HELMHOLTZ ASSOCIATION
Alliance on Systems Biology
HelmholtzZentrum münchen
German Research Center for Environmental Health

TECHNISCHE UNIVERSITÄT MÜNCHEN

Enriching the characterization of complex clinical and molecular phenotypes with deep learning

Gökcen Eraslan

TECHNISCHE UNIVERSITÄT MÜNCHEN

TUM School of Life Sciences

Enriching the characterization of complex clinical and molecular phenotypes with deep learning

Gökcen Eraslan

Vollständiger Abdruck der von der TUM School of Life Sciences der Technischen Universität München zur Erlangung des akademischen Grades eines

Doktors der Naturwissenschaften (Dr. rer. nat.)

genehmigten Dissertation.

Vorsitzender:

Prof. Dr. Dmitrij Frishman

Prüfer der Dissertation:

1. Prof. Dr. Dr. Fabian J. Theis
2. Prof. Dr. Frank Johannes

Die Dissertation wurde am 24.06.2021 bei der Technischen Universität München eingereicht und durch die TUM School of Life Sciences am 15.09.2021 angenommen.

Acknowledgments

Foremost, I am eternally grateful to Nikola Müller and Fabian Theis for their tremendous support and for allowing me to be a part of an excellent team with great diversity, knowledge, and creativity. I would also like to thank all ICB members for the welcoming environment and inspiring discussions in seminars.

I consider being a member of the CCM and ICB a real privilege. I want to thank all members of the institute for making my Ph.D. journey a pleasant memory to remember. I am grateful to QBM organizers, lecturers, and principal investigators for successfully designing and implementing this graduate program focusing on collaborative projects. I would also like to thank HELENA for promoting international networking and lab visits of graduate students with generous financial support.

Finally, I would like to thank Başak, my parents, and my brother for the immense joy, wisdom, and enlightenment they bring into my life.

Abstract

With the recent advances in high-throughput experimental techniques, the volume and diversity of biological data reached the highest level. Today this data regime allows us to accomplish many critical tasks in biology via data-driven, genome-wide models. The recent breakthroughs in neural networks led to highly expressive machine learning models that have become powerful tools in data science. Due to the characteristics of biological data modalities and the domain-specific challenges, the necessity for machine learning algorithms tailored for genomics data is more significant than ever before. In this thesis, we present two novel algorithms we developed to bridge this gap and improve the characterization of biological datasets using expressive deep learning models. These algorithms address major challenges in delineating the sources of variation in molecular and clinical phenotypes in human genetics and single-cell genomics.

The first method is a novel machine learning-based variant prioritization approach to identify non-coding variants that potentially play a critical role in the underlying biology of complex diseases and traits, presumably by modulating the regulatory circuitry. Our approach, DeepWAS, uses deep neural networks to facilitate generating hypotheses about potential cell types and regulatory elements underlying complex diseases and traits. Leveraging pre-trained neural networks for predicting transcription factor (TF) binding sites from DNA sequences allowed us to estimate the regulatory effect of non-coding variants. Subsequently, we used potentially regulatory variants for modeling genotype-phenotype associations with a robust multivariate variable selection method. We applied DeepWAS to complex phenotypes and diseases like multiple sclerosis, major depressive disorder, and height to generate testable hypotheses where potential cell types and regulatory elements are involved.

The second method is an unsupervised machine learning algorithm, DCA, to refine the representations of single cells, which are impaired by the substantial noise in the single-cell RNA-seq measurement process due to technical factors. We utilized autoencoders to capture the structure of the data by compressing the data and representing cells with a few hidden variables. Reconstructing the input data using only these core features omits unimportant patterns in data and produces a denoised output, where the biological signal is accentuated. We tailored the noise model of this scalable denoising method to the characteristics of single-cell data, such as the sparsity and the count structure. We demonstrated that denoising improves typical single-cell downstream tasks using simulated and real datasets. Our method thus facilitates understanding the heterogeneity of cell populations.

In summary, we developed data-driven methods to improve the characterization of clinical and molecular phenotypes from regulatory and single-cell genomics perspectives; at different scales, ranging from cell populations to human populations.

List of publications

This thesis is based on the methods and results reported in the following publications. Detailed author contributions are given at the beginning of each chapter.

Chapter 3

- Janine Arloth*, **Gökcen Eraslan***, Till F. M. Andlauer, Jade Martins, Stella Iurato, Brigitte Kühnel, Melanie Waldenberger, Josef Frank, Ralf Gold, Bernhard Hemmer, Felix Luessi, Sandra Nischwitz, Friedemann Paul, Heinz Wiendl, Christian Gieger, Stefanie Heilmann-Heimbach, Tim Kacprowski, Matthias Laudes, Thomas Meitinger, Annette Peters, Rajesh Rawal, Konstantin Strauch, Susanne Lucae, Bertram Müller-Myhsok, Marcella Rietschel, Fabian J. Theis, Elisabeth B. Binder, Nikola S. Mueller *DeepWAS: Multivariate genotype-phenotype associations by directly integrating regulatory information using deep learning* (PLoS Computational Biology 16, no. 2 (2020): e1007616, `https://doi.org/10.1371/journal.pcbi.1007616`)

- **Gökcen Eraslan**, Nikola S. Mueller, Fabian J. Theis. *Misina: Finding microRNA-mediated effects of genetic variants with an integrative approach* (in preparation)

- Nikola Müller, Ivan Kondofersky, **Gökcen Eraslan**, Karolina Worf, Fabian J. Theis. "Bioinformatics in Psychiatric Genetics." *Psychiatric Genetics: A Primer for Clinical and Basic Scientists* (2018) `https://doi.org/10.1093/med/9780190221973.001.0001`

Chapter 4

- **Gökcen Eraslan***, Lukas M. Simon*, Maria Mircea, Nikola S. Mueller, Fabian J. Theis. *Single-cell RNA-seq denoising using a deep count autoencoder*, Nature Communications, 10 (2019): 390, `https://doi.org/10.1038/s41467-018-07931-2`

Further publications

Further peer-reviewed publications contributed by the author are given below:

- **Gökcen Eraslan***, Ziga Avsec*, Julien Gagneur, Fabian J. Theis, *Deep learning: New computational modelling techniques for genomics*, Nature Reviews Genetics, 20, pages 389–403 (2019), `https://doi.org/10.1038/s41576-019-0122-6`

- Darina Czamara, **Gökcen Eraslan**, Christian Page, Jari Lahti, Marius Lahti-Pulkkinen, Esa Hämäläinen, Eero Kajantie, Hannele Laivuori, Pia Villa, Rebecca Reynolds, Wenche Nystad, Siri Håberg, Stephanie London, Kieran O'Donnell, Elika Garg, Michael Meaney, Sonja Entringer, Pathik Wadhwa, Claudia Buss, Meaghan Jones, David Lin, Julia MacIsaac, Michael Kobor, Nastassja Koen, Heather Zar, Karestan Koenen, Shareefa Dalvie, Dan Stein, Ivan Kondofersky, Nikola Mueller, Fabian Theis, Major Depressive Disorder Working Group of the Psychiatric Genomics Consortium, Katri Räikkönen, and Elisabeth Binder. *Integrated analysis of environmental*

and genetic influences on cord blood DNA methylation in new-borns, Nature Communications 10.1 (2019): 2548. https://doi.org/10.1038/s41467-019-10461-0

- Anat Kreimer, Haoyang Zeng, Matthew D. Edwards, Yuchun Guo, Kevin Tian, Sunyoung Shin, Rene Welch, Michael Wainberg, Rahul Mohan, Nicholas A. Sinnott-Armstrong, Yue Li, **Gökcen Eraslan**, Talal Bin Amin, Ryan Tewhey, Pardis C. Sabeti, Jonathan Goke, Nikola S. Mueller, Manolis Kellis, Anshul Kundaje, Michael A Beer, Sunduz Keles, David K. Gifford, Nir Yosef. *Predicting gene expression in massively parallel reporter assays: A comparative study*, Human mutation, 38(9), pp.1240-1250, (2017), https://doi.org/10.1002/humu.23197

*These authors contributed equally.

Contents

Chapter 1

Introduction

Computational biology is the science of characterizing biological systems through computational modeling of the experimental data (Kitano 2002; Kohl et al. 2010). This line of research ambitiously aims to develop a thorough understanding of the wide range of processes and components of biological systems at different resolutions ranging from the molecular phenotypes like gene expression and transcription factor (TF) binding (Deplancke et al. 2016) to the complex clinical phenotypes like multiple sclerosis (Baranzini and Oksenberg 2017) and major depressive disorder (Wray et al. 2018). In computational biology research, modeling is used as an abstraction technique by distilling the utterly complex world of biology into a few concepts, thereby providing a simpler representation of reality. For example, pseudotime inference, a commonly used modeling technique in single-cell genomics, aims to resolve the order of differentiating cells along the differentiation trajectory using the snapshot of gene expression information of every single cell (Haghverdi, Büttner, et al. 2016). It abstracts the extreme complexity of the cell circuitry, which gives rise to the high-dimensional transcriptomic profiles into a single concept: the position along the differentiation axis. Although the abstract view of models cannot fully encapsulate reality, it provides useful approximations of a specific aspect of the data. As George E. P. Box elegantly said: *"All models are wrong, but some are useful."* (Box 1976).

New technologies have been the drivers of life sciences. Robert Hooke, one of the most prolific scientists of the seventeenth century, described a fly's eye and plant cells for the first time in history (Hooke 1961) and pioneered a new field of study with these phenomenal observations. The development of the technology, in this case the compound light microscopes, was the key to the breakthrough. After Hooke, the advances in technology continued to provide new tools for observing the world of molecular biology and for generating new types of measurements and data. In the information age, the completion of human genome sequence and the advent of inexpensive microarrays enabled genotyping individuals at scale and initiated the next generation of genetic studies (Hoheisel 2006). This technology empowered identifying

the sequence variation in the human genome that covaries with a phenotype of interest with sufficient statistical power. These studies, called genome-wide association studies (GWAS), not only fundamentally changed our understanding of complex traits and diseases (Visscher et al. 2017) but also facilitated the functional annotation and interpretation of the human genome (Buniello et al. 2019). Similarly, the rise of a recent disruptive technology, called single-cell genomics, has been revolutionizing how we observe cellular processes today (Wagner et al. 2016) by providing measurements of molecular phenotypes such as gene expression (Macosko et al. 2015) and chromatin accessibility (Buenrostro et al. 2015; Schwartzman and Tanay 2015) at single-cell resolution. Single-cell genomics has been adding new dimensions to our knowledge of fundamental biological concepts such as cellular identity (Wagner et al. 2016), differentiation (Velten et al. 2017; Bach et al. 2017) and organ development (Jun Ding et al. 2018).

Computers and computational techniques have been reshaping biology and how biological questions are formulated (Markowetz 2017). In his book, Life Out of Sequence (Stevens 2013), Hallam Stevens wrote, *"Biology adapted itself to the computer, not the computer to biology"*. This adaptation is driven by the impact of computational concepts such as simulations and modeling on biology. For example, completion of the human genome in the Human Genome Project was enabled by the application of efficient sequence alignment and scaffolding methods to the shotgun sequencing data (Weber and Myers 1997). After the sequencing effort, one of the lead scientists in the project, Eric S. Lander, famously summed up the results as *"Genome: Bought the book; hard to read"*. Researchers in biology have been seeking new ways to "read the book" through computational, mathematical and statistical methods. The genome-wide association and quantitative trait loci (QTL) studies link many clinical and molecular phenotypes to genomic loci thanks to the efficient implementations of fundamental statistical models like linear and logistic regression that can analyze large-scale genotype and gene expression data (Purcell et al. 2007; Shabalin 2012). Furthermore, machine learning and predictive modeling techniques assist the progress of annotating the human genome by extrapolating our knowledge beyond already characterized genomics regions (Ernst and Kellis 2012; Libbrecht and Noble 2015).

Recent theoretical and practical advances in machine learnin, particularly in predictive modeling with deep neural networks led to remarkable applications in many fields, including computer vision and natural language processing (LeCun et al. 2015). For some prediction tasks like image recognition, neural networks are now the default modeling approaches that can perform beyond human-level accuracy (K. He et al. 2015). Moreover, neural networks have been fundamentally transforming many fields by replacing traditional learning algorithms. In computational biology, where machine learning is already an essential tool (Libbrecht and Noble 2015), the first applications of deep neural networks emerged as promising techniques

for sequence analysis and binding prediction of DNA and RNA binding proteins (J. Zhou and Troyanskaya 2015; Alipanahi et al. 2015; Kelley, Snoek, et al. 2016). Deep neural networks, which form the basis of the two methods we propose, are introduced in the next section in detail.

1.1 Artificial neural networks

In the 1950s, the research field of artificial neural networks set out to investigate the organization, dynamics and information storage mechanisms of the human brain by designing computational models of the biological neurons, so-called "brain models" (Rosenblatt 1958; Rosenblatt 1961). This effort subsequently turned into an ambitious goal to perform specific complex tasks such as language translation similarly to how the human brain would perform them, hence simulating the intelligent behavior known as *artificial intelligence* (AI) (Russell and Norvig 2016). The first notable attempt towards this goal was the Perceptron, a special-purpose hardware designed and built by American psychologist Frank Rosenblatt. The Perceptron was a hardware implementation to perform supervised learning where a task is learned as a mapping from an input to an output using pairs of input and output samples (Alpaydin 2009). It used a simple neuron model introduced earlier by McCulloch and Pitts (1943), where a neuron is represented as a computational unit that takes input signals, e.g. pixels of an image, calculates a weighted sum of input features and finally produces a binary value. "On" state of the binary output produced by an artificial neuron, which indeed represents a linear classifier, can be interpreted as "firing" when the received stimuli exceed a threshold. Furthermore, Rosenblatt proposed a simple learning method, called perceptron learning, which relied on iteratively updating the neuron weights until the difference between the desired output and the model output is lower than a predefined threshold (Rosenblatt 1961). It was also shown that the learning algorithm is guaranteed to find a solution where all samples are classified correctly if the samples are linearly separable.

In parallel with the research on artificial neural networks, the quest of exploring the horizons of computation was shaping up by the key players in the field such as Alan Turing who proposed to consider the question "Can machines think?" (Turing 1950). The efforts following up on this question led to significant developments in AI. In the proposal of the Dartmouth workshop in 1956, a two-month workshop on AI organized by John McCarthy, who is considered one of the founders of AI, the aim of the AI research was summarized as: "*to find how to make machines use language, form abstractions and concepts, solve kinds of problems now reserved for humans, and improve themselves.*". For example, Arthur Samuel, who also coined the term artificial intelligence, implemented a checkers program which was able to play at a strong amateur level and disproved the idea that computer programs cannot

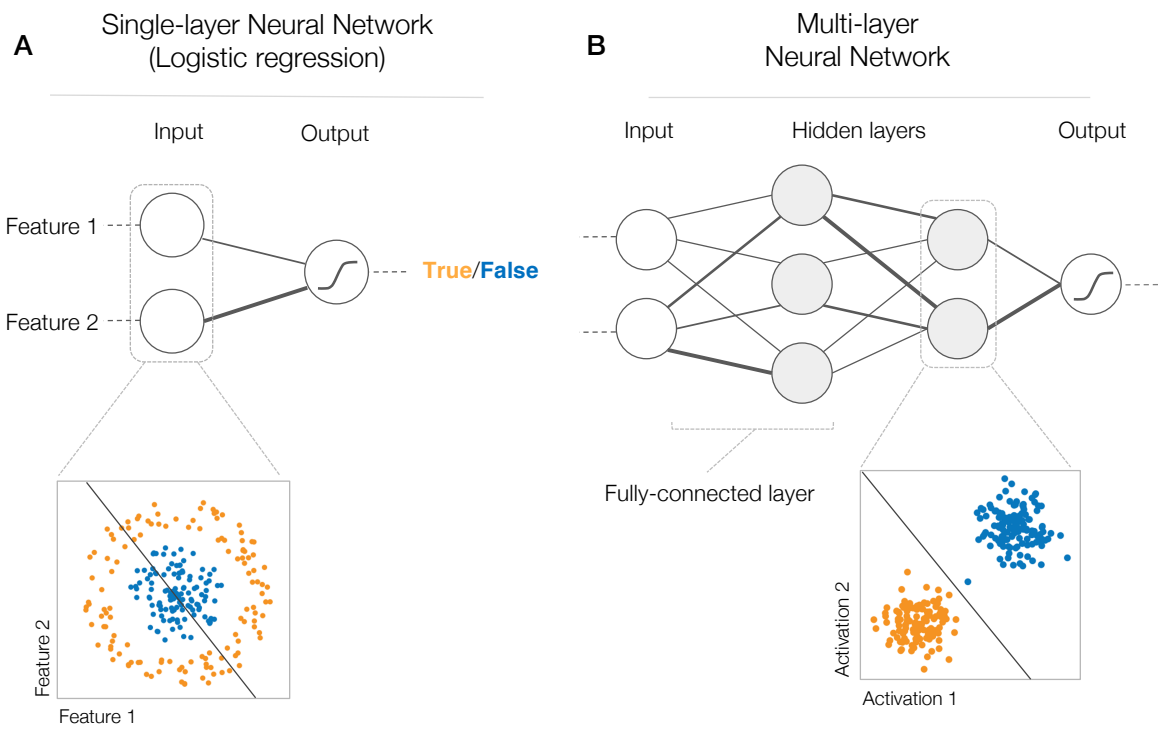


Figure 1.1: Feature learning and nonlinearity aspects of multilayer neural networks compared to logistic regression. **(A)** Logistic regression, which can be interpreted as a neural network with a single layer, fails to classify two linearly non-separable classes. **(B)** Multilayer neural networks, exemplified here with a network with two fully-connected layers, discover new representations of the data where nonlinearities are captured in the hidden layers. This example demonstrates how hidden layer representations are leveraged to make linearly non-separable data separable. The original features of the data and the hidden representation captured in the second hidden layer are shown in scatter plots. The thickness of edges represents edge weights.

play games better than their creators by winning against Samuel (Samuel 1959; Russell and Norvig 2016). Similarly, programs for proving mathematical theorems (Gelernter 1959), and the introduction of genetic algorithms (Friedberg 1958) subsequently raised the expectations of AI and contributed to the momentum of the progress in the field.

Despite the enthusiasm and raised expectations, most of the tasks that were expected to be accomplished soon turned out to fail miserably by the end of the 1970s. Furthermore, the perceptron method was criticized by Marvin Minsky and Seymour Papert on the grounds that perceptrons would fail to learn the tasks involving linearly non-separable classes (Minsky and Papert 1969). Due to the impact of this criticism and the disappointment caused by failing to reach the ultimate aim of performing tasks at the human level, this line of research entered a period of stagnation called the "AI winter" in the 1970s (Lighthill 1973; Russell and Norvig 2016). During this period, the fundings, which the U.S. Department of Defense primarily provided, are dramatically reduced and neural networks could not become a critical method in the field until the mid-1980s.

Over time, the biological neuron interpretation and the overall influence of biology on neural networks faded away and neural networks evolved from an ambitious idea of achieving human-level performance using brain models into a theoretical framework of predictive modeling (Bishop 1995). Recently, the emergence of key elements such as convenient neural network frameworks (Abadi et al. 2015), increasing availability of GPUs (Shi et al. 2016), brilliant heuristics for making training more efficient (Kingma and Ba 2014; Ioffe and Szegedy 2015), and massive amounts of data (Jia Deng et al. 2009) set the scene for a creative and productive era of machine learning. After notable applications in image recognition (Krizhevsky et al. 2012), object detection (Girshick et al. 2014) and image segmentation (Long et al. 2015) where the traditional methods were outperformed by large margins, neural networks found applications in other domains such as machine translation (Y. Wu et al. 2016) as well as audio (A. v. d. Oord et al. 2016) and image synthesis (Radford et al. 2015). Consequently, neural networks not only moved these domains forward but also fundamentally transformed them by replacing key processing steps in their methodology.

The key factor behind the success of recent neural network applications and the deep learning hype is the major methodological and practical improvements over traditional machine learning methods. First of all, multilayer neural networks can capture nonlinearities in the data and effectively exploit this information in classification or regression tasks which might substantially affect performance. Compared to linear regression, which operates on raw input features, multilayer neural networks create hidden features of the data by successively transforming the input features nonlinearly (Figure 1.1). Second, the flexibility of neural networks enables specific structural characteristics of the data to be reflected in the network architecture, which is a concept known as *inductive bias*. For example, convolutional neural networks exploit the locality in the data based on the assumption that proximal features are often dependent. Consequently, this design dramatically reduces the number of parameters and improves generalizability (Goodfellow, Bengio, et al. 2016). Convolutional architectures fit well to the structure of images (Krizhevsky et al. 2012) and DNA sequences (J. Zhou and Troyanskaya 2015; Alipanahi et al. 2015; Kelley, Snoek, et al. 2016) where locality assumption holds. Third, designing neural networks that can perform multiple related prediction tasks simultaneously, called multitasking, is trivial. Such designs may improve overall prediction quality by taking dependencies across prediction tasks into account and discovering representations of the input that can be shared across tasks.

Similarly, multiple modalities of the same data, e.g. an image and a sentence describing the image or DNA sequence and chromatin accessibility of a locus, can be combined either at the input or output level in the neural network architecture conveniently by adding an input or output layer per modality (Eser and Churchman 2016). This bears extra importance for predictive modeling in computational biology since readouts from multiple omic layers

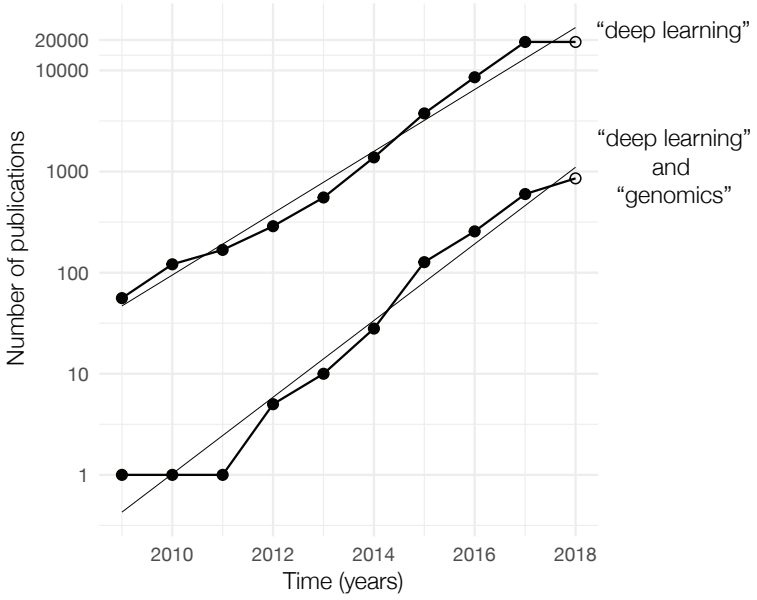


Figure 1.2: The number of publications about deep learning and deep learning in genomics increased exponentially between January 2009 and October 2018. Publication counts are obtained from `app.dimensions.ai` using queries "`deep learning`" and "`deep learning`" `AND` "`genomics`".

in matched samples are commonly measured and modeled (Bersanelli et al. 2016; Stoeckius et al. 2017; Hasin et al. 2017). Finally, neural networks also excel at characterizing unlabeled datasets either by extracting useful hidden variables (G. E. Hinton and Salakhutdinov 2006; Vincent et al. 2008) or by estimating complex multivariate data generating distributions, thereby allowing sampling of new data points in an unsupervised setting (Kingma and Welling 2013; Goodfellow, Pouget-Abadie, et al. 2014).

Machine learning and predictive modeling are used extensively in computational biology for tasks ranging from the prediction of transcription factor (TF) or RNA-binding protein (RBP) binding sites to splicing or cis-regulatory element prediction (Libbrecht and Noble 2015). The success of neural network applications in other fields and increasingly available biological data catalyzed the deep learning applications in our field (Stephens et al. 2015). In three seminal works published in 2015 and 2016 (J. Zhou and Troyanskaya 2015; Alipanahi et al. 2015; Kelley, Snoek, et al. 2016), convolutional neural networks are used for predicting the binding sites of transcription factors and chromatin accessibility where the prediction problem is simply formulated as classification of given biological sequences and existing machine learning methods are outperformed with a large margin. Furthermore, the potential of deep learning-based sequence models has been leveraged to predict regulatory effects of variants, a task known as variant effect prediction (VEP) (J. Zhou and Troyanskaya 2015; Kelley, Snoek, et al. 2016; J. Zhou, Park, et al. 2018). Methods utilizing predicted variant effects for devising

rich hypotheses on the mechanisms of complex traits and diseases have been proposed recently (Arloth et al. 2020; J. Zhou, C. L. Theesfeld, et al. 2018). Genome-wide genotype-phenotype associations, which is discussed in the next section, provides a general framework that is critical to understand in the light of new methods using variant effects predicted by neural networks.

Since the pioneering applications of deep learning in computational biology, the number of published applications continued to increase exponentially (Figure 1.2). Today, neural networks are employed as the method of choice for many tasks in genomics. See Eraslan, Avsec, et al. (2019) for a comprehensive review of deep learning applications in genomics.

1.2 Genome-wide association studies

Curiosity about the inheritance patterns of the observable characteristics of living organisms has been a major driving force in science for centuries. Since Gregor Mendel established the field of genetics in the late 19th century by revealing the basic principles of inheritance in pea plants, genetics expanded beyond inheritance. However, identifying the genetic drivers of traits and diseases is still a critical scientific challenge today.

In the early 2000s, two major developments in genetics marked the start of a new era. First, the sequence of the human genome is identified by the Human Genome Project with the aim of having a complete map of all genes in the human genome (International Human Genome Sequencing Consortium 2001; Venter et al. 2001). Second, the availability of low-cost technologies, such as microarrays, enabled the measurement of genome-wide genetic variation, known as genotyping, in a large number of individuals (Gunderson et al. 2005). Large-scale genotyping projects (International HapMap Consortium 2003; Siva 2008) yielded a deep catalog of human genetic variation that is now essential for genetics research.

The sequence of the human genome, the "big data" of genome-wide genotypes and the scalable methods developed in parallel empowered the widespread comparative analyses of the sequence variations in human populations in order to reveal the genetic basis of complex diseases and traits. These studies, called genome-wide association studies (GWAS), aim to find associations between the genetic variation and the phenotype of interest. The proportion of phenotypic variance explained by the genetic variation provides information about the genetic architecture, complexity and heritability of the disease or trait. Today, GWAS is the most powerful statistical tool for identifying the loci that are potentially relevant for a given phenotype and characterizing the genetic basis of phenotypes of interest.

To date, GWAS revealed over 75,000 unique genome-wide genotype-phenotype associations for over 2000 diseases and traits reported in around 3500 publications (Buniello et al. 2019). These associations not only fundamentally changed our understanding of complex traits and

diseases by narrowing down the loci relevant for phenotypes (Visscher et al. 2017) but also shaped our view of genetic variation at both the individual and the population level. However, it remains a challenge to pinpoint the variants that are causal for the phenotype of interest due to two major reasons. First, there are regions in the genome where the variants are inherited hence co-occur together. This concept, called the linkage disequilibrium (LD), leads to large blocks of highly correlated genetic variants. Therefore, the unit of highest granularity in GWAS results is an LD block which flags most variants in the block as potentially causal. Second, nearly 90% of the genome-wide significant variants are localized in the non-coding regions of genes (Edwards et al. 2013). This poses another layer of complexity since there is no trivial interpretation of non-coding variants as opposed to the coding variants such as missense mutations. Such variants are considered to contribute to the disease mechanism through modulating the regulatory mechanisms.

This thesis revisits the variant prioritization problem by leveraging deep learning-based variant effect prediction approaches, namely DeepSEA (J. Zhou and Troyanskaya 2015). This new sequence-based predictive approach enables us to suggest potentially causal non-coding variants and establish a link between the regulatory mechanism modulated by the non-coding variants and the phenotype of interest.

1.3 Single-cell RNA sequencing

Single-cell RNA sequencing (scRNA-seq) is an experimental technique for profiling transcriptomes of individual cells in the target biological sample. scRNA-seq has been fundamentally changing our understanding of molecular cell biology and diseases by facilitating the characterization of cell populations, the regulatory circuitry of cells, and the organization of cell populations in tissues.

Experimental steps of scRNA-seq comprise the dissociation and isolation of individual cells, followed by the library preparation and sequencing (Figure 1.3). Cells are first dissociated from a solid sample into a suspension of individual cells via enzymatic or mechanical dissociation methods. This step is followed by cell isolation, where cells are isolated into either the wells of a well-plate via FACS, microfluidic devices (Picelli et al. 2013; Jaitin et al. 2014; Soumillon et al. 2014) or emulsion droplets[1] (Macosko et al. 2015; Klein et al. 2015; Zheng et al. 2017). Isolated cells are then uniquely barcoded in the isolated environment where cell lysis, reverse transcription and cDNA amplification processes are performed.

