDX tissues, showing the stability of the signatures defining intra-tumor and inter-tumor variability. **(b)** Gene set enrichment analysis of genes associated with *PDX Signature 1* and *PDX Signature 2*. *Figure taken from Liu et al. (2019) with permission of Springer Nature.*

regulatory signatures. We found the genes linked to these regulatory relationships, effectively building two groups of genes associated with each of the interrogated signatures. Then, we performed gene set enrichment analysis using these groups of genes (**Figure 8.6a**) and found that *PDX Signature 1* was highly enriched for gene sets associated with metastasis, poor survival, and cancer proliferation. Whereas *PDX Signature 2* showed enrichment for gene set related to good survival in lung cancer, these findings pointed us to hypothesize that a fraction of the PDX2 cells (i.e., the cells with high exposure to *PDX Signature 1*) might come from a population that was starting a metastatic process.

In sum, the application of the proposed workflow to infer regulatory signatures from the scCAT-seq PDX tissues dataset, showed us that it is possible to recover signatures explaining inter-tumor and intra-tumor variability, making it a viable methodology to understand tumor development, progression, and subtype identification.

8.3 Understanding early development regulation in human

After evaluating the potential of our workflow to learn regulatory signatures that recapitulate the action of CREs regulating gene expression, we explored how well the regulatory signatures were able to characterize single-cell identities in a continuous developmental process. Therefore, we applied our workflow to the scCAT-seq pre-implantation human embryos dataset (**Figure 8.1c**), which comprised of 29 single cells from the morula stage and 43 from the blastocyst stage.

8.3.1 Regulatory differences in human morula and blastocyst

We inferred regulatory relationships for every single cell in the human embryos dataset and used ButchR to decompose regulatory signatures (**Figure 8.7a**), finding a total of two regulatory signatures ($k = 2$), one corresponding to cells in the morula stage (*Morula Signature*) and the other to blastocyst stage cells (*Blastocyst Signature*).

Then, we extracted the regulatory relationships that showed more contribution towards the definition of the *Morula Signature* and *Blastocyst Signature*. We found pluripotency markers such as NANOG (**Figure 8.7b** left panel) for the *Morula Signature* and trophectoderm markers as GATA3 (**Figure 8.7b** right panel) for the *Blastocyst Signature*, confirming that the regulatory signatures were able to disentangle the regulatory patterns that arise in a continuous developmental process.

8.3.2 Identification of inner cell mass cells with i2NMF

The blastocyst stage consists of the inner cell mass (ICM) and trophectoderm cells. During blastocyst development, a fraction of the ICM cells segregates into pluripotent epiblast (Shahbazi and Zernicka-Goetz 2018). Remarkably, there were three blastocysts among the cells that showed high exposure to the *Morula Signature* (**Figure 8.7a**), revealing pluripotency characteristics in these three single cells, and suggesting that they

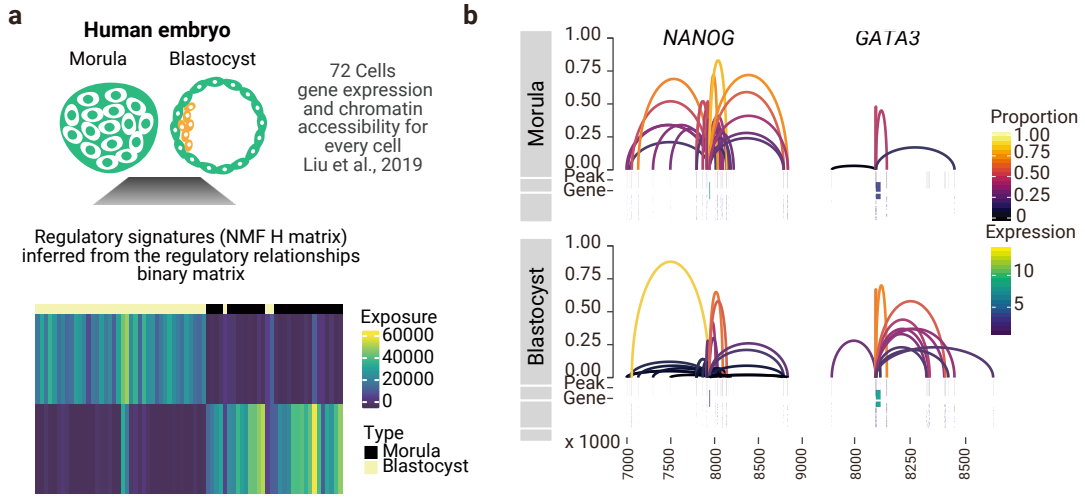


Figure 8.7: NMF analysis of the regulatory relationships of human pre-implantation embryos. **(a)** Heatmap representation of the exposure matrix H_{reg} decomposed from single cells of human pre-implantation embryos. Values represent the exposure of every cell to the regulatory signatures. **(b)** Regulatory relationships for *NANOG* and *GATA3* in morula and blastocyst cells. Tracks are built as in **Figure 8.4b**. *Figure taken from Liu et al. (2019) with permission of Springer Nature.*

could be part of the ICM cells (hereafter referred as ICM-like cells).

As the ICM-like cells shared characteristics with the *Morula Signature* and *Blastocyst Signature*, but not defining a specific signature for them, we hypothesized that the specificity of these cells might be related to the heterogeneous effect explained by the chromatin accessibility or the gene expression layers. Therefore, we used i2NMF (see "i2NMF: An integrative approach to discover dataset-specific effects") to disentangle the homogeneous and the heterogeneous effects in the morula, blastocyst, and ICM-like cells (**Figure 8.8**). In this case, we used the gene matrix expression and the matrix of peak counts and integrated both datasets across the single cells.

We recovered two signatures in the exposure matrix H_S (i.e., shared effect between

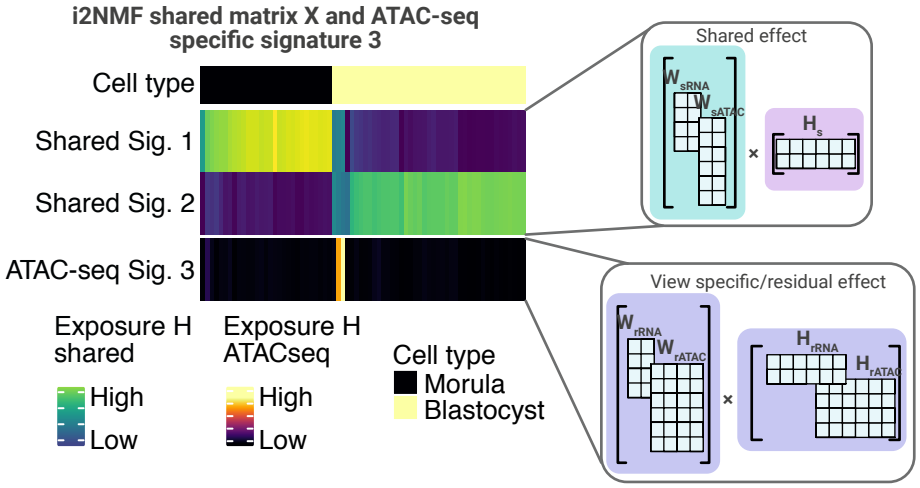


Figure 8.8: i2NMF integrative analysis of scRNA-seq and scATAC-seq data from human pre-implantation embryos. Heatmap representation of the exposure matrix H_S (i.e., shared effect between both omics layers) and the *ATAC-seq Signature 3* from the matrix H_{rATAC} (i.e., the specific signature from the residual effect of the scATAC-seq).

both omics layers), corresponding to morula and blastocyst cells. The ICM-like cells showed mixed exposure to both of these signatures, indicating that the features that uniquely define these cells may come from only one omic layer (i.e., defined either by gene expression or chromatin accessibility). After performing the second stage of i2NMF, we were able to find one signature (i.e., *ATAC-seq Signature 3*) explained only from the scATAC-seq data that was highly specific for two of the ICM-like cells (**Figure 8.8**). Using ButchR, we extracted the *ATAC-seq Signature 3* specific regions and looked at the regulatory relationships (n=372) that linked these regions to ICM markers (i.e., KLF17, NANOG, and OCT4. **Figure 8.9a**), revealing a set of regulatory relationships that showed a unique pattern of activation in the ICM-like cells.

In order to understand if the differences driven by chromatin changes were also detectable at the gene expression layer, we inspected the expression of KLF17, NANOG, and OCT4

in the ICM-like cells that showed a unique pattern of chromatin accessibility regulation, and compared it to the morula, and blastocyst cells (**Figure 8.9b**). It had been previously described that KLF17, NANOG, and OCT4 are expressed in all cells part of the ICM (Shahbazi and Zernicka-Goetz 2018). Remarkably, we found higher relative expression levels of these genes in the evaluated ICM-like cells, pointing that the specific features that define the ICM-like cells from the chromatin accessibility layer are regulating a restricted group of genes, which make them exhibit pluripotency traits.

Taken together, these results confirmed that combining i2NMF with our proposed workflow to infer regulatory signatures, captured specific regulatory traits in morula and blastocyst cells, as well as pointing towards the identification of rare cells with ICM-like properties.

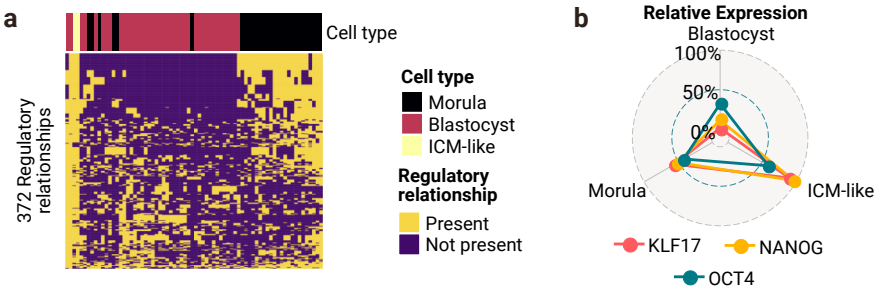


Figure 8.9: Identification of inner cell mass cells. **(a)** Regulatory relationships involving links to the *ATAC-seq Signature 3* specific regions (**Figure 8.8**). **(b)** Relative expression of the genes KLF17, NANOG, and OCT4 in Morula, Blastocyst, and ICM cells.

8.4 Chapter summary

Cell fate and cell state are tightly coupled to the dynamics of the regulatory links between the epigenetic and transcriptomic landscapes of the cell (Spitz and Furlong 2012; Shema, Bernstein, and Buenrostro 2019; Moris, Pina, and Arias 2016). Here, we established a

new workflow to infer regulatory relationships between genes and their *cis*-regulatory elements for individual cells, followed by the identification of regulatory signatures using ButchR. This workflow uses data where simultaneous measurements of chromatin accessibility and gene expression are available for the same single cell. We used three scCAT-seq datasets generated by Liu et al. (2019), where scRNA-seq and scATAC-seq data were retrieved for every cell.

The inference of regulatory relationships at a single cell granularity consists of three main steps: (i) identification of active TFs for every cell, (ii) identification of active accessible regions in the nearby region of gene promoters, and (iii) significance evaluation of the activation of one accessible region in the gene expression of an associated gene.

The regulatory relationships found with the proposed workflow were validated using publicly available ChIA-PET interaction profiles. We also found that the regulatory signatures were able to explain intra- and inter-tumor variability. Furthermore, we paired i2NMF with the prediction of regulatory relationships to explain regulatory differences in cells from human pre-implantation embryos. Taking together, these findings provide new insights into the control of gene regulatory networks, and an effective workflow to understand the regulation of CREs and gene expression.

Chapter 9

Deconvolution of regulon-guided signatures

Regulons are the building blocks of GRNs. They are defined by a module of genes *cis*-regulated by one TF (Van de Sande et al. 2020). It has been shown that regulon composition for the same TF can be different between tissues (i.e., the connections in GRNs show high tissue specificity) and that TF expression is not sufficient to regulate gene expression (Sonawane et al. 2017). Therefore, regulon activity cannot be measured by considering the TF expression alone. Nevertheless, identifying regulons and quantifying their activity are key components to explain the output of a GRN in an individual cell (cell state) (Moris, Pina, and Arias 2016).

Although we showed the potential of our proposed workflow to infer regulatory signatures where chromatin accessibility and gene expression are available for every single cell, the instances where both omic-layers are available are limited at the moment. Therefore, it is also important to consider new methods that help to explain the regulatory differences between cells without using data generated from techniques as scCAT-seq and SHARE-seq.

In this chapter, we propose how using regulon-guided signatures (i.e., signatures decom-

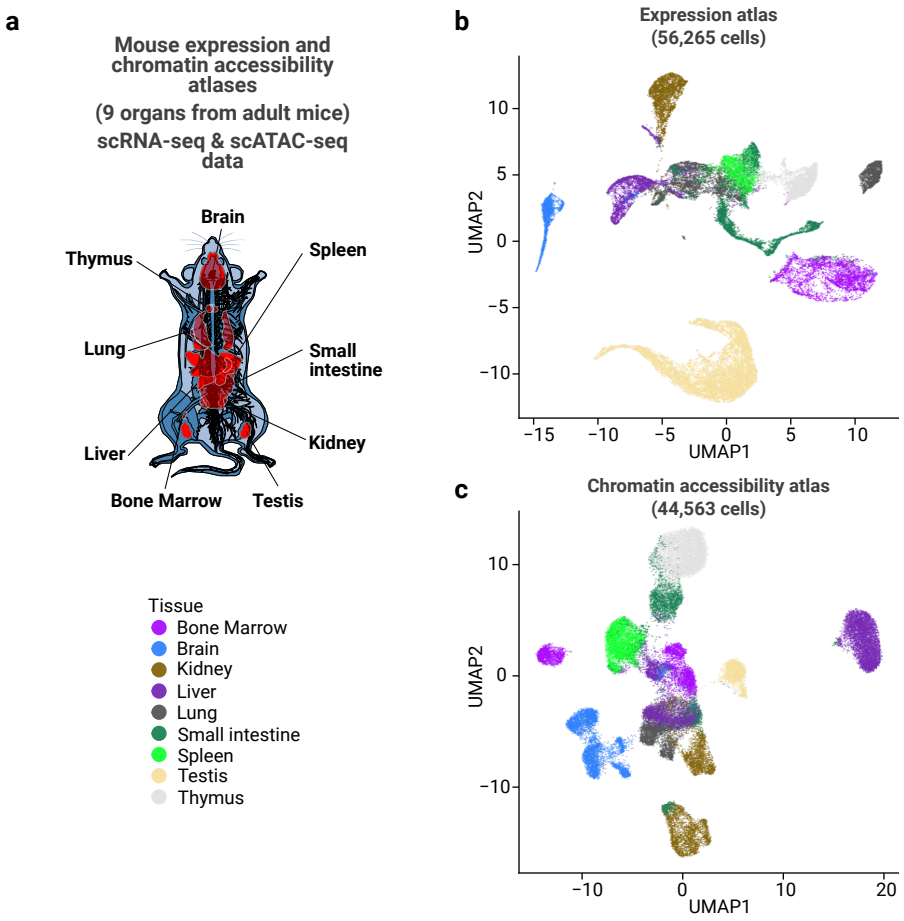


Figure 9.1: Mouse gene expression and chromatin accessibility atlases. **(a)** Schematic representation of the mouse tissues included in gene expression and chromatin accessibility atlases. **(b)** UMAP embedding representation of the 56,265 cells included in the gene expression atlas **(c)** UMAP embedding representation of the 44,563 cells included in the chromatin accessibility atlas.

posed by ButchR from a regulon activity matrix) can be a helpful approach to explain cell state differences between single cells. We also show a new method to infer cell-state-specific regulons using a combination of scRNA-seq and scATAC-seq from different cells. To this end, we used publicly available adult mouse single-cell data from the tissues de-

picted in **Figure 9.1a**, including a gene expression atlas with 56,265 cells (Han et al. 2018) (**Figure 9.1b**) and a chromatin accessibility atlas with 44,563 cells (Cusanovich et al. 2018) (**Figure 9.1c**).

9.1 Regulon activity quantification from scRNA-seq data

SCENIC (Aibar et al. 2017; Van de Sande et al. 2020) is one of the most widely used methods to reconstruct regulons and quantify their activity based on the expression level of single cells. The most important feature implemented for the first time in SCENIC was the possibility of quantifying regulon activity for every single cell, in contrast to methods based on correlations, in which the activity of a TF is predicted across all cells.

The regulon activity quantification strategy implemented in SCENIC consist of the following three steps:

1. **Co-expression modules identification:** starting from a scRNA-seq data count matrix, SCENIC uses GRNBoost2 (Moerman et al. 2019) to infer co-expression modules by performing a random forest nonlinear regression for every possible TF target gene.
2. **Regulon construction:** all the non-*cis*-regulatory connections are pruned from the coexpression modules using cisTarget, which finds enriched TF motifs near every putative target gene by scoring *cis*-regulatory modules. This effectively produces a regulatory module consisting of one TF and all its *cis*-regulated target genes (Janky et al. 2014; Herrmann et al. 2012; Imrichová et al. 2015).
3. **Regulon activity quantification:** every regulon is quantified in every cell by AUCell, which estimates an enrichment score of the genes that constitute the regulatory module.

We hypothesized that decomposing regulon activity matrices with ButchR will provide the possibility of recovering signatures that recapitulate regulation in single cells, e.g., to distinguish different patterns of regulation across cells sharing similar states or fates.

Therefore, we predicted regulon activities using SCENIC for all the 56,265 cells included in the adult mouse single-cell gene expression atlas (Han et al. 2018), and a total of 304 active regulons were found across the cells. Then, we decomposed regulon-guided signatures from the regulon activity quantification using ButchR. The optimal number of signatures was $k = 11$ and 10 of them showed a high correspondence to only one of the original tissues. This showed the specificity of the signatures found from the decomposition of regulon activity in single cells (**Figure 9.2**).

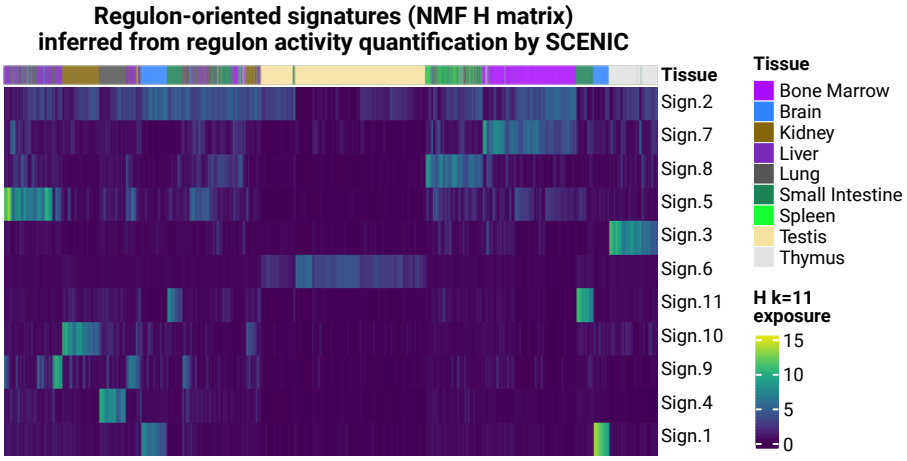


Figure 9.2: NMF matrix H of pySCENIC regulon activity scores. Signatures were learned using the regulon activity quantification by SCENIC. The optimal factorization rank was $k = 11$.

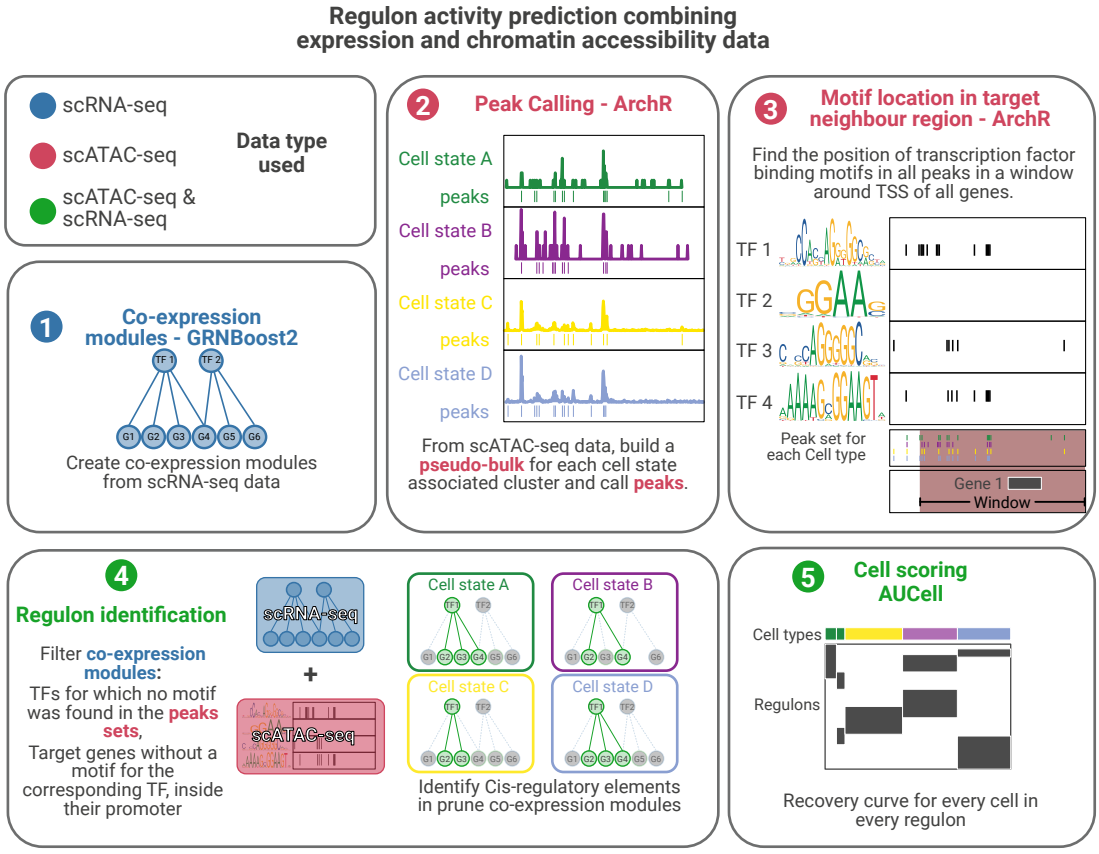


Figure 9.3: Schematic representation of regulon activity prediction using scRNA-seq and scATAC-seq data. Strategy to predict regulon activity by inferring cell state-specific regulons. **(1)** Identification of co-expression modules using GRNBoost2. **(2)** Identification of peaks for every cluster of cells from scATAC-seq that are related to a single cell state **(3)** Identification of the motif position for every TF in every set of peaks in a window around the transcription start site (TSS) of all genes. **(4)** Construction of cell state-specific regulons by pruning the coexpression modules found with GRNBoost2. **(5)** Quantification of regulon activity using AUCell. Steps that involve only scATAC-seq data are colored in red, steps that involve only scRNA-seq data are seen in blue, and steps that integrate scRNA-seq and scATAC-seq data in green.

9.2 Cell state-specific regulon activity quantification integrating scRNA-seq and sc-ATAC-seq data

Despite the accuracy of the signatures found from the SCENIC regulon activity quantification, we hypothesized that the pruning of coexpression modules step in SCENIC (step 2) is not the closest representation of the heterogeneous regulon composition seen across different tissues (i.e., the same TF can be acting on a different set of target genes depending on the tissue). In the pruning step, SCENIC finds one set of regulons for all the cells under interrogation, depending on a whole-genome ranking database of the motifs that are linked to known TFs (https://resources.aertslab.org/cistarget/). Therefore, these databases are built for the whole organism without the option of including associated chromatin accessibility information which could show variation across cells with different cell states. Taking this into account, we propose that defining regulons reflecting the different cell states found in a heterogeneous collection of single cells, will be a closer representation of the underlying GRNs.

We formulated the following workflow to quantify regulon activity by using cell state-specific regulons (cssRegulons) (**Figure 9.3**):

1. **Coexpression modules identification:** starting from a scRNA-seq data count matrix, infer co-expression modules using GRNBoost2 (same as SCENIC step 1).
2. **Position of TF-associated motifs identification:**
    - Starting from a contextually similar scATAC-seq data (i.e., the scRNA-seq and scATAC-seq dataset should be from similar tissues or experimental conditions), the cluster identity in these cells is defined from the scRNA-seq data, using a cell state annotation (e.g., cell type, tissue, or cluster group). This is done by aligning both datasets, using the ArchR package (Granja et al. 2021) implementation of the *FindTransferAnchors* function from the Seurat package (Butler et al. 2018).
    - Call peaks using MACS2 for every cluster of cells in the scATAC-seq data

(every cluster corresponds to one of the cell state definitions in the scRNAseq data).

- Annotate the peak sets by finding the location of TF-associated motifs using the *addMotifAnnotations* function from the ArchR package.

3. **Cell state-specific regulon construction:** every list of motifs locations (every list is associated with one cell state) is used to identify *cis*-regulatory connections in the coexpression modules. Resulting in a collection of regulons (cssRegulons) for every cell state defined in the scRNA-seq data (i.e., for a given TF the resulting regulon composition can differ across cell states).
4. **Cell state-specific regulon activity quantification:** every cssRegulon activity is quantified in every cell by AUCell.

After predicting cssRegulon activities for all 56,265 cells included in the mouse gene expression atlas by using the proposed workflow, we found 669 active regulons across the cells. We decomposed cssRegulon-guided signatures using ButchR. Similar to what was found for the decomposition using the regulon activity quantification by SCENIC, the optimal number of signatures was $k = 11$, which also showed a high correspondence to only one of the original tissues (**Figure 9.4**). This proved that the cssRegulon-guided signatures are also able to show specificity while capturing information for more than twice the amount regulons captured by SCENIC at the same time (304 in SENIC and 669 from cssRegulons).

9.3 Validation of cssRegulon-guided signature specific regulons

In order to validate that the larger number of cssRegulons was not due to capturing false positives, we extracted those cssRegulons that showed high specificity towards only one signature associated with a tissue, we estimated then the mean expression of the regulon TF across cells of every tissue. For instance, we found 25 cssRegulons highly associated

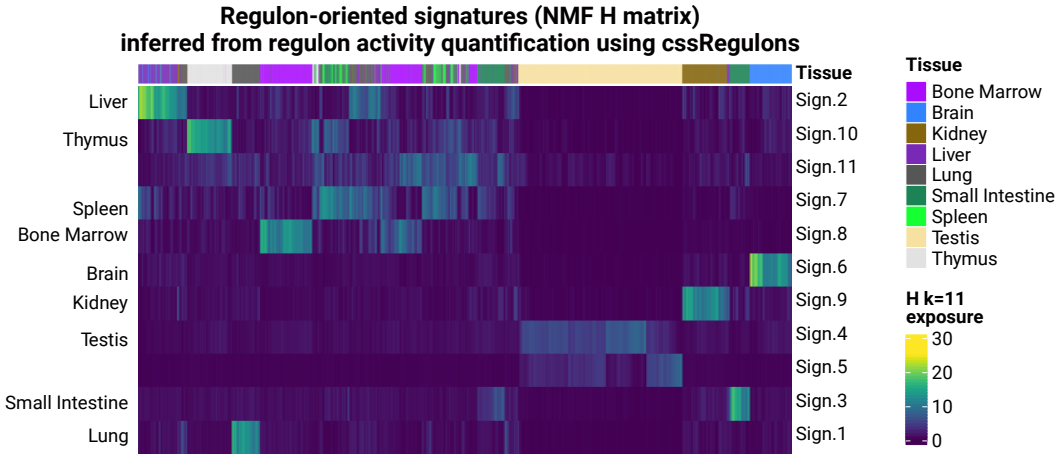


Figure 9.4: NMF matrix H of regulon activity scores found from cssRegulons. Signatures were learned using the cell state-specific regulons activity quantification scores. The optimal factorization rank was $k = 11$.

with the liver signature and the expression of all these regulon TFs was higher in the liver cells than in other tissues (**Figure 9.5a**). This confirmed that the cssRegulons are in fact capturing tissue-specific regulatory patterns.

We also studied the expression of the cssRegulons TFs that were not captured by SCENIC. Interestingly, the expression of those TFs was restricted to a small fraction of cells for the associated tissue. To exemplify, 8 out of the 25 cssRegulons highly associated with the liver were not captured by SCENIC, and the expression of these regulon TFs was restricted to a maximum of 4% of the liver cells (**Figure 9.5b**). Remarkably, after comparing the expression of these TFs across all the cell types found in liver, we found them to be highly expressed only in a few cell types (**Figure 9.6**). Thus, this strongly suggests that the cssRegulons are able to find regulatory modules that are active in smaller cell sub-populations.

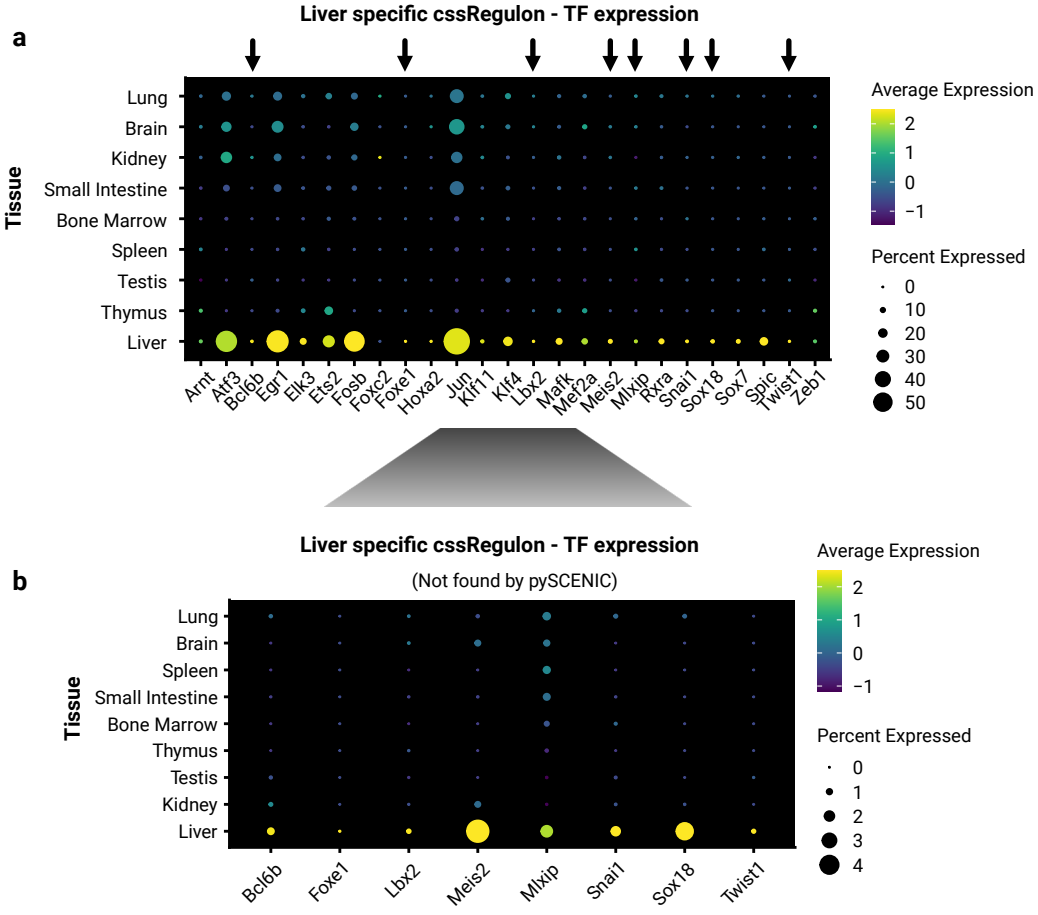


Figure 9.5: Expression of liver-specific TFs found using cssRegulons **(a)** Mean expression of TFs that were highly associated with the liver NMF signature across tissues, arrows indicate TF identified with the cssRegulon approach but not by SCENIC. **(b)** Mean expression of highly associated TFs with the liver NMF signature identified with the cssREgulon approach but not by SCENIC.

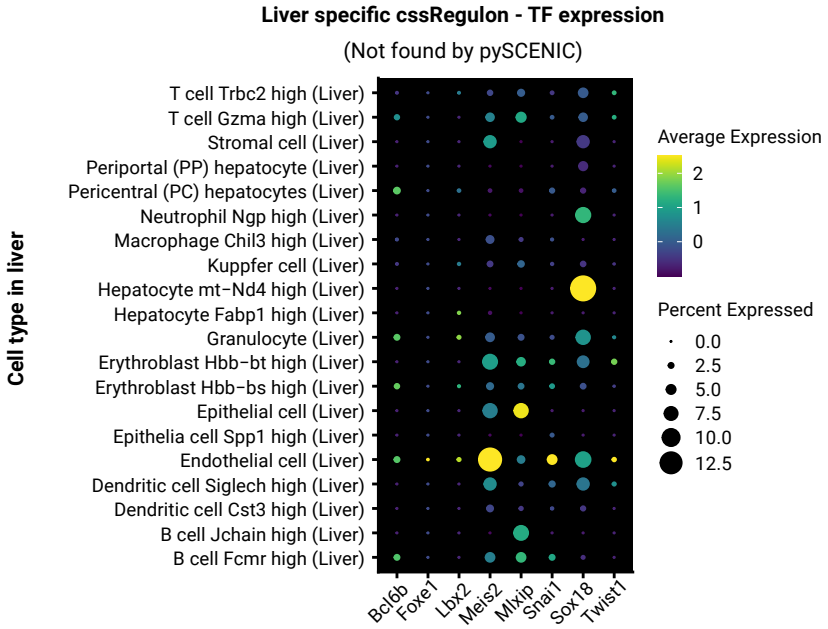


Figure 9.6: Expression of liver-specific TFs uniquely with cssRegulons. Mean expression of TFs across cell types found in mouse liver. Only TFs highly associated with the liver NMF signature, and identified with the cssRegulon approach but not by SCENIC are shown.

9.4 Comparison of regulon composition

ButchR is able to classify each regulon as differentially active for every signature, depending on the regulon contribution to every signature. As we were able to identify one signature associated with every tissue from the mouse gene expression atlas, we explored how the cssRegulons composition could vary across tissues. Therefore, we computed the Jaccard similarity of all cssRegulons that were active in 6 or more of the evaluated tissues (**Figure 9.7a**), revealing that cssRegulons controlled by the same TF are more similar to each other than to cssRegulons from the same tissue. Nevertheless, there is also variability in cssRegulons controlled by the same TF, proving that using the proposed approach is actually capturing regulatory differences across multiple cell states.

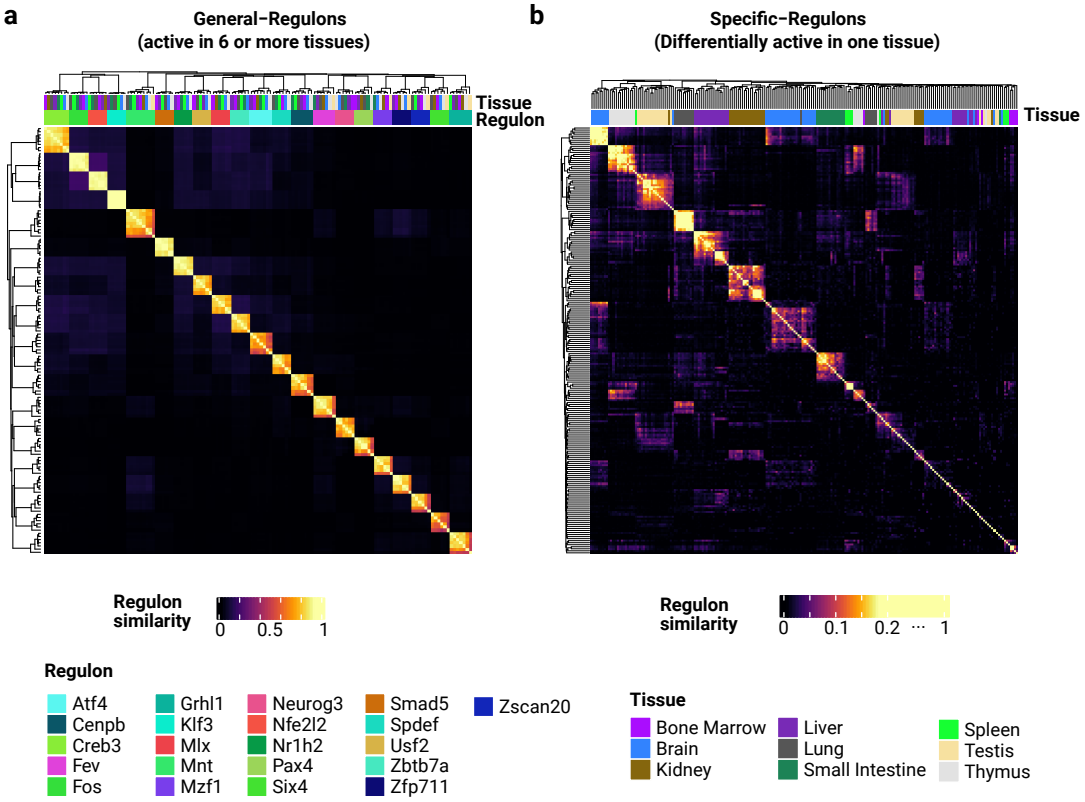


Figure 9.7: Jaccard similarity of cssRegulons. **(a)** Similarity of all cssRegulons that were active in 6 or more of the evaluated tissues. Regulon ID is indicated with the name of the associated TF. The tissue where the regulon is active is shown. **(b)** Similarity of cssRegulons differentially active in only one tissue. The tissue where the regulon is active is shown.

We further tested how similar are cssRegulons differentially active in only one tissue (specific regulons). Therefore, in this case, the comparison included only cssRegulons that were associated with different TFs (i.e., the list of TFs associated with the regulon does not contain duplicates). We expected that specific regulons from the same tissue were controlling a similar group of target genes. While this was true for many regulons, we also saw that there were subclusters of regulons more similar between tissue(s) than within them, reflecting common regulatory pathways across tissues controlled by different TFs (**Figure 9.7b**).

Taken together, the construction of cssRegulons are indeed reflecting the different regulatory links that can be found across different cell types. Moreover, the cssRegulon-guided signatures decomposed by ButchR allow the identification of regulatory patterns that are shared across multiple cell states or that are highly associated with only one of the cell states under consideration.

9.5 Chapter summary

Gene regulatory networks are composed of multiple layers of regulation, a general abstraction of these networks is to group modules of one TF and all its target genes, into a functional unit called regulon. The identification of regulons and quantification of their activity are key components to understand the regulatory differences that translate into defining the fate and state of a cell.

Here, we established a new approach to infer regulon-guided signatures by quantifying and decomposing the activity levels of cell state-specific regulons. This approach uses data from contextually similar scRNA-seq and scATAC-seq datasets. We start by (i) identifying co-expression modules, followed by (ii) identifying the position of TF-associated motifs, (iii) the construction of cell state-specific regulons, (iv) the quantification of regulon activity, and (v) decomposition of the activity levels using ButchR.

We found that this strategy is able to identify patterns of regulation that are related to

a broad cluster of cells and also to regulons that are active in only a selected number of cellular subtypes. Furthermore, we were able to classify regulons into active and inactive for every learned signature by decomposing the regulon activity scores with ButchR. Taking together, these findings suggest that cell state-specific regulons are an accurate reflection of the complex mechanism of regulation found in GRNs.

Part IV. Data Accessibility and Reproducibility

So far, we have shown how ButchR proved to be a reliable tool as we use it in different settings and found its utility to understand complex biological processes. We also proposed many extensions and workflows that can be carried out using ButchR, some of them are an integral part of the package (e.g., i2NMF and projection of transcriptomic data onto a reference atlas) and others have to be executed before using the package to decompose meaningful signatures (e.g., prediction of regulatory relationships and quantification of regulon activity). Nevertheless, being open science one of the backbones of this thesis, all the analyses shown so far have been fully documented and are open to anyone.

In this final part (**Part IV**), we show how we fully committed to open and reproducible research, going from creating pipelines dedicated to reproducing complete research projects (chapter: "About reproducibility"), to create interactive visualization tools that bring our results to everyone in the community and also encourage collaboration and sharing resources (chapter: "About data sharing and visualization").

Chapter 10

About reproducibility

As data analyses become more complex, involving multiple datasets and combining results from dozens of tools, the need for transparent and fully reproducible research is of pressing need. We committed to making fully available and reproducible the complete analysis derived from this study. Thus, comprehensive code repositories and pipelines have been released for all the main parts of this project.

10.1 ButchR and ShinyButchR

Starting with the development of ButchR was a complex, task, involving the integration of different programming languages and testing diverse types of data. Continuous integration (CI) is a software development practice to merge and test small code changes in a frequent manner, aiming to create healthier software. We integrated CI into ButchR using the hosted service *Travis CI* (`https://travis-ci.org/github/hdsu-bioquant/ButchR`) in order to build and test the source code of ButchR hosted on GitHub. The potential of CI is tightly linked to the developers' efforts to create unit tests for the most essential parts of the code. For ButchR, we created a complete array of unit tests with the R package *testthat* (Wickham 2011), including tests for all the functions and

visualization tools included in the package.

Keeping track of the coverage of the unit tests over the complete code is not an easy task. To this end, we also integrated the code coverage tool *Codecov* into ButchR (https://app.codecov.io/gh/hdsu-bioquant/ButchR). So far, every time that a new change is made into ButchR, 97% of the code is tested in order to know everything is working properly. All of this helped to make ButchR a reliable package and made public the complete development process. Therefore, any user that wants to recover the test and code coverage reports for the last build of the package can find the links in the *Travis CI* and *Codecov* badges in the ButchR GitHub repository (https://github.com/wurst-theke/ButchR).

Additionally, as was mentioned before in chapters 3 and 4, we created Docker images for ButchR and ShinyButchR, which will help anyone who wants to reproduce the analyses done with these tools by allowing its use without installing any software dependencies (besides Docker). Besides, any of the matrix decompositions shown in this work can be repeated using the live version of ShinyButchR (https://hdsu-bioquant.shinyapps.io/shinyButchR/).

10.2 Regulatory subtypes in neuroblastoma pipeline

Part of the collaborative effort published in Gartlgruber et al. (2021), involved compiling and analyzing multiple sources and types of data. Keeping track of the analyses done in a project of such extension can become a daunting task without the correct management tools. Thus, in order to structure and organize the different analysis steps, a comprehensive Snakemake-based (Köster and Rahmann 2012) has been made available at https://github.com/hdsu-bioquant/project_NB_SE. This pipeline can be used to reproduce all the results reported in the chapter "Neuroblastoma regulatory subtypes defined by super-enhancers" and part of the results in the chapter "Projection of transcriptomic neuroblastoma data onto a single-cell reference atlas."

10.3 Understanding regulatory heterogeneity with scCAT-seq pipeline

Coordination of international collaborations such as the work presented by us in Liu et al. (2019), requires extensive documentation of all the analysis steps. This is extremely important, not only for other researchers interested in the work but also to complete a successful collaboration. We created a Snakemake pipeline fully documented that can use any human data that were sequenced using multi-omics techniques such as scCAT-seq and SHARE-seq to predict regulatory relationships. This pipeline is hosted in GitHub (`https://github.com/hdsu-bioquant/scCAT`), and can also be used to reproduce the prediction of regulatory signatures as explained in the chapter "Understanding gene expression regulation with scCAT-seq."

10.4 Chapter summary

With the increasing complexity of data analysis in all the fields of life sciences, providing the community with the proper tools to reproduce and validate our findings is of the utmost importance. Here, we describe the approaches we took into account to ensure that all of our analyses were fully reproducible. For instance, ButchR and ShinyButchR were implemented with continuous integration, ensuring their robustness, and also the provided Docker repositories can be used to reproduce the matrix decompositions shown through all this work.

Regarding the reproducibility of the analyses shown in the chapter: "Neuroblastoma regulatory subtypes defined by super-enhancers" and chapter: "Understanding gene expression regulation with scCAT-seq," we have created two Snakemake pipelines to recreate any step of the analysis workflow.

Chapter 11

About data sharing and visualization

Although there is growing pressure coming from other researchers and journals to share data and the files that support a publication, there are still limited initiatives to make these data more friendly and easy to use for other researchers that were not involved in the original design of the experimental setting. Websites like https://descartes.brotmanbaty.org/ from The Brotman Baty Institute, https://www.internationalgenome.org/ from the 1000 Genomes Project (Birney and Soranzo 2015), https://www.gtexportal.org/ from The GTEx Project or https://www.humancellatlas.org/ from the Human Cell Atlas (Rozenblatt-Rosen et al. 2017) are aiming to breach this gap, compiling a wide arrange of resources in an easy to understand way. In the spirit of creating such kinds of resources, we created interactive applications to support our findings.

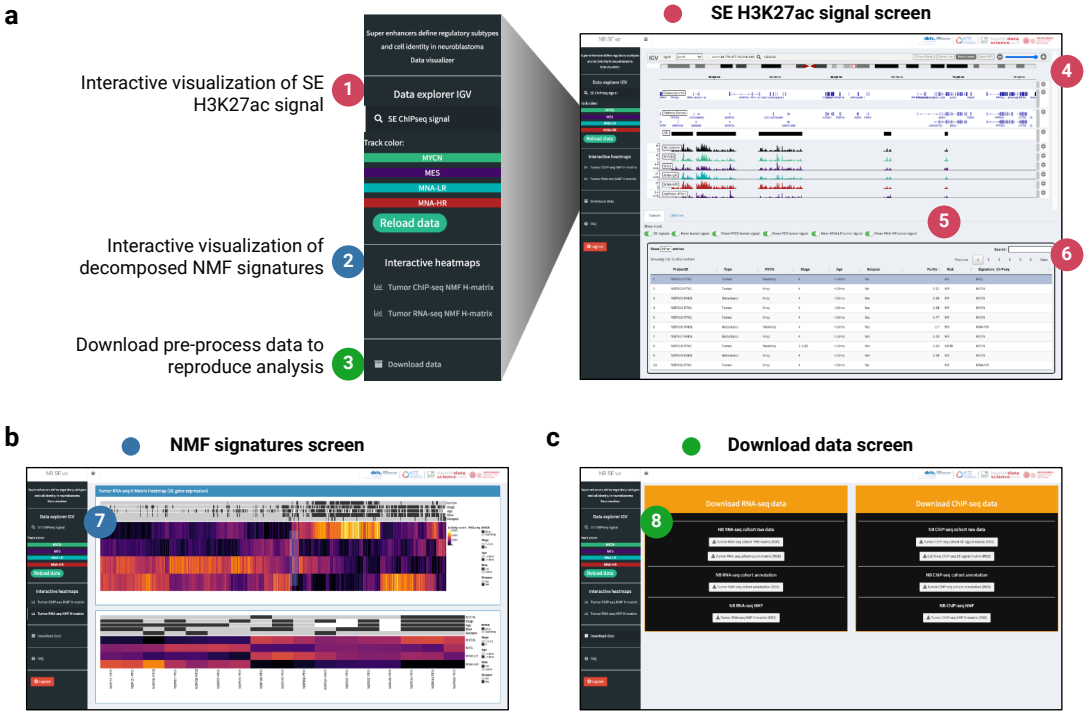


Figure 11.1: The NB-SE-viz Shiny app is composed of three interactive screens with the main menu providing access buttons to them **(1-3)**. **(a)** The SE H3K27ac signal screen allows the visualization of the H3K27ac signal of the SE regions **(4)**, providing controls to show the mean tumor signal **(5)** or the signal for every sample **(6)**. **(b)** The NMF signatures screen contains an interactive heatmap of the H_{SE} and H_{SE-Exp} matrices. **(c)** The Download data screen provides access to all the pre-processed data used in the main analysis.

11.1 Interactive visualization of neuroblastoma super-enhancers data

We created a visualization tool to explore data from neuroblastoma epigenetic subtypes (NB-SE-viz) (**Figure 11.1**) by compiling the results published in Gartlgruber et al. (2021). NB-SE-viz is a Shiny app that allows the exploration of the epigenomic tracks in

an interactive genome viewer, provides an interactive visualization of the neuroblastoma NMF signatures, and includes all the processed data for downloading. NB-SE-viz is available at https://nbseB087.dkfz.de.

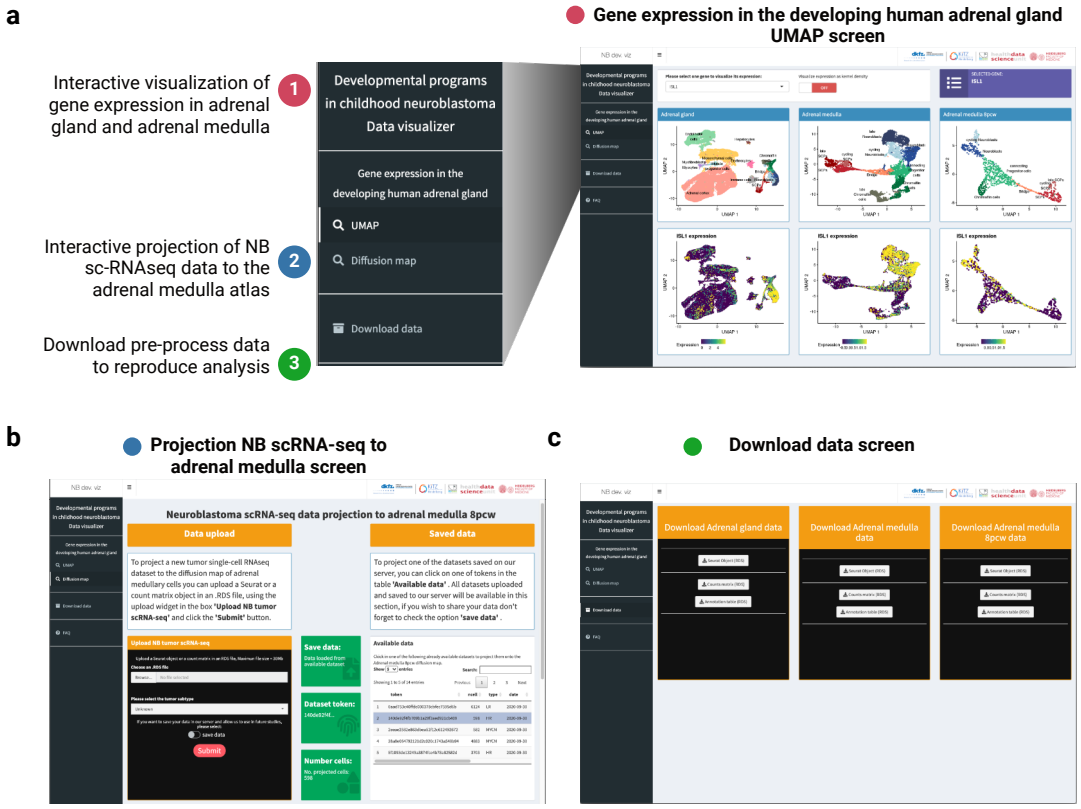


Figure 11.2: The NB-dev-viz Shiny app is composed of three interactive screens with the main menu providing access buttons to them **(1-3)**. **(a)** The "*Gene expression in the developing human adrenal gland*" screen allows to visualize all the cells included in the atlas in a UMAP embedding and colored them according to its relative expression for a selected gene. **(b)** The "*Projection of NB scRNA-seq to adrenal medulla*" screen contains an interactive tool to project single-cell transcriptomic data from neuroblastomas onto the atlas. **(c)** The "*Download data*" screen provides access to all the pre-processed data of the atlas.