Strengths and limitations of library preparation protocols vary (Ziegenhain et al. 2017; Jiarui Ding, Adiconis, et al. 2019). In the plate-based methods, cells are sorted into the wells of a well plate. Such protocols are more sensitive (i.e. the number of genes detected per cell

[1] Although these are the most commonly-used protocols today, many alternative single-cell profiling protocols exist such as those based on nanowell and microfluidic chips (Han et al. 2018; Gierahn et al. 2017).

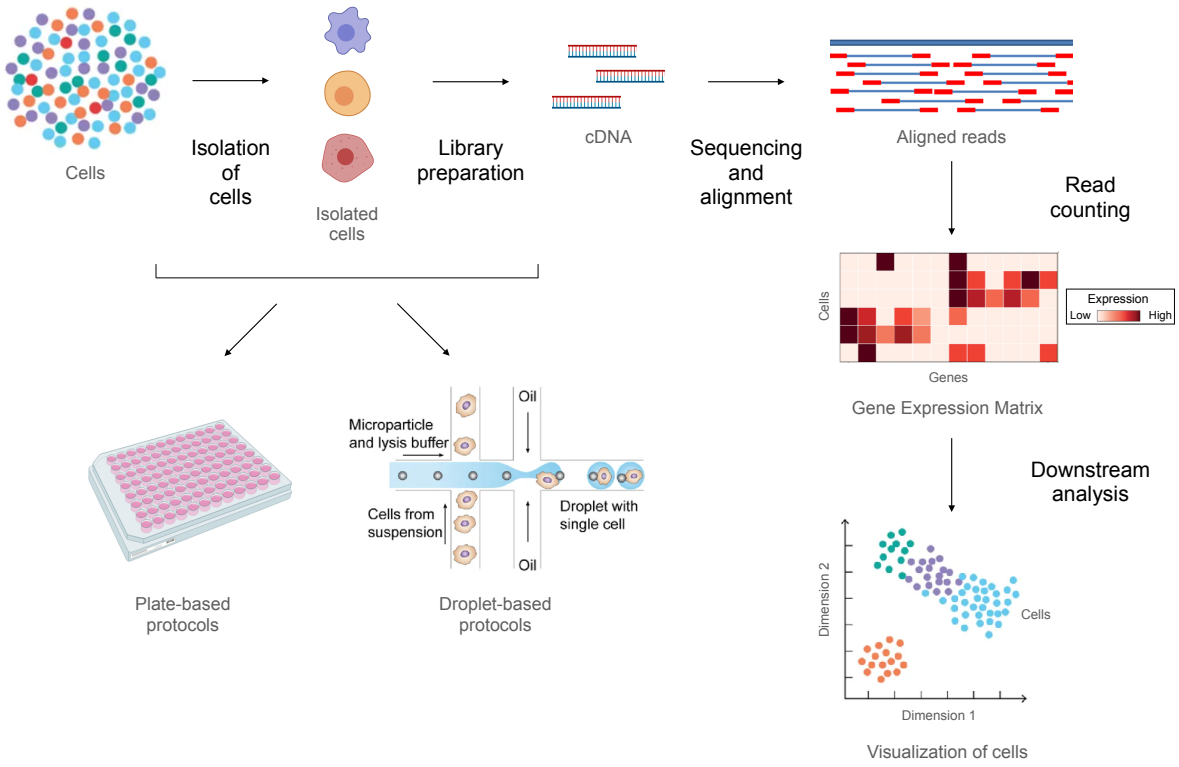


Figure 1.3: Single-cell RNA sequencing workflow. Dissociation of cells is followed by the library preparation where each cell is uniquely barcoded. Two primary library preparation techniques, plate-based and droplet-based protocols, are shown. After sequencing, read alignment and read counting, downstream analysis is performed for characterizing the biological heterogeneity of the sample. Adapted from Hwang et al. (2018).

is higher) and give good read coverage throughout the entire transcript (hence called "full transcript methods"). However, the number of cells being profiled is limited by the size of the well-plate and the number of plates. Therefore these protocols are considered low-throughput[2]. Droplet-based methods provide high-throughput experiments where tens of thousands of cells can be profiled at once. Furthermore, these methods are more cost-effective compared to plate-based protocols. Another advantage of the droplet-based methods is that the unique molecular identifiers (UMIs), an experimental technique where each transcript is labeled with a unique barcode to avoid PCR duplicates, are used as a standard practice (Klein et al. 2015). A major limitation of the droplet-based protocols is relatively lower sensitivity and the capture of only 3'-end or 5'-end of the transcripts. The lack of coverage throughout the entire transcript might hinder conducting certain types of analyses such as allele-specific expression and alternative splicing analysis, which are feasible with full-transcript protocols.

Since the first whole-transcriptomics study with four cells in 2009 (Tang et al. 2009),

[2]SPLiT-seq (Rosenberg et al. 2018) and sci-rna-seq (J. Cao, Packer, et al. 2017) offer elegant high-throughput strategies for plate-based methods.

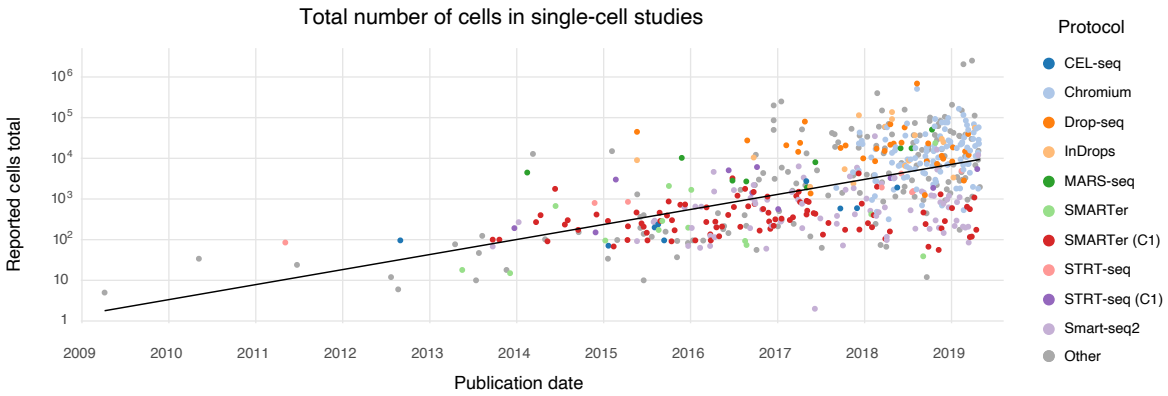


Figure 1.4: The trend in the number of profiled cells in single-cell datasets over time (Svensson et al. 2019). Each dot is a single-cell study. Colors represent the protocol used in the studies. The black line is a linear fit that shows the exponential increase in the number of profiled cells.

numerous variations of single-cell protocols have emerged (Figure 1.4). Each protocol has different pros and cons in throughput, sampling bias, technical variation, sensitivity, data quality and cost-effectiveness (Ziegenhain et al. 2017; Jiarui Ding, Adiconis, et al. 2019). As a low-cost, high-throughput alternative, droplet-based protocols are preferred today, especially in large-scale studies such as atlas projects where the goal is typically to characterize a specific organ or even organism (J. Cao, Packer, et al. 2017; Plass et al. 2018). A well-known example is The Human Cell Atlas project (Regev et al. 2017). This recent large-scale collaborative effort aims to create a comprehensive catalog of all human cell types and states using single-cell genomics.

Similar to experimental methods, computational tools and algorithms are also rapidly evolving and adapting to the advances of the field (G. Chen et al. 2019). The main goal of the downstream analysis and modeling in single-cell genomics is to characterize given biological samples mainly using exploratory data analysis techniques and statistical models. This characterization typically starts with quality control (QC) steps to ensure the validity of the downstream analysis and to have an accurate picture of the biology of a given sample.

There are various QC steps in scRNA-seq, including cell calling, ambient RNA detection, and doublet detection. In droplet-based methods, a major problem called "ambient RNA" stems from the fact that cell death and lysis cause some transcripts of primarily highly expressed genes to contaminate the suspension. These transcripts can then be captured in the droplets. For the droplets containing a cell, ambient RNA introduces spurious expression and shifts the gene expression profiles in random directions. The droplets without a cell might end up in the final gene expression count matrix as regular observations due to non-zero expression originating from the ambient RNA. Cell calling methods aim to distinguish between these two types of observations in the expression count matrix (Lun, Riesenfeld, et al. 2019).

Furthermore, detecting the composition and the amount of ambient RNA in cell-containing droplets and removing the ambient RNA effect from the counts is a more challenging task (Heaton et al. 2019; Fleming et al. 2019). Doublet detection methods aim to find droplets with more than one cell (i.e. multiplets). Doublets with cells from population A and population B might look like a "novel population" or a "transition state" with expression profiles similar to those of A and B. Therefore keeping such observations in the count matrix might give rise to misleading results in the analysis, especially if populations A and B are biologically relevant. Simple cut-off based heuristics applied to the total number of detected transcripts are also widely used to eliminate empty droplets and multiplets partly, assuming that empty droplets contain fewer transcripts than regular cells, whereas multiplets contain higher (Luecken and Theis 2019). Before the downstream analysis, the final preprocessing step that is typically used is the normalization step which accounts for the cell-to-cell differences in sequencing depth. The counts of two cells are not directly comparable without this correction. The simplest normalization method used today is to scale total counts of all cells to a constant number such as ten thousand, which is called TP10k (transcripts per ten thousand), similar to the TPM method, which is often used in bulk RNA-seq where constant is simply one million instead of ten thousand. However, this approach does not entirely remove the correlation between the gene expression profiles and the total counts. Also, it does not consider the biological factors (e.g. cell type-specific effects) behind the variation in total counts. Better methods that take the count structure and biological variability of the data into account are also proposed in the literature (Lun, Bach, et al. 2016; Bacher et al. 2017; Hafemeister and Satija 2019).

Downstream analysis steps of single-cell gene expression data consist of various unsupervised and exploratory data science techniques. Clustering and data visualization are arguably the most crucial steps of such pipelines facilitating the identification of biological variability in the data (Figure 1.3). While clustering identifies groups of cells with relatively low within-group variation compared to between-group variation, differential expression methods aim to find differences in gene expression between such groups and/or between biological conditions using statistical methods. In the cases where data exhibit continuous trends rather than a discrete cluster structure (e.g. differentiation), the techniques tailored for continuous phenotypes such as trajectory inference (Haghverdi, Buettner, et al. 2015; Haghverdi, Büttner, et al. 2016; Trapnell et al. 2014) or RNA velocity (La Manno et al. 2018; Bergen et al. 2019) are typically used. In addition to these exploratory techniques, single-cell data can also be used to interrogate gene-gene relationships and identify modules of genes as a proxy of cellular programs based on correlation or other similarity metrics. However, due to the inherent noise of the single-cell data (e.g. dropout due to the stochasticity in transcript capture and amplification), gene-gene correlations are typically underestimated, which might hinder the downstream analysis. Using reasonable heuristics, denoising techniques aim to recover the

correlation between the genes that is lost due to the noise. In the literature, denoising has been used to improve the identification of gene modules and further functional analyses built on these modules (Smillie et al. 2019).

Many steps of the downstream analysis such as differential expression, denoising and normalization, depend on the noise model assumption of the single-cell count data. A common choice in these models is to first log-transform the count data using a pseudocount of one (i.e. `log(X+1)`) and then assume Gaussian distribution as a simple noise model. This approach ignores specific characteristics of the data, such as the count structure and high sparsity. Other choices for the noise model include count distributions such as Poisson, negative binomial and zero-inflated negative binomial. This thesis introduces a denoising method for scRNA-seq that leverages unsupervised learning concepts from modern machine learning, where we use a noise model that is tailored to the characteristics of single-cell data.

1.4 Research questions

In this thesis, we aim to answer the following questions about the major challenges in two different domains of biology and genetics; GWAS and single-cell genomics:

1. In GWAS, it is challenging to prioritize potentially causal regulatory variants amongst several risk variants identified by GWAS due to the highly correlated and cosegregated variants. Variant prioritization approaches aim to characterize such variants by incorporating new layers of information from alternative sources that might help identify causal variants and their proxies (Figure 1.5). Machine learning-based variant effect prediction methods are getting increasingly available and performant. The first research question we aim to answer is whether we can generate better functional hypotheses on clinical phenotypes by combining genotype data from individuals with the variant effect predictions produced by machine learning models.

2. Single-cell genomics is an indispensable tool in our toolbox of measurement techniques which provides a unique way to investigate cellular processes and phenotypes at single-cell resolution. However, due to the low amount of RNA in cells, single-cell RNA sequencing (scRNA-seq) gene expression readout is corrupted by the characteristic noise of the measurement process. This corruption hinders having accurate representations of cells and poses a challenge in downstream analysis, which needs to be accounted for with proper noise models (Figure 1.5). We aim to use genomics-based deep learning to address this problem by denoising the expression signal which yields a faithful representation of the underlying biology and improves downstream analysis.

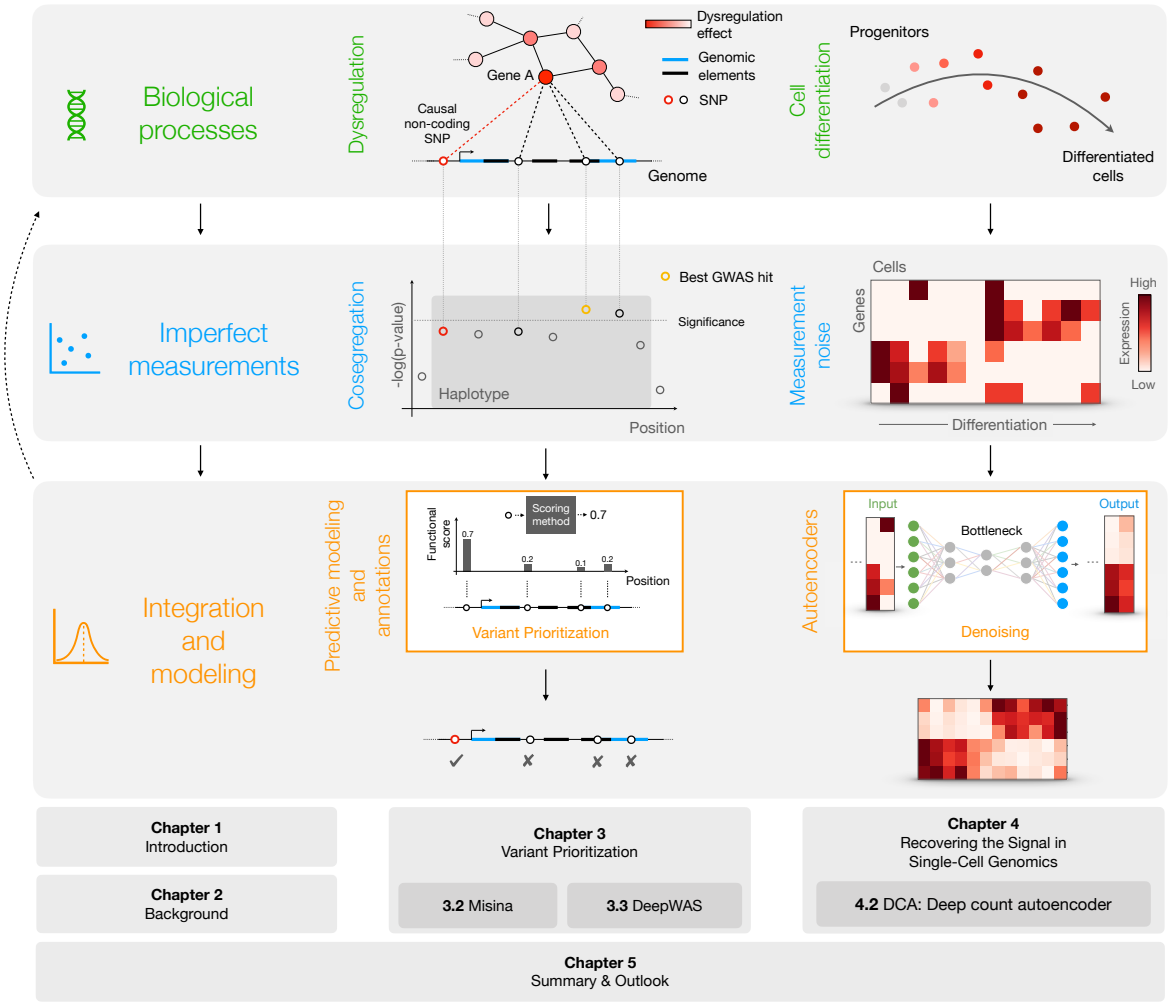


Figure 1.5: Overview of this thesis. Biological processes that we aim to understand, such as gene dysregulation and cell differentiation, are observed as imperfect measurements through biological experiments. In this process, the data act as a corrupted proxy of reality due to several factors like measurement noise or biological confounders such as cosegregation. Modeling provides an abstract representation of reality which can be utilized to improve our understanding of biology. Here, this flow is exemplified in two application domains. First, variant prioritization in population studies (Chapter 3) where we use integration (Section 3.2) and predictive modeling (Section 3.3) for finding potentially causal variants. Second, denoising approaches improve the single-cell gene expression readouts (Chapter 4) and enhance the overall picture of molecular phenotypes such as cell differentiation. Our novel autoencoder-based denoising approach is presented in Section 4.2.

1.5 Overview of this thesis

Chapter 2 introduces an overview of the statistical learning methods and biological data modalities used throughout the thesis. The methods section includes the introduction of the fundamental modeling concepts that form the basis of this thesis, such as generalized linear models, stability selection and neural networks. The chapter concludes with the definitions of the essential data types of various *omic* layers such as transcriptomics and epigenomics, other related concepts like genotyping and microRNAs, as well as the public data sources like the ENCODE (ENCODE Project Consortium 2007) and Roadmap Epigenomics projects (Kundaje et al. 2015).

Chapter 3 discusses variant prioritization methods from different perspectives, focusing on the analysis of non-coding variants. The first section, which was published as a book chapter (Schulze and F. McMahon 2018), provides an overview of the tools and methods commonly used for prioritizing coding and non-coding variants based on positional overlaps. Next, we introduce Misina, our method that facilitates the interrogation of the microRNA-mediated effects of genetic variants by integrating microRNA-binding site predictions, microRNA and target gene expression and eQTL effects of variants (publication in preparation). We aim to understand the regulatory determinants of complex diseases and traits that act through modulating microRNA-target interactions with this integrative approach. We conclude the chapter by introducing another approach, DeepWAS, which identifies the disease- or trait-associated non-coding variants with a potential regulatory and causal role (Arloth et al. 2020). With this approach, we aim to broaden the scope of typical genome-wide association studies by introducing relevant cell lines and regulatory elements like transcription factors.

Chapter 4 starts with the conceptual overview of dropout and the recovery of the biological signal in scRNA-seq datasets. After we present two existing denoising methods, we introduce our deep count autoencoder (DCA) approach, which utilizes unsupervised neural networks to capture the data manifold and subsequently to reconstruct the data with a pronounced biological signal (Eraslan, Simon, et al. 2019). We further demonstrate the benefits of denoising using commonly used downstream methods.

Chapter 5 concludes the thesis with a summary and remarks on future directions. The conceptual and structural overview of this thesis is given in Figure 1.5.

Chapter 2

Background

This chapter briefly introduces the statistical learning techniques, data modalities and sources used in the following chapters. Section 2.1 focuses particularly on linear models and variable selection employed in genotype-phenotype associations in the context of variant prioritization in Chapter 3, as well as the neural network and autoencoder frameworks which serve as a background for the unsupervised machine learning model introduced in Chapter 4. This section is followed by Section 2.2 where brief descriptions of omics modalities such as transcriptomics and epigenomics are presented. Finally, Section 2.3 concludes the chapter with the list of public data sources like ENCODE and GTEx that were mainly utilized in Chapter 3.

2.1 Models

This section gives an overview of commonly used linear and non-linear models from supervised and unsupervised learning perspectives, as well as the complementary technique of variable selection. With a probabilistic (and frequentist) formulation, the objective of the supervised models can be described as maximizing a likelihood function $\mathcal{L}(\theta; y_i)$ of the conditional distribution $P(y_i|\mathbf{x}_i; \theta)$ which acts as a mapping of the inputs $\mathbf{x}_i$ to the outputs y_i parameterized by θ. In the unsupervised case, the distribution of interest can be broadly defined as $P(\mathbf{x}_i; \theta)$ where the aim is to find the parameters that jointly models the inputs $\mathbf{x}_i$ without an explicit mapping by maximizing the likelihood $\mathcal{L}(\theta; \mathbf{x}_i)$. Following subsections summarize different combinations of linear and non-linear functions with different probability distribution assumptions where the inference relies on the maximum likelihood-based parameter estimation (Friedman et al. 2001).

2.1.1 Generalized linear models

Linear regression is a simple and powerful method for modeling the relationship between a target (or dependent) variable and one or more features (or independent variables) from a set of observations using the linear combination of features. The coefficients of the linear combination are called the model parameters. The data matrix with n observations and p features can be written as $\mathbf{X} \in \mathbb{R}^{n\times(p+1)}$ where the additional feature represents constant ones that are used as the intercept term in the model. A single observation, namely a row of the $\mathbf{X}$ matrix, is represented as $\mathbf{x}_i$ where $i = 1, ..., n$ is the observation index. The target variable and the model parameters are represented as $n-$dimensional and $(p+1)-$dimensional vectors $\mathbf{y}$ and $\boldsymbol{\beta}$.

$$\mathbf{x}_i = \begin{pmatrix} 1 \\ x_{i,1} \\ x_{i,2} \\ \vdots \\ x_{i,p} \end{pmatrix} \quad \mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \\ y_3 \\ \vdots \\ y_n \end{pmatrix} \quad \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{pmatrix}$$

A linear regression model can be described as

$$y_i = \mathbf{x}_i^T\boldsymbol{\beta} + \epsilon_i,\ \epsilon_i \sim N(0, \sigma^2)$$

where the independent Gaussian noise term is denoted as ϵ_i. From a probabilistic perspective, the estimation of the model parameters can be formulated as the maximum likelihood estimation of the mean parameter of conditional distribution $P(y_i|\mathbf{x}_i; \boldsymbol{\beta}, \sigma^2)$ using the Gaussian probability density function (PDF):

$$P(y_i|\mathbf{x}_i; \boldsymbol{\beta}, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y_i - \mathbf{x}_i^T\boldsymbol{\beta})^2}{2\sigma^2}\right)$$

The log-likelihood of parameters, $\boldsymbol{\beta}$, can then be defined as follows:

$$\begin{aligned} \log \mathcal{L}(\boldsymbol{\beta}; \mathbf{y}) &= \log\left(\prod_i^n \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y_i - \mathbf{x}_i^T\boldsymbol{\beta})^2}{2\sigma^2}\right)\right) \\ &= n \log \frac{1}{\sqrt{2\pi\sigma^2}} - \frac{1}{2\sigma^2}\sum_i^n (y_i - \mathbf{x}_i^T\boldsymbol{\beta})^2 \end{aligned}$$

It can be shown that optimizing the model parameters with the most commonly used error function in linear regression, mean squared error (MSE), is equivalent to maximizing the log-likelihood of parameters with the assumption that $P(y_i|\mathbf{x}_i)$ is normally distributed. Given that the variance parameter is a non-negative constant, MSE error (L_{MSE}) can be derived as follows:

$$\begin{aligned}\underset{\boldsymbol{\beta}}{\text{argmax}}\ n\log\frac{1}{\sqrt{2\pi\sigma^2}} - \frac{1}{2\sigma^2}\sum_i^n(y_i - \mathbf{x}_i^T\boldsymbol{\beta})^2 &= \underset{\boldsymbol{\beta}}{\text{argmin}}\sum_i^n(y_i - \mathbf{x}_i^T\boldsymbol{\beta})^2 \\ &= \underset{\boldsymbol{\beta}}{\text{argmin}}\,\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 \\ &= \underset{\boldsymbol{\beta}}{\text{argmin}}\,L_{\text{MSE}}(\mathbf{y}, \mathbf{X}\boldsymbol{\beta})\end{aligned}$$

See Chapter 5.5.1 of the deep learning book (Goodfellow, Bengio, et al. 2016) for more details.

Generalized linear models (GLMs) generalizes the concept of linear regression to arbitrary distributions by linking the linear combination of features to the expectation of a distribution using link functions. Therefore, investigating the link between the loss functions that are typically used in classification and regression settings and the maximum likelihood estimates of the parameters of corresponding probability distributions allows us to understand various noise models used in GLMs and neural networks.

Logistic regression

For binary classification, the conditional distribution $P(y_i|\mathbf{x}_i;\boldsymbol{\beta})$ can be modeled with the Bernoulli distribution in the GLM framework. The likelihood function of Bernoulli is given below

$$P(y_i|\mathbf{x}_i;\boldsymbol{\beta}) = \mathcal{L}_{\text{bernoulli}}(\boldsymbol{\beta};\mathbf{y}) = \begin{cases}\hat{y}_i & \text{if } y_i = 1, \\ 1-\hat{y}_i & \text{if } y_i = 0.\end{cases}$$

$$\hat{y}_i = \sigma(x_i^T\boldsymbol{\beta})$$

where the $\hat{y}_i$ parameter in $[0, 1]$ interval is the only parameter of Bernoulli, which also corresponds to the mean of the distribution. The sigmoid link function is defined as $\sigma(x) = 1/(1+e^{-x})$ which converts the linear transformation of the data into the mean of the distribution, is referred to as the *inverse link function* in generalized linear model (GLM) literature, whereas it is called the *activation function* in the neural network literature. The same likelihood function can also be expressed as

$$\mathcal{L}_{\text{bernoulli}}(\boldsymbol{\beta}; y_i) = \hat{y}_i^{y_i} + (1 - \hat{y}_i)^{(1-y_i)}$$
$$\mathcal{L}_{\text{bernoulli}}(\boldsymbol{\beta}; \mathbf{y}) = \prod_i^n \hat{y}_i^{y_i} + (1 - \hat{y}_i)^{(1-y_i)}$$

whereas the negative log-likelihood can be written as

$$-\log \mathcal{L}_{\text{bernoulli}}(\boldsymbol{\beta}; \mathbf{y}) = \sum_i^n -y_i \log \hat{y}_i - (1 - y_i) \log(1 - \hat{y}_i)$$

which yields the error function known as binary cross-entropy (BCE) or *log loss* in statistical learning. This can be further extended to the categorical log-likelihood for multi-class problems. This shows that maximizing the log-likelihood of Bernoulli distribution is equivalent to minimizing the log loss:

$$\underset{\boldsymbol{\beta}}{\text{argmax}} \ \log \mathcal{L}_{\text{bernoulli}}(\boldsymbol{\beta}; \mathbf{y}) = \underset{\boldsymbol{\beta}}{\text{argmin}} \, L_{\text{BCE}}(\hat{\mathbf{y}}, \mathbf{y})$$

Regression with count data

Modeling count data is critical in computational biology as the readout of many experiments, such as high-throughput sequencing assays, exhibits count structure. For count data, the data generating process can be regarded as a process where the number of occurrences of an event is counted. If the interval where we count the occurrences is fixed, then this process is called the Poisson process and the $P(y_i|\mathbf{x}_i; \boldsymbol{\beta})$ distribution can be modeled using the Poisson distribution and the corresponding likelihood function:

$$P(y_i|\mathbf{x}_i; \boldsymbol{\beta}) = \mathcal{L}_{\text{poisson}}(\boldsymbol{\beta}; \mathbf{y}) = \prod_i^n \frac{\hat{y}_i^{y_i} \exp(-\hat{y}_i)}{y_i!}$$
$$\hat{y}_i = \exp(x_i^T \boldsymbol{\beta})$$

where the only parameter $\hat{y}_i$ of Poisson represents both the mean and the variance of the distribution. The exponential link function keeps the estimated mean parameter non-negative. The negative log-likelihood is given as:

$$-\log \mathcal{L}_{\text{poisson}}(\boldsymbol{\beta}; \mathbf{y}) = \sum_i^n \hat{y}_i - y_i \log \hat{y}_i + \log(y_i!)$$

For maximizing the conditional likelihood, we can minimize the model parameters with

respect to the negative log-likelihood as

$$\underset{\boldsymbol{\beta}}{\operatorname{argmin}} \sum_{i}^{n} \hat{y}_i - y_i \log \hat{y}_i$$

Negative binomial regression

The major limitation of Poisson distribution is that the mean and variance are equal and are controlled by a single parameter. This, however, is not realistic in many real-world cases. Negative binomial (NB) is another discrete distribution that can be used when the mean and the variance of the data are not equal. Although the classical textbook definition of NB is the number of successes before a specified number of failures occur in i.i.d Bernoulli trials, the alternative parameterization with mean and dispersion is more intuitive and useful. NB likelihood can be written as follows:

$$P(y_i|\mathbf{x}_i;\boldsymbol{\beta}) = \mathcal{L}_{\mathrm{NB}}(\boldsymbol{\beta};\mathbf{y}) = \prod_{i}^{n} \frac{\Gamma(y_i+\theta^{-1})}{\Gamma(\theta^{-1})\Gamma(y_i+1)} \left(\frac{\theta^{-1}}{\theta^{-1}+\hat{y}_i}\right)^{\theta^{-1}} \left(\frac{\hat{y}_i}{\theta^{-1}+\hat{y}_i}\right)^{y_i}$$
$$\hat{y}_i = \exp(x_i^T\boldsymbol{\beta})$$

where parameters $\hat{y}_i$ and θ represent the mean and dispersion. The variance of the distribution is:

$$\sigma^2 = \mu + \theta^{-1}\mu^2$$

which shows the mean-variance dependence. Note that this distribution reduces to Poisson when $\theta = \infty$.