11.2 Developmental programs in childhood neuroblastoma data visualizer

To support the first single-cell developing human adrenal gland transcriptomic atlas in the world *(Jansky, et al. 2021. Developmental programs in childhood neuroblastoma. Nature Genetics)*, we developed the interactive Shiny app NB-dev-viz (**Figure 11.2**). Besides of providing a portal to visualize the gene expression in cells from the adrenal gland, we also included tools to share and create a global database of single-cell transcriptomic data from neuroblastomas. The app is available at https://adrenal.kitz-heidelberg.de/developmental_programs_NB_viz/.

The main screen in the app allows to interactively select one gene and visualize its expression in a UMAP embedding of all the cells included in the atlas (**Figure 11.3**). This window shows all the different cell types found for the adrenal gland, adrenal medulla, and adrenal medulla at the 8th post-conception week, which makes it easier to find if a selected gene is only expressed at a certain cell state.

Aiming to build a larger collaborative community of researchers fighting against neuroblastoma, we added a second screen that contains tools to understand the tumor composition of neuroblastomas for which scRNA-seq data is available, and also a database of neuroblastoma tumor for which scRNA-seq data is available. After loading a new dataset, the app project the neuroblastoma single cells onto a diffusion map of the adrenal medulla cells. Any group of researchers that want to share their own data can upload it into the app and provide a contact email.

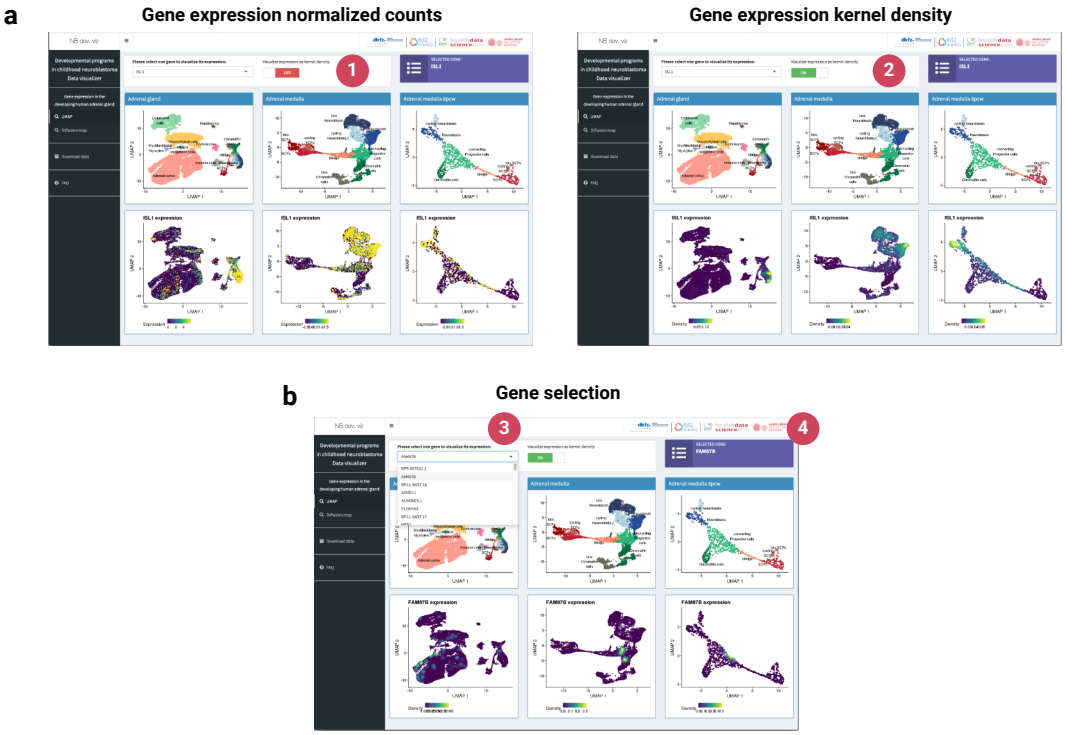


Figure 11.3: Interactive visualization of gene expression in the human adrenal gland. The "*Gene expression in the developing human adrenal gland*" screen contains different elements to enhance and change the visualization. **(a)** The default option is to show the expression of a gene as the normalized counts across all the cells **(1)**. The option "*Visualize expression as kernel density*" can be turned on to show the expression as a kernel density computed for every gene, allowing to found genes expressed in smaller groups of cells **(2)**. **(b)** The controller shown in **(3)** provides a list of all the genes included in the atlas and to select one to show its expression, the currently selected gene is highlighted in the notification window **(4)**.

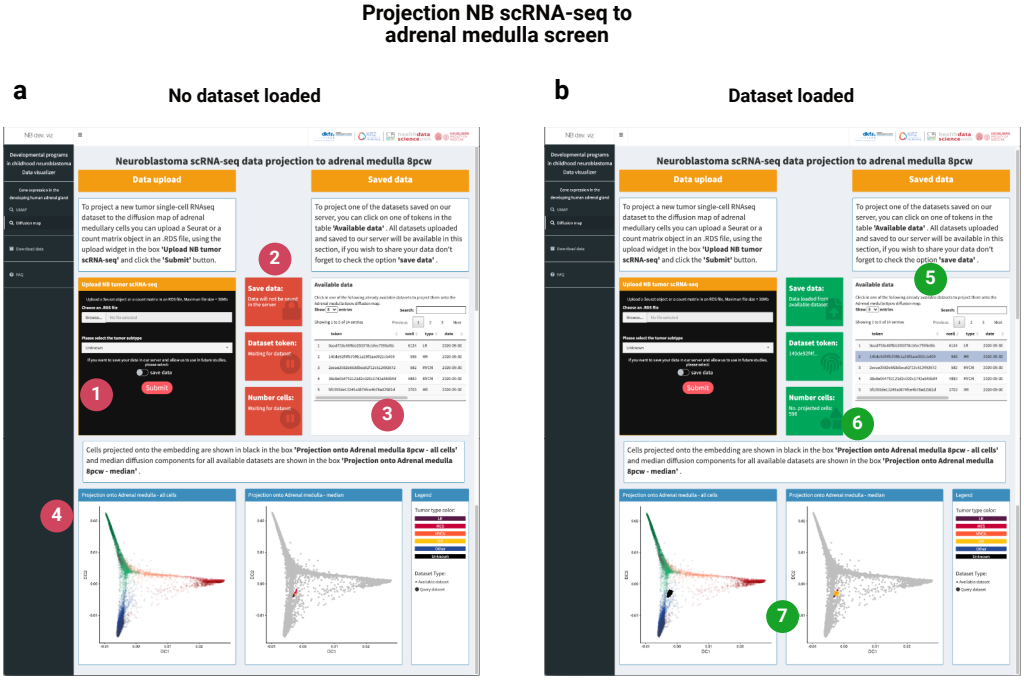


Figure 11.4: Projection of NB scRNA-seq data onto a human adrenal medulla atlas. The "*Projection of NB scRNA-seq to adrenal medulla*" screen provides a complete engine to understand the tumor composition of neuroblastomas for which scRNA-seq data is available. **(a)** Without any data loaded, the app shows a controller to upload a new dataset not included in the app **(1)**, a notification window displays that no data is loaded **(2)**, a list of datasets that are available in the app database **(3)**, and a diffusion map of adrenal medullary cells **(4)**. **(b)** A dataset included in the app can be loaded from the list in **(5)**; after loading a dataset, the notification window **(6)** shows how many cells are included in the dataset and also a token assigned to uniquely identify this dataset. The cells from a loaded dataset are projected onto the diffusion maps **(7)**.

11.3 MapMyCorona

The current SARS-CoV-2 pandemic has shown an impressive speed of spreading throughout the world. One of the main challenges is to understand how the viral sequence is evolving day by day. There are now multiple strains that are active at the same time, and keeping track of the population dynamics is not an easy task.

In order to address these challenges, we joined in the world effort to fight COVID-19 by developing a Shiny app (MapMyCorona) to display in an easy way, the sequence similarity and alterations between a query sequence and a central database of viral SARS-CoV-2 sequences on a world map.

MapMyCorona is publicly hosted at https://hdsu-bioquant.shinyapps.io/mapmycorona/.

MapMyCorona is an intuitive app, which allows easy exploration of the BLAST results with only a few preparation steps (**Figure 11.5**):

- **Step 1:** Upload a query sequence to the server. The sequence can be uploaded from a fasta file or directly pasting it into the provided text box.
- **Step 2:** Select if the query sequence is a nucleotide or protein sequence.
- **Step 3:** The "BLAST options" menu includes multiple parameters to fine-tune the BLAST search.
- **Step 4:** After uploading the query sequence, click on the "Submit" button to perform the BLAST search against the central database of viral SARS-CoV-2 sequences.
- **Step 5:** Once the BLAST search is finished, the app displays the hits in a world map, and also shows the number of hits by month.
- **Step 6:** The hits can be filtered to include only sequences from the desired date range.
- **Step 7:** Total number of hits against the central database.
- **Step 8:** Total number of hits after filtering.
- **Step 9:** Additionally, MapMyCorona also provides an interactive table to visualize

all the hits. The data can be downloaded as a csv file.

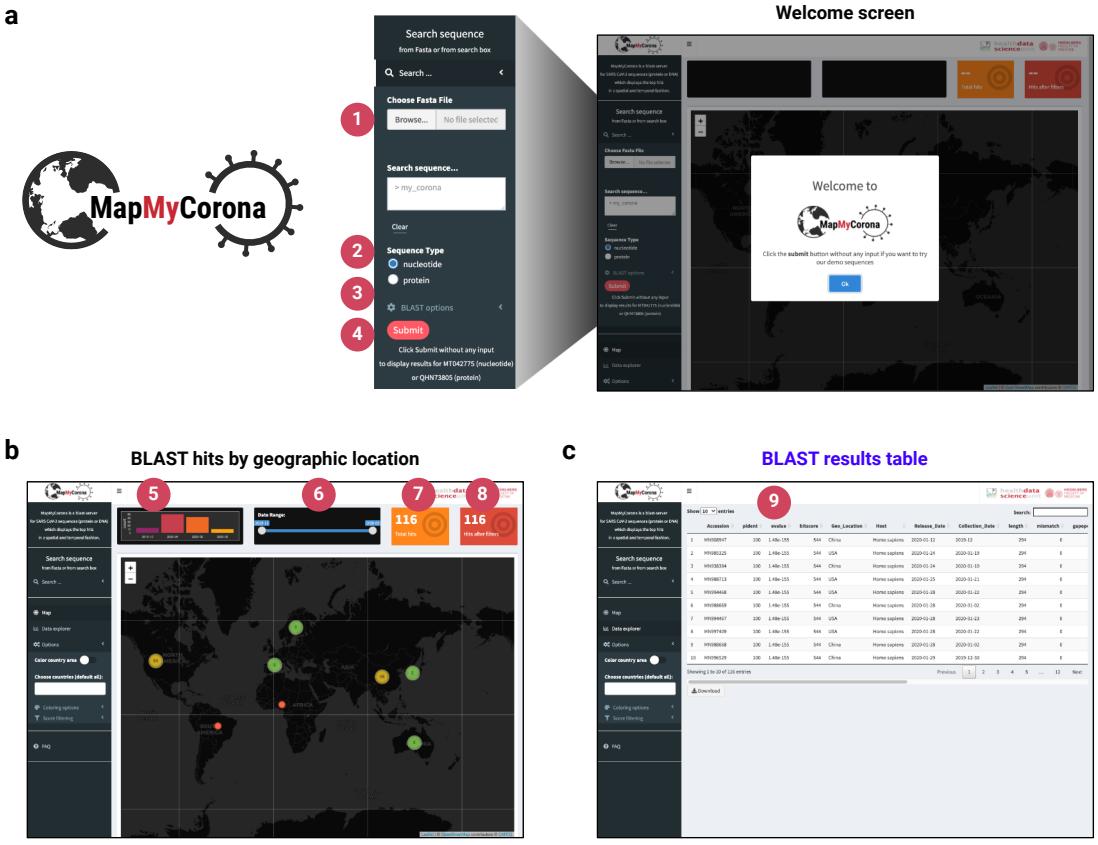
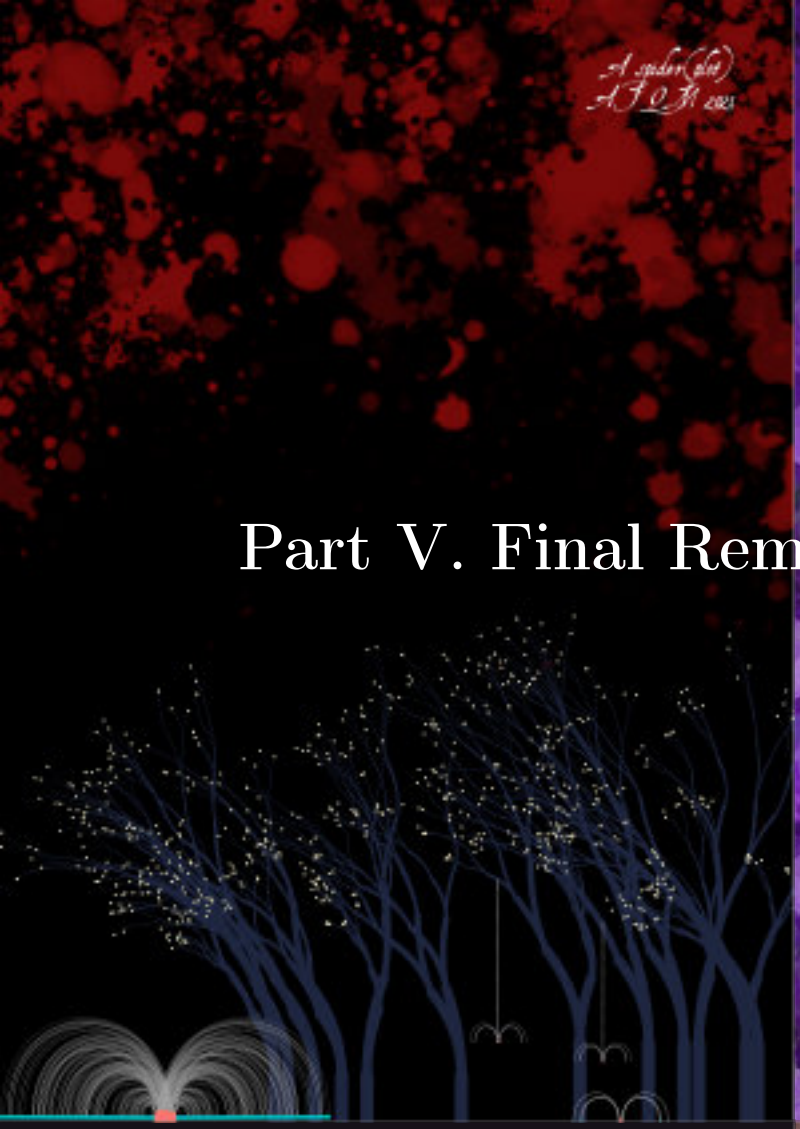


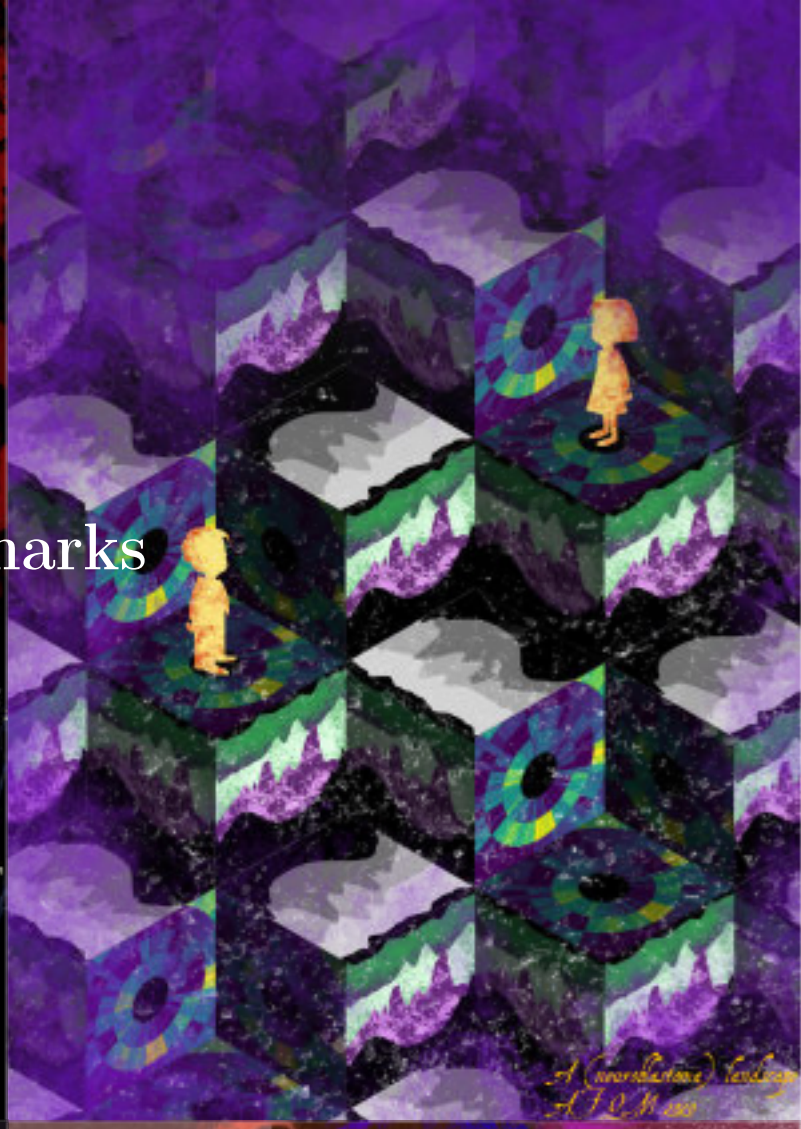
Figure 11.5: The MapMyCorona Shiny app is composed of three main screens, the **(a)** welcome, **(b)** BLAST hits by geographic location, and **(b)** BLAST results table screens. It allows to upload a query sequence **(1-2)**, perform a BLAST against a central database **(3-4)**, visualize and filter the results in a world map **(5-8)**, and download a table with all the resulting hits **(9)**.

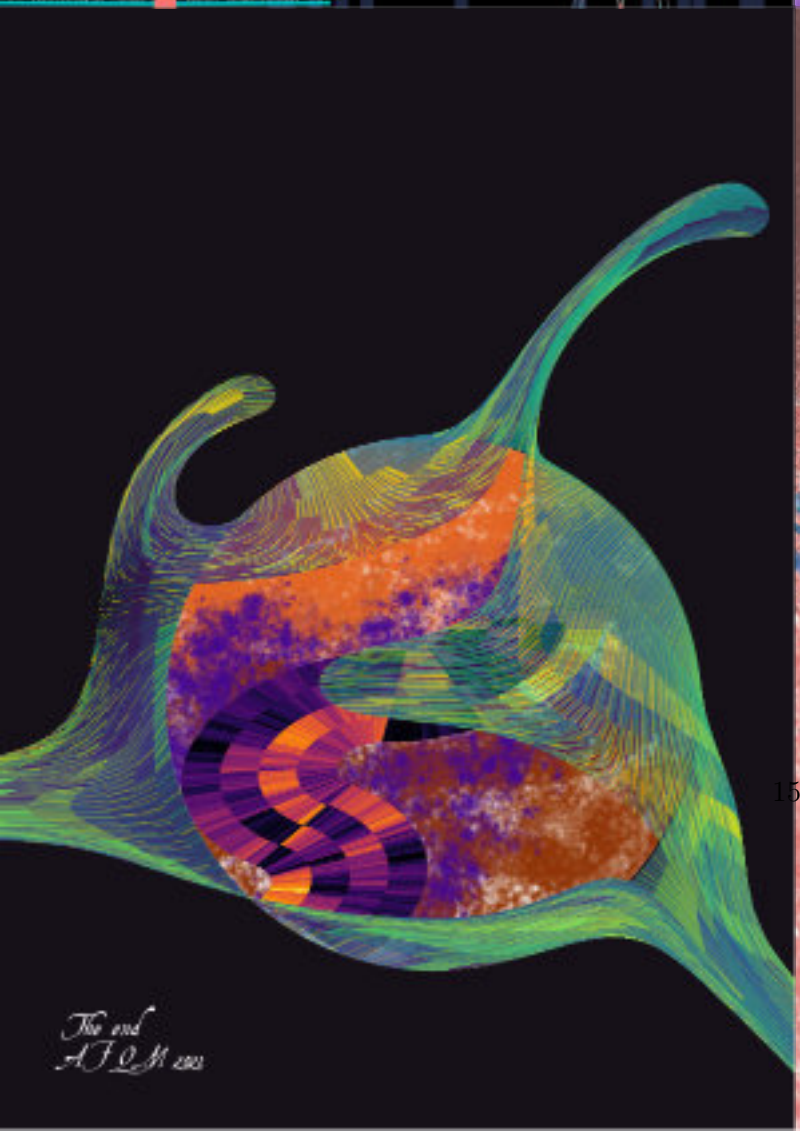
11.4 Chapter summary

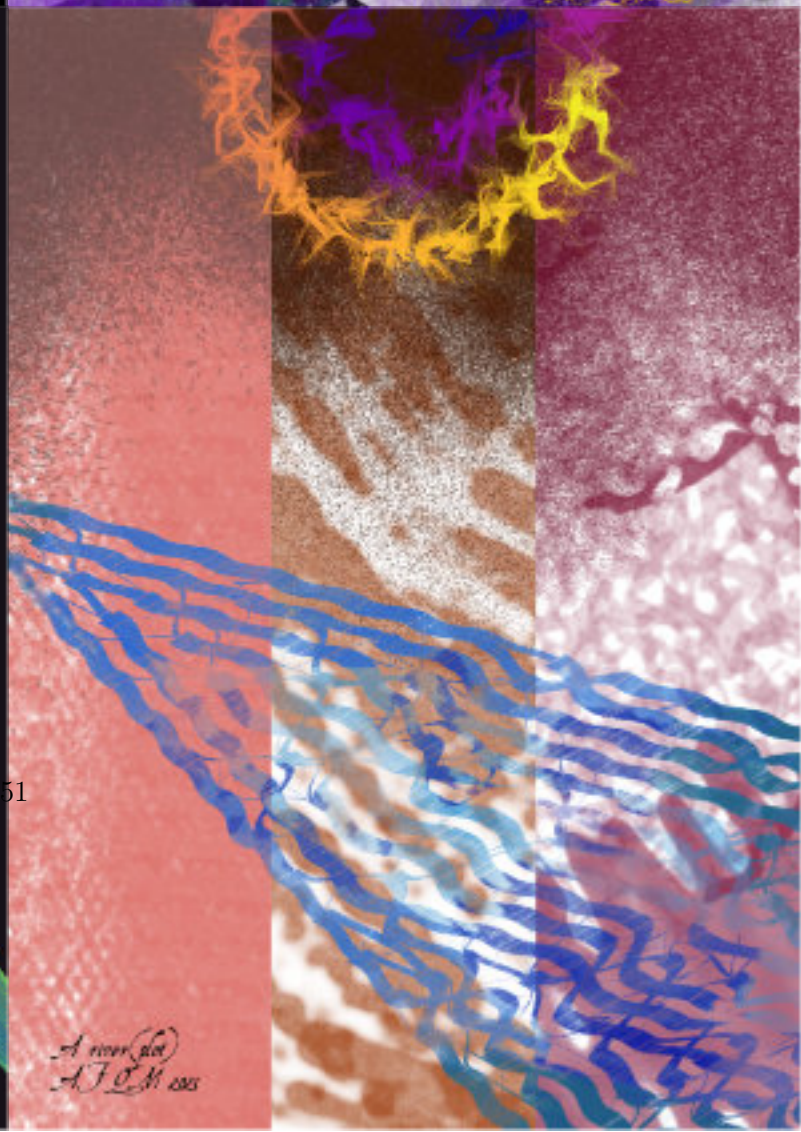
Encouraging toward transparent and reproducible scientific studies, it is nowadays a requirement from almost every journal to share and provide access to the data used to support the findings in a particular study. Going beyond of just sharing the raw data, we developed interactive applications to explore and understand our findings in a deeper way. We created an application to explore the regulatory subtypes in neuroblastoma (https://nbseB087.dkfz.de), another app to understand the developmental programs in neuroblastoma (https://adrenal.kitz-heidelberg.de/developmental_programs_NB_viz/); and also joining into the world effort to fight the current pandemic, we developed MapMyCorona, a tool to display sequence similarity and alterations of a given sequence on a world map (https://hdsu-bioquant.shinyapps.io/mapmycorona/).