The negative log-likelihood of NB is:

$$\begin{aligned} -\log \mathcal{L}_{\mathrm{NB}}(\boldsymbol{\beta};\mathbf{y}) = &-\log\Gamma(y_i+\theta^{-1}) + \log\Gamma(\theta^{-1}) + \log\Gamma(y_i+1) \\ &- \theta^{-1}\left(\log\theta^{-1} - \log(\theta^{-1}+\hat{y}_i)\right) - y_i\left(\log\hat{y}_i - \log(\theta^{-1}+\hat{y}_i)\right) \end{aligned}$$

In the GLM setting, we can interpret this equation as a loss function by dropping the

terms that do not depend on $\hat{y}_i$:

$$\begin{aligned}\underset{\boldsymbol{\beta}}{\operatorname{argmin}} \sum_{i}^{n} \theta^{-1} \log(\theta^{-1} + \hat{y}_i) - y_i \left(\log \hat{y}_i - \log(\theta^{-1} + \hat{y}_i)\right) \\ = \sum_{i}^{n} (\theta^{-1} + y_i) \log(\theta^{-1} + \hat{y}_i) - y_i \log \hat{y}_i\end{aligned}$$

Dispersion parameter can either be optimized as a free parameter or can be conditioned on x_i via another set of parameters $\boldsymbol{\beta}_\theta$ e.g. $\theta_i = \exp(x_i^T \boldsymbol{\beta}_\theta)$.

Zero-inflated negative binomial regression

In some cases, there might be an excess number of zeros in the count data, which can be accounted for using so-called zero-inflated models. In such cases, the conditional distribution $P(y_i|\mathbf{x}_i; \boldsymbol{\beta})$ can be modeled using zero-inflated negative binomial (ZINB), which can be written as a mixture of a point mass at zero and a negative binomial:

$$\mathcal{L}_{\text{ZINB}}(y_i; \pi_i, \hat{y}_i, \theta) = \begin{cases} \pi_i + (1 - \pi_i)\ \mathcal{L}_{\text{NB}}(0; \hat{y}_i, \theta) & \text{if } y_i = 0 \\ (1 - \pi_i)\ \mathcal{L}_{\text{NB}}(y_i; \hat{y}_i, \theta) & \text{if } y_i > 0 \end{cases}$$

where π represents the mixture coefficient between the point mass and NB. Note that the model reduces to negative binomial if $\pi = 0$. It is also important to note that here the mixture parameter π is sample-specific similar to the mean parameter of the negative binomial, μ_i, therefore it is conditioned on the data. This can be achieved by estimating π_i using the sigmoid link function e.g. $\pi_i = \sigma(x_i^T \boldsymbol{\beta}_\pi)$ as in logistic regression.

Negative log-likelihood can be written as:

$$-\log \mathcal{L}_{\text{ZINB}}(\boldsymbol{\beta}; \mathbf{y}) = \begin{cases} -\log\left(\pi_i + (1 - \pi_i)\ \left(\frac{\theta^{-1}}{\theta^{-1} + \hat{y}_i}\right)^{\theta^{-1}}\right) & \text{if } y_i = 0 \\ -\log(1 - \pi_i) - \log \mathcal{L}_{\text{NB}}(\boldsymbol{\beta}; \mathbf{y}) & \text{if } y_i > 0 \end{cases}$$

which can be used as an error function in ZINB regression to estimate model parameters. In Section 4.2, we use ZINB log-likelihood as a noise model within the autoencoder framework to find the underlying manifold of single-cell RNA sequencing data, which is exploited for denoising the expression levels of cells.

2.1.2 Variable selection

Lasso

Lasso (Tibshirani 1996) is a feature selection and regularization method based on the L1 norm of model parameters in regression analysis, often preferred to identify associations in high dimensional datasets. Parameter estimation can be described as follows:

$$\underset{\boldsymbol{\beta}}{\operatorname{argmin}} \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda \|\boldsymbol{\beta}\|_1$$

where the hyperparameter λ represents the strength of the L1-regularization, which shrinks model parameters towards zero. Lasso is used for finding SNP-phenotype associations in this thesis (Section 3.3).

Stability selection

The variation observed in feature selection when the model is fitted to similar datasets might hugely impact reported results. Especially in computational biology, where selected features such as genes are reported as the "signatures" of molecular or clinical phenotypes via guilt-by-association, this aspect of feature selection called stability bears extra importance.

Stability methods seek robust feature selection procedures by taking the uncertainty of feature selection into account using subsets or bootstrap samples of datasets. In stability selection (Meinshausen and Bühlmann 2010), the authors propose a method where Lasso models are fitted to the small random subsets of a given dataset ($\lfloor n/2 \rfloor$) repeatedly N times which yields selection probabilities for all variables. Afterward, a stringent predefined selection probability cutoff (π_{thr}) is applied in order to obtain a stable feature set. However, the novelty of the method lies in the theorem, which under some assumptions establishes an elegant link between the selection cutoff and Type I error rate. In the method, the upper bound of per-family error rate (PFER), $E(V)$, representing the expected number of falsely selected variables is defined in terms of the selection cutoff π_{thr}, the average number of selected variables q_Λ and the total number of variables p:

$$E(V) \leq \frac{1}{2\pi_{thr} - 1} \frac{q_\Lambda^2}{p}$$

With predefined PFER and selection cutoff values, q_Λ is calculated from the formula and the regularization parameters $\lambda \in \Lambda$ are determined accordingly.

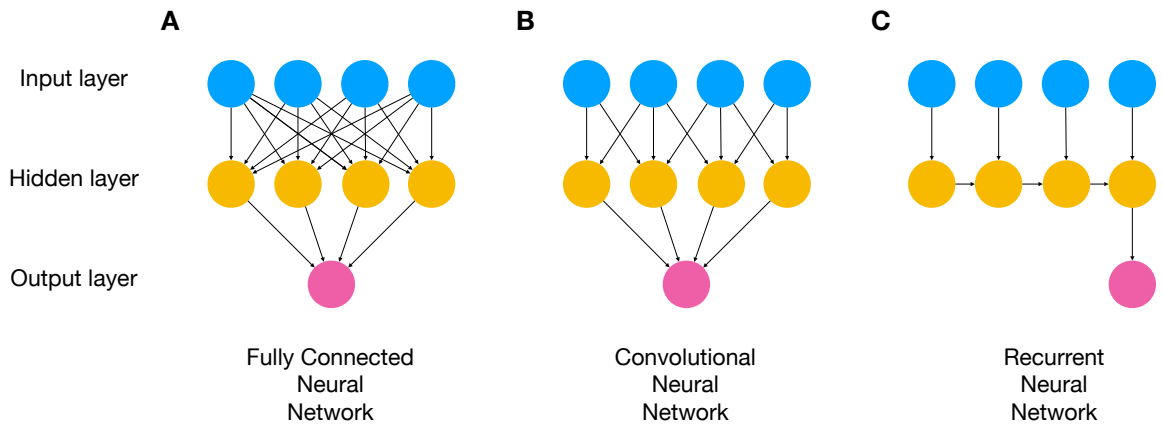


Figure 2.1: Neural network architectures. Layered structures and connection patterns of neurons are shown for each of the major neural network architectures, namely **(A)** fully connected, **(B)** convolutional and **(C)** recurrent neural networks.

2.1.3 Neural networks

There are three major architectures of neural networks (Figure 2.1). Architectures are designed by reflecting a bias into the architecture matching the nature of the data and the network structure, a concept known as the *inductive bias*. Fully connected neural networks (also known as multilayer perceptrons, MLPs) consist of layers in which every neuron is connected to every neuron in the next layer (Figure 2.1A) and are preferred for tabular data without an assumption suggesting a particular connectivity pattern between the features.

Convolutional neural networks (CNNs) assume the locality of input features which fits well to the structure of images, sentences and biological sequences. CNNs can be considered regularized (or sparsified) versions of MLPs (Figure 2.1B). The inductive bias of recurrent neural networks (RNNs) (Figure 2.1C) also favors locality in data, making it suitable for modeling sequence data.

The general structure of neural networks can be summarized as

$$
\begin{aligned}
f(\mathbf{X}_1, \mathbf{W}_1) &= \mathbf{X}_2 \\
f(\mathbf{X}_2, \mathbf{W}_2) &= \mathbf{X}_3 \\
&\dots \\
g(\mathbf{X}_l, \mathbf{W}_l) &= \hat{\mathbf{Y}} \\
\underset{\mathbf{W}_1, \mathbf{W}_2, \dots, \mathbf{W}_l}{\operatorname{argmin}} & \; L(\mathbf{Y}, \hat{\mathbf{Y}})
\end{aligned}
$$

where $\mathbf{X}_1$ represents the input matrix. l and $\mathbf{W}_l$ depict the layer index and the corresponding layer parameters. $f(\mathbf{X}_i, \mathbf{W}_i) = \sigma(\mathbf{X}_i \mathbf{W}_i)$ for feedforward neural networks. σ is a simple nonlinear function, such as tanh or sigmoid, called the activation function, which allows the

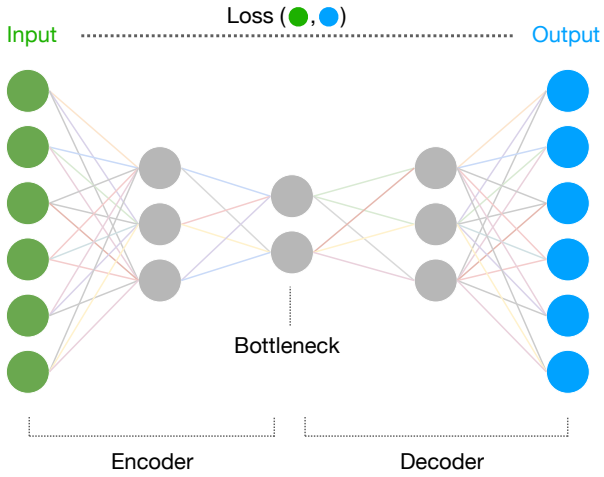


Figure 2.2: Architecture of an autoencoder. Input and output layers of the same size and a bottleneck layer of lower dimension characterize an autoencoder. The loss function quantifies the concordance between the original input and the reconstruction.

neural network to capture nonlinear relationships in data. Final layer activation depends on the type of task that the network is trained for. For example, sigmoid, softmax or linear activations are used for binary, multi-class classifications and regression, respectively. L represents the loss function.

Convolutional neural networks simply replace the matrix multiplication with the convolution operation: $f(\mathbf{X}_i, \mathbf{W}_i) = \sigma(\mathbf{X}_i * \mathbf{W}_i)$. Recurrent neural networks process the data sequentially: $f(\mathbf{X}_i^{(j)}, \mathbf{W}_i) = \sigma(\mathbf{X}_i^{(j)}\mathbf{W}_i + \mathbf{H}_i^{(j-1)}\mathbf{W}_i^{(h)})$ where the superscript j is the time index[1]. $\mathbf{H}^{(j-1)}$ and $\mathbf{W}^{(h)}$ denote hidden state vector from the previous time point and hidden state parameters.

Autoencoders

Unlike supervised learning where the mapping of the inputs $(\mathbf{x}_i)$ to the outputs (y_i) is explicitly defined, unsupervised learning aims to find the parameters of the joint distribution $P(\mathbf{x}_i; \theta)$ without an explicit mapping or labels (e.g. y_i) by maximizing the likelihood $\mathcal{L}(\theta; \mathbf{x}_i)$. An autoencoder is a type of neural network typically used in unsupervised and representation learning (Goodfellow, Bengio, et al. 2016). Autoencoders consist of an input layer, one or more hidden layers including a bottleneck layer, and an output layer (Figure 2.2) where the hidden layers perform nonlinear operations due to the nonlinear activation functions such as rectified linear unit (ReLU) (Goodfellow, Bengio, et al. 2016). Input and output layers are of the same width, which is equal to the number of features of the input data while the bottleneck layer is typically of much lower dimensionality. Layers before and after the bottleneck are considered

[1]Although bidirectional RNNs also exist, here the unidirectional RNNs are described for clarity.

two components of the network and called encoder and decoder, respectively. In the standard autoencoders, the weights of the network are optimized via numerical optimization to obtain the best reconstruction (i.e. predicted output) from the output layer using a loss function that compares the reconstruction with the original input. Although mean-squared error (MSE) is usually used as the default loss function, other likelihood functions that are more suitable for the structure of the data can also be used (Eraslan, Simon, et al. 2019) as in the GLM framework (see Section 2.1.1). The mathematical description of the model in Figure 2.2 is given below:

$$\begin{aligned}\mathbf{E} &= \mathrm{ReLU}(\mathbf{X}\mathbf{W}_E)\\ \mathbf{B} &= \mathrm{ReLU}(\mathbf{E}\mathbf{W}_B)\\ \mathbf{D} &= \mathrm{ReLU}(\mathbf{B}\mathbf{W}_D)\\ \hat{\mathbf{X}} &= f(\mathbf{D}\mathbf{W}_O)\end{aligned}$$

where ReLU is used as the activation function of the hidden layers, whereas the activation of the output function is shown as f, which similar to the link functions in GLMs, depends on the likelihood function. Here E, B, D and $\hat{X}$ represent, the first encoder layer, the bottleneck layer, the first decoder layer and the output, respectively. The optimization objective can be written using the loss function (i.e. likelihood function) as:

$$\underset{\Theta}{\operatorname{argmin}} - \log \mathcal{L}(\mathbf{X}; \hat{\mathbf{X}}, \Theta)$$

where Θ denotes the set of all parameters i.e. $\{\mathbf{W}_E, \mathbf{W}_B, \mathbf{W}_D, \mathbf{W}_O\}$.

Interestingly, there is a connection between autoencoders and principal component analysis (PCA) where the loadings of PCA and the weights of an autoencoder with a single linear layer trained with the MSE loss function span the same subspace (Plaut 2018).

The bottleneck layer acts as a major architectural constraint, which prevents the network from learning the identity function. There are two critical consequences of this constraint. First, neural network learns how to encode the data points in a low dimensional space, also called the latent space, effectively by capturing the underlying manifold from which the high-dimensional measured data is likely originated. This makes autoencoders suitable for dimension reduction and other representation learning tasks. Moreover, in the probabilistic formulations of autoencoders such as variational autoencoders (VAEs) (Kingma and Welling 2013), latent variables, which are represented by the bottleneck neurons, can be used as building blocks and combined with additional latent variables that are needed for specific tasks (Lopez et al. 2018). Second, autoencoders naturally function as denoising algorithms

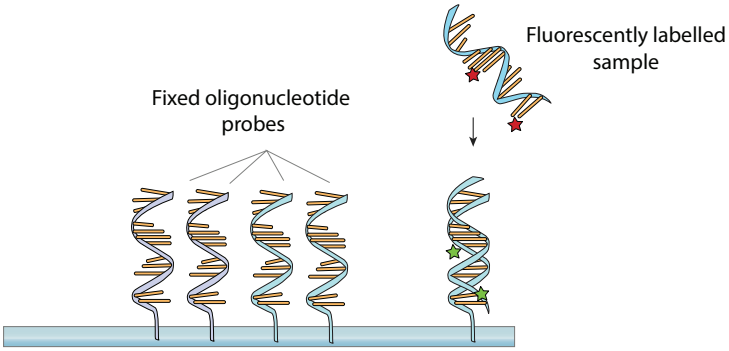


Figure 2.3: Microarrays with fixed DNA oligonucleotide probes allow us to quantify expression levels of particular genes or to determine the genotypes of individuals by measuring the fluorescent signal emitted from the labeled samples. (Figure is adapted from Wikipedia and is in the public domain.)

since the decoder reconstructs the data from its compressed representation, which is likely to capture only essential features of the data. This idea is explored in Chapter 4 in the context of single-cell genomics.

2.2 Omics modalities

The computational methods in biology operate on the measurements from a collection of biological entities. The entire collection of specific entities of the same type, such as genes, transcripts and proteins is typically referred to with the "-ome" suffix i.e. *genome*, *transcriptome* and *proteome*. The concept of "the totality of biological entities" is also applied to many other types of entities that are not part of the central dogma such as *methylome* for the regions of the genome where methylation is observed and *metabolome* for the totality of metabolites in a cell. The field of study focusing on such *omes* is called *omics* (Hasin et al. 2017) e.g. genomics.

Various measuring techniques and protocols in biology allow us to study and understand the biological phenomenon of interest like gene regulation or cell differentiation indirectly through data and modeling. In the following subsections, we briefly describe the data types and data resources used in the analyses throughout the thesis.

2.2.1 Transcriptomics

Gene expression microarrays

Microarrays are screening tools that typically contain DNA oligonucleotide probes (or proteins in protein arrays) attached to a solid surface. Fluorescently labeled sample sequences hybridize with matching probes and emit light (Figure 2.3). For example, the microarrays with the probes designed to match specific transcripts allow us to quantify the expression levels of these transcripts based on the measurement of the emitted fluorescence intensity. It is still commonly

used in bioinformatics because of its relatively low cost compared to the high-throughput sequencing-based methods.

High-throughput sequencing

High-throughput sequencing (HTS) methods substantially changed our understanding of biology and diseases (Reuter et al. 2015). For example, RNA sequencing showed three-quarters of our genome is transcribed (Djebali et al. 2012). HTS also enabled the characterization of various aspects of RNA biology such as splicing (Pan et al. 2008), non-coding RNA functions (Guttman et al. 2009) and genomic sites undergoing RNA editing (J. B. Li et al. 2009). In contrast to microarrays, HTS can also be employed without targeting specific sequences in the genome; thus, it is a better candidate for genome-wide exploratory research. Today, not only gene expression in bulk but also gene expression in single-cells, DNA-protein and RNA-protein interactions, chromatin-chromatin interactions, DNA accessibility and DNA methylation can be studied genome-wide using HTS-based assays.

2.2.2 Epigenomics

Histone modifications

Histones are proteins that form a bead-like tetrameric protein complex around which chromosomal DNA is wrapped in the nuclei of eukaryotic cells. DNA makes approximately 1.7 turns, or about 146 base pairs, around the protein complex. The resulting complex comprising DNA and the histone proteins is called chromatin, where the basic repeating unit is called the nucleosome. This mechanism organizes large genomes into highly-ordered compact structures in the nuclei.

Through post-translational modifications (PTM), histones can undergo various covalent modifications such as acetylation or methylation. Such modifications can activate or repress gene expression by either changing the accessibility of the DNA due to loosening or tightening of the DNA around the histones or by recruiting other factors.

The standard nomenclature for histone modifications is 1) the name of the histone (e.g. H3) 2) single-letter amino acid abbreviation and its position in the protein (e.g. Lysine 4) 3) the type of modification (e.g. *me* for methylation) 4) number of modification (e.g. *me1* for monomethylation, *me2* for dimethylation, etc.). For example, H3K4me1 and H3K4me3 modifications are highly enriched at enhancer and promoter regions in the genome, respectively (Calo and Wysocka 2013).

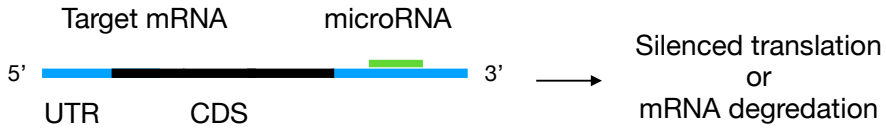


Figure 2.4: microRNAs regulates target genes by mediating translation silencing or mRNA degradation.

ChIP-seq

Chromatin immunoprecipitation sequencing (ChIP-seq) is an *in vivo* assay to identify genome-wide binding sites of DNA-associated proteins such as transcription factors. The protocol starts with cross-linking DNA and the protein complex of interest using formaldehyde. Next, the DNA is fragmented and protein-specific antibodies are used to immunoprecipitate the DNA-protein complex. DNA fragment that is bound to the protein complex is then extracted and sequenced (Furey 2012). The major steps of the computational pipeline are read mapping to the genome and peak calling, which employs statistical methods to distinguish protein binding sites from the noise of the data. ChIP-seq is a widely used assay for studying the binding sites of TFs or the regions enriched with histone modifications.

2.2.3 MicroRNAs

MiRNAs are small (19–24 nucleotide) non-coding RNAs that function in the degradation or silencing of translation. It is predicted that these small single-stranded RNA molecules regulate more than half of the human protein-coding genes (Bartel 2009). MiRNAs act through complementary binding to the untranslated region of a target gene (Figure 2.4A). The seed region of the binding site plays a vital role in target recognition (Lewis et al. 2003) and is often located between base pairs 2–7 on the 5' end of miRNAs.

2.2.4 Genotype data

A *single nucleotide polymorphism* (SNP) is a single base-pair variation in the genome where different alleles are observed in the population in relatively high frequency e.g. at least 1% of the population. Today, around 12.8 million SNPs are cataloged in the human genome (dbSNP, build 128, (Sherry et al. 2001)).The standard method for measuring SNPs, also called genotyping, is SNP arrays. In SNP arrays, allele-specific oligonucleotide probe sets are designed for all SNPs. Segmented, amplified and fluorescently labeled DNA fragments of SNP loci hybridize with the probe sets designed to match a specific allele and emit light which

shows that the allele is present in the fragment (Figure 2.3). Whole exome sequencing (WES) and whole-genome sequencing (WGS) are also used for measuring genetic variation.

The majority of SNPs are in biallelic form, meaning that they are represented with two alleles occurring in the SNP locus e.g. A and T. Genotype of an individual for a SNP represents the combination of two alleles from the copies of DNA from each parent at the SNP locus e.g. AA, AT and TT where AA and TT refer to homozygous genotypes, whereas AT is heterozygous. Genotypes are determined by combining the measurements of each allele. For example, for the samples with homozygous AA genotype, the intensity of the probes with A allele is higher compared to those with heterozygous AT genotype. Classification of the noisy measurement of fluorescent light intensities into genotypes is called *genotype calling*. Measurement noise can be taken into account in genotype calling methods (Carvalho et al. 2009).

Linkage disequilibrium (LD)

Linkage disequilibrium (LD) is defined as the nonrandom association of alleles at two or more loci (Slatkin 2008). This is mainly caused by the lack of recombination between the loci, which leads alleles to be cosegregated. From a probabilistic perspective, LD is a measure for testing the dependence of two random variables. In a given population, for two loci, A and B, the difference between the product of individual allele frequencies, p_A and p_B, and the joint probability, p_{AB}, represents the coefficient of linkage disequilibrium, commonly denoted as D: $D = p_{AB} - p_A p_B$. When $D = 0$, the inheritance of alleles is an independent event and the alleles A and B are in *linkage equilibrium*. The formula of D can also be interpreted as the difference between the observed and expected frequency of *haplotype AB* where haplotype refers to the cooccurrence of alleles A and B on the same chromosome. An alternative metric that is often used is the squared correlation between the allele frequencies p_A and p_B, which can be simply written in terms of D as $r^2 = \frac{D^2}{p_A(1-p_A)p_B(1-p_B)}$.

The LD structure provides the information that enables predicting genetic variants that are not directly measured via genotyping arrays or sequencing-based assays. Variants highly correlated with the measured variants due to LD can be predicted through genotype imputation, where sequencing-based reference panels are used to estimate haplotype (Marchini and Howie 2010). These imputed genotypes can be represented as probabilities of possible genotypes i.e. AA, AT and TT that sum up to one. These probabilities, called dosages, can be converted to "best-guess" genotypes by selecting the genotype with the highest probability. This conversion can be performed more stringently by defining a threshold e.g. 0.9 and setting the predicted genotype to unknown if the maximum probability does not exceed the threshold. However, this may lead to missing values which cause a major problem for models using genotypes as predictors.

GWAS

A *genome-wide association study* tests potential associations between phenotype and genotype. For quantitative phenotypes (e.g. LDL cholesterol and height), univariate linear regression is used for testing the association where phenotype measurement and genotype are used as dependent and independent variables, respectively. In the regression model, genotypes are typically encoded as $\{0, 1, 2\}$ where 0 and 2 represent homozygous genotypes for reference and alternative alleles and 1 represents heterozygous genotype. For dichotomous phenotypes (e.g. multiple sclerosis) χ^2 test is used for identifying SNPs where the frequency of one allele is significantly different between cases and controls. The resulting test statistic (odds ratio in χ^2 test and regression coefficient in linear regression) and p-value are used as the metrics for the association's magnitude and significance level.

The standard genome-wide significance threshold is defined as 5×10^{-8}, therefore the variants that are associated with the phenotype with a higher significance level (hence lower p-value) are defined as genome-wide significant GWAS hits. The rationale behind this threshold is that the number of independent loci is estimated as 150 per 500 kilobase pairs for Central European, Japanese and Chinese populations by the International HapMap Consortium in 2005 (Altshuler et al. 2005) which yields around a million independent loci when extrapolated to the whole genome (∼3.3Gb). Using a Bonferroni correction with a 0.05 significance level then leads to the standard genome-wide threshold i.e. $0.05/10^6 = 5 \times 10^{-8}$.

Quantitative trait loci (QTL) studies

A quantitative trait locus (QTL) is a locus where genetic variation correlates with the variation of a (typically molecular) quantitative phenotype. For example, the association of the expression of a gene with a genotype is called expression QTL (eQTL). There are also other types of QTLs assessing the associations of different molecular readouts, such as DNA methylations and histone modifications with genotype. Such QTLs are called meQTLs[2] and histoneQTLs, respectively. Depending on the distance between the variant and the locus of the molecular phenotype (e.g. gene position), QTLs are characterized as either *cis* and *trans*. *cis*-eQTLs are loci of genetic variation that is close (typically $\leq$ 1Mb) to the target gene, which are sometimes called *eGenes*, and are hypothesized to mediate the expression of a gene locally via altering the chromatin structure or transcription factor binding. *trans*-eQTLs reside far from the variant (either $>$ 1Mb or on a different chromosome) and act through an intermediary factor called a *cis*-mediator such as a transcription factor (Q. Li et al. 2013) (Figure 2.5).

eQTL studies have been useful for understanding gene regulation and interpreting the results of GWAS. Since many GWAS hits are in non-coding regions of the genome, eQTLs

[2]Expression quantitative trait methylation (eQTM) which quantifies the association between the gene expression and methylation is sometimes confused with meQTL.

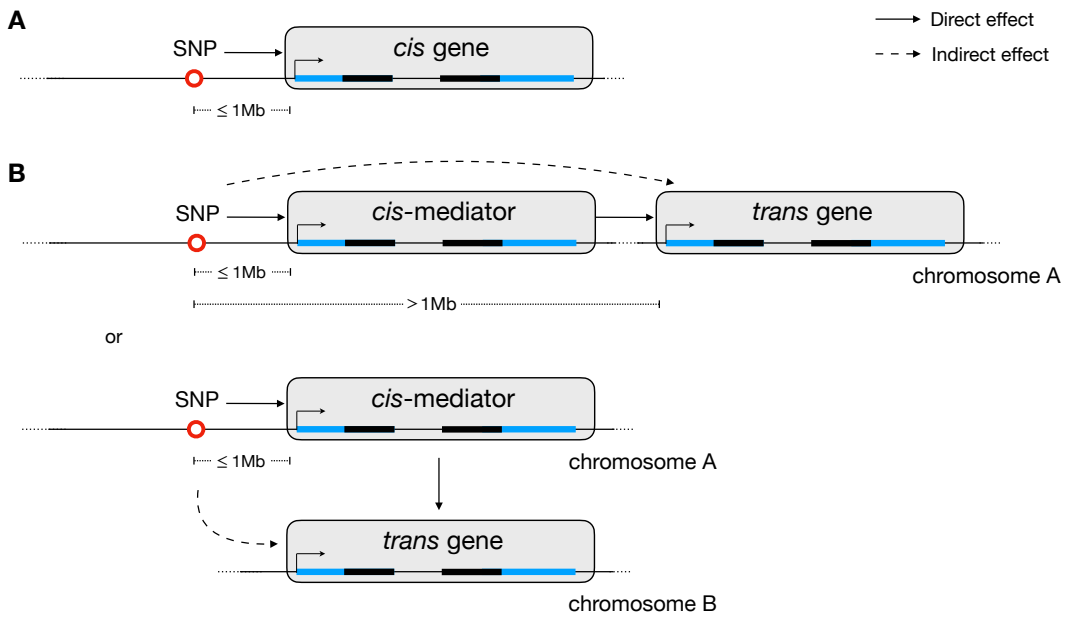


Figure 2.5: *cis*- and *trans*-eQTLs are loci where genetic variation is significantly associated with gene expression. The target gene is typically within 1Mb of the *cis*-eQTLs (**A**). *trans*-eQTLs affect the expression of the target gene, which is more distant than 1Mb on the same chromosome or a different chromosome through other genes called *cis*-mediators (**B**).

facilitate hypothesizing about the mechanism of action for the GWAS hits. For example, overlapping GWAS hits with eQTLs might reveal through which gene a variant is linked to the disease or phenotype (GTEx Consortium 2017).