Part V. Final Remarks









Chapter 12

Overall discussion and conclusion

This work was focused on developing a flexible and user-friendly toolkit to perform Non-negative Matrix Factorization (NMF) on any type of genomic data, and also on how it can be used to answer relevant questions as identification of tumor subtypes and deconvolution of regulatory signatures. NMF is an ever-increasing family of algorithms (M. W. Berry et al. 2007), that decomposes an initial matrix into two matrices of lower dimension W and H. Its main use has been in image analysis to recognize the main parts that constitute an image, as well as in recommendation systems (like movie or product recommendations) to infer scores on a sparse matrix (Benzi et al. 2016; Luo et al. 2014). In genomics and computational biology, NMF has been shown to work in many instances, e.g., to infer mutational signatures (Ludmil B. Alexandrov et al. 2020) and identify cellular subtypes (Shao and Höfer 2017). Despite this, obtaining and interpreting NMF results is not an easy task, because the current software packages and libraries with NMF implementations lack rich visualization tools that help the users to understand and select the optimal parameters for the matrix decomposition. Furthermore, NMF are highly demanding computational algorithms in time and resources. Therefore this work provides a robust NMF-based package called ButchR that implements fast NMF solvers while including diverse innovative visualizations and even an interactive tool. In addition, ButchR is able to use any type of genomics data to find the most important features that

define it.

12.1 Of ButchR and its development

During the development of ButchR, one of the most important decisions was to select R or Python to deploy our package. Currently, R is a widely used platform for the analysis of biological data, and besides its core central repository (Comprehensive R Archive Network CRAN) (R Core Team 2020), it is also supported by the Bioconductor (Huber et al. 2015), an open-source software project compiling a collection of packages and tools for the analysis of high-throughput data. On the other hand, Python has better integration of the open-source machine learning platform TensorFlow (Abadi et al. 2016) at its disposal, which was one of the most critical factors to deploy the NMF solvers based on matrix operations. Considering this, and also in the light of the most recent developments in the analysis of single-cell sequencing data, to which packages like Seurat (Butler et al. 2018) and ArchR (Granja et al. 2021) are becoming the standard for the analysis of such data, we decided to use R as the platform for ButchR. Nevertheless, to exploit the Python/TensorFlow framework, all the matrix decomposition algorithms implemented in ButchR use TensorFlow and connect to the R session using the package Reticulate (Allaire et al. 2017). Furthermore, R is also advantageous as it contains a vast amount of packages designed for the analysis of bulk high-throughput data, one resource that is currently absent in Python.

We also acknowledge the fact that growing Python-based initiatives like Scanpy (Wolf, Angerer, and Theis 2018) and EpiScanpy (Danese et al. 2019) have the potential to become the default analysis platform in the future due to the fast processing times and the popularity of Python among data scientist. Thus, although ButchR is an R package, its core is solely based on Python/TensorFlow, giving us the opportunity to deploy the complete ButchR suite as a Python package in the future.

In order to identify the biological processes captured by the NMF signatures, ButchR

includes multiple feature extraction and visualization functions. Starting from any non-negative matrix, ButchR is able to complete an entire NMF-based analysis returning rich visualizations. The package can be freely installed from GitHub (https://github.com/wurst-theke/ButchR) and we also made available a Docker image (https://hub.docker.com/r/hdsu/butchr) including the package, auxiliary libraries, and test datasets to allow the immediate usage of ButchR without installing any dependency.

12.2 NMF limitations

Despite all their advantages (e.g., enhanced interpretability of the data structure, decomposed signatures can be used for feature extraction, applicability to different types of data, among others), NMF algorithms have some intrinsic limitations. For instance, opposite to methods based on singular vector decomposition (Meyer 2000), there is not a global optimal solution for the NMF algorithms as they all converge to a local minimum (M. W. Berry et al. 2007). Therefore, there is no guarantee that one solution is the optimal factorization for a given matrix. Following one common practice, we overcame this problem in ButchR by comparing the local minima from multiple random initializations and returning the results of the best local minimum found. The number of random initializations n_{init} is one of the hyperparameters in the ButchR functions (e.g., for exploratory analyses $n_{\text{init}} = 5$ is sufficient to recover a good factorization). However, this is also related to another disadvantage of the NMF, which is the long computing time, even without accounting for multiple random initializations. To speed up all the NMF algorithms included in ButchR, we implemented the matrix decomposition steps in Tensorflow, which allows ButchR to parallelize every step and have better memory management, in addition to the possibility of using GPUs and TPUs to accelerate even further all the matrix operations.

Another hyperparameter in the NMF is the optimal factorization rank k (i.e., number of signatures, factors, or basis vectors). The majority of the current packages expect

the user to know beforehand this number (Welch et al. 2019; D. Song et al. 2021; Gaujoux and Seoighe 2010), but it might be difficult to determine if no prior information about the expected number of classes is known or if unknown classes are present. The R package NNLM (X. Lin and Boutros 2020) implemented a method to select an optimal k based on the mean square errors of the reconstruction of missing values from the input matrix; however, this approach relies on the identification of missing values, which is not a trivial task (e.g., identification of true zeroes and missing values). Therefore, we used a different approach with ButchR, where the user can select and perform the decomposition over a wide range of factorization ranks, and one of the implemented functions will automatically recommend an optimal k based on the minimization of the Frobenius error, the coefficient of variation and the mean Amari distance (Wu et al. 2016), and the maximization of the cophenetic correlation coefficient (Brunet et al. 2004). We also included options to manually select and optimal k, and the users are guided in their selection by a visualization of the factorization metrics.

Assessing the importance of one signature decomposed with NMF may be difficult if the metadata associated with the original data is not extensive enough. In such cases, it might not be possible to find a direct link between an increase in the exposure to the signature and the presence of certain annotation. This limiting factor in the interpretability of the NMF could hinder the conclusions driven from it by disregarding signatures with no clear metadata association as irrelevant. To address this problem, we included a riverplot visualization ("Appendix B: How to read a riverplot") that provides a visual guide of signature importance by showing how stable and pure it is even if different factorization ranks are selected. At its core, NMF is a parts-based representation of the data, which means that every learned signature represents one essential part of the data. Therefore, the riverplot visualization is a visual guide to recognize the main parts that constitute the original dataset.

12.3 Other packages with NMF implementations

As mentioned before, there are currently multiple R packages with NMF implementations. For instance, the package NMF was the first R package implementing NMF (Gaujoux and Seoighe 2010), LIGER implements iNMF to integrate multiple scRNA-seq datasets (Welch et al. 2019), scPNMF uses NMF to extract features from single-cell data (D. Song et al. 2021), and NNLM proposes an NMF algorithm using sequential coordinate-wise descent (X. Lin and Boutros 2020), among others. However, these packages are built to use only one algorithm, while ButchR is a toolkit meant to use a complete array of NMF algorithms that can be easily expanded in the future. Furthermore, many packages are now being designed to be used only with single-cell data in mind. A clear design decision from the start of ButchR development was to keep it as versatile and flexible as possible, not only aiming to provide an analysis platform for one type of data, but rather a universal NMF toolkit that can be easily adapted to all types of biological datasets.

Furthermore, none of the packages mentioned before take into account the assessment of signature stability across factorization ranks, which is one of the strengths of ButchR. As we have shown, the identification of relevant signatures can be guided using the visualizations generated with ButchR.

12.4 Why interactive applications

Effective data visualization is fundamental in any exploratory data analysis by guiding researchers into understanding the structure and patterns of any given dataset (Sudarikov, Tyakht, and Alexeev 2017; Moon et al. 2019; B. Lee et al. 2020). In particular, one of the most useful ways of representing data is by interactive visualizations. Although many computational tools are coming out every week, there are just a few examples that are using interactive visualizations to represent complex datasets

(Ovchinnikova and Anders 2020; Butler et al. 2018; Cao et al. 2019). Since the beginning of this project, we made multiple visualization tools trying to create an intuitive and easy-to-use tool, as well as to share resources and help in building a collaborative community. To this end, we developed ShinyButchR, an interactive Shiny application that uses ButchR to execute an NMF-based analysis from start to end. All the visualizations generated by ShinyButchR are fully interactive, allowing easy exploration of the matrix decomposition results. Furthermore, the results obtained from the app can be exported as plain text files to be used in any software, or as RDS files to perform more downstream analyses using R. ShinyButchR is publicly available and free to use at `https://hdsu-bioquant.shinyapps.io/shinyButchR/`, and as well as ButchR, a Docker image is available (`https://hub.docker.com/r/hdsu/shinybutchr`) which allows the execution of the app in any system. The description of ButchR and ShinyButchR was published in Quintero et al. (2020). Besides reaching broader audiences by lowering the technical proficiency levels to perform an NMF-based analysis, ShinyButchR is an initiative to inspire the whole scientific community into sharing resources and promoting open research.

Furthermore, ShinyButchR brings the opportunity to perform and reproduce all the analyses shown in this work (with a certain error rate as a consequence of the randomness in the initialization of the matrices H and W), using the processed data linked to each of the publications.

12.5 Using ButchR for signature identification

In the biological and clinical context, feature extraction and signature identification are two critical steps to understand diverse biological processes. In particular, a signature is defined as a group of features that are sufficient to identify a certain genotype or phenotype. For instance, genomic signatures are strings of DNA and RNA sequences used to determine the identity of a certain genotype (Fernandes and Zhang 2014; Slezak,

Hart, and Jaing 2019), expression signatures link a phenotype to a certain pattern of gene expression (Szymczak et al. 2021; Rahman et al. 2020; Sotiriou and Pusztai 2009). In the context of single-cell data, signatures are groups of genes that help to identify cell states. At its core, the main goal we had with the development of ButchR was to capture meaningful signatures in the form of a lower-dimensional representation of the original data. Following this goal, we used the publicly available data of labeled cell types from the human hematopoietic system (Corces et al. 2016) to show how signatures learned with ButchR were able to extract the biological differences and undergoing developmental processes seen in the hematopoietic system. This system, and in particular the dataset from Corces et al. (2016) has been thoroughly described, which made it a perfect proof of concept study to prove the utility of the package. The results of such proof of concept were positive, proving that ButchR was useful for answering complex biological questions.

Besides ButchR, other tools and workflows have been developed for signature identification and analysis. To exemplify:

- SigProfiler (Bergstrom et al. 2019) and SignatureAnalyzer (Haradhvala et al. 2018) are two tools that create mutational signatures from somatic mutations. Similar to ButchR, SigProfiler and SignatureAnalyzer also use NMF to decompose an input matrix (in this case a matrix of somatic mutations across multiple tumor samples) into signatures. SignatureAnalyzer implements a special Bayesian variant of the NMF described by (Tan and Févotte 2013). Both of these tools have been used by the pan-cancer analysis of whole genomes (PCAWG) consortium to recover 67 consensus signatures from whole-genome sequencing data of 4.645 and whole-exome sequencing data of 19,184 tumor samples (Ludmil B. Alexandrov et al. 2020). Since then, these signatures have been used in numerous studies (Campbell et al. 2020; Gerstung et al. 2020; Moore et al. 2020; Calabrese et al. 2020), showing the utility of signatures decomposed with NMF to explain the characteristic mutations of different types of cancer. SigProfiler and SignatureAnalyzer are highly customized

to recover mutational signatures, and in contrast with ButchR, it is not possible to use them to recover other types of signatures like gene expression or chromatin accessibility signatures.

- YAPSA (yet another package for signature analysis) (Hübschmann et al. 2020) is an R package originally designed for the analysis of mutational signatures. Although YAPSA is not able to find signatures *de novo*, it can find the exposure of any given sample to an existent set of signatures. As the framework in YAPSA is flexible to be also used in other types of signatures, in ButchR we leveraged YAPSA's functionality to find the exposure matrix H for a query dataset from a set of known signatures.

- Hydra (Pfeil et al. 2020) is a tool to identify tumor subtypes using multimodal gene expression signatures. This tool uses a Dirichlet process mixture model to find genes whose expression is a mixture of two or more Gaussian distributions, and then it clusters those genes using a multivariate mixture model, allowing the signatures to be identified by characterizing each cluster. As ButchR, this method is able to find signatures in cancer samples without including matched normal tissue samples. However, in cases where the samples are not forming separate clusters, but rather a continuum (e.g., in developmental processes), this method will not be able to recover signatures that explain such continuous processes. On the other hand, the NMF signatures learned with ButchR are always expressed as a range of exposures.

- Single-cell analysis toolkits like Seurat (Butler et al. 2018; Stuart et al. 2019) and Scanpy (Wolf, Angerer, and Theis 2018) find marker genes (i.e., signatures) for clusters of cells in order to identify cell types or cell states. These tools work by reducing the dimension of the original dataset using PCA, followed by clustering and identification of marker genes by comparing the expression of one cluster to the rest. One advantage of this is that it is not necessary to know the number of clusters beforehand, although this strategy will work better in a dataset with well-defined

clusters. In contrast, ButchR can also be used to find signatures associated with cell states, but it is not necessary to compare clusters of cells. Instead, signature associated genes (i.e., marker genes) can be extracted from the matrix W.

- scGeneFit (Dumitrascu et al. 2019) is a recently published method to find markers in single-cell clusters. This method uses an innovative strategy to incorporate a previously generated hierarchical partition of labels. Therefore, if previous information is known, this method can use the natural graph structure of cell states' similarities to guide the marker discovery (or build a hierarchy by clustering the cells). However, similarly to the strategies followed in Seurat and Scanpy, this method works better in the presence of distinct clusters.

In general, most current methods for signature identification are specially tailored to only one data type. The advantage of this is that all the subroutines and functions can be more specific, providing more pre-processing workflows and options, relevant only for the data type in question. On the other hand, more flexible toolkits like ButchR are easier to integrate at any point of an analysis workflow. For instance, we used ButchR to find regulatory and regulon-guided signatures, finding the most relevant TFs for a given cell state.

In the particular case of mutational signature identification, NMF has been extensively used. However, most of the methods rely on contrasting two conditions or one cluster against others for signatures based on gene expression. This makes it difficult to identify signatures of transitional stages, whereas NMF-based methods can produce signatures with a gradient of exposures.

12.6 Using ButchR to discover a new neuroblastoma subtype

We used ButchR to investigate the epigenomic landscape of Neuroblastoma, a neuroendocrine tumor derived from the neural crest. Two different cell identities (i.e., mesenchymal-type and adrenergic-type) have been described in neuroblastoma cell lines (Van Groningen et al. 2017; Boeva et al. 2017). However, the effect of cell identity on the progression and relapse of neuroblastoma tumors is still unknown. Therefore, our main motivation to study the epigenomic landscape in neuroblastoma was to find whether these two cell identities were also present in human tumors and cell lines and how these were regulated.

Starting from genome-wide profiles for the histone mark H3K27ac across 60 neuroblastomas, covering different clinical and molecular neuroblastoma subtypes, we were able to identify four super-enhancer-driven epigenetic signatures that were also recovered from three different bulk RNA-seq cohorts. In contrast to what was described by Van Groningen et al. (2017) and Boeva et al. (2017), our signatures were able to dissect the adrenergic-type into three regulatory subtypes, namely MYCN-amplified, MYCN non-amplified high-risk, and MYCN non-amplified low-risk. To support the definition of these three signatures, we found that they showed a clear association to known clinical outcomes. On the other hand, in line with the findings in neuroblastoma cell lines, our fourth signature defined a newly described subtype that is associated with cell migration and epithelial-mesenchymal transition (mesenchymal subtype).

In contrast to a classical definition of a tumor subtype (Galon et al. 2012; Tsang and Tse 2020; S. Ackermann et al. 2018), the signatures recovered with the NMF provide a measurement of the exposure of every tumor to a given subtype. This is one of the greatest advantages of using a soft-clustering method like NMF because the signatures do not necessarily represent on/off status. The concept of exposure to a signature helps to understand continuous processes as cell differentiation and in this particular case tumor

development, as interestingly several neuroblastoma tumors showed a high exposure score to multiple NMF signatures. Suggesting that the intratumor heterogeneity seen in some neuroblastomas (Pugh et al. 2013; S. Ackermann et al. 2018; Schramm et al. 2015) could be partially explained by the presence of tumor cells exhibiting characteristics of different epigenetic subtypes.

It has been shown that neuroblastoma tumors can be infiltrated by normal Schwann cell precursors (SCPs) (Ambros et al. 1996; Shimada et al. 1999). Thus, it can be argued that the clear association of the NMF mesenchymal signature to the SCPs is just a reflection of capturing the multipotent traits of normal SCPs (Jessen and Mirsky 2019) that infiltrated the tumors included in our cohort. However, neuroblastomas infiltrated by SCPs are found more often in cases with favorable outcomes than in cases with unfavorable outcomes. As the mesenchymal signature showed enrichment of relapse samples, indicates that the group of tumors with higher exposures to the mesenchymal signatures are actually related to unfavorable outcomes. This suggests that the identification of the mesenchymal signature is not related to the infiltration of SCPs.

To understand the possible cell of origin of the neuroblastoma mesenchymal subtype, we developed a new method for ButchR to project any bulk or single-cell transcriptomic data onto a reference single-cell atlas. This method consists of computing NMF signatures from the scRNA-seq data of the reference atlas, followed by finding the exposure of all query data samples/cells to the atlas signatures, this effectively bring the query samples/cells to the same space of the atlas. Using this method, we projected bulk RNA-seq data from a neuroblastoma cohort of 579 tumor samples onto an atlas of developing adrenal gland for mouse, and also onto an atlas of developing human adrenal gland. We found that the mesenchymal subtype shared similar regulatory landscapes with multipotent Schwann cell precursors. These findings were also validated by projecting scRNA-seq data for three neuroblastoma cell lines, supporting our results for the mesenchymal cell lines SK-N-AS and SK-N-SH. The description of the epigenomic subtypes in neuroblastoma was published in Gartlgruber et al. (2021).

12.7 Projection of transcriptomic data onto a single-cell reference atlas

One of the main research focus since the emergence of single-cell transcriptomics has been the development of methods that allow the projection or integration of new data onto an existent reference atlas. Besides our proposed workflow to compute the exposure of individual samples or cells to the NMF signatures learned from a single-cell reference atlas (effectively bringing them into the same reference space), other methods have been published that seek to accomplish similar goals. For instance:

- scmap (Kiselev, Yiu, and Hemberg 2018) finds clusters in the reference atlas and then projects new cells onto it by an exhaustive search of the similarity between every cell and the centroid of the clusters. One limitation of this method is that it can only map every cell to the cluster, but not to points in the hyperplane that lay between clusters. In contrast, the ButchR-based method uses linear combination decomposition (LCD) to estimate the matrix H_{query} (exposure of new samples to the atlas signatures) from the exposure values of the matrix W_{Atlas} (atlas signatures), which allows the identification of states that do not perfectly align with one of the clusters in the atlas.

- ProjecTILs (Andreatta et al. 2020) computes the PCA rotation matrix from a reference atlas, and then uses it to transform the gene expression of a query dataset into PCA loadings that will be in the same reference space with the atlas; however, this method requires a previous integration step to align both datasets using Seurat (Butler et al. 2018). On the other hand, our proposed workflow does not require the previous alignment step, the only requirement is to normalize the columns of the input matrix.

- scArches (Lotfollahi et al. 2020) uses a transfer learning approach based on conditional variational autoencoders or conditional generative adversarial networks. After training a model from the atlas, scArches uses the query data to fine-tune

the model and integrate both datasets into a new space. In comparison to scArches, our workflow does not change the space (signatures) of the reference atlas but instead finds the position of the new cells into it. In cases in which the main goal is to understand the position of new cells in the atlas distribution without changing it, we consider that our approach will be more adequate.

As it can be seen, there are several methods useful for projecting single-cell data onto a reference atlas. However, the biggest strength of our workflow is that it was conceived to project any type of transcriptomic data onto the single-cell reference. Therefore, we are not limited to use other single-cell datasets as queries, as the workflow can also work with bulk RNA-seq and microarray data. This flexibility was deciding to investigate the cell of origin of the neuroblastoma mesenchymal subtype.

12.8 Decomposition of regulatory signatures

The reconstruction of gene regulatory networks (GRNs) is one the most studied field nowadays (Chai et al. 2014; Fiers et al. 2018; Thompson, Regev, and Roy 2015; Moris, Pina, and Arias 2016). Deciphering the interactions between active transcription factors (TFs) and the *cis*-regulatory elements (CREs) of their target genes is one of the keys to explaining different cell states and differentiation processes (Moris, Pina, and Arias 2016; Spitz and Furlong 2012). Therefore, techniques like scCAT-seq (Liu et al. 2019), SNARE-seq (Chen, Lake, and Zhang 2019), Paired-seq (Zhu et al. 2019), and SHARE-seq (Ma et al. 2020) that co-profile expression and chromatin accessibility in individual cells are crucial to model GRNs at a single-cell resolution. We established a new workflow to use such data to infer regulatory relationships between genes and their CREs for individual cells, and coupled with ButchR to infer regulatory signatures. We validated the regulatory relationships predicted from data of three cell lines (K562, HeLa-S3, and HCT116) using publicly available ChIA-PET interaction profiles, finding that a large fraction of the predicted regulatory links were also present in the interaction profiles.

Furthermore, we extracted regulatory signatures from data of two lung patient-derived xenografts, these signatures were able to capture the intra- and inter-tumor variability between both tissues, resulting in a clear regulatory difference and groups of TFs that acted only in one group of the cells. This method was also applied to data from human pre-implantation embryos, finding one signature associated with cells from the morula stage, and another signature associated with blastocyst cells. The regulatory differences found in three cells from the blastocyst stage pointed us to identify these cells as part of the inner cell mass. This method was implemented for the study published in Liu et al. (2019).

Our approach (Liu et al. 2019) to model GRNs using data generated from techniques such as SNARE-seq, Paired-seq, and SHARE-seq is opposite from the current approaches. For instance, in the recent publication for SHARE-seq, Ma et al. (2020) showed a new method to determine groups of peaks regulating one gene (in what they called domains of regulatory chromatin), that is based on correlating peak signal and gene expression across all the cells, and standardizing using the mean and standard deviation estimated from a background model using chromVAR. Although we did not account for the correction of the background model, we found that using correlation-based models will recover fewer true regulatory interactions than our strategy to predict regulatory links for every single cell (**Figure 8.3**).

Due to the sparse nature of single-cell data where open regions and expressed genes are not detected by technical variation (i.e., dropouts), aiming to find all the true regulatory relationships present in a cell is not technically feasible with the current technologies. Accordingly, finding common regulatory patterns (i.e., regulatory signatures) among groups of cells using NMF helps to mitigate the effect of this technical variation (**Figure 8.4a**).

One interesting concept that we did not explore with scCAT-seq was the occurrence of cell states that are defined only from changes in the chromatin accessibility landscape, which define poised or primed cells that can undergo a differentiation process in the

future (Ma et al. 2020; Lara-Astiaso et al. 2014; Bernstein et al. 2006; Rada-Iglesias et al. 2011). From our definition of regulatory relationships, this can be explored in futures studies by comparing the exposure of chromatin accessibility signatures to regulatory signatures in individual cells.

Our findings showed that the creation of one new feature space that model GRNs at the single-cell level (i.e., predicted regulatory relationships for every cell) helped to find common patterns of regulation between cells, and understand the variation of regulation in different models (e.g., cell lines, tumors, and pre-implantation human embryos).

12.9 Combining scRNA-seq and scATAC-seq to define regulon-guided signatures

As most of the single-cell studies are generating scRNA-seq data nowadays, many tools have been developed to reconstruct GRNs using only such datasets (Holland et al. 2020; Fiers et al. 2018) However, more studies are starting to produce data to measure the epigenome of single-cells using techniques such as scATAC-seq (Buenrostro et al. 2015). SCENIC is a tool to quantify the activity modules formed by one TF and its target genes (Aibar et al. 2017; Van de Sande et al. 2020), such modules are called regulons, and they constitute the building blocks of GRNs. Despite using only scRNA-seq data to infer regulons, we showed that using ButchR to find regulon-guided signatures allows the recovery of regulatory differences at the tissue level in adult mice. On the other hand, SCENIC relies on a database of binding sites for TF motifs to construct the regulons, which does not take the variation in chromatin accessibility that can be seen across different cell states. Therefore, we created a new approach to model GRNs leveraging contextually similar scRNA-seq and scATAC-seq data (i.e., same conditions and same organism). With this approach, we were able to infer and quantify regulons that are related to specific cell states. This method consists of the identification of coexpression modules, the determination of the position of TF associated motifs, construction and

quantification of cell state-specific regulons, followed by recovering regulon-guided signatures using ButchR. We applied this workflow using publicly available data for nine tissues of adult mice and found that it can recover patterns of regulation shared across all the cells from the same tissue, but also it can recover TFs that are active in only a few cellular subtypes. Taking together, this finding showed that the usage of these two new workflows provided new insights into the regulation of gene expression and are a valid strategy to model gene regulatory networks. we created a new method to construct regulons leveraging contextually similar scRNA-seq and scATAC-seq data. With this approach, we were able to create regulons that are related to cell state.

A recent benchmark study (Holland et al. 2020) demonstrated how SCENIC consistently recovers less TFs than other tools such as DoRothEA and metaVIPER. By using the cell state-specific regulons, we were able to consistently recover more than twice the number of TFs that SCENIC does, showing how some TFs are only active in small cell populations. Supporting this, we found groups of TFs that were associated with only a few types of cells in a particular tissue, when analyzing the signatures recovered using the quantification of the state-specific regulons.

12.10 Limitations of ButchR

From the beginning of the project, ButchR was conceived to be used with any type of genomic data. However, this also comes with compromises, such as pre-processing steps (e.g., data normalization and filtering low-quality features) are expected to be done before using ButchR. In addition, special objects returned by different packages need to be converted to a regular matrix. Keeping this in mind, we have also included multiple vignettes to guide on how to do these steps.

Although ButchR has implementations for different NMF algorithms to decompose single or multiple matrices, other NMF algorithms can be more suitable to use under specific circumstances. However, the source code of ButchR is highly modular, and new NMF

solvers can be added easily in the future.

12.11 Final remarks

Finally, one of the central pillars of this thesis was to conduct open and reproducible research. Keeping this spirit in mind, all the analyses shown in this work have been made available in different GitHub repositories (linked in their respective publications), the source code of ButchR and ShinyButchR is also available. Furthermore, we also developed three interactive applications that help to understand our findings and allow other researchers to easily obtain the data used in this work. NB-SE-viz can be used to explore the regulatory subtypes in neuroblastoma (https://nbseB087.dkfz.de), NB-dev-viz is an explorer of the developmental programs in neuroblastoma (https://adrenal.kitz-heidelberg.de/developmental_programs_NB_viz/), and MapMyCorona is our contribution to the world effort to fight the current pandemic, which displays the sequence similarity and alterations of one sequence against a database on a world map (https://hdsu-bioquant.shinyapps.io/mapmycorona/).

The findings of this study allowed us to publish one original paper describing ButchR and ShinyButchR, two original papers describing new biological insights supported by signatures found using ButchR, and to present our results in oral and poster presentations.

We presented in this study ButchR, a new toolkit to infer signatures and extract relevant features associated with genotypes and phenotypes using NMF. We demonstrated how ButchR is useful for analyzing multiple types of data, and how its signatures are able to capture relevant biological information. The accompanying app ShinyButchR can be effectively used to perform a complete ButchR-based analysis in an interactive fashion. This toolkit is a new valuable resource to the scientific community, and it can be used to understand complex biological processes.

Appendix A: Data description

Through the development of this thesis, diverse datasets were used to perform the analyses shown in this work. This appendix compiles all the sources of the data used in every chapter.

Table S1: Datasets produced by other groups in collaborative projects

Dataset description	Data type	Chapter	Reference
H3K27ac NB cohort 60 tumors	bulk ChIP-seq	Neuroblastoma	Gartlgruber et al. (2021)
H3K27ac NB 25 cell lines	bulk ChIP-seq	Neuroblastoma	Gartlgruber et al. (2021)
NB cohort 579 tumor samples	bulk RNA-seq	Neuroblastoma	Gartlgruber et al. (2021)
SK-N-AS, CLB-GA, and KELLY	bulk ATAC-seq	Neuroblastoma	Gartlgruber et al. (2021)
Chromatin interaction SK-N-AS	bulk HiChIP	Neuroblastoma	Gartlgruber et al. (2021)
Chromatin interaction CLB-GA	bulk HiChIP	Neuroblastoma	Gartlgruber et al. (2021)
Human adrenal medulla 6,249 cells	scRNA-seq	Neuroblastoma	Jansky et al. (2021)
Two NB tumors 1,812+1,742 cells	scRNA-seq	Neuroblastoma	Jansky et al. (2021)
K562 74 cells	scCAT-seq	scCAT-seq	G. Li et al. (2010)
HeLa-S3 42 cells	scCAT-seq	scCAT-seq	G. Li et al. (2010)
HCT116 90 cells	scCAT-seq	scCAT-seq	G. Li et al. (2010)
Two lung PDX tissues 157+176 cells	scCAT-seq	scCAT-seq	G. Li et al. (2010)
Human morula/blastocyst 72 cells	scCAT-seq	scCAT-seq	G. Li et al. (2010)

Table S2: Publicly available datasets used in this work

Dataset description	Data type	Chapter	Reference
Human hematopoietic system 45 samples	bulk RNA-seq	ButchR	Corces et al. (2016)
Molecular signatures collection MSigDB	Signatures	ButchR	Subramanian et al. (2005)
Mouse substantia nigra 51,912 cells	scRNA-seq	i2NMF	Saunders et al. (2018)
Human substantia nigra 40,453 cells	scRNA-seq	i2NMF	Welch et al. (2019)
NRC NB cohort 162 samples	bulk RNA-seq	Neuroblastoma	Rajbhandari et al. (2018)
TARGET NB cohort 283 samples	bulk RNA-seq	Neuroblastoma	Pugh et al. (2013)
3D-Genome structure 8 human cell types	in situ Hi-C	Neuroblastoma	Rao et al. (2014)
Mouse organogenesis cell atlas (MOCA)	scRNA-seq	Neuroblastoma	Cao et al. (2019)
Mouse adrenal medulla 384 cells	scRNA-seq	Neuroblastoma	Furlan et al. (2017)
MGI human-mouse homologous master list	Annotation	Neuroblastoma	Bult et al. (2019)
Interaction data K562, HeLa-S3, HCT116	ChIA-PET	scCAT-seq	G. Li et al. (2010)
Mouse scRNA-seq atlas 56,265 cells	scRNA-seq	Regulons	Han et al. (2018)
Mouse scATAC-seq atlas 44,563 cells	scATAC-seq	Regulons	Cusanovich et al. (2018)

Appendix B: How to read a riverplot

The riverplot or Sankey diagram is a powerful representation to depict the flow rate of one group of values to another. In terms of the NMF, the riverplot is a helpful visualization of the degree of similarity between signatures at consecutive factorization ranks. We have included a new method in ButchR to produce a riverplot of the matrix decomposition results, this visualization is made using the R package *riverplot* (Weiner 2017).

The riverplot is a tree-like representation where nodes represent the NMF signatures, it can also be interpreted as an acyclic directed graph arranged in rows. The rows of the NMF riverplot are always sorted from the minimum factorization rank to the maximum factorization rank originally used to run the matrix decomposition (**Figure S1a**). Every row of the riverplot contains a number of nodes equal to the equivalent factorization rank k (e.g., the row corresponding to k=3 will contain three nodes representing the signatures learned with k=3) (**Figure S1b**).

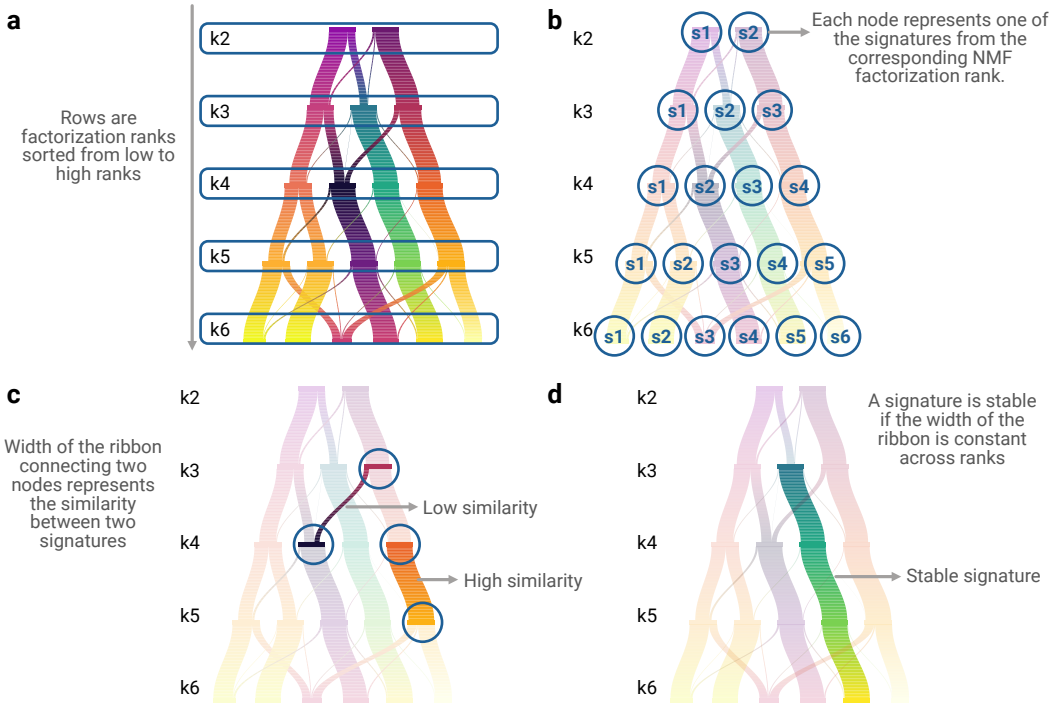


Figure S1: How to read a river plot? Paired with ButchR, the riverplot or Sankey diagram visualization is a powerful tool to understand the stability of the learned signatures, as well as identifying the flux of information between them. **(a)** The riverplot is organized as an acyclic directed graph arranged in rows. The rows are sorted in ascending order of factorization rank. **(b)** Every row of the riverplot is composed by nodes representing the signatures for the indicated factorization ran, i.e., the signatures derived from a same factorization rank are lay side by side, occupying one row of the graph. **(c)** The similarity between two signatures at consecutive factorization ranks is encoded in the width of the edge connecting them. **(d)** Stable signatures will appear as a ribbon of constant width across factorization ranks.

The relative similarity between signatures can be computed using the matrices H or the matrices W extracted from all factorization ranks. In the case of a riverplot built from the matrices W, the similarity between one signature s_{k_i} and all signatures in the next factorization rank $k+1$ is estimated in the following way:

1. Extract the exposure values e_{k_i} of the signature s_{k_i} from column i of the matrix W_k.

2. Fit the exposure values of matrix W_{k+1} to e_{k_i} using non-negative least squares (nnls).
3. Extract the coefficient values of the nnls solution which represent the relative similarity between signature s_{k_i} and all signatures in the next factorization rank s_{k+1_j} ($j \in \mathbb{Z} : k \in [1, k+1]$).

This procedure is repeated for all signatures s_{k_i} ($i \in \mathbb{Z} : i \in [1, k]$) across every factorization rank $k \in \mathbb{Z} : k \in [k_{min}, k_{max}]$.

The estimated similarities are encoded in the width on the edge that connects two signatures in the riverplot (**Figure S1c**). A signature is stable if the widths of the edges remain constant across factorization ranks forming a ribbon (**Figure S1d**). In terms of the flux of information this means that there is no influx from several signatures to define the signature, and also the outflux of information from the signature is low, remaining constant even if different factorization ranks are selected.

Appendix C: ShinyButchR tutorial

In this appendix, we show the steps to use ShinyButchR, and how to produce interactive visualizations of the NMF results.

Disclosure: The results presented in this chapter have been published in Quintero et al. (2020) and reproduced here with the permission of Oxford University Press, license number 5011370897521.

Data loading and NMF parameter selection

In order to use ShinyButchR, the data need to be uploaded into the app, and the NMF parameters have to be set according to the user's needs. The following steps describe how to complete these requirements, and also provide some clues on how to set the NMF parameters to obtain a good decomposition:

Step 1. Load setup screen:

Click on the [Data and annotation upload] tab, lo load the [Setup screen] (**Figure 4.1a**). It contains the [Matrix upload] and [Annotation upload] boxes, to upload new data; the [NMF params] box to tune the NMF parameters, and the [Start NMF] box to begin the analysis.

Step 2. Upload a non-negative matrix in an RDS or a CSV file:

Click on the [Browse...] button of the [Matrix upload] box to browse files in your local system and upload an RDS or CSV file containing a non-negative matrix. The file limit is 30 MB (corresponding to a numeric matrix of approximately 600 columns and 5000 features), but it can be changed with the local distribution of the app.

Step 3. Upload annotation as a CSV table or an RDS file containing an R data frame:

To perform the signature association analysis and produce a more informative heatmap of the matrix H, a file with associated biological/clinical information can be uploaded as well. Click on the [Browse...] button of the [Annotation upload] box to browse files in your local system and upload an RDS or CSV file containing an annotation table, the first column of this table should match the sample/cell identification names stored in the column names of the uploaded matrix.

Step 4. Selection of factorization rank range:

Select the range of the factorization ranks used to decompose the input Matrix by changing the values of the input boxes [Minimum factorization rank] and ["Maximum factorization rank]. The minimum number of ranks allowed is 2 and the maximum should be less than the total number of samples/cells (i.e., the number of columns in the input matrix).

Step 5. Selection of factorization method and number of iterations:

Click on the [Select factorization method] option from the [NMF params] box to select the algorithm desired to run the matrix decomposition. The options available in ShinyButchR are NMF (Seung and Lee 1999) and GRNMF-SC (C. Lin and Pang 2015).

The number of random initialization to use can be set in the [Number of initializations] parameter box, as well as the convergence threshold in the [Convergence threshold] box. We suggest using at least 2 random initialization and a convergence threshold of 40 to

obtain more stable results.

Step 6. Run matrix decomposition using NMF:

The NMF decomposition can be executed after uploading the data and selecting the desired parameters. Click on the [Submit] button inside the [Start NMF] box. A waiting screen will appear while the decomposition is performed, the total computation time depends on the size of the input matrix and the NMF parameters. For instance, decomposing a matrix with 22,000 genes and 45 samples with a range of factorization ranks from 5 to 8, and 10 random initializations will take about one minute.

Interactive exploration of NMF results

After the matrix decomposition is completed, the results can be explored using the wide arrange of interactive visualizations included in ShinyButchR. The following steps describe how to browse and export the results for a downstream analysis:

Step 7. Load output results visualization screen:

The [Results screen] (**Figure 4.1a**) is the main screen to explore the matrix decomposition results. All the visualizations provided by ShinyButchR can be found on this screen. Depending on the availability of a valid annotation table, the resulting figures will be more informative and help the user to identify the biological nature of the NMF signatures.

Step 8. Selection of optimal factorization rank k:

The decomposition diagnostic plot provides a guide to select the optimal factorization rank. This plot will contain informative statistics for each factorization rank and can be found in the [Optimal factorization rank] box.

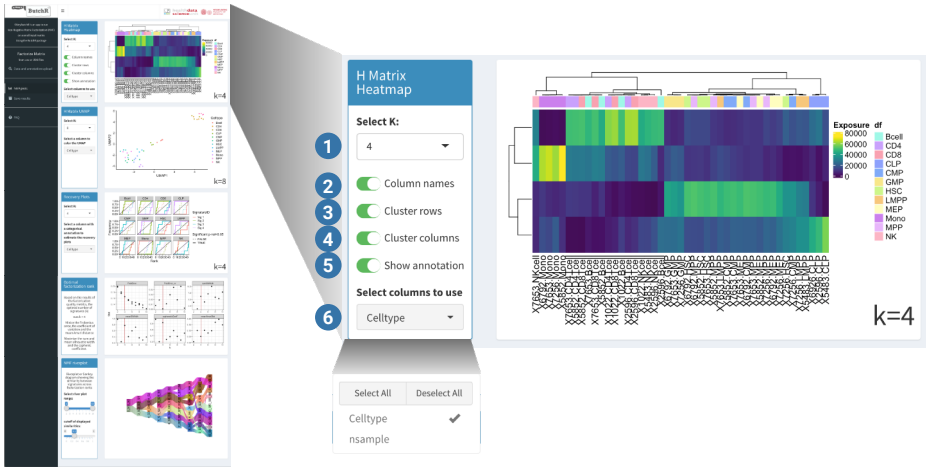


Figure S2: ShinyButchR interactive H matrix visualization. (1) For every factorization rank a heatmap can be generated, (2) the column names of the original input matrix can be displayed, (3) the signatures can be clustered by similarity between them, (4) as well as the samples/cells, (5-6) if a valid annotation file was uploaded into the app the selected annotation tracks can be displayed.

Step 9. Visualization of the matrix H heatmap:

For every factorization rank, ShinyButchR provides a heatmap visualization of the exposure values from the matrix H, using the R package ComplexHeatmap (Gu, Eils, and Schlesner 2016). The rows of this matrix are the signatures learned from the input matrix and can be used to soft cluster the samples/cells. The visualization results can be enhanced by uploading a table with biological and clinical metadata associated with the samples/cells of the input matrix. The heatmap representation is available in the [H Matrix Heatmap] box (**Figure S2**).

Step 10. Uniform manifold approximation and projection (UMAP) visualization based on the matrix H:

The [H matrix UMAP] box shows the UMAP embedding (Diaz-Papkovich et al. 2019) built from the selected factorization rank. Similarly, as with the Matrix H heatmap, the UMAP embedding can be enhanced by using metadata associated with the samples/cells of the input matrix. In this case, the color can be changed based on the selected variable (**Figure S3**).

Step 11. Visualization of recovery plots:

As described in "ButchR: NMF suit to slice genome-scale datasets," the association of the NMF signatures with biological and clinical variables can be measured and visualized using a recovery curve. This visualization found in the [Recovery plots] box, is a powerful tool to assign an identity to the recovered signatures. A recovery curve can be constructed for every categorical annotation variable included in the annotation file (**Figure S4**).

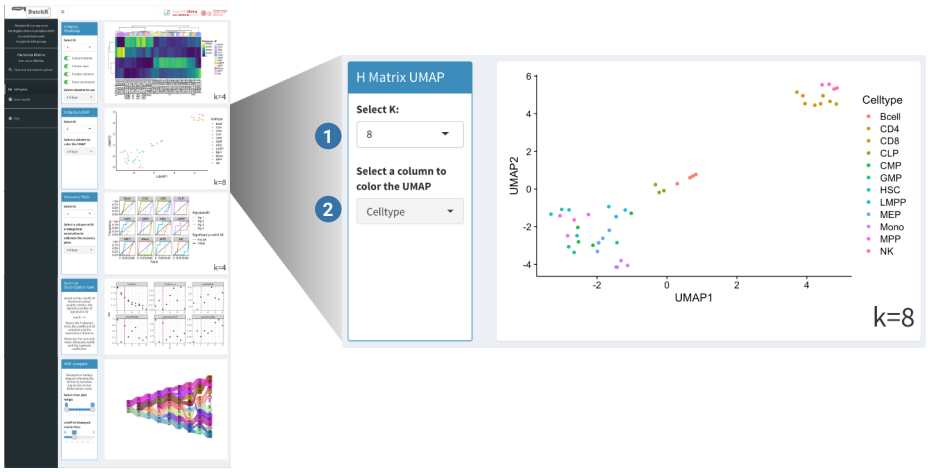


Figure S3: ShinyButchR exposure UMAP embedding. (1) For every factorization rank a UMAP embedding can be constructed, (2) and colored according to one selected annotation variable.

Step 12. Visualization of signature stability with a Riverplot:

The last visualization included in ShinyButchR is a riverplot depicting the similarity between linked signatures (see "Appendix B: How to read a riverplot" for a detailed explanation of the riverplot visualization). The riverplot plot can be found in the [NMF riverplot] box. It provides a visual inspection of the stability of the signatures learned using NMF. Two sliders are included to change the cutoff of the displayed similarities, and the range of factorization ranks to include in the visualization (**Figure S5**).

Step 13. Export results and post-processing:

ShinyButchR also includes the functionality to save and export the results of the workflow, either as a CSV file or a native R RDS file. Click on the [Save results] tab, lo load the [Save results screen], and save the results of the current experiment.

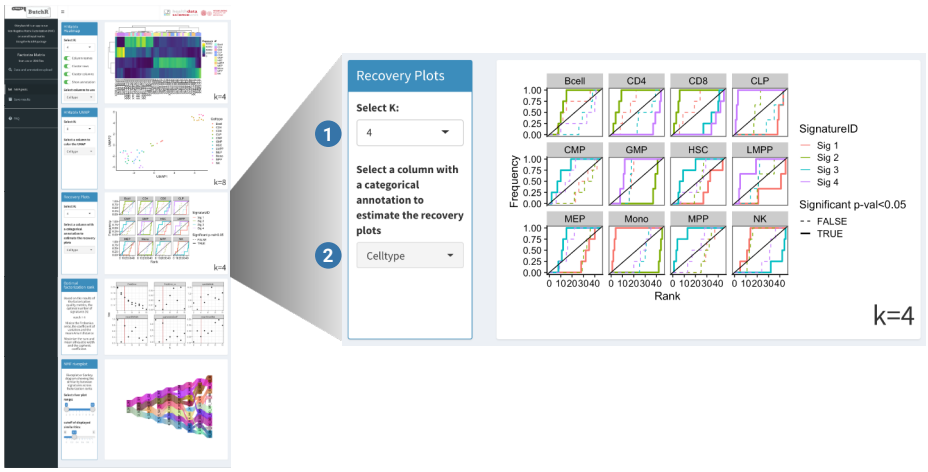


Figure S4: ShinyButchR recovery plots. A recovery curve is displayed (1) for every signature and (2) every class of a selected categorical annotation variable.

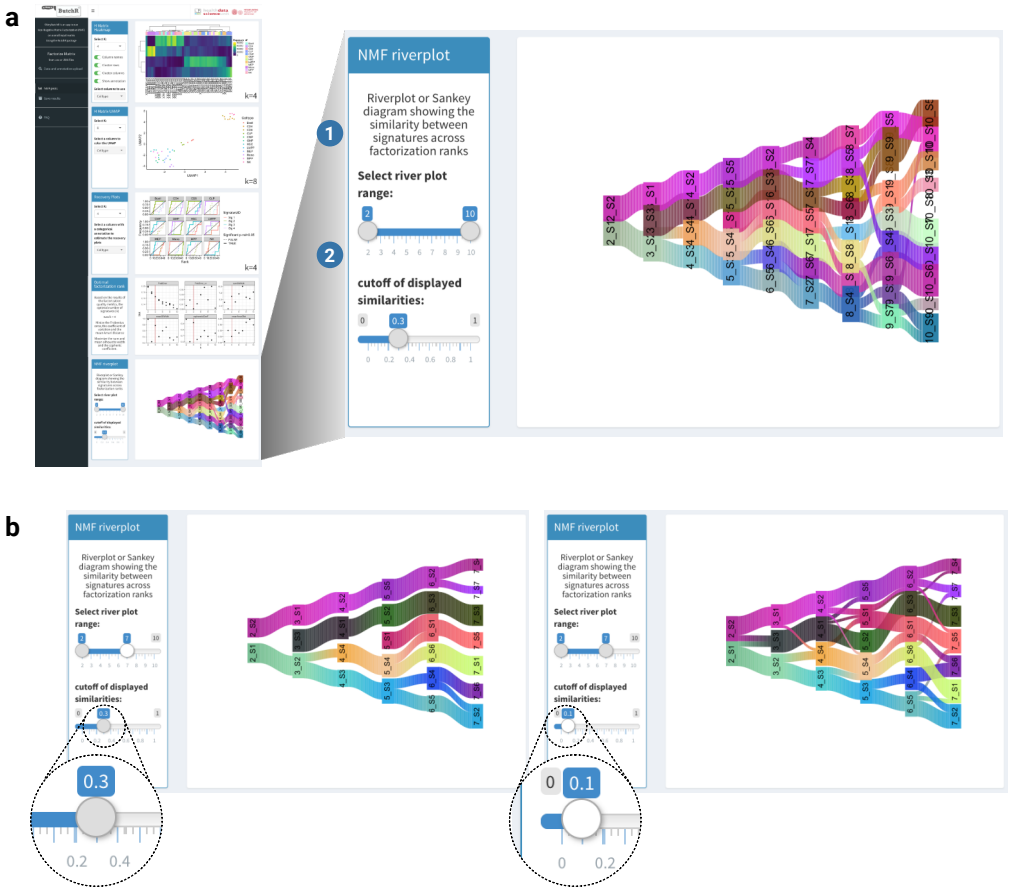


Figure S5: ShinyButchR interactive riverplot. **(a)** A riverplot is constructed (1) using the range of factorization ranks selected in the "Select river plot range" slider, (2) and low similarities can be removed from the visualization using the "cutoff of displayed similarities" slider. **(b)** Only the most stable connections will be displayed if a large similarity cutoff is selected, or **(c)** all minor connections will be included in the visualization if a low large similarity cutoff is selected.

References

10 Abadi, Martín, Paul Barham, Jianmin Chen, Zhifeng Chen, Andy Davis, Jeffrey Dean, Matthieu Devin, et al. 2016. "TensorFlow: A system for large-scale machine learning." In *Proceedings of the 12th USENIX Symposium on Operating Systems Design and Implementation, OSDI 2016*. http://arxiv.org/abs/1605.08695.

Ackermann, Amanda M., Zhiping Wang, Jonathan Schug, Ali Naji, and Klaus H. Kaestner. 2016. "Integration of ATAC-seq and RNA-seq identifies human alpha cell and beta cell signature genes." *Molecular Metabolism*. https://doi.org/10.1016/j.molmet.2016.01.002.