2.3 Public data sources

2.3.1 ENCODE Project

The Encyclopedia of DNA Elements (ENCODE) is a collaborative effort aiming to characterize all functional elements in the non-protein-coding regions of the genome (ENCODE Project Consortium 2007) which are initially thought to be "junk". Various types of assays are applied to several human cell lines and tissues as a part of the project to investigate. As a part of the project, gene expression, DNA-protein interactions, RNA-DNA interactions, DNA accessibility and chromatin-chromatin interactions are investigated mainly in human cell lines via thousands of experiments using various types of assays. ENCODE provides invaluable data resources, especially for the researchers studying regulatory genomics by assessing sequence specificities of TF binding sites or histone modifications using simple motif-based approaches or deep sequence models.

2.3.2 Roadmap Epigenomics

Roadmap Epigenomics set out to produce publicly available human reference epigenomes with tissue-specific DNA methylation, histone modifications and chromatin accessibility datasets (Kundaje et al. 2015). Currently, the project portal provides 127 epigenomes (111 from Roadmap Epigenomics and 16 from ENCODE) with 31 histone modifications as well as DNase-seq and DNA methylation tracks from many primary human tissues, embryonic stem cells and various cell lines.

2.3.3 GTEx

The Genotype-Tissue Expression (GTEx) project provides a reference bulk RNA-seq data resource and a tissue bank to study tissue-specific gene expression, regulation and eQTLs in non-diseased human tissues collected from postmortem donors (Lonsdale et al. 2013). The latest version of GTEx consists of 7,051 samples from 449 donors across 44 human tissues (31 solid-organ tissues, 10 brain regions, whole blood, two cell lines and skin samples), each from at least 70 donors (GTEx Consortium 2017). In addition to the publicly available RNA-seq datasets, the resource provides *cis*- and *trans*-eQTL for all tissues, eQTL-disease associations, allele-specific expression and tissue-specific alternative splicing information.

Chapter 3

Functional genotype-phenotype associations and variant prioritization

Identifying the genetic determinants of complex human traits and diseases is one of the primary goals in human genetics. Towards this goal, genome-wide association studies (GWAS) emerged as a powerful tool for linking the sequence variation in the genomes of human populations to various phenotypes (J. J. Lee et al. 2018). Potential determinants in the form of phenotype-genotype associations are valuable as they narrow down the search space of the entire genome to fewer loci. However, due to the *linkage disequilibrium (LD)*, a phenomenon that gives rise to non-random associations of variants in genomic blocks, many variants in proximity are highly correlated. Such variants may exhibit statistical associations at similar levels making it highly challenging to disentangle the causal effects of each variant within a given block of correlated variants solely using the GWAS statistics (Figure 3.1A-B). Therefore, the resolution of GWAS results is rather limited to such LD blocks *(haplotypes*). Typically, only the variant with the strongest association with phenotype, called the *lead variant*, is reported for each independent locus instead of the variants that are highly correlated in the proximity of the lead variant called the *proxy variants*.

At this haplotype-level resolution, the difficulty of pinpointing causal variants poses a major roadblock towards the ultimate goal of revealing the mechanisms underlying the diseases and traits (Tak and Farnham 2015). This problem is exacerbated by the fact that most variants are single-nucleotide polymorphisms (SNPs) in non-coding regions according to the catalogs of disease-associated variants that were identified through GWAS (Leslie et al. 2014). This adds another layer of complexity as it is often more challenging to characterize non-coding variants than protein-coding variants.

Variant prioritization methods aim to identify variants that are potentially causal for a phenotype by scoring them or predicting their effects using additional sources of information.

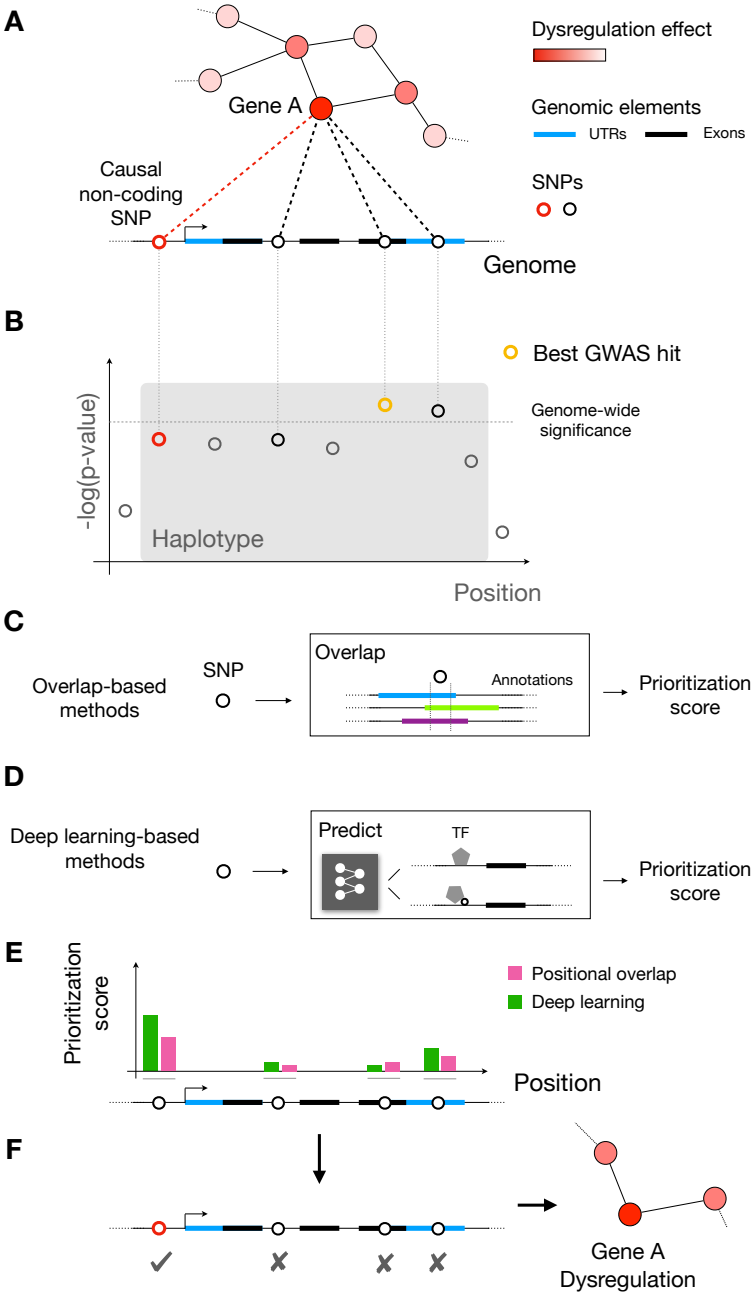


Figure 3.1: Visualization of variant prioritization. The dysregulation of *Gene A* due to a non-coding variant is depicted as the cause of a disease. The effect of the dysregulation propagates in the hypothetical regulatory network **(A)**. Typical GWAS identifies the correct haplotype. However, the causal non-coding variant does not reach genome-wide significance **(B)**. Variant prioritization approaches take a variant as input, infer its regulatory effect, and output a prioritization score. Two major types of methods are overlap-based **(C)** and deep learning-based **(D)** approaches. The comparison of the prioritization scores of given variants **(E)** suggests the correct regulatory variant as causal. Thanks to the prioritization methods, it can be hypothesized that the dysregulation of *Gene A* through the identified regulatory variant has a vital role in disease etiology **(F)**.

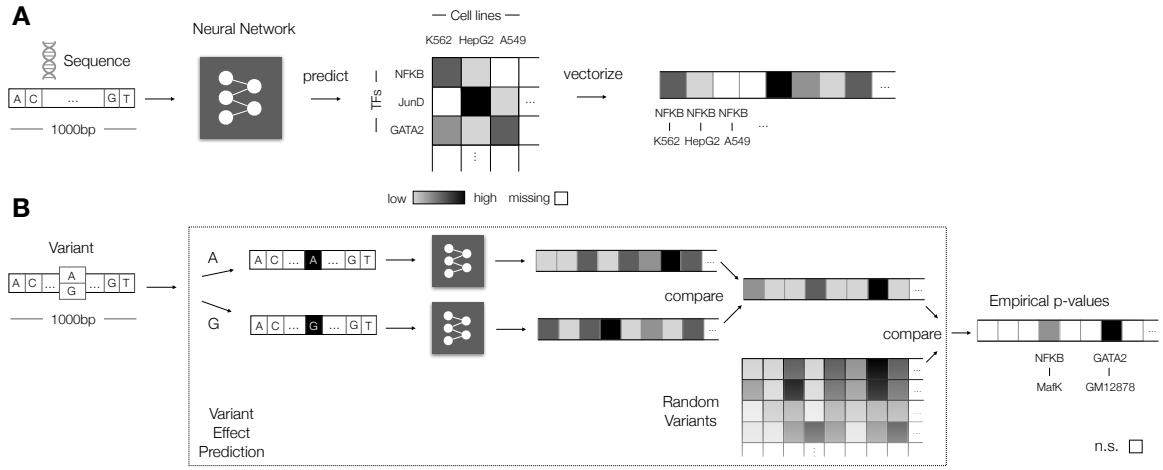


Figure 3.2: Variant prioritization using neural networks. **(A)** A convolutional neural network is trained for predicting the binding of several transcription factors (TF) in various cell lines from the given DNA sequences using available experimental datasets (e.g. ChIP-seq). **(B)** 1kb sequence where the variant in question is at the center is fed through the pre-trained neural network with two different alleles. A regulatory score for the given variant is calculated via comparing predictions of two alleles, which is compared to a group of randomly selected background variants. This yields the final significance of the regulatory impact of the variant.

Two major types of variant prioritization methods are overlap-based and deep neural network-based approaches. Overlap-based methods predict the potential effect of given variants by evaluating the positional overlap between SNPs and various annotations of genomic elements such as the transcription factor (TF) binding sites or microRNAs (Figure 3.1C). A major drawback of methods based on positional overlap is that the actual impact of each variant on regulatory elements is not assessed because the association is simply based on the positional overlap. For example, two SNPs that colocalize in the same ChIP-seq peak of a TF might have opposing effects or no functional effects at all. *In silico* approaches that predict the disruption of TF binding motifs (Thomas-Chollier et al. 2011; S. G. Coetzee, G. A. Coetzee, et al. 2015) have been proposed to resolve this shortcoming. However, our knowledge of TF binding motifs as well as the mechanistic understanding and contribution of individual motif elements to the binding event are still incomplete, limiting the success of such methods.

The second group consists of convolutional neural networks, which are increasingly preferred for variant prioritization (Figure 3.1D, Figure 3.2). In this approach, first, the regulatory effect of each allele of a given variant is predicted using a neural network that is pre-trained to predict TF binding events. Second, the comparison of these predicted effects yields an overall regulatory effect of the given variant. Unlike the positional overlaps, neural networks can estimate the importance of a base pair change in a given locus using the DNA sequence patterns of TF binding sites learned from the data. The step following the prioritization typically aims to hypothesize about the mechanism of action in the light of new associations and predictions (Figure 3.1E-F).

Towards developing better hypotheses of complex diseases and traits, we developed two variant prioritization approaches. Misina, a novel overlap-based variant prioritization approach, provides a new angle on the microRNA-mediated effects of variants in microRNA regulation loci in the context of complex diseases. Misina allows rapid exploration and prioritization of user-provided or disease-associated SNPs residing in microRNA binding sites of target genes by querying an integrated framework of seven SNP- and miRNA-related data sources. Misina provides an easy-to-use analysis interface for non-expert researchers seeking to investigate microRNA-mediated effects of genetic variants by automating many tedious steps required for complex queries such as LD proxy search, consideration of microRNA seed type and relative position of SNP in the binding site, matching tissue-specific eQTL genes with the microRNA target genes, and the exploration of microRNA expression in various tissues.

Our second method DeepWAS is a multivariate genotype-phenotype association approach that utilizes the predictions of the cell type-specific regulatory effect of single variants from a pre-trained deep neural network (J. Zhou and Troyanskaya 2015). As a result, DeepWAS associates variants not only with a trait or disease but also proposes a regulatory mechanism, for example, a variant effect on altered transcription factor binding or DNA accessibility in a specific cell type. We analyzed genotype data from three different cohorts, namely multiple sclerosis (MS), major depressive disorder (MDD) and height, to reveal regulatory determinants of these phenotypes. Using publicly available data, we underpinned the functionality of identified putatively regulatory SNPs in specific tissues and disease contexts. We demonstrated the potential of the DeepWAS method to generate testable hypotheses from genotype data, even for small sample sizes.

Although these two methods use very different techniques and data sources, they both provide new perspectives on the regulatory determinants of complex phenotypes in a complementary manner by covering different components of the regulome. In addition to these two approaches which constitute most of this chapter, we present other approaches developed to analyze coding and non-coding variants in the literature in Section 3.1.

The results reported in this chapter are part of the following publications and/or book chapters. The contributions of the author are given below each publication.

- **Section 3.1**: Nikola Müller, Ivan Kondofersky, **Gökcen Eraslan**, Karolina Worf, Fabian J. Theis. "Bioinformatics in Psychiatric Genetics." *Psychiatric Genetics: A Primer for Clinical and Basic Scientists* (2018) `https://doi.org/10.1093/med/9780190221973.001.0001`

  Contribution of the author: Literature review of different types of variant prioritization tools used for analyzing coding and noncoding variants and writing

- **Section 3.2**: **Gökcen Eraslan**, Nikola S. Mueller, Fabian J. Theis. *Misina: Finding microRNA-mediated effects of genetic variants with an integrative approach* (in preparation)

  Contributions of the author:

  - Design and implementation of the variant prioritization method
  - Integration of various databases of SNPs, LD proxies, eQTLs, microRNA targets, microRNA expression and genotype-phenotype associations
  - Investigation of variants associated with Alzheimer's disease
  - Interpretation of results
  - Generating figures and writing

- **Section 3.3**: Janine Arloth*, **Gökcen Eraslan***, Till F. M. Andlauer, Jade Martins, Stella Iurato, Brigitte Kühnel, Melanie Waldenberger, Josef Frank, Ralf Gold, Bernhard Hemmer, Felix Luessi, Sandra Nischwitz, Friedemann Paul, Heinz Wiendl, Christian Gieger, Stefanie Heilmann-Heimbach, Tim Kacprowski, Matthias Laudes, Thomas Meitinger, Annette Peters, Rajesh Rawal, Konstantin Strauch, Susanne Lucae, Bertram Müller-Myhsok, Marcella Rietschel, Fabian J. Theis, Elisabeth B. Binder, Nikola S. Mueller *DeepWAS: Multivariate genotype-phenotype associations by directly integrating regulatory information using deep learning* (PLoS Computational Biology 16, no. 2 (2020): e1007616, `https://doi.org/10.1371/journal.pcbi.1007616`) *These authors contributed equally

  Contributions of the author:

  - Design and implementation of the variant prioritization method
  - Prediction of the regulatory effects of non-coding variants
  - Application of the method to MS, MDD and height phenotypes
  - Functional characterization of the hits
  - Integration of binding QTLs, GTEx bulk RNA-seq (GTEx Consortium 2017) and graph database-based (Neo4j) visualizations
  - Interpretation of results
  - Generating figures and writing

3.1 Literature review of variant prioritization

This subsection gives a brief overview of existing methods for the prioritization of coding and non-coding variants.

3.1.1 Coding variants

Many tools exist for annotating coding variants. Examples include ANNOVAR (K. Wang et al. 2010) and VEP (Ensembl's Variant Effect Predictor) (McLaren et al. 2016) which are widely-used tools overlapping genetic variants with regions of the genome and providing many filtering and annotation options. Another prominent tool in the field is GEMINI (Paila et al. 2013) which integrates publicly available annotations such as ENCODE (ENCODE Project Consortium 2007), UCSC (Hsu et al. 2006), KEGG (Kanehisa and Goto 2000), and user-defined annotations as well as cohort genotype and phenotype data. Its efficient storage

and query framework facilitates the downstream analysis. The key benefit of these tools is that they save researchers from making manual queries to different sources of variant annotation and prediction using alternative transcript sets e.g. ENSEMBL (Hubbard et al. 2002), RefSeq (Pruitt et al. 2006). The use of different transcript sets can have drastic effects on annotation-based queries, as McCarthy et al. (2014) have demonstrated. According to the study, there is only a 44% agreement between the results of RefSeq and ENSEMBL annotations when ANNOVAR is used to identify putative loss-of-function variants with these two transcript sets.

The evaluation of gene enrichment for groups of variants is an alternative approach to variant prioritization. VEGAS (J. Z. Liu et al. 2010) is an example of enrichment tests per gene to determine the genes that harbor variants associated with the disease. Similarly, other tools such as DEPICT (Pers et al. 2015), INRICH (P. H. Lee et al. 2012) and GRAIL (Raychaudhuri et al. 2009) allow the genes linked to GWAS hits to be tested for pathway enrichment. Gene-based scoring initiatives such as the Residual Variation Intolerance Score (RVIS) are also worth mentioning where the genes are ranked based on how well they can tolerate mutations (Petrovski et al. 2013). The objective of the EXAC project is to consolidate exome-sequencing data from many researchers to provide a comprehensive resource on coding variants (Lek et al. 2016). Along with the genic intolerance scores in RVIS, annotations of rare variants in EXAC are useful resources that can be used to prioritize rare coding variants.

Coding variants are often evaluated according to their effect on the protein structure. SIFT (Ng and Henikoff 2003) and PolyPhen (Adzhubei et al. 2010) are popular tools that can predict whether changes in amino acid sequence alter the structure and function of proteins. In their outputs, most of the variant annotation tools also include SIFT and PolyPhen scores.

3.1.2 Non-coding variants

Prioritization of non-coding variants is a more difficult task since less evidence is available regarding the function of non-coding regions. The assessment of the enrichment of cis-regulatory elements such as promoters and enhancers for specified non-coding variants is a common practice in many published GWAS (Tak and Farnham 2015). Comprehensive epigenetic and regulatory annotations reported by ENCODE (ENCODE Project Consortium 2007) and Roadmap Epigenomics (Kundaje et al. 2015) projects are commonly used by researchers for employing such enrichment tests (Schizophrenia Working Group of the Psychiatric Genomics Consortium 2014).

The investigation of non-coding variants that putatively affect the function of small non-coding RNAs is also used to shed light on disease etiology (Hauberg et al. 2016). Other integrative approaches that can be used to query the overlap of the given variants with regulatory elements include Haploreg (L. D. Ward and Kellis 2016), RegulomeDB (Boyle et al. 2012) and FunciSNP (S. G. Coetzee, Rhie, et al. 2012). Machine learning methods such as

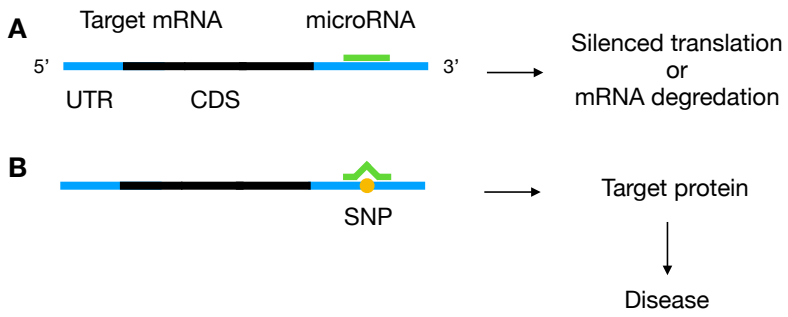


Figure 3.3: microRNA-mediated dysregulation effects of variants. microRNAs regulates target genes by mediating translation silencing or mRNA degradation **(A)**. Genetic variation within the microRNA binding site might impair microRNA binding leading to the dysregulation of the target gene and the disease **(B)**.

GWAVA (Ritchie et al. 2014) and CADD (Kircher et al. 2014) prioritize variants based on variant pathogenicity predicted from various variant characteristics. Likewise, FunSeq2 (Fu et al. 2014) uses a scoring scheme where variant features are weighted for the prioritization task.

A major problem in the analysis of GWAS hits and their proxies is that the variants that do not meet the genome-wide significance, so-called sub-threshold variants, might play a role in the development of the disease via various biological mechanisms (Figure 3.1). Examination of such variants revealed new loci associated with cardiac traits where several variants were found to be enriched in enhancer regions (Xinchen Wang et al. 2016). A comprehensive review of functional analysis and follow-up studies of GWAS can be found in Edwards et al. (2013), L. Hou and H. Zhao (2013) and Tak and Farnham (2015).

3.2 Identifying microRNA-mediated effects of genetic variants

miRNA binding sites are a key component of the regulome since miRNAs play a critical role in post-transcriptional gene regulation. miRNA-mediated dysregulation effect of SNPs was associated with diseases in previous studies (Chin et al. 2008; Esteller 2011) (Figure 3.3). Therefore, the interrogation of SNPs in miRNA binding sites provides a means for identifying functional variants, which might shed light on the pathogenesis of a disease. In this section, with our overlap-based variant prioritization approach called Misina, we focus on the interrogation and prioritization of non-coding variants in miRNA gene regulation loci.

The key aspects of our approach compared to existing methods are as follows:

1. We integrate the linkage disequilibrium (LD) data of given phenotype-associated SNPs to identify all potential SNPs that might lead to miRNA-mediated dysregulation of the target gene. This is necessary because phenotype-associated SNPs identified in GWAS

only establish a potential link between the phenotype and the haplotype block which consists of several SNPs in addition to reported phenotype-associated SNP.

2. There are several miRNA target prediction and validation approaches with different assumptions and strengths. Reporting miRNA target genes from several miRNA target datasets enables researchers to either investigate a consensus of different approaches or focus only on preferred datasets.

3. Misina uses key pieces of miRNA binding information such as miRNA seed type and variant position within the binding site in order to perform variant prioritization and scoring.

4. The experimental evidence of the expression of both the target gene and the miRNA is important to hypothesize miRNA-mediated dysregulation. Matching the tissues where miRNA and target gene are expressed may further strengthen the dysregulation hypothesis.

We designed an integrative approach using a user-defined set of input SNPs or already cataloged SNPs. Interactive web interface of Misina aims to provide an easy-to-use analysis tool for non-expert users seeking to investigate miRNA-mediated effects in GWAS results. We included an LD proxy search and overlapped SNPs with data from experimentally validated miRNA-target interactions (starBase (J.-H. Li et al. 2013)) and two target prediction databases (miranda (Betel et al. 2010) and TargetScan (V. Agarwal et al. 2015)). We chose two criteria indicating that a SNP was likely to impair a miRNA binding site, namely 7- or 8-mer seed type (strong initial binding) and relative position of SNP in the binding site. The resulting miRNA-SNP pairs were enriched with known eQTLs providing the third scoring criteria suggesting mechanistic effects if the same SNP-gene pairs were associated in the eQTL studies. Furthermore, the human tissues where the listed miRNA are expressed are displayed. The novel yet simple and practical way to prioritized resulting SNPs aimed to guide non-expert users through the results to reveal high-confidence pairs.

3.2.1 SNP Prioritization

We prioritized SNPs to guide users in pinpointing disruptive miRNA/SNPs pairs, which may be functional. Using expert knowledge, we implemented the following miRNA/SNP scoring scheme:

- miRNA seed type: SNPs are prioritized if they fall into the binding site of a miRNA whose seed type is of either 7– or 8–mer. In the literature, it was reported that four types of canonical sites correlate with targeting efficacy such that 8mer $>$ 7mer–m8 $>$ 7mer-A1 $>$ 6mer (Bartel 2009). Therefore, we prioritized 8– and 7–mer seed types.

- Relative SNP position: SNPs within the miRNA binding site (SNPs in 1–12 bp from 5' end of miRNA) are prioritized. Although it is proposed that target recognition occurs based on the match in the seed region (i.e. 2–8nt of 5' end), it has been reported that off-seed SNPs can also have a major impact on target regulation (Xiuchao Wang et al. 2016; Dorn et al. 2012).

- miRNA target – eGene match: SNPs are prioritized if miRNA target gene match eQTL gene (miRNA gene identical to eGene). It is known that miRNAs play an essential role in the regulation of gene expression. The SNPs within the miRNA binding site might alter the mRNA-miRNA binding and lead to dysregulation of target genes. In this case, investigating the expression of the target gene and, more importantly, how significantly it changes due to the alterations in the binding site is worthwhile. The integration of GTEx dataset enables us to detect such significant changes in the expression of target genes. Therefore, SNPs are prioritized if the miRNA target is reported as an eGene (significant eQTL gene).

SNPs that satisfy these rules are specified on the web interface and are sorted by total scores (ranging from 0 to 3). In addition, the expression of the listed miRNAs in human tissues that are potentially affected by SNPs can be examined through miRmine (Panwar et al. 2017) and miR Tissue Atlas (Ludwig et al. 2016).

3.2.2 Design of the Misina framework

Seven SNP- and miRNA-associated data sources were integrated for Misina (Figure 3.4). First, a user-defined set of (risk) SNPs (dbSNP identifiers) was used for analysis. As an important feature of Misina, users can select any phenotype of interest cataloged by GRASP Project version 2.0 (Eicher et al. 2015) spanning 8.87 million SNPs from 2082 studies. Second, all SNPs in high LD with the given risk SNPs were identified. The most recent LD information was automatically retrieved from SNiPA (Arnold et al. 2014). LD r^2 cutoff and 1000G population can be both configured. Third, hg19 (Human Genome version 19) genomic coordinates of all SNPs were retrieved from dbSNP build 142 (Sherry et al. 2001) and SNPs that fall into the miRNA binding sites were determined by overlapping genomic positions. Resulting miRNA-gene-pairs consisted of experimentally validated and/or predicted miRNA targets from TargetScan v7.0 (V. Agarwal et al. 2015), miranda (Betel et al. 2010) and starBase 2.0 (J.-H. Li et al. 2013). Finally, identified SNPs were searched in GTEx v6 eQTL dataset (Lonsdale et al. 2013) and GWAS catalog (Buniello et al. 2019). Hits that were also eSNPs in GTEx dataset and SNPs or genes reported in the GWAS catalog were annotated. miRNA expression in human tissues is also available thanks to miRmine and miRNA Tissue Atlas datasets (Panwar et al. 2017; Ludwig et al. 2016).

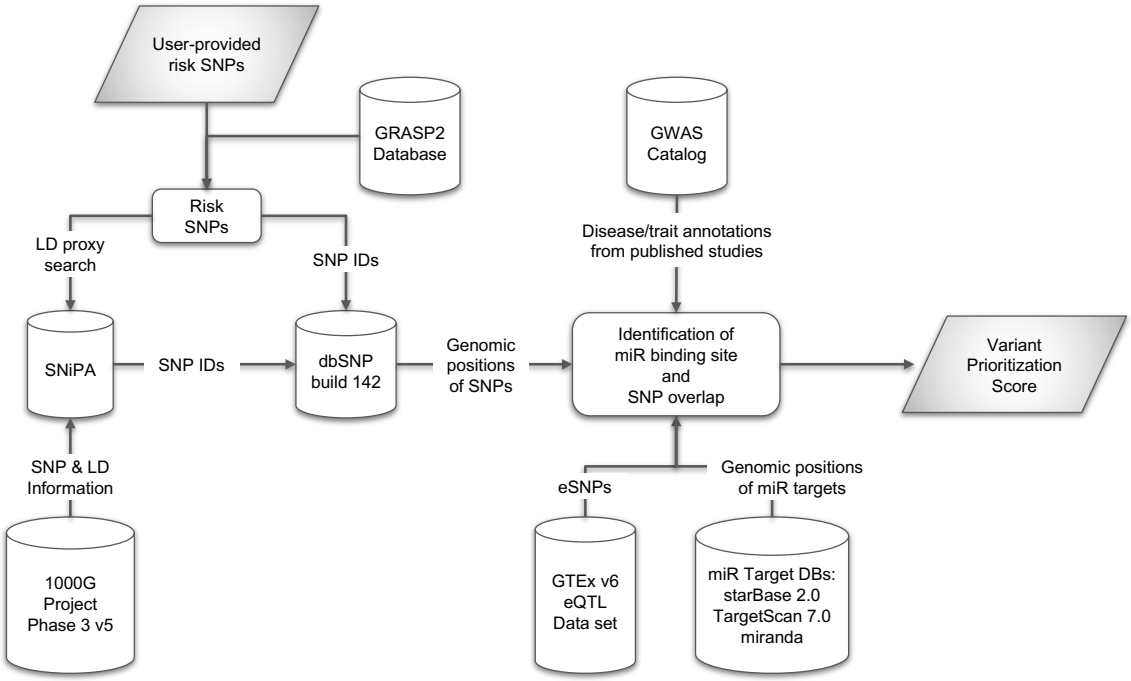


Figure 3.4: Workflow of the integrative approach. Misina integrates seven SNP- and miRNA-related data sources to provide a user-friendly analysis interface.