Ackermann, Sandra, Maria Cartolano, Barbara Hero, Anne Welte, Yvonne Kahlert, Andrea Roderwieser, Christoph Bartenhagen, et al. 2018. "A mechanistic classification of clinical phenotypes in neuroblastoma." *Science*. https://doi.org/10.1126/science.aat6768.

Aibar, Sara, Carmen Bravo González-Blas, Thomas Moerman, Vân Anh Huynh-Thu, Hana Imrichova, Gert Hulselmans, Florian Rambow, et al. 2017. "SCENIC: Single-cell regulatory network inference and clustering." *Nature Methods*. https://doi.org/10.1038/nmeth.4463.

Alexandrov, Ludmil B., Jaegil Kim, Nicholas J. Haradhvala, Mi Ni Huang, Alvin Wei Tian Ng, Yang Wu, Arnoud Boot, et al. 2020. "The repertoire of mutational signatures in human cancer." *Nature*. https://doi.org/10.1038/s41586-020-1943-3.

Alexandrov, Ludmil B, Serena Nik-Zainal, David C Wedge, Peter J Campbell, and Michael R Stratton. 2013. "Deciphering signatures of mutational processes operative in human cancer." *Cell Reports* 3 (1): 246–59. https://doi.org/10.1016/j.celrep.2012.12.008.

Allaire, J J, Kevin Ushey, Yuan Tang, and Dirk Eddelbuettel. 2017. *reticulate: R Interface to Python*. https://github.com/rstudio/reticulate.

Ambros, Ingeborg M., Andrea Zellner, Borghild Roald, Gabriele Amann, Ruth Ladenstein, Dieter Printz, Helmut Gadner, and Peter F. Ambros. 1996. "Role of Ploidy, Chromosome 1p, and Schwann Cells in the Maturation of Neuroblastoma." *New England Journal of Medicine*. https://doi.org/10.1056/nejm199606063342304.

Andreatta, Massimo, Jesus Corria-Osorio, Sören Müller, Rafael Cubas, George Coukos, and Santiago J. Carmona. 2020. "Projecting single-cell transcriptomics data onto a reference t cell atlas to interpret immune responses." https://doi.org/10.1101/2020.06.23.166546.

Argelaguet, Ricard, Damien Arnol, Danila Bredikhin, Yonatan Deloro, Britta Velten, John C. Marioni, and Oliver Stegle. 2020. "MOFA+: A statistical framework for comprehensive integration of multi-modal single-cell data." *Genome Biology*. https://doi.org/10.1186/s13059-020-02015-1.

Argelaguet, Ricard, Britta Velten, Damien Arnol, Sascha Dietrich, Thorsten Zenz, John C Marioni, Florian Buettner, Wolfgang Huber, and Oliver Stegle. 2018. "Multi-Omics Factor Analysis—a framework for unsupervised integration of multi-omics data sets." *Molecular Systems Biology*. https://doi.org/10.15252/msb.20178124.

Attaf, Noudjoud, Iñaki Cervera-Marzal, Chuang Dong, Laurine Gil, Amédée Renand, Lionel Spinelli, and Pierre Milpied. 2020. "FB5P-seq: FACS-Based 5-Prime End Single-Cell RNA-seq for Integrative Analysis of Transcriptome and Antigen Receptor Repertoire in B and T Cells." *Frontiers in Immunology*. https://doi.org/10.3389/fimmu.2020.00216.

Attiyeh, Edward F., Wendy B. London, Yael P. Mossé, Qun Wang, Cynthia Winter, Deepa Khazi, Patrick W. McGrady, et al. 2005. "Chromosome 1p and 11q Deletions and Outcome in Neuroblastoma." *New England Journal of Medicine*. https://doi.org/10.1056/nejmoa052399.

Bannister, Andrew J., and Tony Kouzarides. 2011. "Regulation of chromatin by histone modifications." https://doi.org/10.1038/cr.2011.22.

Baron, Chloé S., Aditya Barve, Mauro J. Muraro, Reinier van der Linden, Gitanjali Dharmadhikari, Anna Lyubimova, Eelco J. P. de Koning, and Alexander van Oudenaarden. 2019. "Cell Type Purification by Single-Cell Transcriptome-Trained Sorting." *Cell*. https://doi.org/10.1016/j.cell.2019.08.006.

Barrett, Tanya, Stephen E. Wilhite, Pierre Ledoux, Carlos Evangelista, Irene F. Kim, Maxim Tomashevsky, Kimberly A. Marshall, et al. 2013. "NCBI GEO: Archive for functional genomics data sets - Update." *Nucleic Acids Research*. https://doi.org/10.1093/nar/gks1193.

Bartenhagen, Christoph, Hans Ulrich Klein, Christian Ruckert, Xiaoyi Jiang, and Martin Dugas. 2010. "Comparative study of unsupervised dimension reduction techniques for the visualization of microarray gene expression data." *BMC Bioinformatics*. https://doi.org/10.1186/1471-2105-11-567.

Bellman, Richard. 1966. "Dynamic programming." *Science*. https://doi.org/10.1126/science.153.3731.34.

Bendall, Sean C., Kara L. Davis, El Ad David Amir, Michelle D. Tadmor, Erin F. Simonds, Tiffany J. Chen, Daniel K. Shenfeld, Garry P. Nolan, and Dana Pe'Er. 2014. "Single-cell trajectory detection uncovers progression and regulatory coordination in human b cell development." *Cell*. https://doi.org/10.1016/j.cell.2014.04.005.

Benzi, Kirell, Vassiiis Kalofolias, Xavier Bresson, and Pierre Vandergheynst. 2016. "Song

recommendation with non-negative matrix factorization and graph total variation." In *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*. Vol. 2016–May. https://doi.org/10.1109/ICASSP.2016.7472115.

Berglund, Anders E., Eric A. Welsh, and Steven A. Eschrich. 2017. "Characteristics and Validation Techniques for PCA-Based Gene-Expression Signatures." *International Journal of Genomics*. https://doi.org/10.1155/2017/2354564.

Bergstrom, Erik N., Mi Ni Huang, Uma Mahto, Mark Barnes, Michael R. Stratton, Steven G. Rozen, and Ludmil B. Alexandrov. 2019. "SigProfilerMatrixGenerator: A tool for visualizing and exploring patterns of small mutational events." *BMC Genomics*. https://doi.org/10.1186/s12864-019-6041-2.

Bernstein, Bradley E., Tarjei S. Mikkelsen, Xiaohui Xie, Michael Kamal, Dana J. Huebert, James Cuff, Ben Fry, et al. 2006. "A Bivalent Chromatin Structure Marks Key Developmental Genes in Embryonic Stem Cells." *Cell*. https://doi.org/10.1016/j.cell.2006.02.041.

Berry, Michael W., and Murray Browne. 2005. "Email surveillance using non-negative matrix factorization." *Computational and Mathematical Organization Theory* 11 (3): 249–64. https://doi.org/10.1007/s10588-005-5380-5.

Berry, Michael W., Murray Browne, Amy N. Langville, V. Paul Pauca, and Robert J. Plemmons. 2007. "Algorithms and applications for approximate nonnegative matrix factorization." *Computational Statistics and Data Analysis* 52 (1). https://doi.org/10.1016/j.csda.2006.11.006.

Berry, Teeara, William Luther, Namrata Bhatnagar, Yann Jamin, Evon Poon, Takaomi Sanda, Desheng Pei, et al. 2012. "The ALKF1174L Mutation Potentiates the Oncogenic Activity of MYCN in Neuroblastoma." *Cancer Cell*. https://doi.org/10.1016/j.ccr.2012.06.001.

Biedler, June L., Lawrence Helson, and Barbara A. Spengler. 1973. "Morphology and Growth, Tumorigenicity, and Cytogenetics of Human Neuroblastoma Cells in Continuous Culture." *Cancer Research*.

Birney, Ewan, and Nicole Soranzo. 2015. "Human genomics: The end of the start for population sequencing." https://doi.org/10.1038/526052a.

Biton, Anne, Isabelle Bernard-Pierrot, Yinjun Lou, Clémentine Krucker, Elodie Chapeaublanc, Carlota Rubio-Pérez, Nuria López-Bigas, et al. 2014. "Independent Component Analysis Uncovers the Landscape of the Bladder Tumor Transcriptome and Reveals Insights into Luminal and Basal Subtypes." *Cell Reports*. https://doi.org/10.1016/j.celrep.2014.10.035.

Blum, Matthias, Hsin Yu Chang, Sara Chuguransky, Tiago Grego, Swaathi Kandasaamy, Alex Mitchell, Gift Nuka, et al. 2021. "The InterPro protein families and domains database: 20 years on." *Nucleic Acids Research*. https://doi.org/10.1093/nar/gkaa977.

Boeva, Valentina, Caroline Louis-Brennetot, Agathe Peltier, Simon Durand, Cecile Pierre-Eugène, Virginie Raynal, Heather C. Etchevers, et al. 2017. "Heterogeneity of neuroblastoma cell identity defined by transcriptional circuitries." *Nature Genetics* 49 (9): 1408–13. https://doi.org/10.1038/ng.3921.

Bosse, Kristopher R., and John M. Maris. 2016. "Advances in the translational genomics of neuroblastoma: From improving risk stratification and revealing novel biology to identifying actionable genomic alterations." https://doi.org/10.1002/cncr.29706.

Bown, Nick, Simon Cotterill, Maria Łastowska, Seamus O'Neill, Andrew D. J. Pearson, Dominique Plantaz, Mounira Meddeb, et al. 1999. "Gain of Chromosome Arm 17q and Adverse Outcome in Patients with Neuroblastoma." *New England Journal of Medicine*. https://doi.org/10.1056/nejm199906243402504.

Bresler, Scott C., Daniel A. Weiser, Peter J. Huwe, Jin H. Park, Kateryna Krytska, Hannah Ryles, Marci Laudenslager, et al. 2014. "ALK Mutations Confer Differential Oncogenic Activation and Sensitivity to ALK Inhibition Therapy in Neuroblastoma." *Cancer Cell*. https://doi.org/10.1016/j.ccell.2014.09.019.

Brunet, Jean-Philippe, Pablo Tamayo, Todd R Golub, and Jill P Mesirov. 2004. "Metagenes and molecular pattern discovery using matrix factorization." *Proceedings of the National Academy of Sciences of the United States of America* 101 (12): 4164–69. https://doi.org/10.1073/pnas.0308531101.

Buenrostro, Jason D., Beijing Wu, Ulrike M. Litzenburger, Dave Ruff, Michael L. Gonzales, Michael P. Snyder, Howard Y. Chang, and William J. Greenleaf. 2015. "Single-cell chromatin accessibility reveals principles of regulatory variation." *Nature*. https://doi.org/10.1038/nature14590.

Buettner, Florian, Kedar N. Natarajan, F. Paolo Casale, Valentina Proserpio, Antonio Scialdone, Fabian J. Theis, Sarah A. Teichmann, John C. Marioni, and Oliver Stegle. 2015. "Computational analysis of cell-to-cell heterogeneity in single-cell RNA-sequencing data reveals hidden subpopulations of cells." *Nature Biotechnology*. https://doi.org/10.1038/nbt.3102.

Bult, Carol J., Judith A. Blake, Cynthia L. Smith, James A. Kadin, Joel E. Richardson, A. Anagnostopoulos, R. Asabor, et al. 2019. "Mouse Genome Database (MGD) 2019." *Nucleic Acids Research*. https://doi.org/10.1093/nar/gky1056.

Burge, Sarah, Elizabeth Kelly, David Lonsdale, Prudence Mutowo-Muellenet, Craig McAnulla, Alex Mitchell, Amaia Sangrador-Vegas, Siew Yit Yong, Nicola Mulder, and Sarah Hunter. 2012. "Manual GO annotation of predictive protein signatures: the InterPro approach to GO curation." *Database : The Journal of Biological Databases and Curation*. https://doi.org/10.1093/database/bar068.

Butler, Andrew, Paul Hoffman, Peter Smibert, Efthymia Papalexi, and Rahul Satija. 2018. "Integrating single-cell transcriptomic data across different conditions, technolo-

gies, and species." *Nature Biotechnology*. https://doi.org/10.1038/nbt.4096.

Calabrese, Claudia, Natalie R. Davidson, Deniz Demircioglu, Nuno A. Fonseca, Yao He, André Kahles, Kjong Van Lehmann, et al. 2020. "Genomic basis for RNA alterations in cancer." *Nature*. https://doi.org/10.1038/s41586-020-1970-0.

Campbell, Peter J., Gad Getz, Jan O. Korbel, Joshua M. Stuart, Jennifer L. Jennings, Lincoln D. Stein, Marc D. Perry, et al. 2020. "Pan-cancer analysis of whole genomes." *Nature*. https://doi.org/10.1038/s41586-020-1969-6.

Cantini, Laura, Pooya Zakeri, Celine Hernandez, Aurelien Naldi, Denis Thieffry, Elisabeth Remy, and Anaïs Baudot. 2021. "Benchmarking joint multi-omics dimensionality reduction approaches for the study of cancer." *Nature Communications*. https://doi.org/10.1038/s41467-020-20430-7.

Cao, Junyue, Diana R. O'Day, Hannah A. Pliner, Paul D. Kingsley, Mei Deng, Riza M. Daza, Michael A. Zager, et al. 2020. "A human cell atlas of fetal gene expression." *Science*. https://doi.org/10.1126/science.aba7721.

Cao, Junyue, Malte Spielmann, Xiaojie Qiu, Xingfan Huang, Daniel M. Ibrahim, Andrew J. Hill, Fan Zhang, et al. 2019. "The single-cell transcriptional landscape of mammalian organogenesis." *Nature* 566 (7745): 496–502. https://doi.org/10.1038/s41586-019-0969-x.

Carbon, Seth, Eric Douglass, Benjamin M. Good, Deepak R. Unni, Nomi L. Harris, Christopher J. Mungall, Siddartha Basu, et al. 2021. "The Gene Ontology resource: Enriching a GOld mine." *Nucleic Acids Research*. https://doi.org/10.1093/nar/gkaa1113.

Chai, Lian En, Swee Kuan Loh, Swee Thing Low, Mohd Saberi Mohamad, Safaai Deris, and Zalmiyah Zakaria. 2014. "A review on the computational approaches for gene regulatory network construction." https://doi.org/10.1016/j.compbiomed.2014.02.011.

Chalise, Prabhakar, and Brooke L. Fridley. 2017. "Integrative clustering of multi-level 'omic data based on non-negative matrix factorization algorithm." *PLoS ONE*. https://doi.org/10.1371/journal.pone.0176278.

Chang, Winston, Joe Cheng, J J Allaire, Yihui Xie, and Jonathan McPherson. 2020. *shiny: Web Application Framework for R*. https://cran.r-project.org/package=shiny.

Chappell, Lia, Andrew J. C. Russell, and Thierry Voet. 2018. "Single-Cell (Multi)omics Technologies." https://doi.org/10.1146/annurev-genom-091416-035324.

Chen, Song, Blue B. Lake, and Kun Zhang. 2019. "High-throughput sequencing of the transcriptome and chromatin accessibility in the same cell." *Nature Biotechnology*. https://doi.org/10.1038/s41587-019-0290-0.

Cho, Jae Yong, Jae Yun Lim, Jae Ho Cheong, Yun Yong Park, Se Lyun Yoon, Soo Mi Kim, Sang Bae Kim, et al. 2011. "Gene expression signature-based prognostic risk score in gastric cancer." *Clinical Cancer Research*. https://doi.org/10.1158/1078-0432.CCR-10-2180.

Ciccarone, Valentina, Barbara A. Spengler, Marian B. Meyers, June L. Biedler, and Robert A. Ross. 1989. "Phenotypic Diversification in Human Neuroblastoma Cells: Expression of Distinct Neural Crest Lineages." *Cancer Research*.

Clark, Stephen J., Ricard Argelaguet, Chantriolnt-Andreas Kapourani, Thomas M. Stubbs, Heather J. Lee, Celia Alda-Catalinas, Felix Krueger, et al. 2018. "scNMT-seq enables joint profiling of chromatin accessibility DNA methylation and transcription in single cells." *Nature Communications* 9 (1): 781. https://doi.org/10.1038/s41467-018-03149-4.

Collado-Torres, Leonardo, Abhinav Nellore, Kai Kammers, Shannon E. Ellis, Margaret A. Taub, Kasper D. Hansen, Andrew E. Jaffe, Ben Langmead, and Jeffrey T. Leek. 2017. "Reproducible RNA-seq analysis using recount2." https://doi.org/10.1038/

nbt.3838.

Colomé-Tatché, M., and F. J. Theis. 2018. "Statistical single cell multi-omics integration." https://doi.org/10.1016/j.coisb.2018.01.003.

Corces, M. Ryan, Jason D. Buenrostro, Beijing Wu, Peyton G. Greenside, Steven M. Chan, Julie L. Koenig, Michael P. Snyder, et al. 2016. "Lineage-specific and single-cell chromatin accessibility charts human hematopoiesis and leukemia evolution." *Nature Genetics*. https://doi.org/10.1038/ng.3646.

Corces, M. Ryan, Jeffrey M. Granja, Shadi Shams, Bryan H. Louie, Jose A. Seoane, Wanding Zhou, Tiago C. Silva, et al. 2018. "The chromatin accessibility landscape of primary human cancers." *Science*. https://doi.org/10.1126/science.aav1898.

Cusanovich, Darren A., Andrew J. Hill, Delasa Aghamirzaie, Riza M. Daza, Hannah A. Pliner, Joel B. Berletch, Galina N. Filippova, et al. 2018. "A Single-Cell Atlas of In Vivo Mammalian Chromatin Accessibility." *Cell*. https://doi.org/10.1016/j.cell.2018.06.052.

Czerwinska, Urszula. 2018. "DeconICA: DeconICA first release." https://zenodo.org/record/1250070%7B/#%7D.YGHsTuaxVTZ.

Danese, Anna, Maria L. Richter, David S. Fischer, Fabian J. Theis, and Maria Colomé-Tatché. 2019. "EpiScanpy: Integrated single-cell epigenomic analysis." https://doi.org/10.1101/648097.

Davidson, Eric H. 2010. "Emerging properties of animal gene regulatory networks." https://doi.org/10.1038/nature09645.

De Vito, Roberta, Ruggero Bellio, Lorenzo Trippa, and Giovanni Parmigiani. 2019. "Multi-study factor analysis." *Biometrics*. https://doi.org/10.1111/biom.12974.

Diaz-Papkovich, Alex, Luke Anderson-Trocmé, Chief Ben-Eghan, and Simon Gravel.

2019. "UMAP reveals cryptic population structure and phenotype heterogeneity in large genomic cohorts." *PLoS Genetics*. https://doi.org/10.1371/journal.pgen.1008432.

Ding, Hongxu, Andrew Blair, Ying Yang, and Joshua M. Stuart. 2019. "Biological process activity transformation of single cell gene expression for cross-species alignment." *Nature Communications*. https://doi.org/10.1038/s41467-019-12924-w.

Domcke, Silvia, Andrew J. Hill, Riza M. Daza, Junyue Cao, Diana R. O'Day, Hannah A. Pliner, Kimberly A. Aldinger, et al. 2020. "A human cell atlas of fetal chromatin accessibility." *Science*. https://doi.org/10.1126/science.aba7612.

Duan, Yuzhu, Daniel S. Evans, Richard A. Miller, Nicholas J. Schork, Steven R. Cummings, and Thomas Girke. 2020. "Signature search: Environment for gene expression signature searching and functional interpretation." *Nucleic Acids Research*. https://doi.org/10.1093/nar/gkaa878.

Duda, R O, P E Hart, and D G Stork. 2001. "Pattern classification." *New York: John Wiley, Section*.

Dumitrascu, Bianca, Soledad Villar, Dustin G. Mixon, and Barbara E. Engelhardt. 2019. "Optimal marker gene selection for cell type discrimination in single cell analyses." https://doi.org/10.1101/599654.

Duren, Zhana, Xi Chen, Mahdi Zamanighomi, Wanwen Zeng, Ansuman T. Satpathy, Howard Y. Chang, Yong Wang, and Wing Hung Wong. 2018. "Integrative analysis of single-cell genomics data by coupled nonnegative matrix factorizations." *Proceedings of the National Academy of Sciences of the United States of America*. https://doi.org/10.1073/pnas.1805681115.

Eppert, Kolja, Katsuto Takenaka, Eric R. Lechman, Levi Waldron, Björn Nilsson, Peter Van Galen, Klaus H. Metzeler, et al. 2011. "Stem cell gene expression programs influence clinical outcome in human leukemia." *Nature Medicine* 17 (9): 1086–94.

https://doi.org/10.1038/nm.2415.

Fernandes, H., and P. Zhang. 2014. "Overview of Molecular Diagnostics in CLinical Pathology." In *Pathobiology of Human Disease: A Dynamic Encyclopedia of Disease Mechanisms*. https://doi.org/10.1016/B978-0-12-386456-7.06306-1.

Fiers, Mark W. E. J., Liesbeth Minnoye, Sara Aibar, Carmen Bravo González-Blas, Zeynep Kalender Atak, and Stein Aerts. 2018. "Mapping gene regulatory networks from single-cell omics data." *Briefings in Functional Genomics*. https://doi.org/10.1093/bfgp/elx046.

Fodor, Imola K. 2002. "A survey of dimension reduction techniques." *Library*. https://doi.org/10.2172/15002155.

Forcato, Mattia, Oriana Romano, and Silvio Bicciato. 2021. "Computational methods for the integrative analysis of single-cell data." *Briefings in Bioinformatics*. https://doi.org/10.1093/bib/bbaa042.

Franzén, Oscar, Li Ming Gan, and Johan L. M. Björkegren. 2019. "PanglaoDB: A web server for exploration of mouse and human single-cell RNA sequencing data." *Database*. https://doi.org/10.1093/database/baz046.

Fröhlich, Holger, Rudi Balling, Niko Beerenwinkel, Oliver Kohlbacher, Santosh Kumar, Thomas Lengauer, Marloes H. Maathuis, et al. 2018. "From hype to reality: Data science enabling personalized medicine." *BMC Medicine*. https://doi.org/10.1186/s12916-018-1122-7.

Furlan, Alessandro, Vyacheslav Dyachuk, Maria Eleni Kastriti, Laura Calvo-Enrique, Hind Abdo, Saida Hadjab, Tatiana Chontorotzea, et al. 2017. "Multipotent peripheral glial cells generate neuroendocrine cells of the adrenal medulla." *Science* 357 (6346). https://doi.org/10.1126/science.aal3753.

Galon, Jérôme, Franck Pagès, Francesco M. Marincola, Helen K. Angell, Magdalena Thurin, Alessandro Lugli, Inti Zlobec, et al. 2012. "Cancer classification using the Im-

munoscore: A worldwide task force." https://doi.org/10.1186/1479-5876-10-205.

Gao, Zhen, Yu Tian Wang, Qing Wen Wu, Jian Cheng Ni, and Chun Hou Zheng. 2020. "Graph regularized L 2,1-nonnegative matrix factorization for miRNA-disease association prediction." *BMC Bioinformatics* 21 (1): 61. https://doi.org/10.1186/s12859-020-3409-x.

Gartlgruber, Moritz, Ashwini Kumar Sharma, Andrés Quintero, Daniel Dreidax, Selina Jansky, Young Gyu Park, Sina Kreth, et al. 2021. "Super enhancers define regulatory subtypes and cell identity in neuroblastoma." *Nature Cancer* 2 (1). https://doi.org/10.1038/s43018-020-00145-w.

Gaujoux, Renaud, and Cathal Seoighe. 2010. "A flexible R package for nonnegative matrix factorization." *BMC Bioinformatics*. https://doi.org/10.1186/1471-2105-11-367.

Gerstung, Moritz, Clemency Jolly, Ignaty Leshchiner, Stefan C. Dentro, Santiago Gonzalez, Daniel Rosebrock, Thomas J. Mitchell, et al. 2020. "The evolutionary history of 2,658 cancers." *Nature*. https://doi.org/10.1038/s41586-019-1907-7.

Goldman, Mary J., Brian Craft, Mim Hastie, Kristupas Repečka, Fran McDade, Akhil Kamath, Ayan Banerjee, et al. 2020. "Visualizing and interpreting cancer genomics data via the Xena platform." https://doi.org/10.1038/s41587-020-0546-8.

Golub, T. R., D. K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J. P. Mesirov, H. Coller, et al. 1999. "Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring." *Science*. https://doi.org/10.1126/science.286.5439.531.

Graham, Jeffrey, Daniel Y. C. Heng, James Brugarolas, and Ulka Vaishampayan. 2018. "Personalized Management of Advanced Kidney Cancer." *American Society of Clinical Oncology Educational Book*. https://doi.org/10.1200/edbk_201215.

Granja, Jeffrey M, M Ryan Corces, Sarah E Pierce, S Tansu Bagdatli, Hani Choudhry, Howard Y Chang, and William J Greenleaf. 2021. "ArchR is a scalable software

package for integrative single-cell chromatin accessibility analysis." *Nature Genetics*. https://doi.org/10.1038/s41588-021-00790-6.

Gu, Zuguang, Roland Eils, and Matthias Schlesner. 2016. "Complex heatmaps reveal patterns and correlations in multidimensional genomic data." *Bioinformatics* 32 (18): 2847–49. https://doi.org/10.1093/bioinformatics/btw313.