3.2.3 Implementation

Misina was implemented in R programming language (R Core Team 2015) using the Shiny web framework (Chang et al. 2015) and Bioconductor packages (Huber et al. 2015). To handle time-consuming LD queries, a simple parallelized asynchronous job management system was implemented. Genomic regions were overlapped using the GRanges/Bioconductor package (Lawrence et al. 2013). Source code of Misina is available at `https://github.com/cellmapslab/misina`.

3.2.4 miRNA-mediated determinants of Alzheimer's disease

Potential role of miRNAs in complex human diseases including Alzheimer's disease (Delay et al. 2012; Femminella et al. 2015), Parkinson's disease (G. Wang et al. 2008), diabetes (Lv et al. 2008; X. Zhao et al. 2013), has been previously studied. However, miRNA-mediated determinants of these diseases along with functional mechanisms have not been described yet. Here we used Misina to prioritize Alzheimer's disease risk SNPs that are potentially contributing to the disease mechanism by inducing a regulatory dysfunction through disrupting or enhancing the binding of miRNAs.

We queried Misina with the Alzheimer's disease-associated SNPs from the GRASP2 catalog

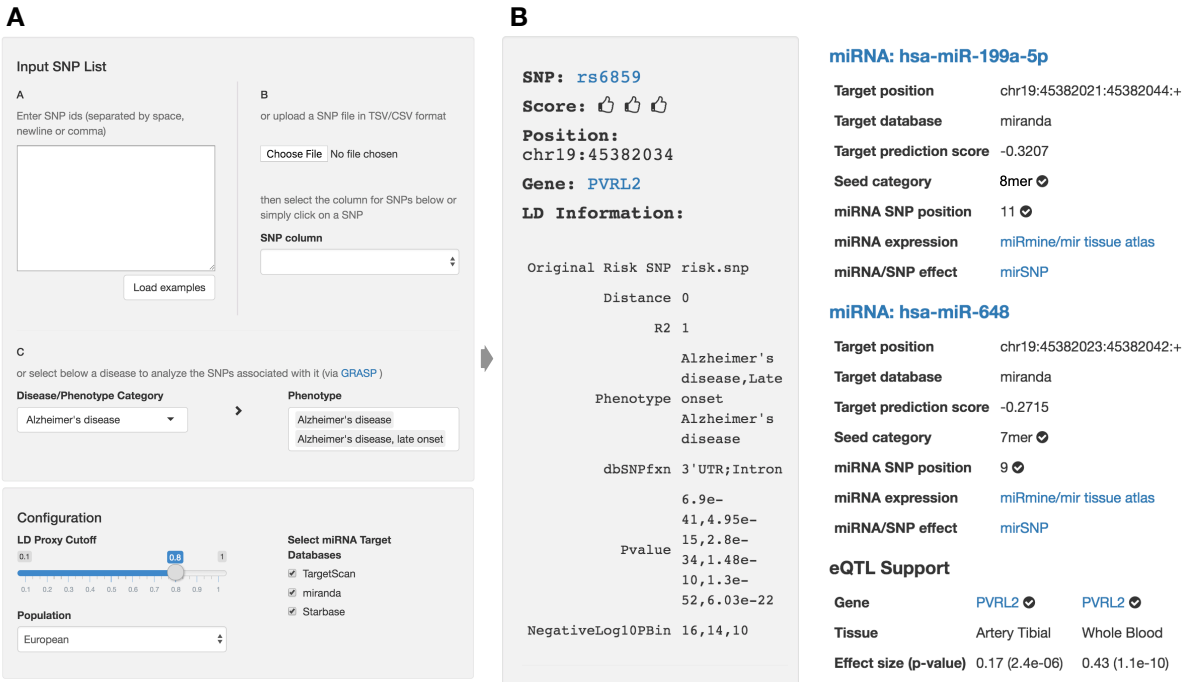


Figure 3.5: Investigation of the genetic factors of Alzheimer's disease using Misina web interface. **(A)** The input screen where Alzheimer's disease-associated SNPs are entered to the input area for the analysis. **(B)** The result of the Alzheimer's analysis. The top hit, rs6859, has the triple score as it overlaps with miRNAs with 8-mer seed type, relative SNP position within miRNA binding site is in 1-12 bp from 5' end of miRNA and miRNA target gene, *PVRL2* (*NECTIN2*), is also identified as an eQTL gene. This result is displayed along with the information about SNP and LD, miRNAs with overlapping binding sites and genes and tissues that rs6859 is associated with as an eSNP in eQTL studies.

(GRASP2 categories: Alzheimer's disease and Alzheimer's disease late-onset, Eicher et al. 2015) (Figure 3.5A). Our query yielded rs6859, a multiallelic risk SNP (A>G, A>T) located in the 3' UTR of PVRL2 (NECTIN2) gene (isoform NECTIN2-202, ENST00000252485), as the top hit with maximum score and the candidate miRNAs (e.g. hsa-miR-199a-5p and hsa-miR-648) targeting NECTIN2 with 8-mer or 7-mer seed types (Figure 3.5B). Among these miRNAs, hsa-miR-199a-5p binds to the 3' UTR of NECTIN2 where rs6859 resides in the 11th position of the 5' end of the miRNA. Moreover, eQTL information provided by Misina showed that the expression of NECTIN2 was significantly associated with rs6859 in the arteries and whole blood samples of GTEx.

Characterization of hsa-miR-199a-5p, rs6859 and NECTIN2 relationship

To characterize the relationship between the three major components of our hypothesis of a miRNA-mediated determinant of Alzheimer's disease, namely rs6859, hsa-miR-199a-5p and NECTIN2 (Figure 3.6A), we sought to examine further experimental evidence from publicly available data sources. As a part of the validation of this hypothesis, miRNA-mRNA

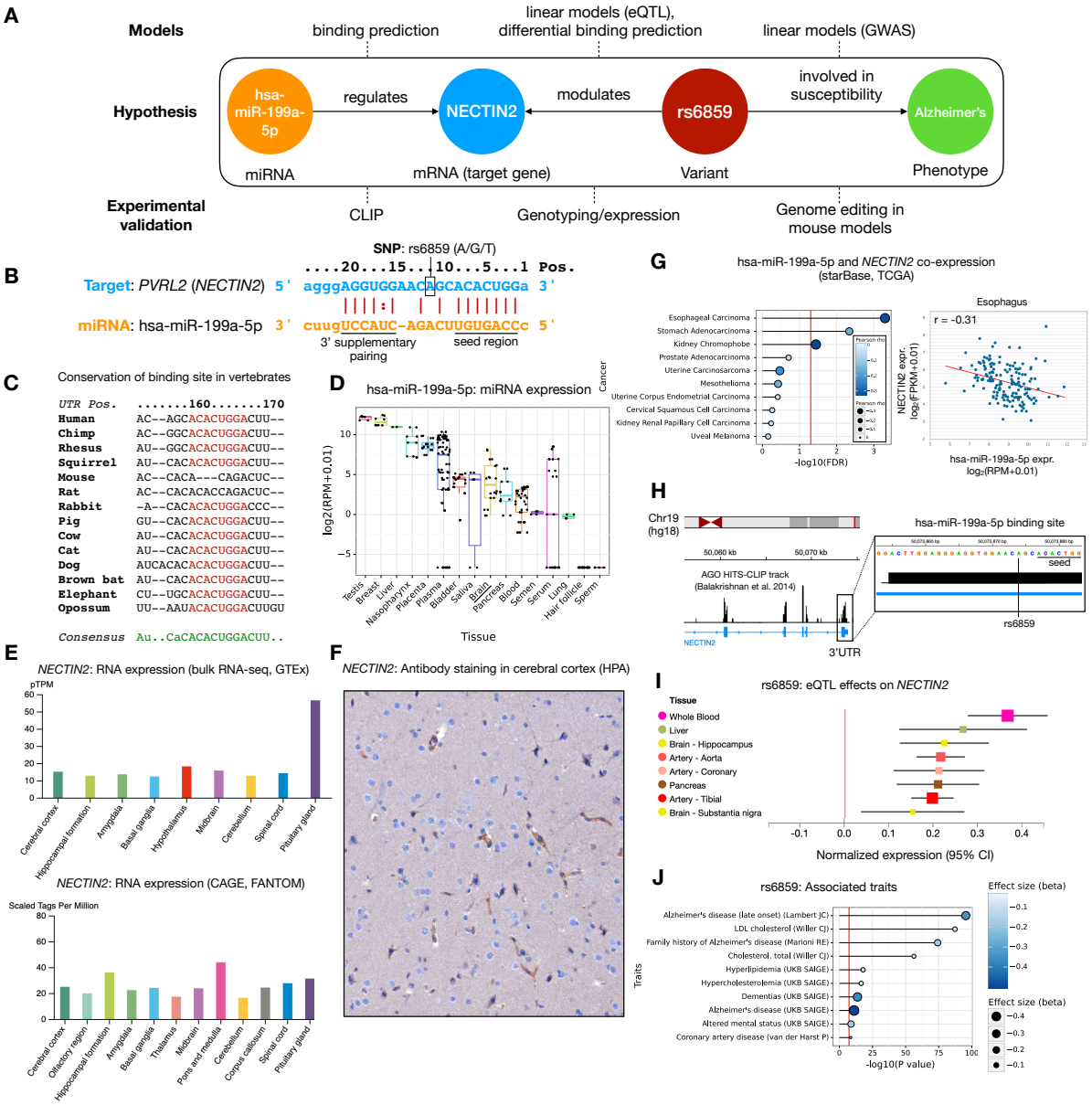


Figure 3.6: Experimental evidence to characterize the links between the variant, gene, miRNA and phenotype. **A)** Diagram of the hypothesis with computational and experimental tools used to construct it. **B)** Alignment of hsa-miR-199a-5p and its binding site in the 3' UTR of the target gene *NECTIN2*. G-U wobble is shown as ":". **C)** Conservation of hsa-miR-199a-5p binding sites among vertebrates. Fully conserved seed region is shown in red. **D)** Expression of hsa-miR-199a-5p in the tissues. **E)** Bulk RNA-seq and CAGE data showing *NECTIN2* expression in various CNS samples. **F)** Antibody staining (DAB) of *NECTIN2* showing protein expression in the cerebral cortex. Brown staining indicates cells expressing *NECTIN2* protein. **G)** Top 10 cancer types where hsa-miR-199a-5p and *NECTIN2* expression are highly anti-correlated (TCGA) and the co-expression of each sample in esophagus. Red line denotes the significance threshold (0.05, left) and regression line (right). **H)** HITS-CLIP coverage track showing the binding sites of Argonaute protein in HUVEC cell lines indicating the miRNA binding sites in *NECTIN2* 3'UTRs. Coverage signal in the hsa-miR-199a-5p binding site is magnified. **I)** eQTL effects of rs6859 on *NECTIN2* expression in various tissues including the brain. **J)** Traits and diseases significantly associated with rs6859 in GWAS. Red line indicates genome-wide significance threshold. **Acronyms:** TCGA: The Cancer Genome Atlas, HPA: Human Protein Atlas, eQTL: expression quantitative loci, CAGE: Cap analysis gene expression, UKB: UK Biobank, DAB: 3,3'-diaminobenzidine, FPKM: Fragments per kilobase of transcript per million mapped reads, RPM: Reads per million mapped reads, pTPM: transcripts per million protein coding genes, CLIP: Cross-Linking and Immuno-Precipitation.

interactions predicted with computational methods can be experimentally validated via CLIP (Cross-Linking and Immuno-Precipitation) experiments where the cross-linked miRNA-mRNA complex is captured by the immunoprecipitation of the Argonaute (AGO) protein (Chi et al. 2009). Linear models are typically applied to genotype-phenotype data to test variant-gene expression and variant-phenotype associations. For the experimental validation, genetic mouse models for Alzheimer's disease can be leveraged to test whether this variant has a causal effect on the disease etiology (Onos et al. 2016). Although the direct experimental validation of this hypothesis is not available, we were able to compile a list of findings below to support the hypothesis.

First, we inspected the miRNA-RNA binding pattern by aligning miRNA and the binding site (Figure 3.6B). In addition to the seed pairing, the alignment showed a secondary site with a G:U wobble in position 15-20 similar to a 3'-supplementary pairing, which is typically observed in non-canonical binding sites (Bartel 2009). Moreover, we observed that the hsa-miR-199a-5p binding site in NECTIN2 is highly conserved among the vertebrates (Figure 3.6C) indicating the conservation of the miRNA-mediated regulatory mechanism across species.

Second, we inspected the expression of both hsa-miR-199a-5p and the target gene NECTIN2 in brain samples available from public data sources. miRmine (Panwar et al. 2017) showed the expression of hsa-miR-199a-5p in the tissues including the brain, which is relevant to Alzheimer's etiology (Figure 3.6D). The inspection of GTEx, FANTOM and Human Protein Atlas (HPA) data sources revealed the ubiquitous RNA and protein expression of NECTIN2 in the brain via bulk RNA-seq, CAGE and antibody staining experiments (Figure 3.6E-F). According to the annotation of the antibody staining image by the HPA project, NECTIN2 is expressed at moderate levels in endothelial cells in the cortex (Figure 3.6F).

Third, we queried starBase v2.0 (J.-H. Li et al. 2013) to find out the co-expression of hsa-miR-199a-5p and NECTIN2, as well as further experimental evidence for the miRNA binding. We evaluated the miRNA-mRNA co-expression in 32 cancer types from the TCGA pan-cancer network (`https://www.cancer.gov/tcga`) where samples were profiled via miRNA-seq and RNA-seq experiments allowing us to assess co-expression. hsa-miR-199a-5p and NECTIN2 were significantly anti-correlated in 3 out of 32 cancer types, namely esophageal carcinoma ($n = 162$, Pearson $\rho = -0.31$, FDR$= 4.98e-4$), stomach adenocarcinoma ($n = 372$, Pearson $\rho = -0.162$, FDR$= 1.73e-3$) and kidney chromophobe ($n = 65$, Pearson: $\rho = -0.32$, FDR$= 3.62e-2$) (Figure 3.6G). Furthermore, starBase showed Argonaute (AGO) binding signal in this region using the HITS-CLIP (HIgh-Throughput Sequencing of RNAs from in vivo Cross-Linking and Immuno-Precipitation) experiment in HUVEC cell line produced by Balakrishnan et al. (2014) indicating miRNA activity in the hsa-miR-199a-5p binding site (Figure 3.6H).

Finally, we investigated the effects of rs6859 on gene expression and complex traits and

diseases via eQTL and GWA studies. eQTL data from the GTEx project identified a significant association of rs6859 with the expression of NECTIN2 in many tissues, including the brain (Figure 3.6I), which supports our hypothesis. Moreover, in addition to Alzheimer's disease, rs6859 is associated with lipid-related and cognitive traits and diseases, including dementia, cholesterol, hypercholesterolemia, and coronary artery disease (Figure 3.6J), which indicates that rs6859 locus might affect the regulatory circuitry linked to many complex phenotypes.

In conclusion, we proposed Misina, an integrative approach with expert scoring of GWAS risk SNPs to identify miRNA-mediated genetic factors. Misina is a resource primarily for epidemiologists who will benefit from the easy-to-use interface to analyze the non-coding effects of GWAS results. Since Misina consolidated many up-to-date SNP and most important miRNA-related data sources, it allows for interrogations involving any SNP datasets. Consideration of multiple data sources supporting miRNA-mediated dysregulation such as the effect of SNPs on gene expression, expression of target gene and miRNA in relevant tissues, the potential strength of miRNA binding via seed type, and inclusion of LD proxies might reveal new mechanisms underlying the phenotype of interest. Moreover, as a proof-of-concept, we deeply characterized an Alzheimer's disease risk SNP, rs6859, which might have critical effects on the regulatory mechanisms related to lipid metabolism and cognitive traits via the dysregulation of NECTIN2.

3.3 Variant prioritization with deep learning

Deep learning-based variant prioritization, which was pioneered by J. Zhou and Troyanskaya (2015) in DeepSEA, is getting increasingly popular in the field. DeepSEA predicts cell type-specific molecular modalities such as histone marks, TF binding, and DNA accessibility from the DNA sequence alone (J. Zhou and Troyanskaya 2015). Trained on the genome-wide sequences obtained from the publicly available ChIP-seq and DNase-seq datasets provided by the ENCODE (ENCODE Project Consortium 2012) and the Roadmap Epigenomics (Kundaje et al. 2015), the model learns complex sequence patterns driving the sequence specificities of these modalities. Importantly, this black box prediction method is then utilized to estimate how variations in the DNA sequence might alter each modality, allowing DeepSEA to prioritize the variants with significant functional impact. Notably, the method exploits cell-type specific regulatory effects of variants under different treatment conditions, adding additional layers to our understanding of context-specific disease mechanisms.

Deep convolutional sequence models like DeepSEA bear a high potential to outperform positional overlap-based variant prioritization approaches due to consideration of not only simple sequence motifs but also higher-order patterns like motifs of motifs (i.e. motif syntax) during the prediction of regulatory effects of variants. Thus, these approaches can distinguish

variants with putative functional effects from those who just reside by chance within an annotated element for a given cell type.

So far, deep learning-based variant effect predictions have only been used as a follow-up step in GWAS. Therefore, there is a disconnect between genotype-phenotype associations and sequence-based effective variant effect prediction methods that can complement each other when used jointly.

3.3.1 DeepWAS: Multivariate genotype-phenotype associations by integrating regulatory information using deep learning

Here we introduce a new strategy harmonizing classical GWAS and the follow-up functional analysis step. In GWAS, typically, single SNPs are individually analyzed and then filtered and prioritized in a post hoc functional analysis step where regulatory information is incorporated. Instead of this follow-up analysis, here we predict the regulatory effects of SNPs in various cell lines with different treatments. Sets of SNPs with similar effects are then jointly tested for association with a disease or trait of interest by using regularized regression models. The advantages of this strategy, called DeepWAS, are two-fold. First, it limits the multiple testing burden of typical GWAS by performing fewer tests. Second, it provides regulatory information at the SNP level without needing a second stage of analysis.

By applying DeepWAS to previously published datasets, we constructed a comprehensive landscape of potential regulatory drivers of multiple sclerosis (MS) (T. F. Andlauer et al. 2016), major depressive disorder (MDD) (Muglia et al. 2010; Rietschel et al. 2010) and height (Wichmann et al. 2005) as a proof of concept application. Moreover, the comparison of the DeepWAS results from cohorts with small sample sizes(n=15k, 3k and 6k for MS, MDD and height) to the results from GWAS meta-analyses (n=116k, 807k and 184k) (Patsopoulos 2018; Howard et al. 2019; H. L. Allen et al. 2010) allowed us to verify that our method can identify existing and novel disease- or trait-associated variants as well as the relevant molecular modalities, cell lines and treatments which facilitates generating novel functional hypotheses on the determinants of these phenotypes.

DeepWAS algorithm

DeepWAS algorithm consists of two steps. The first step of the pipeline obtains functional importance scores from the pre-trained DeepSEA neural network for each given variant (Figure 3.7A). The primary outcome of this step is a grouping of variants where each group comprises variants that are predicted to affect the same regulatory element in the same cell line, called functional units (e.g. all variants modulating the NFKB binding event in GM12878 cell lines) (Figure 3.7B). The second step associates these putative regulatory variants with the phenotype of interest using a multivariate Lasso model for each variant group inferred in the previous

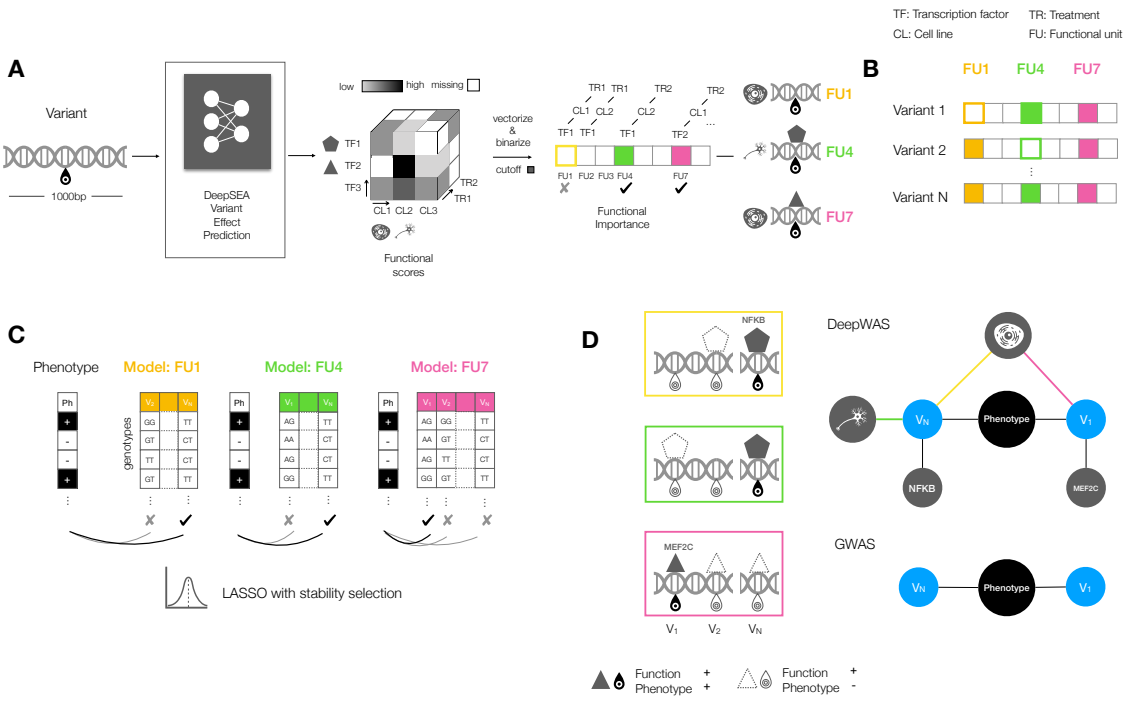


Figure 3.7: Workflow of DeepWAS. **(A)** For a given variant, the DeepSEA model (J. Zhou and Troyanskaya 2015) predicts whether the probabilities of chromatin features (i.e. TF binding, DNase-I hypersensitivity and histone marks) are affected by the variant for a given 1,000 base-pair DNA sequence around the variant. The potential functional effects of a variant are determined by binarizing the functional scores using a cutoff. Functional units (FU) represent chromatin feature, cell line and treatment combinations e.g. TF1 / Cell line 1/ Treatment 2. **(B)** Repeating this process for all genotyped variants leads to a large matrix of variant effect predictions. **(C)** For each FU, the genotype-phenotype association is tested using Lasso regression with stability selection. **(D)** In contrast to GWAS, DeepWAS suggests a regulatory process as well as the cell lines and TFs that are potentially relevant for the phenotype of interest.

step. The independent variables of these models are the elements of the variant group (Figure 3.7C). Since the variant groups represent the regulatory context where the variant is "active", the final genotype-phenotype associations can be interpreted in a context-specific way which distinguishes DeepWAS from the typical GWAS approach (Figure 3.7D). The following two subsections present the details of these two steps in more detail.

1. Variant effect predictions with DeepSEA

To identify the SNPs that, through modifying regulatory elements, may play a critical role in human diseases or traits, we used the pre-trained DeepSEA model (J. Zhou and Troyanskaya 2015) . For a given 1kb DNA sequence, this model predicts the probability that a molecular event happens within the given sequence for each of the 919 predefined events. These events are of three major types of events measured in the ENCODE project (ENCODE Project Consortium 2007): TF binding, DNase-I hypersensitivity (DHS) and histone modifications.

These three event types, called chromatin features (e.g. NFKB binding), are measured (and predicted) in cell lines (e.g. K562) under different treatment conditions (e.g. TNF). To have a compact representation of these events with three components, i.e. chromatin feature, cell line and treatment, we named them "functional units" (FUs, e.g. NFKB:K562:TNF) here. 919 FUs comprised binding sites of 160 different TFs (690 TF profiles in total), 125 DHS, and 104 histone mark tracks across 17 treatment conditions and 31 cell lines.

We used the variant effect prediction methodology described in DeepSEA (Figure 3.7A-B):

- For each of the SNPs in the set of measured genotypes for all three phenotypes, MS, MDD and height, the 1000 bp reference genome sequence centered at a SNP position is retrieved. The sequences for both reference and alternative alleles are generated simply by replacing the base at the center of the sequence with the corresponding allele.
- We generated the predictions for all sequences using the pre-trained DeepSEA network v0.94 (Figure 3.2A), which was downloaded from `http://deepsea.princeton.edu/help/`. Events that likely happen in sequences with reference and alternative alleles are obtained.
- The chromatin effect of each SNP was calculated by comparing the event probabilities of two alleles (Figure 3.2B, see DeepSEA J. Zhou and Troyanskaya 2015 for more details).
- To distinguish predicted high-effect SNPs from those occurring just by chance (i.e. to control the false positive rates), an empirical p-value procedure is employed. To calculate the empirical p-values (named "e-values" by the authors of DeepSEA) for each FU, the chromatin effects of one million random variants from the 1,000 Genomes Project (1000 Genomes Project Consortium 2015) are used as a null distribution and the proportion of random variants that have a greater impact than that of the observed variants are calculated (Figure 3.2B).
- The SNPs with an e-value smaller than 5×10^{-5} were considered potentially regulatory and used in the regression models for testing genotype-phenotype associations.

2. Regularized regression models for testing associations

Regularized regression models like Lasso offer means to examine the associations in high dimensional data. In the context of genetics, this is useful to test the associations between a group of SNPs and the response variable, which is either a trait or a disease. In DeepWAS, for each FU, we fit Lasso models with stability selection to test genotype-phenotype associations where only the SNPs with a significant effect on the FU of interest are included as covariates (Figure 3.7C). This grouped testing approach represents the hypothesis that the accumulated downstream effects of many variants acting on a specific FU might alter a specific cellular

function that is critical for the phenotype of interest. This improves the power to detect sets of regulatory variants with a potential role in disease etiology.

Unlike GWAS, this approach implicates the relevant chromatin feature, cell type and treatment for the phenotype in addition to the genotype-phenotype associations (Figure 3.7D). For a continuous response (e.g. height), the Lasso model and how the regularization is applied are given below:

$$y_i = \sum_{j \in S_k} \beta_{jk} \mathbf{X}_{ij} + \beta_{sex,k}\text{sex}_i + \beta_{age,k}\text{age}_i + \beta_{coh,k}\text{coh}_i + \left(\sum_l \beta^l_{anc,k}\text{anc}^l_i \right) + \beta_{0k} + \epsilon$$

$$\underset{\beta}{\text{argmin}} \sum_{i=1}^{M} \left[y_i - \beta_{0k} - \sum_{j \in S_k} \beta_{jk} \mathbf{X}_{ij} + \beta_{sex,k}\text{sex}_i + \beta_{age,k}\text{age}_i + \beta_{coh,k}\text{coh}_i + \left(\sum_l \beta^l_{anc,k}\text{anc}^l_i \right) \right]^2 + \lambda \left\| \beta \right\|_1$$

$$\underset{\beta}{\text{argmin}} \sum_{i=1}^{M} (y_i - \hat{y}_i)^2 + \lambda \left\| \beta \right\|_1$$

Phenotypes with a binary response variables (i.e. MS and MDD) are modeled with logistic models:

$$\text{logit } P(y_i = 1) = \beta_{0k} + \sum_{j \in S_k} \beta_{jk} \mathbf{X}_{ij} + \beta_{sex,k}\text{sex}_i + \beta_{age,k}\text{age}_i + \beta_{coh,k}\text{coh}_i + \left(\sum_l \beta^l_{anc,k}\text{anc}^l_i \right)$$

$$\hat{y}_{ik} = \text{sigmoid} \left(\sum_{j \in S_k} \beta_{jk} \mathbf{X}_{ij} + \beta_{sex,k}\text{sex}_i + \beta_{age,k}\text{age}_i + \beta_{coh,k}\text{coh}_i + \left(\sum_l \beta^l_{anc,k}\text{anc}^l_i \right) \right)$$

$$\underset{\beta}{\text{argmax}} \sum_{i=1}^{M} y_i \log(\hat{y}_{ik}) + (1 - y_i) \log(1 - \hat{y}_{ik}) + \lambda \left\| \beta \right\|_1$$

Variables in the equations represent:

- i,j,k: Subscripts representing individual, SNP and functional units, respectively.
- M: Number of individuals.
- sex, age, coh, anc: Sex, age, cohort membership and the multidimensional scaling (MDS) components of the genotypes representing the ancestry of the individual. l is used as an index for the MDS components.
- $\mathbf{X}$: Genotype matrix where the genotypes are encoded using the probabilities of three possible genotypes in an additive manner e.g. $\mathbf{X}_{ij} = 2P(\text{AA}_{ij}) + P(\text{Aa}_{ij})$. $P(\text{AA}_{ij})$ and $P(\text{Aa}_{ij})$ represent the homozygous and heterozygous genotype probabilities. Therefore $\mathbf{X}_{ij}$ takes on a continuous value in the range $[0, 2]$.