Gulati, Gunsagar S., Shaheen S. Sikandar, Daniel J. Wesche, Anoop Manjunath, Anjan Bharadwaj, Mark J. Berger, Francisco Ilagan, et al. 2020. "Single-cell transcriptional diversity is a hallmark of developmental potential." *Science*. https://doi.org/10.1126/science.aax0249.

Gundersen, Gregory W., Matthew R. Jones, Andrew D. Rouillard, Yan Kou, Caroline D. Monteiro, Axel S. Feldmann, Kevin S. Hu, and Avi Ma'Ayan. 2015. "GEO2Enrichr: Browser extension and server app to extract gene sets from GEO and analyze them for biological functions." *Bioinformatics*. https://doi.org/10.1093/bioinformatics/btv297.

Guyon, Isabelle, Jason Weston, Stephen Barnhill, and Vladimir Vapnik. 2002. "Gene selection for cancer classification using support vector machines." *Machine Learning*. https://doi.org/10.1023/A:1012487302797.

Haddad, Rima, Philippe Guardiola, Brigitte Izac, Christelle Thibault, Jerry Radich, Anne Lise Delezoide, Claude Baillou, et al. 2004. "Molecular characterization of early human T/NK and B-lymphoid progenitor cells in umbilical cord blood." *Blood* 104 (13): 3918–26. https://doi.org/10.1182/blood-2004-05-1845.

Han, Xiaoping, Renying Wang, Yincong Zhou, Lijiang Fei, Huiyu Sun, Shujing Lai, Assieh Saadatpour, et al. 2018. "Mapping the Mouse Cell Atlas by Microwell-Seq." *Cell*. https://doi.org/10.1016/j.cell.2018.02.001.

Haradhvala, N. J., J. Kim, Y. E. Maruvka, P. Polak, D. Rosebrock, D. Livitz, J. M. Hess, et al. 2018. "Distinct mutational signatures characterize concurrent loss of

polymerase proofreading and mismatch repair." *Nature Communications*. https://doi.org/10.1038/s41467-018-04002-4.

Hastie, T; Tibshirani. 2017. "The Elements of Statistical Learning Second Edition." *Math. Intell.* http://arxiv.org/abs/arXiv:1011.1669v3.

Hänzelmann, Sonja, Robert Castelo, and Justin Guinney. 2013. "GSVA: Gene set variation analysis for microarray and RNA-Seq data." *BMC Bioinformatics*. https://doi.org/10.1186/1471-2105-14-7.

Heintzman, Nathaniel D., and Bing Ren. 2009. "Finding distal regulatory elements in the human genome." https://doi.org/10.1016/j.gde.2009.09.006.

Heinz, Sven, Casey E. Romanoski, Christopher Benner, and Christopher K. Glass. 2015. "The selection and function of cell type-specific enhancers." https://doi.org/10.1038/nrm3949.

Herrmann, Carl, Bram Van De Sande, Delphine Potier, and Stein Aerts. 2012. "i-cisTarget: An integrative genomics method for the prediction of regulatory features and cis-regulatory modules." *Nucleic Acids Research*. https://doi.org/10.1093/nar/gks543.

Hnisz, Denes, Brian J. Abraham, Tong Ihn Lee, Ashley Lau, Violaine Saint-André, Alla A. Sigova, Heather A. Hoke, and Richard A. Young. 2013. "XSuper-enhancers in the control of cell identity and disease." *Cell* 155 (4): 934. https://doi.org/10.1016/j.cell.2013.09.053.

Holland, Christian H., Jovan Tanevski, Javier Perales-Patón, Jan Gleixner, Manu P. Kumar, Elisabetta Mereu, Brian A. Joughin, et al. 2020. "Robustness and applicability of transcription factor and pathway analysis tools on single-cell RNA-seq data." *Genome Biology* 21 (1). https://doi.org/10.1186/s13059-020-1949-z.

Hon, Gary C., R. David Hawkins, and Bing Ren. 2009. "Predictive chromatin signatures in the mammalian genome." *Human Molecular Genetics*. https://doi.org/

10.1093/hmg/ddp409.

Horak, Peter, Barbara Klink, Christoph Heining, Stefan Gröschel, Barbara Hutter, Martina Fröhlich, Sebastian Uhrig, et al. 2017. "Precision oncology based on omics data: The NCT Heidelberg experience." *International Journal of Cancer*. https://doi.org/10.1002/ijc.30828.

Hotelling, H. 1933. "Analysis of a complex of statistical variables into principal components." *Journal of Educational Psychology*. https://doi.org/10.1037/h0071325.

Huang, Miller, and William A. Weiss. 2013. "Neuroblastoma and MYCN." *Cold Spring Harbor Perspectives in Medicine*. https://doi.org/10.1101/cshperspect.a014415.

Huang, Sijia, Kumardeep Chaudhary, and Lana X. Garmire. 2017. "More is better: Recent progress in multi-omics data integration methods." https://doi.org/10.3389/fgene.2017.00084.

Huber, Wolfgang, Vincent J. Carey, Robert Gentleman, Simon Anders, Marc Carlson, Benilton S. Carvalho, Hector Corrada Bravo, et al. 2015. "Orchestrating high-throughput genomic analysis with Bioconductor." *Nature Methods*. https://doi.org/10.1038/nmeth.3252.

Hübschmann, Daniel, Lea Jopp-Saile, Carolin Andresen, Stephen Krämer, Zuguang Gu, Christoph E. Heilig, Simon Kreutzfeldt, et al. 2020. "Analysis of mutational signatures with yet another package for signature analysis." *Genes Chromosomes and Cancer*. https://doi.org/10.1002/gcc.22918.

Hyvärinen, A., and E. Oja. 2000. "Independent component analysis: Algorithms and applications." *Neural Networks*. https://doi.org/10.1016/S0893-6080(00)00026-5.

Imrichová, Hana, Gert Hulselmans, Zeynep Kalender Atak, Delphine Potier, and Stein Aerts. 2015. "I-cisTarget 2015 update: Generalized cis-regulatory enrichment analysis in human, mouse and fly." *Nucleic Acids Research*. https://doi.org/10.1093/nar/gkv395.

Jaatinen, Taina, Heidi Hemmoranta, Sampsa Hautaniemi, Jari Niemi, Daniel Nicorici, Jarmo Laine, Olli Yli-Harja, and Jukka Partanen. 2006. " Global Gene Expression Profile of Human Cord Blood-Derived CD133 + Cells ." *Stem Cells* 24 (3): 631–41. https://doi.org/10.1634/stemcells.2005-0185.

Janky, Rekin's, Annelien Verfaillie, Hana Imrichová, Bram van de Sande, Laura Standaert, Valerie Christiaens, Gert Hulselmans, et al. 2014. "iRegulon: From a Gene List to a Gene Regulatory Network Using Large Motif and Track Collections." *PLoS Computational Biology*. https://doi.org/10.1371/journal.pcbi.1003731.

Janoueix-Lerosey, Isabelle, Delphine Lequin, Laurence Brugières, Agnès Ribeiro, Loïc De Pontual, Valérie Combaret, Virginie Raynal, et al. 2008. "Somatic and germline activating mutations of the ALK kinase receptor in neuroblastoma." *Nature*. https://doi.org/10.1038/nature07398.

Jansky, Selina, Ashwini Kumar Sharma, Verena Körber, Andrés Quintero, Umut H. Toprak, Elisa M. Wecht, Moritz Gartlgruber, et al. 2021. "Single-cell transcriptomic analyses provide insights into the developmental origins of neuroblastoma." *Nature Genetics*, March, 1–11. https://doi.org/10.1038/s41588-021-00806-1.

Jassal, Bijay, Lisa Matthews, Guilherme Viteri, Chuqiao Gong, Pascual Lorente, Antonio Fabregat, Konstantinos Sidiropoulos, et al. 2020. "The reactome pathway knowledgebase." *Nucleic Acids Research*. https://doi.org/10.1093/nar/gkz1031.

Jessen, Kristjan R., and Rhona Mirsky. 2019. "Schwann cell precursors; multipotent glial cells in embryonic nerves." https://doi.org/10.3389/fnmol.2019.00069.

Jollife, Ian T., and Jorge Cadima. 2016. "Principal component analysis: A review and recent developments." https://doi.org/10.1098/rsta.2015.0202.

Jones, Philip, David Binns, Hsin Yu Chang, Matthew Fraser, Weizhong Li, Craig McAnulla, Hamish McWilliam, et al. 2014. "InterProScan 5: Genome-scale protein function classification." *Bioinformatics*. https://doi.org/10.1093/bioinformatics/

btu031.

Kairov, Ulykbek, Laura Cantini, Alessandro Greco, Askhat Molkenov, Urszula Czerwinska, Emmanuel Barillot, and Andrei Zinovyev. 2017. "Determining the optimal number of independent components for reproducible transcriptomic data analysis." *BMC Genomics*. https://doi.org/10.1186/s12864-017-4112-9.

Kavitha, K. R., Aiswarya V. Ram, S. Anandu, S. Karthik, Sreeja Kailas, and N. M. Arjun. 2018. "PCA-based gene selection for cancer classification." In *2018 IEEE International Conference on Computational Intelligence and Computing Research, ICCIC 2018*. https://doi.org/10.1109/ICCIC.2018.8782337.

Keogh, Eamonn, and Abdullah Mueen. 2017. "Curse of Dimensionality." In *Encyclopedia of Machine Learning and Data Mining*. https://doi.org/10.1007/978-1-4899-7687-1_192.

Kiselev, Vladimir Yu, Andrew Yiu, and Martin Hemberg. 2018. "Scmap: Projection of single-cell RNA-seq data across data sets." *Nature Methods*. https://doi.org/10.1038/nmeth.4644.

Knudson, A. G., and L. C. Strong. 1972. "Mutation and cancer: neuroblastoma and pheochromocytoma." *American Journal of Human Genetics*.

Kolodziejczyk, Aleksandra A., Jong Kyoung Kim, Valentine Svensson, John C. Marioni, and Sarah A. Teichmann. 2015. "The Technology and Biology of Single-Cell RNA Sequencing." https://doi.org/10.1016/j.molcel.2015.04.005.

Kopka, Joachim, Nicholas Schauer, Stephan Krueger, Claudia Birkemeyer, Björn Usadel, Eveline Bergmüller, Peter Dörmann, et al. 2005. "GMD@CSB.DB: The Golm metabolome database." *Bioinformatics*. https://doi.org/10.1093/bioinformatics/bti236.

Köster, Johannes, and Sven Rahmann. 2012. "Snakemake-a scalable bioinformatics workflow engine." *Bioinformatics*. https://doi.org/10.1093/bioinformatics/bts480.

Lal, Samir, Amy E. McCart Reed, Xavier M. de Luca, and Peter T. Simpson. 2017. "Molecular signatures in breast cancer." https://doi.org/10.1016/j.ymeth.2017.06.032.

Langmead, Ben, and Steven L. Salzberg. 2012. "Fast gapped-read alignment with Bowtie 2." *Nature Methods*. https://doi.org/10.1038/nmeth.1923.

Lara-Astiaso, David, Assaf Weiner, Erika Lorenzo-Vivas, Irina Zaretsky, Diego Adhemar Jaitin, Eyal David, Hadas Keren-Shaul, et al. 2014. "Immunogenetics. Chromatin state dynamics during blood formation." *Science (New York, N.Y.)*. https://doi.org/10.1126/science.1256271.

Lawson, Charles L., and Richard J. Hanson. 1995. *Solving Least Squares Problems*. https://doi.org/10.1137/1.9781611971217.

Lee, Bongshin, Eun Kyoung Choe, Petra Isenberg, Kim Marriott, and John Stasko. 2020. "Reaching Broader Audiences with Data Visualization." *IEEE Computer Graphics and Applications*. https://doi.org/10.1109/MCG.2020.2968244.

Lee, Myeong Sup, Kristina Hanspers, Christopher S. Barker, Abner P. Korn, and Joseph M. McCune. 2004. "Gene expression profiles during human CD4+ T cell differentiation." Int Immunol. https://doi.org/10.1093/intimm/dxh112.

Lee, Su In, and Serafim Batzoglou. 2003. "Application of independent component analysis to microarrays." *Genome Biology*. https://doi.org/10.1186/gb-2003-4-11-r76.

Lenhard, Boris, Albin Sandelin, and Piero Carninci. 2012. "Metazoan promoters: Emerging characteristics and insights into transcriptional regulation." https://doi.org/10.1038/nrg3163.

Lenoir, Tim, and Eric Giannella. 2006. "The emergence and diffusion of DNA microarray technology." *Journal of Biomedical Discovery and Collaboration*. https://doi.org/10.1186/1747-5333-1-11.

Li, Fei, Yang Cao, Lu Han, Xiuliang Cui, Dafei Xie, Shengqi Wang, and Xiaochen Bo. 2013. "Geneexpressionsignature: An r package for discovering functional connections using gene expression signatures." *OMICS A Journal of Integrative Biology*. https://doi.org/10.1089/omi.2012.0087.

Li, Guoliang, Melissa J. Fullwood, Han Xu, Fabianus Hendriyan Mulawadi, Stoyan Velkov, Vinsensius Vega, Pramila Nuwantha Ariyaratne, et al. 2010. "ChIA-PET tool for comprehensive chromatin interaction analysis with paired-end tag sequencing." *Genome Biology*. https://doi.org/10.1186/gb-2010-11-2-r22.

Li, Tao, Jiandong Wang, Huiping Chen, Xinyu Feng, and Feiyue Ye. 2006. "A NMF-based collaborative filtering recommendation algorithm." In *Proceedings of the World Congress on Intelligent Control and Automation (WCICA)*. https://doi.org/10.1109/WCICA.2006.1714249.

Li, Xueqing, Yuanzheng Chen, Cong Fu, Hongwen Li, Kun Yang, Jianhai Bi, and Ran Huo. 2020. "Characterization of epigenetic and transcriptional landscape in infantile hemangiomas with ATAC-seq and RNA-seq." *Epigenomics*. https://doi.org/10.2217/epi-2020-0060.

Li, Yang Eric, Mu Xiao, Binbin Shi, Yu Cheng T. Yang, Dong Wang, Fei Wang, Marco Marcia, and Zhi John Lu. 2017. "Identification of high-confidence RNA regulatory elements by combinatorial classification of RNA-protein binding sites." *Genome Biology* 18 (1): 169. https://doi.org/10.1186/s13059-017-1298-8.

Libbrecht, Maxwell W., and William Stafford Noble. 2015. "Machine learning applications in genetics and genomics." https://doi.org/10.1038/nrg3920.

Liberzon, Arthur, Chet Birger, Helga Thorvaldsdóttir, Mahmoud Ghandi, Jill P. Mesirov, and Pablo Tamayo. 2015. "The Molecular Signatures Database Hallmark Gene Set Collection." *Cell Systems*. https://doi.org/10.1016/j.cels.2015.12.004.

Liberzon, Arthur, Aravind Subramanian, Reid Pinchback, Helga Thorvaldsdóttir, Pablo

Tamayo, and Jill P. Mesirov. 2011. "Molecular signatures database (MSigDB) 3.0." *Bioinformatics*. https://doi.org/10.1093/bioinformatics/btr260.

Lim, Elgene, Di Wu, Bhupinder Pal, Toula Bouras, Marie Liesse Asselin-Labat, François Vaillant, Hideo Yagita, Geoffrey J. Lindeman, Gordon K. Smyth, and Jane E. Visvader. 2010. "Transcriptome analyses of mouse and human mammary cell subpopulations reveal multiple conserved genes and pathways." *Breast Cancer Research* 12 (2). https://doi.org/10.1186/bcr2560.

Lin, Chuang, and Meng Pang. 2015. "Graph Regularized Nonnegative Matrix Factorization with Sparse Coding." *Mathematical Problems in Engineering* 2015. https://doi.org/10.1155/2015/239589.

Lin, Xihui, and Paul C. Boutros. 2020. "Optimization and expansion of non-negative matrix factorization." *BMC Bioinformatics*. https://doi.org/10.1186/s12859-019-3312-5.

Liu, Longqi, Chuanyu Liu, Andrés Quintero, Liang Wu, Yue Yuan, Mingyue Wang, Mengnan Cheng, et al. 2019. "Deconvolution of single-cell multi-omics layers reveals regulatory heterogeneity." *Nature Communications* 10 (1): 470. https://doi.org/10.1038/s41467-018-08205-7.

Lock, Eric F, Katherine A Hoadley, J S Marron, and Andrew B Nobel. 2013. "JOINT AND INDIVIDUAL VARIATION EXPLAINED (JIVE) FOR INTEGRATED ANALYSIS OF MULTIPLE DATA TYPES." *The Annals of Applied Statistics* 7 (1): 523–42. https://doi.org/10.1214/12-AOAS597.

Lotfollahi, Mohammad, Mohsen Naghipourfar, Malte D. Luecken, Matin Khajavi, Maren Büttner, Ziga Avsec, Alexander V. Misharin, and Fabian J. Theis. 2020. "Query to reference single-cell integration with transfer learning." https://doi.org/10.1101/2020.07.16.205997.

Lu, Yan, William Lemon, Peng Yuan Liu, Yijun Yi, Carl Morrison, Ping Yang, Zhifu Sun, et al. 2006. "A gene expression signature predicts survival of patients with stage

I non-small cell lung cancer." *PLoS Medicine*. https://doi.org/10.1371/journal.pmed.0030467.

Luo, Xin, Mengchu Zhou, Yunni Xia, and Qingsheng Zhu. 2014. "An efficient non-negative matrix-factorization-based approach to collaborative filtering for recommender systems." *IEEE Transactions on Industrial Informatics* 10 (2). https://doi.org/10.1109/TII.2014.2308433.

Ma, Sai, Bing Zhang, Lindsay M. LaFave, Andrew S. Earl, Zachary Chiang, Yan Hu, Jiarui Ding, et al. 2020. "Chromatin Potential Identified by Shared Single-Cell Profiling of RNA and Chromatin." *Cell*. https://doi.org/10.1016/j.cell.2020.09.056.

Matthay, Katherine K., John M. Maris, Gudrun Schleiermacher, Akira Nakagawara, Crystal L. Mackall, Lisa Diller, and William A. Weiss. 2016. "Neuroblastoma." *Nature Reviews Disease Primers*. https://doi.org/10.1038/nrdp.2016.78.

Mayr, Lorenz M., and Dejan Bojanic. 2009. "Novel trends in high-throughput screening." https://doi.org/10.1016/j.coph.2009.08.004.

Melo, Jeane C. B., George D. C. Cavalcanti, and Katia S. Guimarães. 2003. "PCA Feature Extraction for Protein Structure Prediction." In *Proceedings of the International Joint Conference on Neural Networks*. https://doi.org/10.1109/ijcnn.2003.1224040.

Meng, Chen, Oana A. Zeleznik, Gerhard G. Thallinger, Bernhard Kuster, Amin M. Gholami, and Aedín C. Culhane. 2016. "Dimension reduction techniques for the integrative analysis of multi-omics data." *Briefings in Bioinformatics*. https://doi.org/10.1093/bib/bbv108.

Meyer, Carl. 2000. *Matrix Analysis and Applied Linear Algebra*. https://doi.org/10.1137/1.9780898719512.

Mirza, Bilal, Wei Wang, Jie Wang, Howard Choi, Neo Christopher Chung, and Peipei

Ping. 2019. "Machine learning and integrative analysis of biomedical big data." https://doi.org/10.3390/genes10020087.

Moerman, Thomas, Sara Aibar Santos, Carmen Bravo González-Blas, Jaak Simm, Yves Moreau, Jan Aerts, and Stein Aerts. 2019. "GRNBoost2 and Arboreto: Efficient and scalable inference of gene regulatory networks." *Bioinformatics*. https://doi.org/10.1093/bioinformatics/bty916.

Moffitt, Richard A, Raoud Marayati, Elizabeth L Flate, Keith E Volmar, S Gabriela Herrera Loeza, Katherine A Hoadley, Naim U Rashid, et al. 2015. "Virtual microdissection identifies distinct tumor- and stroma-specific subtypes of pancreatic ductal adenocarcinoma." *Nature Genetics* 47 (10): 1168–78. https://doi.org/10.1038/ng.3398.

Moignard, Victoria, Steven Woodhouse, Laleh Haghverdi, Andrew J. Lilly, Yosuke Tanaka, Adam C. Wilkinson, Florian Buettner, et al. 2015. "Decoding the regulatory network of early blood development from single-cell gene expression measurements." *Nature Biotechnology*. https://doi.org/10.1038/nbt.3154.

Moon, Kevin R., David van Dijk, Zheng Wang, Scott Gigante, Daniel B. Burkhardt, William S. Chen, Kristina Yim, et al. 2019. "Visualizing structure and transitions in high-dimensional biological data." *Nature Biotechnology*. https://doi.org/10.1038/s41587-019-0336-3.

Moore, Luiza, Daniel Leongamornlert, Tim H. H. Coorens, Mathijs A. Sanders, Peter Ellis, Stefan C. Dentro, Kevin J. Dawson, et al. 2020. "The mutational landscape of normal human endometrial epithelium." *Nature*. https://doi.org/10.1038/s41586-020-2214-z.

Moris, Naomi, Cristina Pina, and Alfonso Martinez Arias. 2016. "Transition states and cell fate decisions in epigenetic landscapes." https://doi.org/10.1038/nrg.2016.98.

Mossé, Yaël P., Marci Laudenslager, Luca Longo, Kristina A. Cole, Andrew Wood, Edward F. Attiyeh, Michael J. Laquaglia, et al. 2008. "Identification of ALK as a major familial neuroblastoma predisposition gene." *Nature*. https://doi.org/10.1038/nature07261.

Mramor, Minca, Gregor Leban, Janez Demšar, and Blaž Zupan. 2005. "Conquering the curse of dimensionality in gene expression cancer diagnosis: Tough problem, simple models." In *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*. https://doi.org/10.1007/11527770_68.

Nakatsukasa, Yuji, and Nicholas J. Higham. 2013. "Stable and efficient spectral divide and conquer algorithms for the symmetric eigenvalue decomposition and the SVD." *SIAM Journal on Scientific Computing*. https://doi.org/10.1137/120876605.

Nevins, Joseph R., and Anil Potti. 2007. "Mining gene expression profiles: Expression signatures as cancer phenotypes." https://doi.org/10.1038/nrg2137.

Newman, Aaron M., Chih Long Liu, Michael R. Green, Andrew J. Gentles, Weiguo Feng, Yue Xu, Chuong D. Hoang, Maximilian Diehn, and Ash A. Alizadeh. 2015. "Robust enumeration of cell subsets from tissue expression profiles." *Nature Methods*. https://doi.org/10.1038/nmeth.3337.

Oldridge, Derek A., Andrew C. Wood, Nina Weichert-Leahey, Ian Crimmins, Robyn Sussman, Cynthia Winter, Lee D. McDaniel, et al. 2015. "Genetic predisposition to neuroblastoma mediated by a LMO1 super-enhancer polymorphism." *Nature*. https://doi.org/10.1038/nature15540.

Olivier, Michael, Reto Asmis, Gregory A. Hawkins, Timothy D. Howard, and Laura A. Cox. 2019. "The need for multi-omics biomarker signatures in precision medicine." https://doi.org/10.3390/ijms20194781.

Ovchinnikova, Svetlana, and Simon Anders. 2020. "Exploring dimension-reduced embed-

dings with Sleepwalk." *Genome Research*. https://doi.org/10.1101/gr.251447.119.

Pal, Sharmistha, Yingtao Bi, Luke MacYszyn, Louise C. Showe, Donald M. O'Rourke, and Ramana V. Davuluri. 2014. "Isoform-level gene signature improves prognostic stratification and accurately classifies glioblastoma subtypes." *Nucleic Acids Research* 42 (8): e64. https://doi.org/10.1093/nar/gku121.

Pearson, Karl. 1901. " LIII. On lines and planes of closest fit to systems of points in space ." *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*. https://doi.org/10.1080/14786440109462720.

Peifer, Martin, Falk Hertwig, Frederik Roels, Daniel Dreidax, Moritz Gartlgruber, Roopika Menon, Andrea Krämer, et al. 2015. "Telomerase activation by genomic rearrangements in high-risk neuroblastoma." *Nature*. https://doi.org/10.1038/nature14980.

Pfeil, Jacob, Lauren M. Sanders, Ioannis Anastopoulos, A. Geoffrey Lyle, Alana S. Weinstein, Yuanqing Xue, Andrew Blair, et al. 2020. "Hydra: A mixture modeling framework for subtyping pediatric cancer cohorts using multimodal gene expression signatures." *PLoS Computational Biology*. https://doi.org/10.1371/journal.pcbi.1007753.

Pugh, Trevor J., Olena Morozova, Edward F. Attiyeh, Shahab Asgharzadeh, Jun S. Wei, Daniel Auclair, Scott L. Carter, et al. 2013. "The genetic landscape of high-risk neuroblastoma." *Nature Genetics*. https://doi.org/10.1038/ng.2529.

Qiu, Xiaojie, Qi Mao, Ying Tang, Li Wang, Raghav Chawla, Hannah A. Pliner, and Cole Trapnell. 2017. "Reversed graph embedding resolves complex single-cell trajectories." *Nature Methods*. https://doi.org/10.1038/nmeth.4402.