- S_k: Set of SNPs with significant predicted regulatory effect on FU k. The regression coefficients (β) are also indexed with k and thus are FU-specific.
- y: Response variable representing the disease or trait. It is binary in the case of MS and MDD and continuous for height. $\hat{y}$ represents the predictions in both models.
- λ: Regularization strength in Lasso which determines the amount of shrinkage applied to the model parameters (β).

The equations above consist of two components, the likelihood term ($\sum_{i=1}^{M}(y_i - \hat{y}_i)^2$ in the first equation and $\sum_{i=1}^{M} y_i \log(\hat{y}_{ik}) + (1 - y_i)\log(1 - \hat{y}_{ik})$ in the second equation) and the regularization term ($\lambda \|\beta\|_1$), which can also be interpreted as Laplace priors over the model parameters β. As described in Section 2.1.2, L1-regularization term improves generalizability and interpretability of the model by shrinking the model parameters towards zero and hence leading to sparse solutions. Coefficients of the SNPs with non-zero values are considered informative for the prediction and used for associating SNPs with the phenotype of interest.

Stability selection and accounting for the covariates

The resulting associations of Lasso models might exhibit high variation across different runs, particularly when applied to datasets of small sizes. We used the stability selection approach proposed by Meinshausen and Bühlmann (Meinshausen and Bühlmann 2010) to improve the stability of the association results. This approach simply tests associations multiple times for a given dataset via resampling of data points which yields the uncertainty estimation of variable selections. These uncertainty estimates are then used for controlling false positive rates (e.g. per-family error rate–PFER). See Section 2.1.2 for more details on the method.

To avoid cohort, sex, age and ancestry-specific effects in genotype-phenotype associations, we used cohort membership, sex, age, and selected MDS ancestry components as additional covariates in Lasso models. The stability selection procedure implemented in the "stabsel" function of the R package "stabs" was used to test genotype-phenotype associations. The type of Lasso implementation, selection probability cutoff and per-family error rate parameters, namely "fitfun", "cutoff" and "PFER" were set to "glmnet.lasso", 0.7 and 1.0, respectively. The subsampling strategy proposed by Shah and Samworth (2012) is used with $n = 100$ subsample replicates and a subsample size of $\lfloor n/2 \rfloor$. The resulting SNPs which represent stable FU-specific associations are named "dSNPs".

Functional annotation of dSNPs

The cell line and tissue data for the ENCODE tracks were downloaded from `https://genome.ucsc.edu/encode/cellTypes.html`. The genome segmentation files of the 15-state

ChromHMM model representing the annotations of cis-regulatory elements were downloaded from the web portal of the Roadmap Epigenomics Project (Kundaje et al. 2015)(`http://egg2.wustl.edu/roadmap/data/byFileType/chromhmmSegmentations/ChmmModels/coreMarks/jointModel/final/all.mnemonics.bedFiles.tgz`). Roadmap segmentations are first collapsed for each tissue into broader groups and then overlapped with the dSNPs using the genomic positions.

To further characterize the regions where dSNPs are located, we overlapped the SNP positions with various genomic annotations including promoters, 3' and 5' UTRs, intergenic regions, downstream and intronic/exonic regions using the ChIPseeker (`https://www.bioconductor.org/packages/release/bioc/html/ChIPseeker.html`) (Yu et al. 2015) Bioconductor R package and the UCSC hg19 "known gene" transcript set. Moreover, we used SNPsea to identify the tissue and cell types that are likely affected by dSNPs. SNPsea version 1.0.3 (Slowikowski et al. 2014) was downloaded from `https://github.com/slowkow/snpsea/` and the enrichment p-values are calculated using the default annotation files via the SNPsea command line interface.

Graph visualizations

The associations between dSNPs and the FU components (i.e. cell types, chromatin features and treatments) as well as those between QTL-gene, QTL-tissue and QTL-cohort are visualized with graphs. The dSNPs, FU information and links to external QTL datasets were added into a local Neo4j graph database (version 3.4.0, `https://neo4j.com`). We used dummy nodes (small gray nodes in graph visualizations) to avoid ambiguity in the dSNP-FU and dSNP-QTL links. In this scheme, dSNPs are connected to the dummy nodes, which are connected to the FU components instead of a direct dSNP-FU connection. An ambiguity would arise when a dSNP with a regulatory effect on two FUs, e.g. JUND:K562 and NFKB:A549, were connected to all four elements where one cannot distinguish whether the FUs are JUND:K562 and NFKB:A549 or JUND:A549 and NFKB:K562. Similarly, dNPs are connected to QTL components (genes, tissues and cohorts) via dummy nodes.

3.3.2 Application of DeepWAS

We used DeepSEA variant effect predictions to filter from the measured SNPs only those with significant cell-type-specific regulatory effects (e-value $< 5 \times 10^{-5}$) (J. Zhou and Troyanskaya 2015). This process yielded around 40,000 SNPs. For each of the 919 FUs, we obtained a list of likely functional and functionally similar SNPs. Next, using multivariate L1-regularized regression models (Lasso) with stability selection (Tibshirani 1996; Meinshausen and Bühlmann 2010; Hofner et al. 2015), we tested the associations of these putatively regulatory variants with the phenotype of interest for each FU individually using specifically the sets of SNPs

functional in a FU. The SNPs with significant associations in 919 regression models, named "dSNPs", and the FU information linked to these models comprise the primary outcome of DeepWAS (Figure 3.7).

We used DeepWAS to characterize the regulatory determinants of three complex phenotypes from previous studies. These consist of two case-control studies, namely MS and MDD where patients are compared to healthy controls, and a cohort study of human body height. In our MS application, we analyzed the KKNMS GWAS dataset consisting of two independent MS case-control cohorts with 15,283 participants in total (T. F. Andlauer et al. 2016). In total, out of 36,409 predicted regulatory variants in 25,000 independent loci, DeepWAS identified 53 MS-associated dSNPs[1] in 16 independent loci that are potentially altering 120 chromatin features in 133 cell lines in 38 independent loci ($r^2 \geq 0.5$). Moreover, this analysis revealed 637 out of 919 FU models with at least one variant association (Figure 3.8). While there was a single SNP association with MS in most regression models, 148 models resulted in more than two significant associations jointly affecting a FU.

We further applied DeepWAS to relatively small GWAS datasets for MDD (n=3,514) (Muglia et al. 2010; Rietschel et al. 2010) and height (n=5,866) (Wichmann et al. 2005) which yielded 61 dSNPs in 237 FUs for MDD and 43 dSNPs in 381 FUs for height.

Comparison of dSNPs with GWAS associations is an important point that is relevant for evaluating the results of our approach. In the following sections, we investigated whether the dSNPs overlap with 1) the results of typical GWAS analysis of the same datasets that we characterized with DeepWAS 2) the results of the larger GWA studies or meta-analyses of the identical phenotypes.

Clinical Samples

We analyzed the genotypes and phenotypes of the MS patients (including the patients at prodromal phase, called CIS–clinically isolated syndrome) in the DE1 and DE2 cohorts of KKNMS (Kompetenznetz Multiple Sklerose) network and the genotypes for the control group are obtained from T. F. Andlauer et al. (2016) (n=15,283 total individuals). For further information about the cohort and genotype data processing, refer to T. F. Andlauer et al. (2016). For MDD, we analyzed the genotypes and phenotypes from two cohorts, recMDD and BoMa (n=3,514), collectively named MDDC. recMDD individuals are recruited at the Max-Planck Institute of Psychiatry (MPIP) in Munich, Germany, and two hospitals BKH Augsburg and Klinikum Ingolstadt, which is previously described in Muglia et al. (2010). BoMa comprised MDD patients described in Rietschel et al. (2010). For further information about the cohorts and genotype data processing, please refer to the corresponding publication (Muglia et al. 2010;

[1] These 53 dSNPs were outside of the major histocompatibility complex (MHC) region. 111 dSNPs residing in the MHC are excluded from the results.

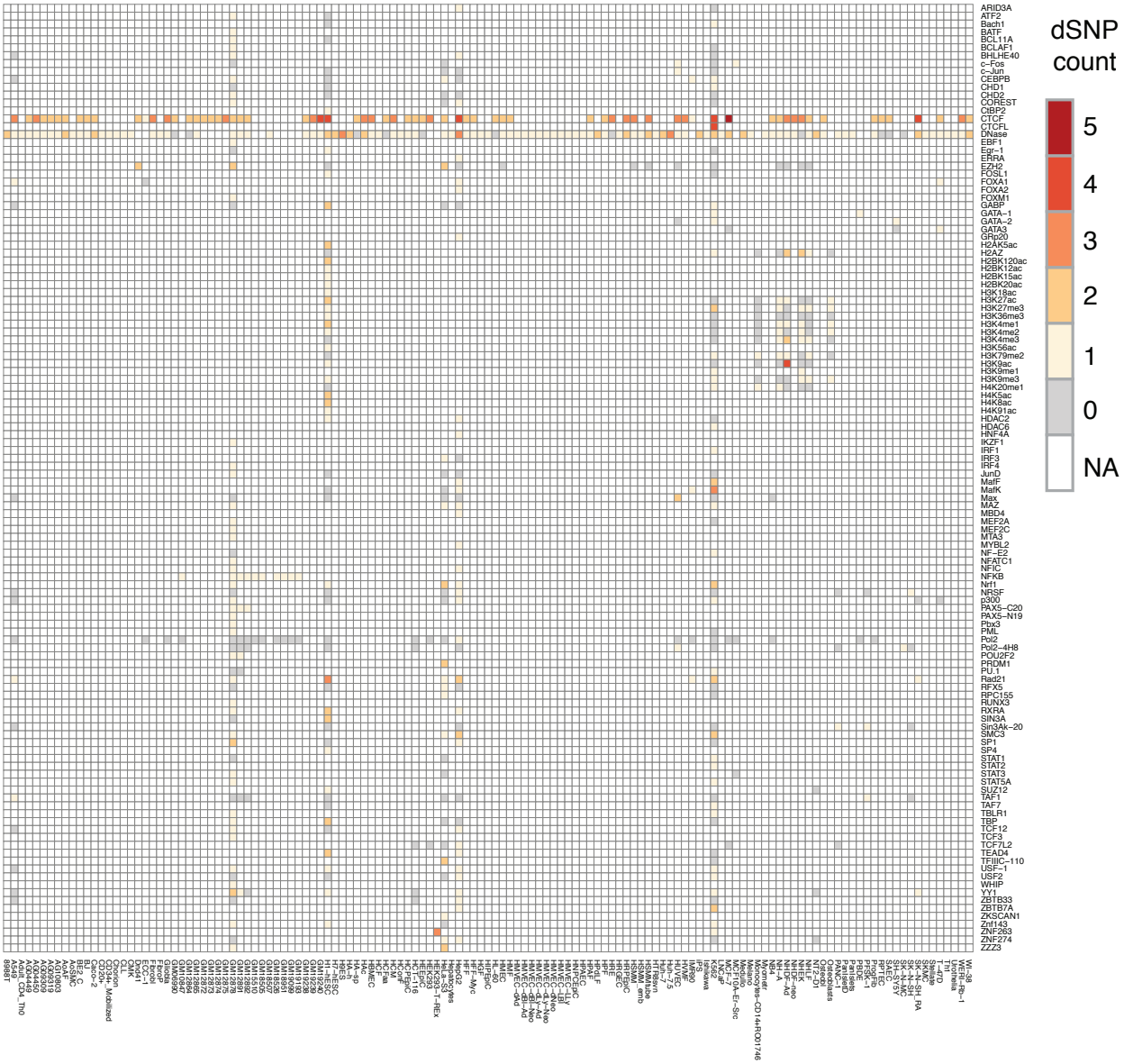


Figure 3.8: Heatmap showing the regulatory components (rows: chromatin features, columns: cell lines) and the counts of the MS dSNPs. In total, 637 out of 919 FUs were associated with at least one variant. 53 MS-specific dSNPs potentially affecting 120 chromatin features in 133 cell lines were identified by DeepWAS. Missing values are shown in white, indicate the FUs for which the experimental data were not available. The cell lines and chromatin features that are well-studied in ENCODE (e.g. GM12878, HepG2, K562 and CTCF, DNase, Pol2) exhibit vertical and horizontal stripes. Adapted from Arloth et al. (2020).

Phenotype	DeepWAS datasets			Func. units	dSNPs	Overlap with cohort-matched GWAS		Largest GWAS		Overlap with largest GWAS
	Dataset	Cohorts	Total ind.			Nominal	Genome-wide	Cohort	Total ind.	Genome-wide
MS	KKNMS	DE1, DE2	15,283	637	53	42	11	IMSGC	115,803	15
MDD	MDDC	BoMa, recMDD	3,514	237	61	60	0	PGC, UKBB	807,553	0
Height	KORA	S3, S4	5,866	381	43	42	1	GIANT	183,727	8

Table 3.1: Results of MS, MDD and height applications of DeepWAS and overlap with cohort-matched and larger GWAS results. MS: Multiple sclerosis, MDD: Major depressive disorder, UKBB: UK Biobank, PGC: Psychiatric Genomics Consortium, IMSGC: International Multiple Sclerosis Genetics Consortium.

Rietschel et al. 2010). GWAS analysis for MDD was conducted separately on recMDD and BoMa cohorts by Janine Arloth. See the DeepWAS publication (Arloth et al. 2020) for more details on methods. For the analysis of height, we used the genotypes and phenotypes from the participants of S3 and S4 cohorts of the KORA (Kooperative Gesundheitsforschung in der Region Augsburg) study (Wichmann et al. 2005) (n=5,866). See Wichmann et al. (2005) for more details on the cohort and genotype data processing. For the comparison with larger GWAS results, we used the variant lists from IMSGC (n=115,803, International Multiple Sclerosis Genetics Consortium 2019), PGC (n=807,553, Psychiatric Genomics Consortium, Howard et al. 2019) and GIANT (n=183,727, H. L. Allen et al. 2010) cohorts for MS, MDD and height respectively.

DeepWAS results are in accordance with the cohort-matched GWAS

We compared the dSNPs with the genome-wide and nominally significant associations identified in the same MS, MDD and height datasets via a typical GWAS approach to assess the agreement between the two approaches (Table 3.1). 11 out of 53 MS dSNPs (or their LD proxies, $r^2 \geq 0.5$) were genome-wide significant in published KKNMS GWAS (T. F. Andlauer et al. 2016). These dSNPs were located in six independent loci near genes EVI5 (lead GWAS SNP: rs6689470), CD58 (rs2300747), CLEC16A (rs6498168), MAZ (rs34286592), SHMT1 (rs4925166), and an intergenic region (rs1891621). The remaining 42 dSNPs were nominally significant in the MS GWAS (p-values $\leq 5.13 \times 10^{-4}$).

60 out of 61 MDD dSNPs (or their proxies, $r^2 \geq 0.5$) were nominally significant in the GWAS of the MDDC cohort, while all 43 height dSNPs reached significance at the nominal

level in the GWAS of the KORA cohort (Wichmann et al. 2005) (p-values $\leq 7.7 \times 10^{-3}$) in addition to a single genome-wide significant SNP. For more detailed results, see Arloth and Eraslan et al. (2020).

These results indicate that DeepWAS was able to identify both novel and existing risk loci with the expected Type I error rate (e.g. PFER). Interestingly, many dSNPs were sub-threshold variants in GWAS which might be due to two reasons. The first one is the lack of sufficient statistical power, which is discussed in the next section where we compared our results with larger GWAS. Second, the multivariate nature of DeepWAS which estimates the phenotypic effects of variants conditioned on other regulatory variants, unlike GWAS, might lead to this difference.

DeepWAS results are in line with larger GWAS

We compared our MS dSNPs with the GWAS results of the International Multiple Sclerosis Genetics Consortium (2019) (IMSGC, n=47,429 cases and n=68,374 controls) where 200 genome-wide risk loci were identified outside of the MHC locus. 39 out of 200 MS risk loci harbored at least one dSNP. 15 out of 53 MS dSNPs were genome-wide significant in ten independent loci, including five loci where the nearest genes were EVI5, CD58, CLEC16A, EPS15L1, LINC00271 as well as five intergenic loci on chromosomes 5, 6, 10, 11 and 22. Eight dSNPs near EVI5, CD58 and CLEC16A genes also reached genome-wide significance in the cohort-matched GWAS.

None of the MDD dSNPs were genome-wide significant in the largest MDD GWAS (Howard et al. 2019) (n=246,363 cases and n=561,190 controls). The highest p-value (i.e. the least significant) of MDD dSNPs in this GWAS was 2.8×10^{-4}. Eight out of 43 height dSNPs reached genome-wide significance in seven independent loci (nearest genes: DIS3L2, HABP4, LCORL, PDLIM4, PXMP4, ZBTB38 and ZNF311) in the latest height GWAS (GIANT Consortium, H. L. Allen et al. 2010, n=183,727).

Except for the MDD, we see a better agreement with the GWAS results as cohort sizes and statistical power increase. This suggests that DeepWAS can detect sub-threshold GWAS variants which can potentially govern critical regulatory components of disease or trait mechanisms.

DeepWAS generates novel hypotheses of disease mechanisms

We next demonstrated how DeepWAS could be leveraged to derive testable hypotheses on regulatory mechanisms that potentially contribute to MS pathology with two cases.

First, we highlighted a group of MS dSNPs (rs62420820, rs12768537 and rs137969) with regulatory effects targeting four functionally-related TFs, namely MafF, MafK, Bach1 and NF-E2 (Figure 3.9A). The biological relevance of this finding in the context of MS pathology is

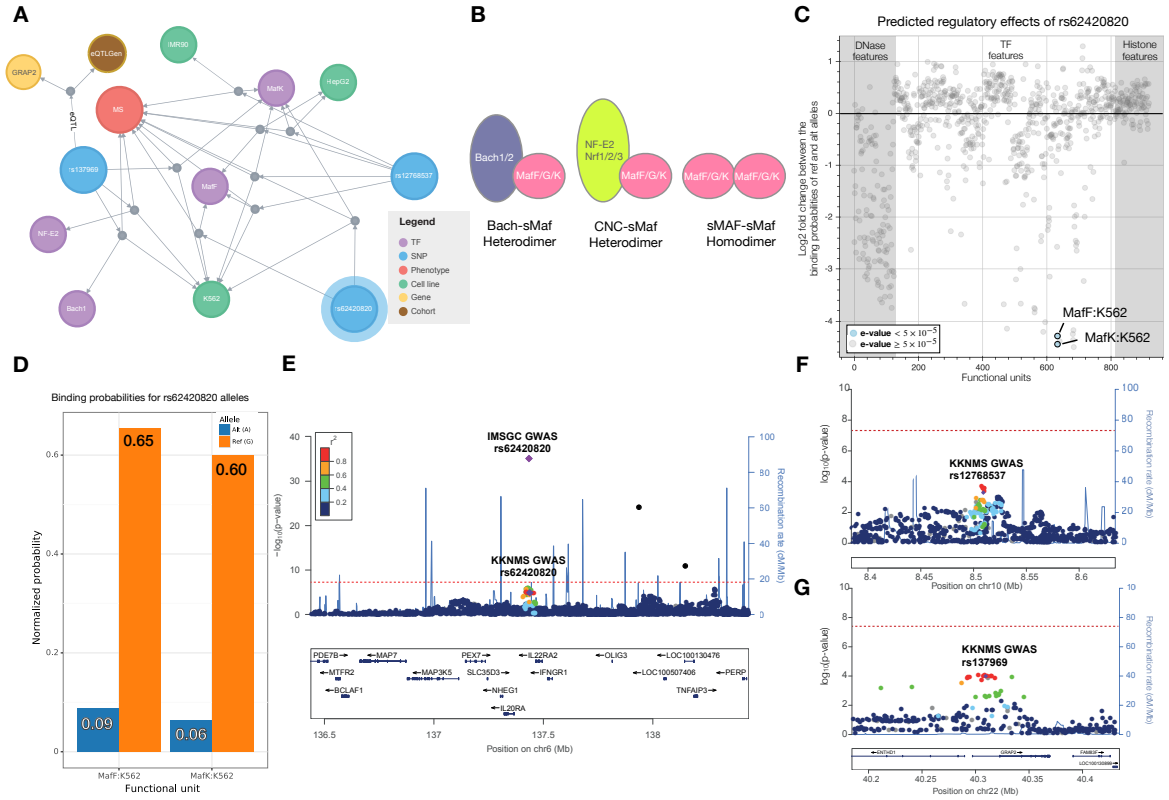


Figure 3.9: **(A)** Network visualization demonstrating the potential regulatory effects of three related MS dSNPs. Edges represent predicted associations of dSNPs, chromatin features and cell lines as well as the statistical associations from the eQTLGen study. **(B)** Homodimer and heterodimer protein complexes formed by small MAF proteins (sMafs) which are predicted to be affected by MS dSNPs and potentially play a role in disease etiology. **(C)** Log2 fold changes between the DeepSEA probabilities of reference and alternative alleles of rs62420820 variant for 919 functional units. Significant effects are highlighted. **(D)** Normalized probabilities of MafF-K562 and MafK-K562 binding events for both alleles. Alternative allele causes significantly lower binding probability compared to the reference allele. **(E-G)** LocusZoom plots of the MS dSNP rs62420820, rs12768537 and rs137969 from small (KKNMS) and large (IMSGC) GWAS results. rs62420820 dSNP, highlighted in blue, is also a genome-wide significant variant in the MS GWAS (IMSGC (International Multiple Sclerosis Genetics Consortium 2019)), but sub-threshold in the cohort-matched GWAS (KKNMS). Dots represent the GWAS p-values. The color of the dots represents LD (r^2) with the lead variant. Gray dots indicate the variants without the LD information. Adapted from Arloth et al. (2020).

twofold. First, it was reported that GRAP2, an eQTL gene for rs137969 (Võsa et al. 2018), is differentially expressed in CD4 T cells purified from MS patients compared to healthy controls and is an MS susceptibility gene (Berge et al. 2019). Second, small Maf protein family (sMafs) are known to form heterodimers with Bach1 and NF-E2, which are also involved in the same regulatory module by DeepWAS, as well as homodimers with other sMafs (e.g. MafG-MafF) (Katsuoka and Yamamoto 2016) (Figure 3.9B). Furthermore, Katsuoka, Motohashi, et al. (2003) reported that mutations in sMafs might lead to neuromuscular dysfunction and neuronal degeneration. Although MafG is not a part of the predictive DeepSEA model (therefore DeepWAS), this finding supports the hypothesis that the Maf family protein complexes are likely involved in the MS pathology through the regulatory mechanisms identified by DeepWAS based on the significant predicted effects of rs62420820 on sMafs (Figure 3.9C-D). Importantly, all three variants were nominally significant in cohort-matched GWAS analyses (Figure 3.9E-G), while rs62420820 was a genome-wide significant variant in the IMSGC GWAS (p-value = 9.26×10^{-36}, Figure 3.9E). Consequently, DeepWAS not only brings the pieces of the puzzle of MS etiology together by placing GRAP2 and sMafs in the same context, but it also uncovers key regulatory non-coding variants that are not easily detectable without performing large GWAS analyses and/or meta-analyses.

Second, we identified rs1985372, a variant on chromosome 16 within an intron of CLEC16A, which is a gene previously associated with MS (T. F. Andlauer et al. 2016). This dSNP was genome-wide significant in both cohort-matched (T. F. Andlauer et al. 2016) and in IMSGC GWAS (International Multiple Sclerosis Genetics Consortium 2019). According to the GWAS results, the minor allele (T) decreases the MS risk (OR=0.853). rs1985372 is also a known eSNP for CLEC16A gene (GTEx Consortium 2017) which further supports the functional role of the variant. DeepWAS now adds to that a testable hypothesis that the T allele of rs1985372 potentially creates a binding site for multiple TFs including GABP, GATA1, GATA2, p300, STAT1, STAT2, STAT5A, and TBLR1 (Figure 3.10), which in turn potentially plays a role in MS pathology via the dysregulation of CLEC16A.

Characterization of regulatory dSNPs via colocalization

We sought to further characterize the dSNPs of three phenotypes by investigating the genomic regions where the dSNPs are located using the UCSC annotations (Casper et al. 2018). 63–87% of the dSNPs were located within non-coding DNA elements. The fraction of intronic dSNPs was higher in MS and height (32% and 33%, respectively) than in MDD (13%). Conversely, the ratio of distal intergenic dSNPs at least 3kb away from the downstream of a gene was higher in MDD (53%) than those in MS and height (36% and 37%, respectively). No MS or MDD dSNPs were observed in coding regions (Figure 3.11A).

We next investigated the tissue-level specificity of dSNPs using the known tissue annotations

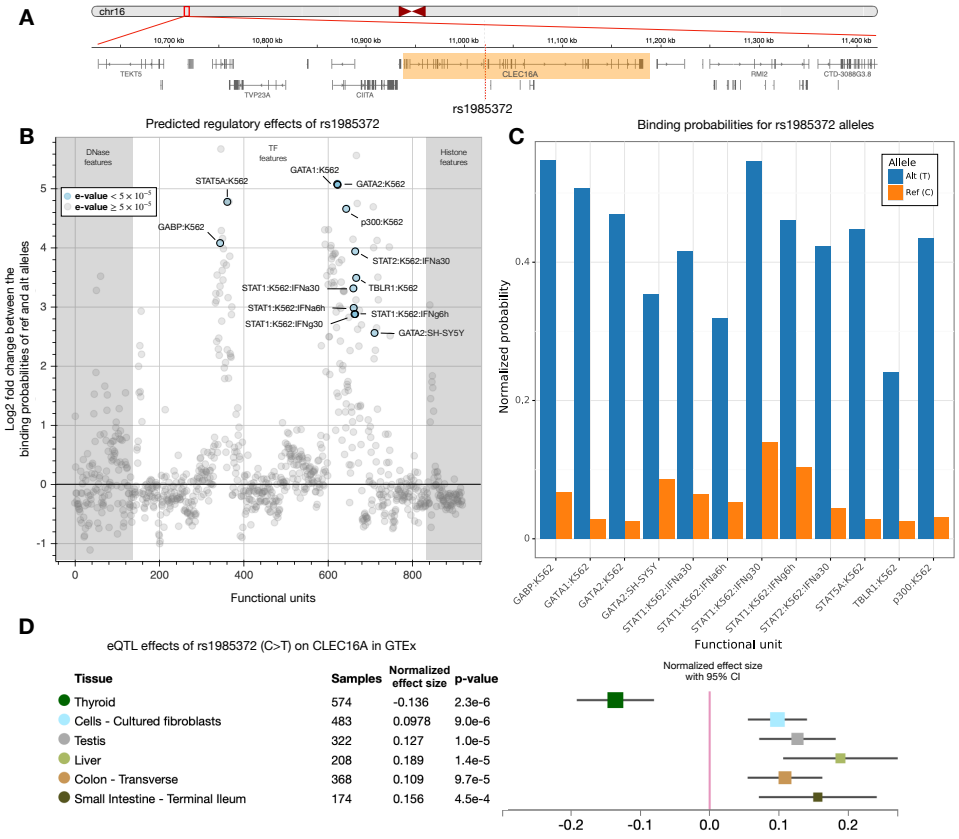


Figure 3.10: **(A)** CLEC16A locus and the position of rs1985372. **(B)** Log2 fold changes between the DeepSEA probabilities of reference and alternative alleles of rs1985372 variant for 919 functional units. Significant effects are highlighted in blue. **(C)** Normalized probabilities of significantly affected binding events for both alleles. **(D)** rs1985372 is significantly associated with the expression of CLEC16A in multiple tissues. Normalized effect sizes and p-values of the associations from the GTEx portal are shown.

of cell lines that are linked to dSNPs to have a bigger picture of the results for each phenotype (Figure 3.11B). Although most of the dSNPs fell into five major categories (i.e. blood, cervix, embryonic stem cells, liver and skin), the fact that a different number of cell lines and/or tissues contributed to these categories (e.g. 79 in the blood whereas 14 in the brain) influenced the number of dSNPs for each category. Interestingly, a higher number of height dSNPs were observed in the pancreas category compared to other phenotypes. Of note, Aune et al. (2012) reported that higher height is associated with increased pancreatic cancer risk among both men and women in a meta-analysis involving twelve cohort studies .

Colocalization of the identified variants with the key regulatory states of the genome, such as enhancers, was another critical step in characterizing dSNPs. We overlapped dSNP positions with the genomic regions whose epigenomic states are inferred by the 15-state ChromHMM model (Ernst and Kellis 2012) (Figure 3.11C). Furthermore, we grouped these states into active and repressive state groups for simplicity and reported aggregate counts for active

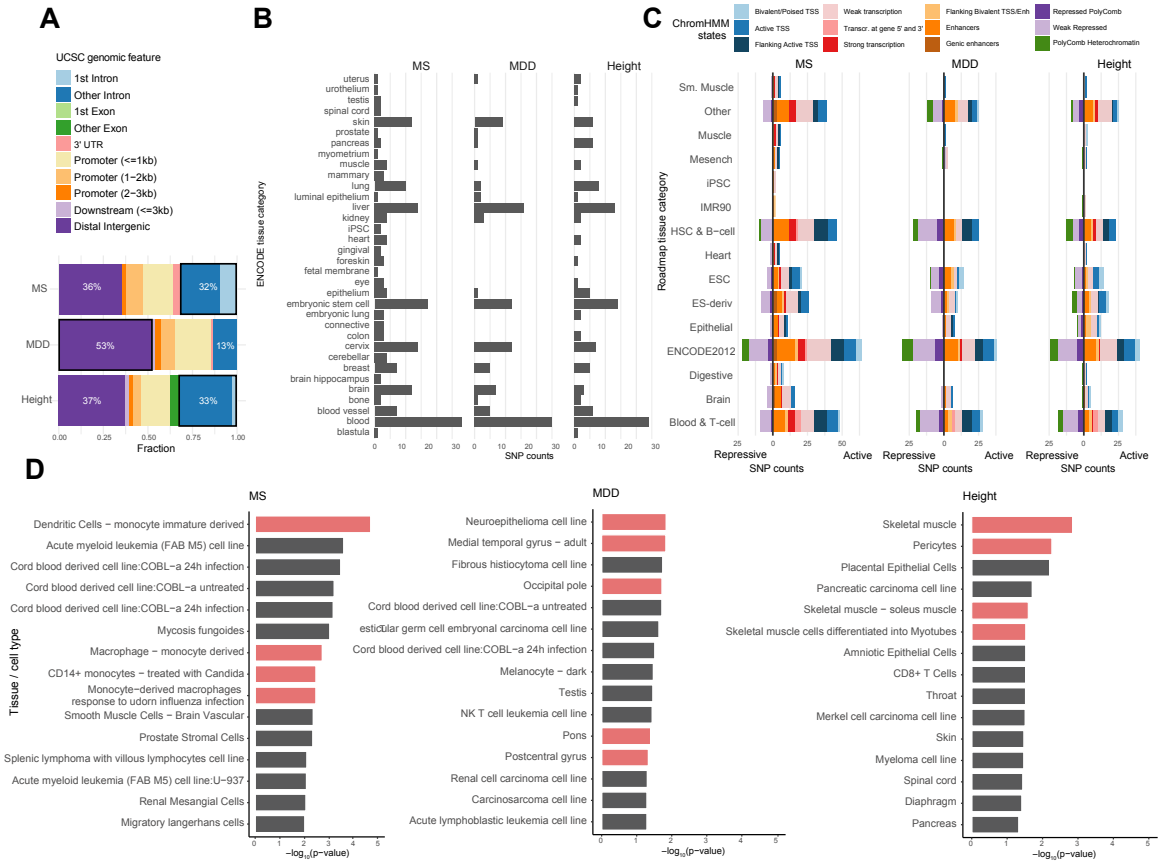


Figure 3.11: **(A)** Overlap of dSNPs with the UCSC annotations of genomic regions. **(B)** The number of dSNPs for each ENCODE tissue category corresponding to the cell line where dSNPs are identified. **(C)** The number of dSNPs overlapping with the ChromHMM states which are grouped into repressive and active for each tissue category. 12 out of 15 ChromHMM states which overlapped with at least one dSNP were shown. **(D)** Top 15 dSNP-enriched tissues and cell types based on FANTOM gene expression data analyzed with SNPsea (Slowikowski et al. 2014) (p-values ≤ 0.05). Relevant tissue/cell types are highlighted for each phenotype. Adapted from Arloth et al. (2020).

Figure 3.12: Bar plot depicting the overlap of MS, MDD and height dSNPs with GTEx eSNPs identified in brain, skeletal muscle and blood samples. Percentages, the number of overlapping dSNPs and the total number of dSNPs are shown on the bars.

and repressive states. We observed that most MS and height dSNPs colocalized with active chromatin states (82% and 86%, respectively) while the fraction of dSNPs falling into the repressive regions was higher in MDD dSNPs (43%) than the other phenotypes. This might indicate that the MS and height variants play a role in the regulation of active transcription while MDD dSNPs are involved in silencing gene activities.

Similarly to the positional overlap with genomic annotations, we investigated whether the loci of dSNPs and their proxies ($r^2 \geq 0.5$) were enriched with the genes with cell type and/or tissue-specific expression using Fantom CAGE data (Andersson et al. 2014). We observed significant associations with many phenotype-related cell types and tissues for all three phenotypes. For instance, immune cell types such as macrophages, dendritic cells and CD14+ monocytes were associated with MS, whereas the genes with skeletal muscle and brain-specific expression patterns gave rise to the associations with height and MDD, respectively (p-values ≤ 0.05, Figure 3.11D).

Effects of dSNPs on gene expression

We expect the predicted regulatory effects of dSNPs to be reflected on the gene expression which can be examined using previously known SNP-gene expression associations, namely eQTLs. We utilized cis-eQTLs detected in three tissues of the GTEx resource (GTEx Consortium 2017) (blood, skeletal muscle and brain) to investigate the overlap between dSNPs and their proxies ($r^2 \geq 0.5$). The percentage of MS and height dSNPs overlapping with GTEx cis-eQTLs was higher in whole blood and skeletal muscle, respectively compared to other phenotypes (36% and 42%, Figure 3.12). 19 MS dSNPs which are also eSNPs in the GTEx whole blood samples, as well as the affected chromatin features and eGenes, are given in Table 3.2. Interestingly, AHI1, IQCB1 and PSAP genes, which were previously associated with MS (International

Tissue	MS dSNP	Proxy SNP	R^2	GTEx Variant Id	P-value	eGene symbol	Affected chromatin feature(s)
Whole Blood	rs11164608	rs4970702	1	1_92944994_A_G_b37	2.44981E-08	EVI5	EZH2,SUZ12
Whole Blood	rs7542867	rs10874726	0.995795	1_93103099_A_T_b37	4.01786E-07	EVI5	Histone marks
Whole Blood	rs1034919	rs10924108	1	1_117062474_T_C_b37	2.60892E-05	RP5-1086K13.3	CTCF,CTCFL,DNase,Egr-1,GABP,Rad21,SIN3A,TBP,ZNF263,ZNF274,ZZZ3
Whole Blood	rs10924104	rs10924108	0.861657	1_117062474_T_C_b37	2.60892E-05	RP5-1086K13.3	Histone marks
Whole Blood	rs35737776	rs10934565	1	3_121664661_G_A_b37	8.79109E-14	**IQCB1**	PU.1
Whole Blood	rs13197384			6_135818897_C_A_b37	1.63426E-19	**AHI1**	CTCFL,RXRA,ZKSCAN1
Whole Blood	rs7797030	rs112311344	0.51365	7_5759119_G_C_b37	2.58714E-06	RP11-527E14.1	BHLHE40,HDAC2
Whole Blood	rs793102			10_31391564_C_T_b37	3.40424E-07	RP11-330O11.3	FOXA1
Whole Blood	rs11000015			10_73571883_C_T_b37	1.81549E-05	**PSAP**	MAZ,RXRA,ZBTB7A
Whole Blood	rs59410994			11_65490939_TTTTTA_T_b37	6.1428E-17	MAP3K11	Histone marks
Whole Blood	rs593525			11_65727799_T_C_b37	9.47338E-06	BANF1	CEBPB
Whole Blood	rs9603589			13_40229744_C_T_b37	6.78673E-11	COG6	TAF1
Whole Blood	rs17214656	rs7171079	1	15_80202643_C_T_b37	5.09342E-11	ST20	IRF3
Whole Blood	rs1057452	rs4788187	0.914983	16_29845685_T_C_b37	4.63003E-05	MVP	Max,Pol2-4H8,Sin3Ak-2
Whole Blood	rs2075657			17_18061528_T_G_b37	1.59206E-05	SMCR8	IRF1,SP4
Whole Blood	rs7207666	rs28880370	1	17_18182720_A_G_b37	4.18938E-08	RP1-178F10.3	Pol2,RPC155,TFIIIC-110,ZNF274
Whole Blood	rs2273030	rs4925160	1	17_18185599_A_G_b37	2.33275E-06	TOP3A	GRp20,NRSF,TAF7,YY1
Whole Blood	rs4925172	rs4925160	1	17_18185599_A_G_b37	2.33275E-06	TOP3A	CTCF
Whole Blood	rs1000329	rs12972942	0.567754	19_16577647_G_A_b37	7.49011E-06	EPS15L1	Max

Table 3.2: MS dSNPs significantly associated with gene expression (i.e. eSNPs) in GTEx cis-eQTL data from the whole blood samples. AHI1, IQCB1 and PSAP genes, shown in bold, were previously associated with MS (International Multiple Sclerosis Genetics Consortium 2019; Berge et al. 2019). R^2 represents the measure of LD between dSNP and the proxy.

Multiple Sclerosis Genetics Consortium 2019; Berge et al. 2019) were observed in the list of eGenes. In the brain samples, MS showed the highest ratio of overlapped dSNPs compared to height and MDD dSNPs[2]. Overall, our results suggest that dSNPs are disease-associated potentially regulatory variants that play a role in the regulation of gene expression in various tissues, including those that are phenotype-related such as blood and brain for MS and skeletal muscle for height.

DeepWAS identifies potential key regulators of MS and MDD

We next sought to find dSNPs active in unusually high numbers of functional units, which might indicate a critical regulatory role. Given that 72% of total dSNPs (113 out of 157) are linked to at most three functional units, dSNPs very rarely affect multiple FUs. However, some outliers, such as MS dSNP rs175714 and MDD dSNP rs7839671, which are linked to 214 and 27 FUs respectively, are worth investigating as potential key regulators (Figure 3.13).

[2]For the overlaps with other sources of eQTLs and methylationQTLs, see the DeepWAS publication (Arloth et al. 2020).

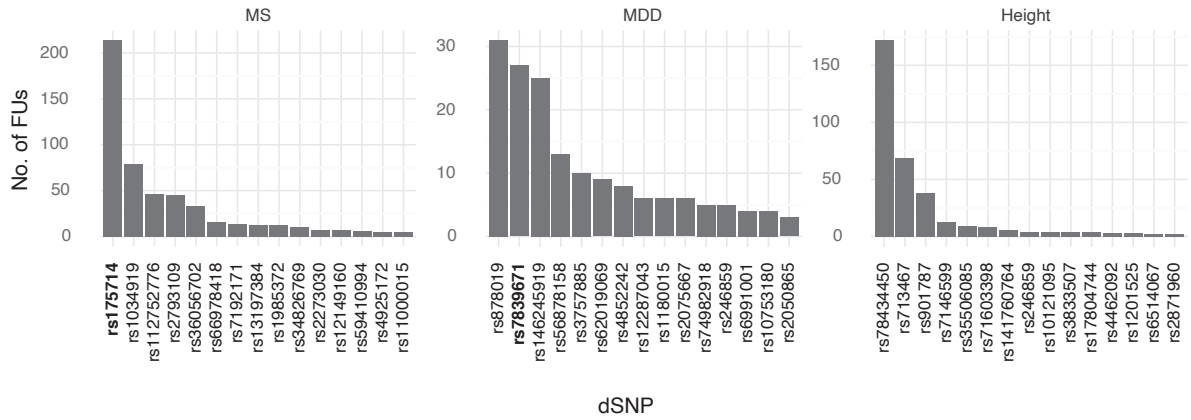


Figure 3.13: Top 15 dSNPs with the highest number of functional unit associations are shown for three phenotypes. Potential key regulators of MS and MDD (rs175714 and rs7839671), which are described in this section, are highlighted.

The first candidate, rs175714, was predicted to affect 29 chromatin features in 116 cell lines (Figure 3.14). One of the affected TFs, MAZ, was previously associated with MS by T. F. Andlauer et al. (2016) via a genome-wide significant SNP rs34286592 (p-value=4.58×10^{-8}), but the underlying mechanism linking MAZ to MS is yet to be described. We identified that two dSNPs which are associated with MS in our multivariate models, rs175714 and rs11000015, were also predicted to affect the binding of MAZ TF (Figure 3.15A-C). Moreover, rs11000015 is a significant cis-eQTL for the Prosaposin (PSAP) gene in the GTEx (Lonsdale et al. 2013) samples collected from thyroid, tibial nerve, blood and others (Figure 3.15D). Berge et al. (2019) reported that Prosaposin protein is differentially expressed in the CD4+ T cells collected from the MS patients compared to those from the healthy controls (p-value=0.004366).

The second key regulator candidate, rs7839671, is an intergenic MDD dSNP potentially affecting 24 chromatin features in 5 cell lines (Figure 3.16A). Together with another MS dSNP rs7661078, rs7839671 was predicted to significantly affect the binding of MEF2C TF (Figure 3.16B-C) which was shown to be an important risk gene for MDD (Howard et al. 2019). Furthermore, there is growing evidence suggesting a critical role of MEF2 gene family in synaptic plasticity under stress (S. X. Chen et al. 2012), memory formation and dendritic spine growth (Barbosa et al. 2008).

3.3.3 Conclusion

Associations between the genetic and phenotypic variation are examined individually for each variant in the existing GWAS practices. This view fails to account for a critical point. Putative determinants of many complex diseases are low-penetrance, non-coding variants with potential regulatory role (Tak and Farnham 2015), some of which are detected below the

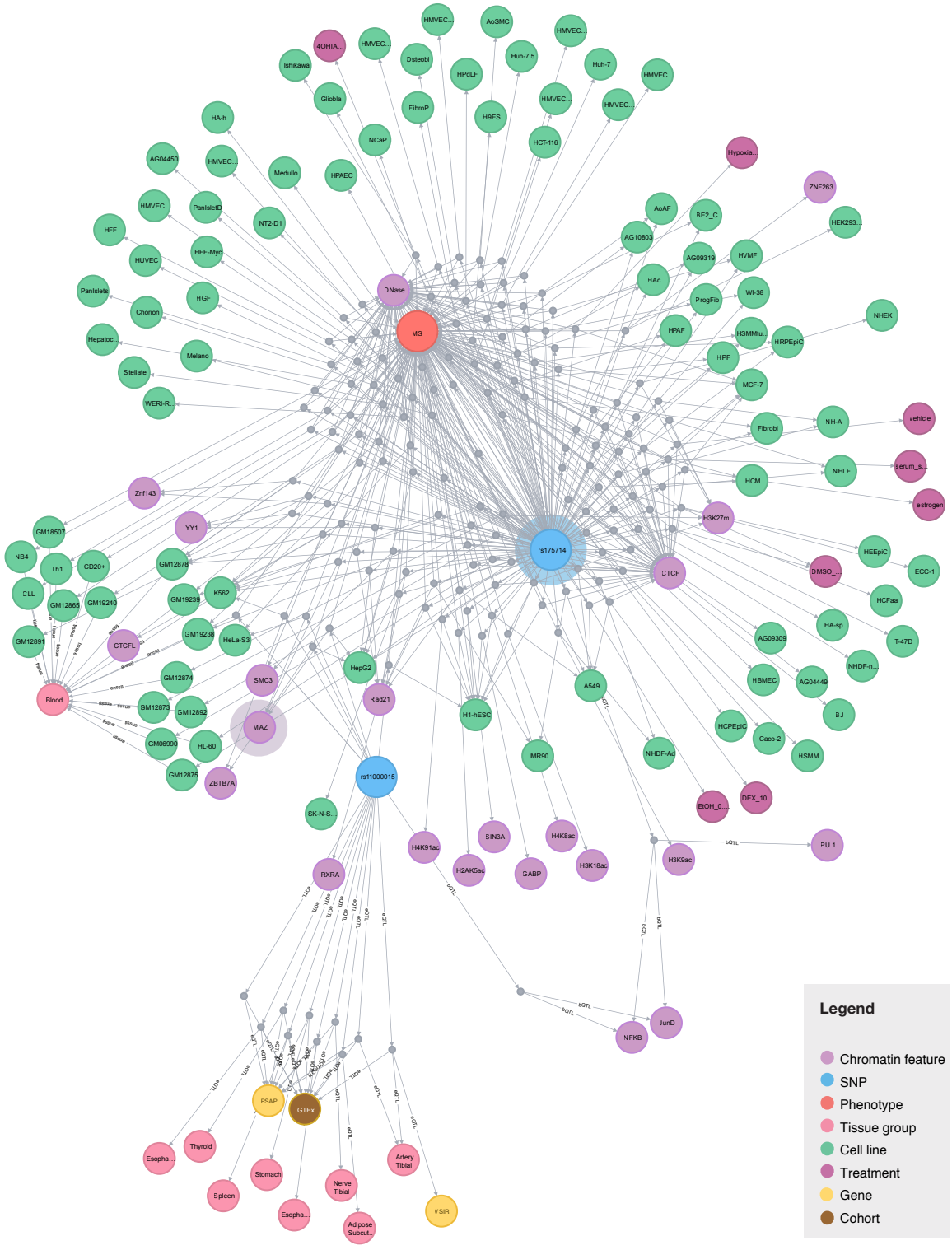


Figure 3.14: Graph visualization of a putative key regulator for MS, dSNP rs175714 (highlighted). One of the critical TFs affected by rs175714 is MAZ (highlighted), which is also among the most significant loci in the cohort-matched GWAS results. dSNP-functional unit relationships are represented as dSNP, chromatin feature, cell line and treatment nodes connected with edges through the dummy nodes (small gray nodes). Similarly, eQTLs are depicted via connected dSNP, gene, source tissue and cohort nodes. Dummy nodes bundle all components of dSNP and eQTL associations together to avoid the ambiguity that might arise when they are directly connected. Adapted from Arloth et al. (2020).

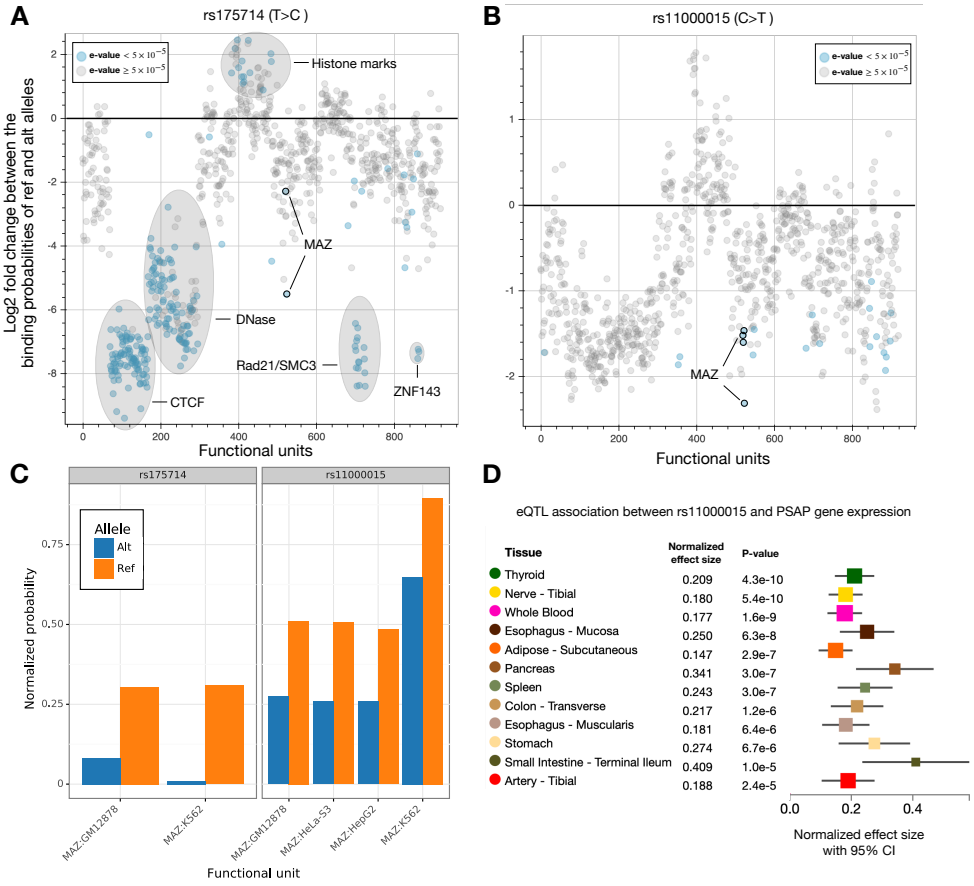


Figure 3.15: Log2 fold changes between the DeepSEA probabilities of reference and alternative alleles of rs175714 **(A)** and rs11000015 **(B)** for 919 functional units. Significant effects are highlighted in blue. **(C)** Normalized probabilities of significantly affected binding events for the alleles of both variants. **(D)** rs11000015 is significantly associated with the expression of PSAP in multiple tissues. Normalized effect sizes and p-values of the associations from the GTEx portal are shown.

genome-wide significance threshold (Xinchen Wang et al. 2016). Therefore, variants with the potential regulatory roles can be prioritized in the association tests to gain power and to develop hypotheses on the regulatory mechanisms that contribute to the causal factors of the phenotype. We implement this idea in our approach, DeepWAS, by fusing current GWAS practices with the functional characterization of the variants into a single pipeline (Figure 3.7). This pipeline starts with predicting the regulatory effects of the given variants on various chromatin features in many cell lines using the predictive power of the deep learning model DeepSEA. Next, multivariate L1-regularized (Lasso) regression models jointly test the association of groups of variants playing similar regulatory roles (i.e. active in the same FU) with the phenotype of interest. Using this testing approach, DeepWAS aims to model the polygenic architecture of complex diseases and traits. The resulting variants, so-called dSNPs, not only indicate a set of variants jointly affecting the phenotype but also suggest hypotheses

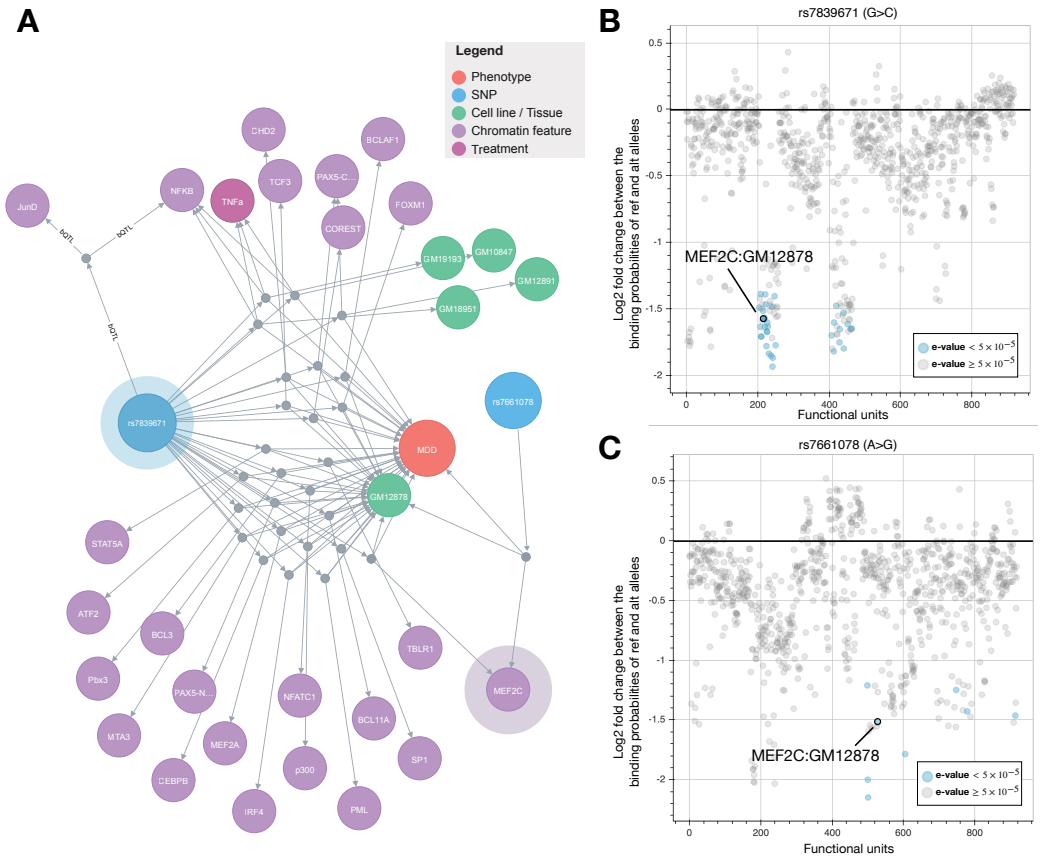


Figure 3.16: **(A)** Graph visualization of MDD dSNPs rs7839671, rs7661078 and the functional units where they are potentially active. dSNP-functional unit relationships are represented as dSNP, chromatin feature, cell line and treatment nodes connected with edges through the dummy nodes (small gray nodes). rs7839671 and TF MEF2C are highlighted. Log2 fold changes between reference and alternative allele DeepSEA probabilities of variants rs7839671 **(B)** and rs7661078 **(C)** for 919 functional units. Significant effects are highlighted in blue. Adapted from Arloth et al. (2020).

on the underlying regulatory mechanisms contributing to the phenotype. We showed that DeepWAS can generate relevant mechanistic hypotheses and potentially increase statistical power by pre-selecting putative regulatory variants, which is useful especially for underpowered cohorts.

In our DeepWAS applications, we identified 53 non-MHC candidate risk SNPs for MS (n=15,283 total individuals), 61 SNPs for MDD (n=3,514) and 43 SNPs for height (5,866). We compared these results to well-powered GWA studies of the same phenotype as well as the GWAS results of the same datasets (i.e. cohort-matched GWAS). 38 out of 53 dSNPs were new putative MS risk SNPs that were sub-threshold in well-powered MS GWAS. All 61 MDD dSNPs and 35 out of 42 height dSNPs were not genome-wide significant in well-powered GWAS of MDD and height, respectively. Although these findings remain to be experimentally validated (both the regulatory mechanisms like TF binding and the disease associations), the

results suggest that DeepWAS is able to generate novel hypotheses of disease mechanisms as well as novel risk loci.

We used the sources of prior biology for the functional characterization of dSNPs, which suggests that the identified variants potentially play a role in the disease-related pathways and are likely to contribute to the phenotypes (Figure 3.9–3.16). First, we observed that three MS dSNPs are also eQTLs affecting disease-associated genes GRAP2, CLEC16A and PSAP (Figure 3.9-3.10, 3.14-3.15) (Berge et al. 2019; T. F. Andlauer et al. 2016). Second, the vast majority of MS dSNPs were predicted to function in hematopoietic cell lines (47%, n=35) and in the brain or spinal cord samples (30%, n=16), which are relevant in the context of MS (Figure 3.11B). Similarly, the overlap with the GTEx cis-eQTL SNPs also showed meaningful patterns such as high overlap with blood and brain eQTLs for MS and skeletal muscle for height (Figure 3.12). Third, the positions of dSNPs significantly overlapped with the genes expressed in a tissue- and cell type-specific manner in the relevant cell types and tissues e.g. immune cells for MS, brain-related cell lines for MDD and the skeletal muscle for height (Figure 3.11D). Last, dSNPs were predicted to be involved in the binding events of TFs that were already linked to diseases, such as MAZ for MS (Figure 3.14) and MEF2C for MDD (Figure 3.16) (T. F. Andlauer et al. 2016; Howard et al. 2019).

In conclusion, DeepWAS couples the concept of deep learning-based variant effect prediction by estimating joint effects of regulatory variants moderating a complex phenotype. This approach is a powerful tool that reveals regulatory mechanisms underlying diseases and traits even for small cohorts and a method for identifying groups of risk variants jointly contributing to the causes of the phenotype of interest through modulation of a common FU.