Qiu, Yixuan, Jiebiao Wang, Jing Lei, and Kathryn Roeder. 2020. "Identification of cell-type-specific marker genes from co-expression patterns in tissue samples." https://doi.org/10.1101/2020.11.07.373043.

Quintero, Andres, Daniel Hübschmann, Nils Kurzawa, Sebastian Steinhauser, Philipp Rentzsch, Stephen Krämer, Carolin Andresen, et al. 2020. "ShinyButchR: Interactive NMF-based decomposition workflow of genome-scale datasets." *Biology Methods and Protocols* 5 (1). https://doi.org/10.1093/biomethods/bpaa022.

R Core Team. 2020. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.r-project.org/.

Rada-Iglesias, Alvaro, Ruchi Bajpai, Tomek Swigut, Samantha A. Brugmann, Ryan A. Flynn, and Joanna Wysocka. 2011. "A unique chromatin signature uncovers early developmental enhancers in humans." *Nature*. https://doi.org/10.1038/nature09692.

Rahman, R., Y. Xiong, J. G. C. van Hasselt, J. Hansen, E. A. Sobie, M. R. Birtwistle, E. Azeloglu, R. Iyengar, and A. Schlessinger. 2020. "Protein structure-based gene expression signatures." https://doi.org/10.1101/2020.06.03.133066.

Rajbhandari, Presha, Gonzalo Lopez, Claudia Capdevila, Beatrice Salvatori, Jiyang Yu, Ruth Rodriguez-Barrueco, Daniel Martinez, et al. 2018. "Cross-cohort analysis identifies a TEAD4–MYCN positive feedback loop as the core regulatory element of high-risk neuroblastoma." *Cancer Discovery*. https://doi.org/10.1158/2159-8290.CD-16-0861.

Rao, Suhas S. P., Miriam H. Huntley, Neva C. Durand, Elena K. Stamenova, Ivan D. Bochkov, James T. Robinson, Adrian L. Sanborn, et al. 2014. "A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping." *Cell*. https://doi.org/10.1016/j.cell.2014.11.021.

Rezanejad bardaji, Hajar, Malek Hossein Asadi, and Mohammad Mehdi Yaghoobi. 2018. "Long noncoding RNA VIM-AS1 promotes colorectal cancer progression and metastasis by inducing EMT." *European Journal of Cell Biology* 97 (4). https://doi.org/10.1016/j.ejcb.2018.04.004.

Rinaudo, Philippe, Samia Boudah, Christophe Junot, and Etienne A. Thévenot. 2016. "biosigner: A new method for the discovery of significant molecular signatures from Omics data." *Frontiers in Molecular Biosciences*. https://doi.org/10.3389/fmolb.2016.00026.

Ross, Robert A., Barbara A. Spengler, and June L. Biedler. 1983. "Coordinate Morphological and Biochemical Interconversion of Human Neuroblastoma Cells." *Journal of the National Cancer Institute*. https://doi.org/10.1093/jnci/71.4.741.

Rouillard, Andrew D., Gregory W. Gundersen, Nicolas F. Fernandez, Zichen Wang, Caroline D. Monteiro, Michael G. McDermott, and Avi Ma'ayan. 2016. "The harmonizome: a collection of processed datasets gathered to serve and mine knowledge about genes and proteins." *Database : The Journal of Biological Databases and Curation*. https://doi.org/10.1093/database/baw100.

Rousseeuw, Peter J. 1987. "Silhouettes: A graphical aid to the interpretation and validation of cluster analysis." *Journal of Computational and Applied Mathematics*. https://doi.org/10.1016/0377-0427(87)90125-7.

Rozenblatt-Rosen, Orit, Michael J. T. Stubbington, Aviv Regev, and Sarah A. Teichmann. 2017. "The Human Cell Atlas: From vision to reality." https://doi.org/10.1038/550451a.

Saei, Amir Ata, Christian Michel Beusch, Alexey Chernobrovkin, Pierre Sabatier, Bo Zhang, Ülkü Güler Tokat, Eleni Stergiou, Massimiliano Gaetani, Ákos Végvári, and Roman A. Zubarev. 2019. "ProTargetMiner as a proteome signature library of anticancer molecules for functional discovery." *Nature Communications*. https://doi.org/10.1038/s41467-019-13582-8.

Sagar, and Dominic Grün. 2020. "Deciphering Cell Fate Decision by Integrated Single-Cell Sequencing Analysis." *Annual Review of Biomedical Data Science*. https://doi.org/10.1146/annurev-biodatasci-111419-091750.

Salazar, Ramon, Paul Roepman, Gabriel Capella, Victor Moreno, Iris Simon, Christa Dreezen, Adriana Lopez-Doriga, et al. 2011. "Gene expression signature to improve prognosis prediction of stage II and III colorectal cancer." *Journal of Clinical Oncology*. https://doi.org/10.1200/JCO.2010.30.1077.

Sarkans, Ugis, Mikhail Gostev, Awais Athar, Ehsan Behrangi, Olga Melnichuk, Ahmed Ali, Jasmine Minguet, et al. 2018. "The BioStudies database-one stop shop for all data supporting a life sciences study." *Nucleic Acids Research*. https://doi.org/10.1093/nar/gkx965.

Saunders, Arpiar, Evan Z. Macosko, Alec Wysoker, Melissa Goldman, Fenna M. Krienen, Heather de Rivera, Elizabeth Bien, et al. 2018. "Molecular Diversity and Specializations among the Cells of the Adult Mouse Brain." *Cell* 174 (4): 1015–1030.e16. https://doi.org/10.1016/j.cell.2018.07.028.

Schneider, Robert, Andrew J. Bannister, Fiona A. Myers, Alan W. Thorne, Colyn Crane-Robinson, and Tony Kouzarides. 2004. "Histone H3 lysine 4 methylation patterns in higher eukaryotic genes." *Nature Cell Biology*. https://doi.org/10.1038/ncb1076.

Schramm, Alexander, Johannes Köster, Yassen Assenov, Kristina Althoff, Martin Peifer, Ellen Mahlow, Andrea Odersky, et al. 2015. "Mutational dynamics between primary and relapse neuroblastomas." *Nature Genetics*. https://doi.org/10.1038/ng.3349.

Schumann, Franziska, Eric Blanc, Clemens Messerschmidt, Thomas Blankenstein, Antonia Busse, and Dieter Beule. 2019. "SigsPack, a package for cancer mutational signatures." *BMC Bioinformatics*. https://doi.org/10.1186/s12859-019-3043-7.

Schwab, Manfred, Kari Alitalo, Karl Heinz Klempnauer, Harold E. Varmus, J. Michael Bishop, Fred Gilbert, Garrett Brodeur, Milton Goldstein, and Jeffrey Trent. 1983. "Amplified DNA with limited homology to myc cellular oncogene is shared by human neuroblastoma cell lines and a neuroblastoma tumour." *Nature*. https://doi.org/

10.1038/305245a0.

Schwab, Manfred, Frank Westermann, Barbara Hero, and Frank Berthold. 2003. "Neuroblastoma: Biology and molecular and chromosomal pathology." https://doi.org/10.1016/S1470-2045(03)01166-5.

Seijo, Luis M., Nir Peled, Daniel Ajona, Mattia Boeri, John K. Field, Gabriella Sozzi, Ruben Pio, et al. 2019. "Biomarkers in Lung Cancer Screening: Achievements, Promises, and Challenges." https://doi.org/10.1016/j.jtho.2018.11.023.

Seung, H. Sebastian, and Daniel D. Lee. 1999. "Learning the parts of objects by non-negative matrix factorization." *Nature* 401 (6755): 788–91. https://doi.org/10.1038/44565.

Shafer, Maxwell E. R. 2019. "Cross-Species Analysis of Single-Cell Transcriptomic Data." https://doi.org/10.3389/fcell.2019.00175.

Shahbazi, Marta N., and Magdalena Zernicka-Goetz. 2018. "Deconstructing and reconstructing the mouse and human early embryo." https://doi.org/10.1038/s41556-018-0144-x.

Shao, Chunxuan, and Thomas Höfer. 2017. "Robust classification of single-cell transcriptome data by nonnegative matrix factorization." *Bioinformatics* 33 (2): 235–42. https://doi.org/10.1093/bioinformatics/btw607.

Shema, Efrat, Bradley E. Bernstein, and Jason D. Buenrostro. 2019. "Single-cell and single-molecule epigenomics to uncover genome regulation at unprecedented resolution." https://doi.org/10.1038/s41588-018-0290-x.

Shi, Xiaoxi. 2020. "A Hybrid Slope One Collaborative Filtering Algorithm Based on Nonnegative Matrix Factorization." In *ACM International Conference Proceeding Series*. https://doi.org/10.1145/3396474.3396496.

Shimada, Hiroyuki, Inge M. Ambros, Louis P. Dehner, Jun-ichi Hata, Vijay V. Joshi, and Borghild Roald. 1999. "Terminology and morphologic criteria of neuroblas-

tic tumors." *Cancer*. https://doi.org/10.1002/(sici)1097-0142(19990715)86:2%3C349::aid-cncr20%3E3.3.co;2-p.

Shimada, Hiroyuki, Jane Chatten, William A. Newton, Nancy Sachs, Ala B. Hamoudi, Tsuneo Chiba, Henry B. Marsden, and Kazuaki Misugi. 1984. "Histopathologic prognostic factors in neuroblastic tumors: Definition of subtypes of ganglioneuroblastoma and an age-linked classification of neuroblastomas." *Journal of the National Cancer Institute*. https://doi.org/10.1093/jnci/73.2.405.

Slezak, Tom, Bradley Hart, and Crystal Jaing. 2019. "Design of genomic signatures for pathogen identification and characterization." In *Microbial Forensics*. https://doi.org/10.1016/B978-0-12-815379-6.00020-9.

Sompairac, Nicolas, Petr V. Nazarov, Urszula Czerwinska, Laura Cantini, Anne Biton, Askhat Molkenov, Zhaxybay Zhumadilov, et al. 2019. "Independent component analysis for unraveling the complexity of cancer omics datasets." https://doi.org/10.3390/ijms20184414.

Sonawane, Abhijeet Rajendra, John Platig, Maud Fagny, Cho Yi Chen, Joseph Nathaniel Paulson, Camila Miranda Lopes-Ramos, Dawn Lisa DeMeo, John Quackenbush, Kimberly Glass, and Marieke Lydia Kuijjer. 2017. "Understanding Tissue-Specific Gene Regulation." *Cell Reports*. https://doi.org/10.1016/j.celrep.2017.10.001.

Song, Dongyuan, Kexin Aileen Li, Zachary Hemminger, Roy Wollman, and Jingyi Jessica Li. 2021. "scPNMF: sparse gene encoding of single cells to facilitate gene selection for targeted gene profiling." *bioRxiv*, February, 2021.02.09.430550. https://doi.org/10.1101/2021.02.09.430550.

Song, Wei, and Xuesong Li. 2019. "A Non-Negative Matrix Factorization for Recommender Systems Based on Dynamic Bias." In *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*. https://doi.org/10.1007/978-3-030-26773-5_14.

Sotiriou, Christos, and Martine J. Piccart. 2007. "Taking gene-expression profiling to the clinic: When will molecular signatures become relevant to patient care?" https://doi.org/10.1038/nrc2173.

Sotiriou, Christos, and Lajos Pusztai. 2009. "Gene-Expression Signatures in Breast Cancer." *New England Journal of Medicine* 360 (8). https://doi.org/10.1056/nejmra0801289.

Spitz, François, and Eileen E. M. Furlong. 2012. "Transcription factors: From enhancer binding to developmental control." https://doi.org/10.1038/nrg3207.

Stark, R, and J Norden. 2020. *SigCheck: Check a gene signature's prognostic performance against random signatures, known signatures, and permuted data/metadata.* https://www.bioconductor.org/packages/release/bioc/html/SigCheck.html.

Stegle, Oliver, Sarah A. Teichmann, and John C. Marioni. 2015. "Computational and analytical challenges in single-cell transcriptomics." https://doi.org/10.1038/nrg3833.

Strazar, Martin, Marinka Zitnik, Blaz Zupan, Jernej Ule, and Tomas Curk. 2016. "Orthogonal matrix factorization enables integrative analysis of multiple RNA binding proteins." *Bioinformatics*. https://doi.org/10.1093/bioinformatics/btw003.

Stuart, Tim, Andrew Butler, Paul Hoffman, Christoph Hafemeister, Efthymia Papalexi, William M. Mauck, Yuhan Hao, Marlon Stoeckius, Peter Smibert, and Rahul Satija. 2019. "Comprehensive Integration of Single-Cell Data." *Cell*. https://doi.org/10.1016/j.cell.2019.05.031.

Stuart, Tim, and Rahul Satija. 2019. "Integrative single-cell analysis." https://doi.org/10.1038/s41576-019-0093-7.

Subramanian, Aravind, Pablo Tamayo, Vamsi K Mootha, Sayan Mukherjee, Benjamin L Ebert, Michael A Gillette, Amanda Paulovich, et al. 2005. "Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles."

Proceedings of the National Academy of Sciences of the United States of America 102 (43): 15545–50. https://doi.org/10.1073/pnas.0506580102.

Sudarikov, Konstantin, Alexander Tyakht, and Dmitry Alexeev. 2017. "Methods for The Metagenomic Data Visualization and Analysis." https://doi.org/10.21775/cimb.024.037.

Sung, Jaeyun, Yuliang Wang, Sriram Chandrasekaran, Daniela M. Witten, and Nathan D. Price. 2012. "Molecular signatures from omics data: From chaos to consensus." https://doi.org/10.1002/biot.201100305.

Szymczak, F., M. L. Colli, M. J. Mamula, C. Evans-Molina, and D. L. Eizirik. 2021. "Gene expression signatures of target tissues in type 1 diabetes, lupus erythematosus, multiple sclerosis, and rheumatoid arthritis." *Science Advances* 7 (2). https://doi.org/10.1126/sciadv.abd7600.

Tam, Patrick P. L., and Joshua W. K. Ho. 2020. "Cellular diversity and lineage trajectory: Insights from mouse single cell transcriptomes." *Development (Cambridge)*. https://doi.org/10.1242/dev.179788.

Tan, Vincent Y. F., and Cédric Févotte. 2013. "Automatic relevance determination in nonnegative matrix factorization with the (β)-divergence." *IEEE Transactions on Pattern Analysis and Machine Intelligence*. https://doi.org/10.1109/TPAMI.2012.240.

Teng, Li, Bing He, Jiahui Wang, and Kai Tan. 2015. "4DGenome: A comprehensive database of chromatin interactions." *Bioinformatics*. https://doi.org/10.1093/bioinformatics/btv158.

Thompson, Dawn, Aviv Regev, and Sushmita Roy. 2015. "Comparative Analysis of Gene Regulatory Networks: From Network Reconstruction to Evolution." *Annual Review of Cell and Developmental Biology*. https://doi.org/10.1146/annurev-cellbio-100913-012908.

Townes, F. William, Stephanie C. Hicks, Martin J. Aryee, and Rafael A. Irizarry. 2019.

"Feature selection and dimension reduction for single-cell RNA-Seq based on a multinomial model." *Genome Biology*. https://doi.org/10.1186/s13059-019-1861-6.

Trapnell, Cole. 2015. "Defining cell types and states with single-cell genomics." https://doi.org/10.1101/gr.190595.115.

Trochet, Delphine, Franck Bourdeaut, Isabelle Janoueix-Lerosey, Anne Deville, Loïc De Pontual, Gudrun Schleiermacher, Carole Coze, et al. 2004. "Germline Mutations of the Paired-Like Homeobox 2B (PHOX2B) Gene in Neuroblastoma." *American Journal of Human Genetics*. https://doi.org/10.1086/383253.

Tsang, Julia Y. S., and Gary M. Tse. 2020. "Molecular Classification of Breast Cancer." https://doi.org/10.1097/PAP.0000000000000232.

Uhlen, Mathias, Cheng Zhang, Sunjae Lee, Evelina Sjöstedt, Linn Fagerberg, Gholamreza Bidkhori, Rui Benfeitas, et al. 2017. "A pathology atlas of the human cancer transcriptome." *Science*. https://doi.org/10.1126/science.aan2507.

Valentijn, Linda J., Jan Koster, Danny A. Zwijnenburg, Nancy E. Hasselt, Peter Van Sluis, Richard Volckmann, Max M. Van Noesel, et al. 2015. "TERT rearrangements are frequent in neuroblastoma and identify aggressive tumors." *Nature Genetics*. https://doi.org/10.1038/ng.3438.

Van de Sande, Bram, Christopher Flerin, Kristofer Davie, Maxime De Waegeneer, Gert Hulselmans, Sara Aibar, Ruth Seurinck, et al. 2020. "A scalable SCENIC workflow for single-cell gene regulatory network analysis." *Nature Protocols*. https://doi.org/10.1038/s41596-020-0336-2.

Van Der Maaten, Laurens, and Geoffrey Hinton. 2008. "Visualizing data using t-SNE." *Journal of Machine Learning Research*.

Van Groningen, Tim, Jan Koster, Linda J. Valentijn, Danny A. Zwijnenburg, Nurdan Akogul, Nancy E. Hasselt, Marloes Broekmans, et al. 2017. "Neuroblastoma is composed of two super-enhancer-associated differentiation states." *Nature Genetics*.

https://doi.org/10.1038/ng.3899.

Vijver, Marc J. van de, Yudong D. He, Laura J. van 't Veer, Hongyue Dai, Augustinus A. M. Hart, Dorien W. Voskuil, George J. Schreiber, et al. 2002. "A Gene-Expression Signature as a Predictor of Survival in Breast Cancer." *New England Journal of Medicine*. https://doi.org/10.1056/nejmoa021967.

Vivian, John, Arjun Arkal Rao, Frank Austin Nothaft, Christopher Ketchum, Joel Armstrong, Adam Novak, Jacob Pfeil, et al. 2017. "Toil enables reproducible, open source, big biomedical data analyses." https://doi.org/10.1038/nbt.3772.

Wagner, Florian. 2015. "GO-PCA: An unsupervised method to explore gene expression data using prior knowledge." *PLoS ONE*. https://doi.org/10.1371/journal.pone.0143196.

Wang, Chuan Yuan, Jin Xing Liu, Na Yu, and Chun Hou Zheng. 2019. "Sparse Graph Regularization Non-Negative Matrix Factorization Based on Huber Loss Model for Cancer Data Analysis." *Frontiers in Genetics*. https://doi.org/10.3389/fgene.2019.01054.

Wang, Lipo, Yaoli Wang, and Qing Chang. 2016. "Feature selection methods for big data bioinformatics: A survey from the search perspective." https://doi.org/10.1016/j.ymeth.2016.08.014.

Wang, Mei Neng, Zhu Hong You, Li Ping Li, Leon Wong, Zhan Heng Chen, and Cheng Zhi Gan. 2020. "GNMFLMI: Graph Regularized Nonnegative Matrix Factorization for Predicting LncRNA-MiRNA Interactions." *IEEE Access*. https://doi.org/10.1109/ACCESS.2020.2974349.

Wang, Zishuai, Xikang Feng, and Shuai Cheng Li. 2019. "SCDevDB: A database for insights into single-cell gene expression profiles during human developmental processes." *Frontiers in Genetics*. https://doi.org/10.3389/fgene.2019.00903.

Weiner, January. 2017. *riverplot: Sankey or Ribbon Plots*. https://cran.r-project.

org/package=riverplot.

Weiss, William A., Ken Aldape, Gayatry Mohapatra, Burt G. Feuerstein, and J. Michael Bishop. 1997. "Targeted expression of MYCN causes neuroblastoma in transgenic mice." *EMBO Journal*. https://doi.org/10.1093/emboj/16.11.2985.

Welch, Joshua D., Velina Kozareva, Ashley Ferreira, Charles Vanderburg, Carly Martin, and Evan Z. Macosko. 2019. "Single-Cell Multi-omic Integration Compares and Contrasts Features of Brain Cell Identity." *Cell* 177 (7): 1873–1887.e17. https://doi.org/10.1016/j.cell.2019.05.006.

Whyte, Warren A., David A. Orlando, Denes Hnisz, Brian J. Abraham, Charles Y. Lin, Michael H. Kagey, Peter B. Rahl, Tong Ihn Lee, and Richard A. Young. 2013. "Master transcription factors and mediator establish super-enhancers at key cell identity genes." *Cell*. https://doi.org/10.1016/j.cell.2013.03.035.

Wickham, Hadley. 2011. "Testthat: Get started with testing." *R Journal* 3 (1). https://doi.org/10.32614/rj-2011-002.

Wildey, Mary Jo, Anders Haunso, Matthew Tudor, Maria Webb, and Jonathan H. Connick. 2017. "High-Throughput Screening." In *Annual Reports in Medicinal Chemistry*. https://doi.org/10.1016/bs.armc.2017.08.004.

Wilson, Kevin W., Bhiksha Raj, Paris Smaragdis, and Ajay Divakaran. 2008. "Speech denoising using nonnegative matrix factorization with priors." In *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*. https://doi.org/10.1109/ICASSP.2008.4518538.

Wishart, David S., Dan Tzur, Craig Knox, Roman Eisner, An Chi Guo, Nelson Young, Dean Cheng, et al. 2007. "HMDB: The human metabolome database." *Nucleic Acids Research*. https://doi.org/10.1093/nar/gkl923.

Wolf, F. Alexander, Philipp Angerer, and Fabian J. Theis. 2018. "SCANPY: Large-scale single-cell gene expression data analysis." *Genome Biology*. https://doi.org/10.

1186/s13059-017-1382-0.

Wong, Ka Chun, Yue Li, and Zhaolei Zhang. 2016. "Unsupervised learning in genome informatics." In *Unsupervised Learning Algorithms*. https://doi.org/10.1007/978-3-319-24211-8_15.

Wu, Siqi, Antony Joseph, Ann S. Hammonds, Susan E. Celniker, Bin Yu, and Erwin Frise. 2016. "Stability-driven nonnegative matrix factorization to interpret Spatial gene expression and build local gene networks." *Proceedings of the National Academy of Sciences of the United States of America*. https://doi.org/10.1073/pnas.1521171113.

Xiao, Qiu, Jiawei Luo, Cheng Liang, Jie Cai, and Pingjian Ding. 2018. "A graph regularized non-negative matrix factorization method for identifying microRNA-disease associations." *Bioinformatics*. https://doi.org/10.1093/bioinformatics/btx545.

Xu, Chunming, and Scott A. Jackson. 2019. "Machine learning and complex biological data." https://doi.org/10.1186/s13059-019-1689-0.

Xu, Weifeng, Beibei Chen, Dianshan Ke, and Xiaobing Chen. 2020. "TRIM29 mediates lung squamous cell carcinoma cell metastasis by regulating autophagic degradation of E-cadherin." *Aging*. https://doi.org/10.18632/aging.103451.

Yang, Zi, and George Michailidis. 2015. "A non-negative matrix factorization method for detecting modules in heterogeneous omics multi-modal data." *Bioinformatics* 32 (1): btv544. https://doi.org/10.1093/bioinformatics/btv544.

Yao, Lijing, Benjamin P. Berman, and Peggy J. Farnham. 2015. "Demystifying the secret mission of enhancers: Linking distal regulatory elements to target genes." *Critical Reviews in Biochemistry and Molecular Biology*. https://doi.org/10.3109/10409238.2015.1087961.

Yu, Na, Ying Lian Gao, Jin Xing Liu, Juan Wang, and Junliang Shang. 2019. "Robust hypergraph regularized non-negative matrix factorization for sample clustering and

feature selection in multi-view gene expression data." *Human Genomics*. https://doi.org/10.1186/s40246-019-0222-6.

Zhang, Sheng, Weihong Wang, James Ford, and Fillia Makedon. 2006. "Learning from incomplete ratings using non-negative matrix factorization." In *Proceedings of the Sixth SIAM International Conference on Data Mining*. https://doi.org/10.1137/1.9781611972764.58.

Zhang, Shihua, Chun-Chi Liu, Wenyuan Li, Hui Shen, Peter W Laird, and Xianghong Jasmine Zhou. 2012. "Discovery of multi-dimensional modules by integrative analysis of cancer genomic data." *Nucleic Acids Research* 40 (19): 9379–91. https://doi.org/10.1093/nar/gks725.

Zhang, Yong, Tao Liu, Clifford A. Meyer, Jérôme Eeckhoute, David S. Johnson, Bradley E. Bernstein, Chad Nussbaum, et al. 2008. "Model-based analysis of ChIP-Seq (MACS)." *Genome Biology*. https://doi.org/10.1186/gb-2008-9-9-r137.

Zhang, Ze, Danni Luo, Xue Zhong, Jin Huk Choi, Yuanqing Ma, Stacy Wang, Elena Mahrt, et al. 2019. "SCINA: A Semi-Supervised Subtyping Algorithm of Single Cells and Bulk Samples." *Genes*.

Zhu, Chenxu, Miao Yu, Hui Huang, Ivan Juric, Armen Abnousi, Rong Hu, Jacinta Lucero, M. Margarita Behrens, Ming Hu, and Bing Ren. 2019. "An ultra high-throughput method for single-cell joint analysis of open chromatin and transcriptome." *Nature Structural and Molecular Biology*. https://doi.org/10.1038/s41594-019-0323-x.