Chapter 4

Recovering the expression signal in single-cell genomics using representation learning

Single-cell genomics is revolutionizing molecular biology. Novel computational and experimental developments in single-cell have enabled in-depth exploration of the transcriptome landscape via the applications in a spectrum ranging from discrete types and states (Lake et al. 2018) to continuous phenotypes of malleable populations such as differentiation trajectories (Haghverdi, Büttner, et al. 2016; Moignard et al. 2015; Herring et al. 2018) as well as applications to disease biology (Keren-Shaul et al. 2017; Stephenson et al. 2018; Gladka et al. 2018). High-quality spatial profiling techniques, multimodal measurements, large-scale perturbation experiments and drug screens are only a few promising directions that will greatly expand the repertoire of single-cell genomics and increase the popularity of single-cell even more in the next few years.

Due to the exponentially growing volume (Figure 1.4) and increasing complexity of single-cell experiments, applying supervised learning techniques, as discussed in the previous chapter, is highly impractical. Therefore, there is an expected shift towards unsupervised (and self-supervised) machine learning in single-cell. As a branch of machine learning, unsupervised representation learning provides a set of techniques to model high-dimensional, large-scale, unlabeled data using the representations of the data points in a new feature space, which makes it a perfect fit for this domain. This chapter will focus on applying representation learning techniques to single-cell RNA-seq (scRNA-seq) datasets, mainly for improving cell representations via denoising the data.

Technical sources of variation such as low RNA capture rate (Kharchenko et al. 2014), amplification bias and varying library sizes (Vallejos, Risso, et al. 2017) contribute to the gene expression readout of cells in scRNA-seq measurements. Another technical factor that

might severely affect the measurements in scRNA-seq is called the dropout (Kharchenko et al. 2014). In the droplet-based single-cell sequencing techniques, shallow sequencing of single cells exacerbates the problem of "detection failure" where no transcripts of some expressed genes are detected, which results in "false" zeros in the expression matrix (Klein et al. 2015; Zheng et al. 2017; Lopez et al. 2018; Lähnemann et al. 2020). Notably, this phenomenon creates a distinction between the "true" zeros (i.e. biological non-expression) and the "false" zeros (i.e. dropout events).

The technical sources of variation can potentially give rise to challenges in the downstream analysis and hinder proper interpretation of the biological signal. For example, analysis of gene-gene relationships (e.g. regulatory network inference, linking inferred gene modules to complex diseases and biological processes) is a fundamental part of computational biology and is gaining popularity in single-cell genomics (Aibar et al. 2017; Fiers et al. 2018; Iacono et al. 2019). Such analyses typically rely on the gene-gene correlations which are inferred from the data. However, due to the critical effects of technical factors, including the dropout noise, these correlations can be underestimated, which might impair the inference of accurate relationships between genes and regulatory network representations.

We further demonstrate the dropout noise and underestimation of gene-gene correlation using a real scRNA-seq dataset, 20,031 CD4+ T cells from Zheng et al. (2017), where the expression of CD3D (a general T cell marker), CD4 (marker of a major T cell subtype, called T helper cells) and CD8 (marker of another major T cell subtype called Cytotoxic T cells) are shown (Figure 4.1). Although all cells likely express CD3 and CD4 genes in this population, CD3 and CD4 are not detected in 25% and 88% of the population, respectively (CD3-CD4-: 4374, CD3-CD4+:509, CD3+CD4-:13397, CD3+CD4+:1751). This sparsity is also visible in the 2D visualization of cells (Figure 4.1C). Horizontal and vertical bars in the scatterplot where CD3D and CD4 are plotted indicate the dropout (Figure 4.1D). The Pearson correlation between CD3 and CD4 genes is 0.0042 and the p-value of Fisher's exact test measuring the dependence between these genes is 0.03. However, we expect to see a much higher concordance between the two genes owing to prior biology. One possible explanation of sparse CD4 expression is the misannotation of CD8 cells as CD4 cells in this population, but there are only 41 cells where CD8 expression is detected[1] (Figure 4.1E-F). In summary, the downstream analyses that are built on the relationship between these two genes estimated from the noisy data can be suboptimal or even misleading.

Denoising is a common task in imaging that aims to distinguish visually meaningful patterns (i.e. the signal) from the noise (Shao et al. 2014). This distinction allows denoising methods to enhance the image quality by removing the noise from the image and increasing

[1]There are T cell subpopulations that might lack CD4 and CD8 expression (i.e. CD3+CD4−CD8−). However, these rare cells are not detected in this population.

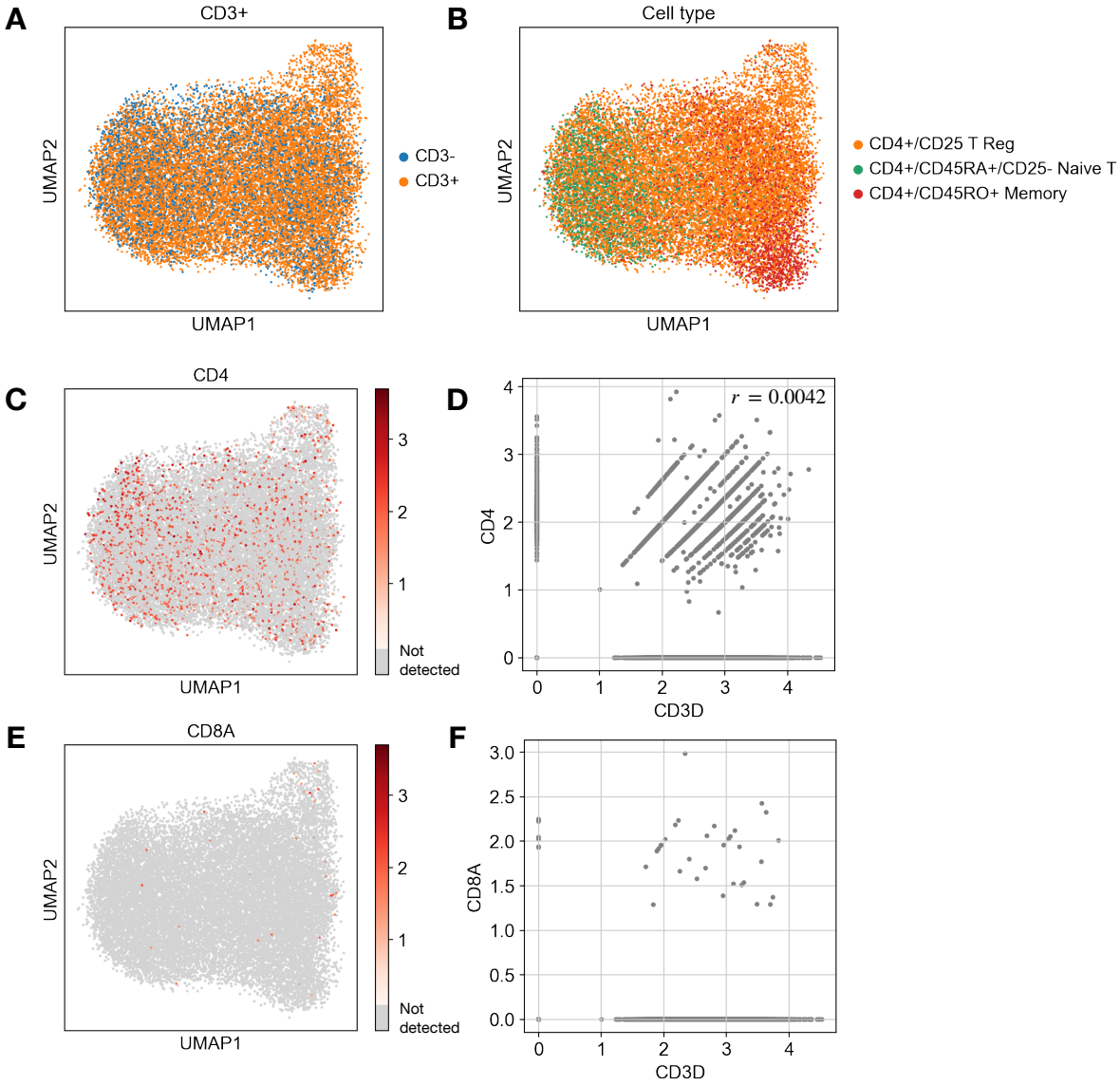


Figure 4.1: Consequence of dropout in CD4+ T cells. UMAP (a nonlinear dimension reduction method) visualization of CD3 positive and negative CD4 cells **(A)** from Zheng et al. (2017) 68k PBMC dataset and subtype/state annotations **(B)** are shown. Although all CD4+ cells here likely express the CD3 gene, the detected CD3 expression is sparse due to dropout **(C)**. Plotting CD3 against CD4 expression shows the dropout pattern as vertical (CD3 dropout) and horizontal (CD4 dropout) lines **(D)**. Cells that are not expressing CD4 are not CD8+ T cells **(E-F)**. log(TP10k+1) normalized expression is shown in (C-F). Gray dots in panels C and E represent cells with no measured expression.

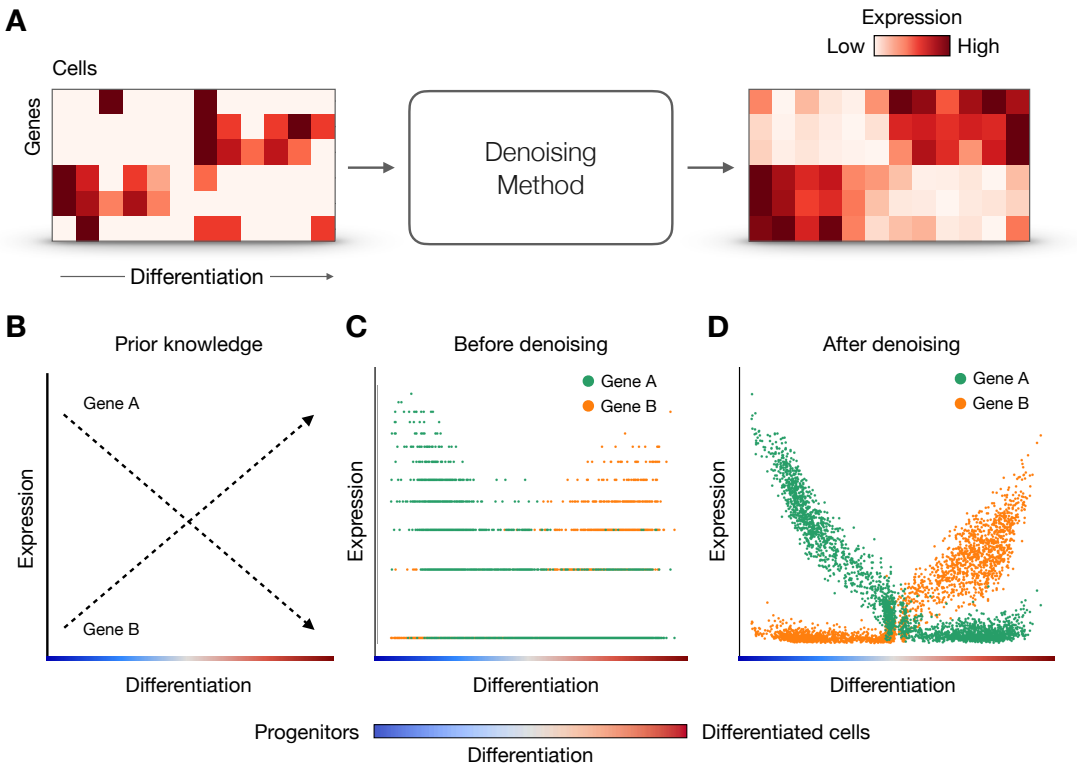


Figure 4.2: The recovery of gene expression trend of hypothetical differentiating cells. **(A)** The denoising process is exemplified with a hypothetical single-cell RNA-sequencing data (gene by cell heatmap) where the cells are sorted by their differentiation status from progenitor to differentiated cells. The denoised output, which is also represented as a heatmap, exhibits an arguably smoother and improved differentiation pattern after denoising. **(B)** Based on our prior knowledge, we expect Gene A and Gene B to be active only in progenitors (e.g. beginning of the differentiation process) and differentiated cells (e.g. later stages of the differentiation process), respectively. The relationship between the hypothetical genes A and B before **(C)** and after **(D)** denoising. Each dot represents a cell. Genes show higher anti-correlation after denoising, which is concordant with the prior knowledge.

the signal-to-noise ratio. Similarly, in biology, the denoising approaches aim to recover the biological signal that is impaired due to the corrupting factors mentioned previously and improve the representation of cells and genes compared to those obtained from the original form of the data (Figure 4.2A). Importantly, denoising applications are typically tailored to the structure of the data in order to reduce the hypothesis space and generalize well. For example, the image denoising methods specialize by making locality assumptions (Buades et al. 2005). Similarly in single-cell, denoising models, or essentially probabilistic algorithms for any downstream task, have to account for the count structure of the data, because certain properties like sparsity and overdispersion distinguish them from other datasets. This leads to the debate on the correct noise models to use with scRNA-seq datasets, which is discussed in the following sections.

The denoising process may improve various downstream analysis tasks such as cell type

identification, visualization, pseudotime, and hence might enhance the interpretation of the data. For example, analyses that are built on the gene-gene similarities can be improved by denoising where the "missing" correlation, which is lost due to the noise, is recovered. Such improvements can be evaluated by assessing whether the denoised data is more concordant with prior knowledge (Figure 4.2B-D). Overview of denoising methods and the effects of denoising on downstream tasks are given in the following sections.

The results reported in Section 4.2 are part of the following peer-reviewed publication. The contributions of the author are given below.

- **Gökcen Eraslan***, Lukas M. Simon*, Maria Mircea, Nikola S. Mueller, Fabian J. Theis. Single cell RNA-seq denoising using a deep count autoencoder, Nature Communications, 10 (2019): 390, `https://doi.org/10.1038/s41467-018-07931-2` *These authors contributed equally

  Contributions of the author:

  - Design and implementation of the method
  - Single-cell RNA-seq simulations using Splatter R package (Zappia et al. 2017)
  - Zero-inflation analysis of five datasets including the simulated data
  - Denoising and dimension reduction applications to Zheng et al. (2017) PBMC 68k and Paul et al. (2015) early blood development datasets
  - Comparison of diffusion pseudotime and gene-gene correlations of Paul et al. (2015) dataset before and after denoising
  - Scalability analysis of five denoising/imputations methods including GPU version of our method with 1.3M mouse brain cell dataset (10X Genomics 2017)
  - Hyperparameter selection and comparison
  - Interpretation of results
  - Generating figures and writing

4.1 Overview of imputation methods

Due to the challenges given in the previous section, imputation and/or denoising methods for scRNA-seq are getting increasingly available in the literature (Azizi et al. 2017; Ronen and Akalin 2018; Dijk et al. 2018; Huang et al. 2018; W. V. Li and J. J. Li 2018). A common theme in the existing approaches is to explicitly exploit cell-cell and/or gene-gene similarities using the correlation structure to generate "corrected" (or "smoothed") expression values. For example, Li and Li proposed an approach, named scImpute, which first groups similar cells into clusters and then identifies likely dropout events using a Gamma-Normal mixture model, and finally substitutes the expression values that are predicted as dropouts by borrowing information from similar cells (W. V. Li and J. J. Li 2018). SAVER is another mixture model-based approach which leverages the similarities of genes using Lasso models (Huang et al. 2018). MAGIC is a

global denoising approach based on graph diffusion, which propagates gene expression through similar cells in the kNN graph (Dijk et al. 2018). In this section, an overview of these two methods is given.

4.1.1 SAVER

SAVER (single-cell analysis via expression recovery) (Huang et al. 2018) is a statistical method for predicting the true expression values of cells in UMI-based scRNA-seq datasets. SAVER models the gene expression using a Poisson-Gamma mixture:

$$x_{ij} \sim Poisson(s_i \lambda_{ij})$$
$$\lambda_{ij} \sim Gamma(\alpha_{ij}, \beta_{ij})$$

Here, x_{ij} represents the observed UMI count of gene j in cell i which is assumed to be sampled from a Poisson distribution with mean $s_i \lambda_{ij}$. s_i represents size normalization factor to account for library size differences and is defined as total UMI counts per cell divided by the mean of total counts, whereas λ_{ij} represents the normalized true expression. A gamma prior with shape and rate parameters α_{ij} and β_{ij} is placed on the true expressions λ_{ij}. μ_{ij} and v_{ij} are the reparameterized mean and variance parameters of the gamma prior such that $\mu_{ij} = \alpha_{ij}/\beta_{ij}$ and $v_{ij} = \alpha_{ij}/\beta_{ij}^2$. The prior mean μ_{ij} is predicted by a Poisson LASSO regression:

$$\log E(x_{ij}/s_i | x_{ik}) = \log \mu_{ij} = \gamma_{j0} + \sum_{k \neq j} \gamma_{jk} \log \left[\frac{x_{ik} + 1}{s_i} \right]$$

where the logarithmized and size factor normalized UMI count of each gene is regressed on the other genes. The LASSO coefficients are denoted as γ_{j0} and γ_{jk} in the equation.

For the prior variance parameter v_{ij}, the authors considered three estimates with different modeling assumptions:

- Constant variance which assumes that the variance is constant for all cells and independent of the mean: $v_{ij} = v_j$
- Constant Fano which assumes that the variance scales linearly with the mean: $v_{ij} = F_j \mu_{ij}$ where $F_j = \frac{1}{\beta_{ij}}$.
- Constant CV^2 which assumes that the variance scales quadratically with the mean: $v_{ij} = CV_j^2 \mu_{ij}^2$ where $CV^2 = \frac{1}{\alpha_{ij}}$.

The estimated parameter with the highest maximum marginalized likelihood of x_{ij} given μ_{ij} and v_{ij} is selected for the final prior variance.

The true expression estimates in SAVER correspond to the mean of the posterior distribution

$$\lambda_{ij}|x_{ij}, \hat{\alpha}_{ij}, \hat{\beta}_{ij} \sim Gamma(x_{ij} + \hat{\alpha}_{ij}, s_i + \hat{\beta}_{ij})$$

which can be written as

$$\hat{\lambda}_{ij} = \frac{Y_{ij} + \hat{\alpha}_{ij}}{s_i + \hat{\beta}_{ij}} = \frac{Y_{ij} + \hat{\beta}_{ij}\hat{\mu}_{ij}}{s_i + \hat{\beta}_{ij}} = \frac{s_i}{s_i + \hat{\beta}_{ij}}\frac{x_{ij}}{s_i} + \frac{\hat{\beta}_{ij}}{s_i + \hat{\beta}_{ij}}\hat{\mu}_{ij}$$

where the shape parameter $\hat{\alpha}_{ij}$ is written in terms of the prior mean $\hat{\mu}_{ij}$ and the rate parameter $\hat{\beta}_{ij}$. This formula can be interpreted as a weighted sum of normalized UMI counts x_{ij}/s_i and the prior mean estimated with Poisson LASSO regression $\hat{\mu}_{ij}$. For the cells with high coverage i.e. high s_i value, the posterior relies more on the observed UMI counts whereas for the LASSO predictions with low uncertainty i.e. high $\hat{\beta}_{ij}$ values, true count estimation move towards the predicted mean. Therefore, the gene expression recovery model of SAVER relies on the information sharing between similar genes through LASSO models.

4.1.2 MAGIC

MAGIC (Markov affinity-based graph imputation of cells) (Dijk et al. 2018) is another approach designed to recover the scRNA-seq signal, which shares information across similar cells using graph diffusion. The method first employs two preprocessing steps performed on the raw counts, followed by graph construction and graph diffusion steps. An overview of these three steps is given below:

1. Preprocessing

   (a) $\mathbf{X}$ matrix, which represents the n-by-p raw count matrix with n cells and p genes, is row-wise normalized over all genes so that after the normalization. After the normalization, each cell has the same total count value, which is equal to the median of total count values:

$$s_i = \sum_{k=1}^{p} x_{ik}$$

$$x_{ij}^{norm} = \frac{x_{ij}}{s_i} median(\mathbf{s}) \text{ where } \mathbf{s} = (s_1, s_2, ..., s_n)$$

$$\mathbf{X}^{norm} = \begin{bmatrix} x_{11}^{norm} & \dots & x_{1p}^{norm} \\ \vdots & \ddots & \vdots \\ x_{n1}^{norm} & \dots & x_{np}^{norm} \end{bmatrix}$$

   (b) The dimensionality of the is reduced by applying principal component analysis on

the normalized data. The number of components is determined by 70% cutoff on the explained variance:

$$\mathbf{X}^{pca} = \text{PCA}(\mathbf{X}^{norm}, 0.70)$$

2. Graph construction

   (a) Cell-cell distance matrix $\mathbf{D}$ is computed using the PCs of the normalized count matrix, $\mathbf{X}^{pca}$. Using k-nearest neighbors

   (b) Distance matrix $\mathbf{D}$ is converted to an affinity matrix $\mathbf{A}$ via an adaptive Gaussian kernel:

   $$A_{ij} = \exp\left(-\frac{D_{ij}^2}{\sigma^2}\right)$$

   where σ represents the width of the adaptive kernel. With the Gaussian kernel, the affinity between cells decreases exponentially with the distance. This is a familiar trick used before in t-SNE (Maaten and G. Hinton 2008) and diffusion maps (Haghverdi, Büttner, et al. 2016) to capture more of the local structure information of the data compared to the global structure.

   (c) The affinity matrix is symmetrized additively:

   $$\mathbf{A}^{sym} = \mathbf{A} + \mathbf{A}^T$$

   (d) Finally, the symmetric affinity matrix $\mathbf{A}^{sym}$ is row-wise normalized so that every row sum up to 1:

   $$M_{ij} = \frac{A_{ij}^{sym}}{\sum_{k=1}^{n} A_{ik}^{sym}}$$

   which represents Markov transition matrix $\mathbf{M}$, where every element M_{ij} denotes the probability of cell i transitioning to cell j.

3. Diffusion

   (a) $\mathbf{M}$ is raised to the power of t, where the elements of the matrix M_{ij}^t represent the transition probabilities from cell i to cell j with a random walk of length t. Exponent t is defined as a hyperparameter which the user defines in advance. Typically, data is imputed with different t values from 1 to 8 and the outcomes are compared heuristically.

   (b) Finally, the data diffusion is performed by multiplying $\mathbf{M}^t$ by the original data matrix $\mathbf{X}$, which results in the smoothed (i.e. imputed) data matrix:

   $$\mathbf{X}^{imputed} = \mathbf{M}^t\mathbf{X}$$

Comprehensive benchmarks of the single-cell imputation/denoising methods including MAGIC, SAVER and many others are available in the literature (W. Hou et al. 2020; Vieth et al. 2019).

4.2 DCA: Deep count autoencoder

Characteristics of the scRNA-seq data, including the count structure, sparsity and nonlinear gene-gene relationships, are often not taken into consideration in the existing scRNA-seq denoising/imputation approaches. Moreover, the exponential increase in the number of profiled cells in single-cell studies (see Figure 1.4) strongly necessitates performant algorithms and implementations that scale up to millions of cells. Here, we propose a novel autoencoder-based[2] denoising approach, called deep count autoencoder (DCA), which addresses these shortcomings and improves the results of various downstream analyses. DCA is tailored to model sparse, count-structured scRNA-seq data with two different loss functions, which are formulated as the negative log-likelihood of the distributions that are commonly used in single-cell genomics (Risso et al. 2018; Lopez et al. 2018), namely negative binomial (NB) and zero-inflated negative binomial (ZINB) distributions. Similar to the generalized linear models (GLMs), the loss function takes the data and the distribution parameters predicted by the network (e.g. mean and dispersion) as input and measures the goodness of fit of these parameters to the data using the likelihood function in an unsupervised manner (Figure 4.3A).

DCA leverages two major advantages of autoencoders that are relevant for denoising scRNA-seq data. First, the ability of autoencoders to capture the manifold underlying the data (e.g. differentiation process of cells) enables DCA to potentially map the data points lying near the manifold due to the measurement noise back onto the manifold (Figure 4.3B). This process also implicitly shares information across genes and takes nonlinear gene-gene dependencies into account in denoising. Second, since autoencoders scale linearly with the number of data points, DCA is highly performant and is able to process up to millions of cells. Furthermore, the flexibility to use different noise models in DCA (i.e. NB and ZINB) is critical for the downstream analysis, mainly because the noise models and distributional assumptions suitable for the single-cell data are still under active debate. For example, Wenan Chen et al. (2018) argued that zero-inflation is less likely to occur in the data generated by the protocols using unique molecular identifiers (UMIs), compared to the scRNA-seq technologies using reads. Moreover, methods using other distributions including Poisson, Beta-Poisson and Normal distribution with zero-inflation were also previously proposed (Kharchenko et al. 2014; Vu et al. 2016; Risso et al. 2018; Pierson and Yau 2015; Finak et al. 2015).

Although the evaluation of denoising methods proves difficult due to the lack of ground truth for real single-cell datasets, we comprehensively evaluated our method on simulated and

[2]See Section 2.1.3 for a detailed description of autoencoders.

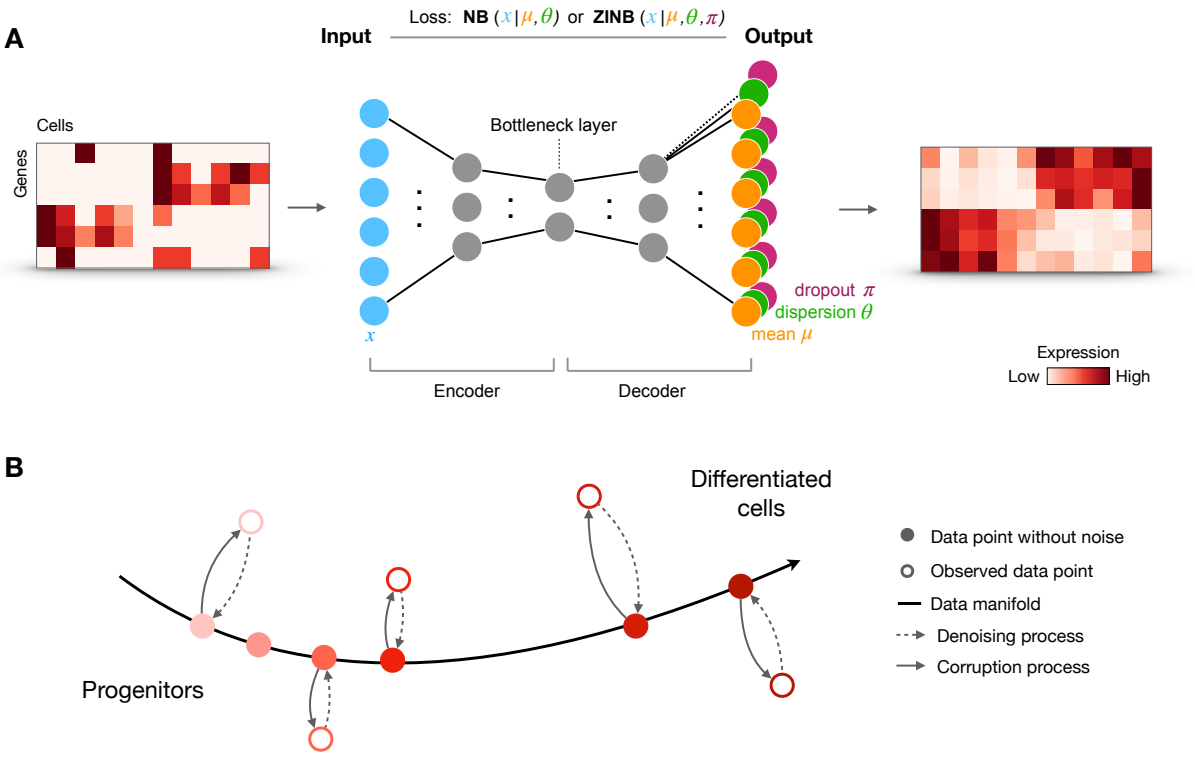


Figure 4.3: **(A)** The architecture of DCA, which takes the raw count matrix (gene by cell heatmap) as input through the input layer (blue circles i.e. genes), and estimates the parameters of the negative binomial (NB) or zero-inflated negative binomial (ZINB) distribution that produce the best reconstruction (output heatmap) according to the loss function. These parameters are mean (μ, orange circles), dispersion (θ, green circles) and dropout probabilities (π, purple circles), if zero-inflation is preferred. **(B)** Illustration of corruption and denoising processes affecting measured transcriptomes of differentiating cells. Ideal data points without noise (filled circles) lie on the differentiation manifold (black curve). The corruption process (solid arrows) moves these points away from the manifold, whereas the corrupted data points (empty circles) which lie near the data manifold are mapped back to the manifold by the denoising method (dotted arrows). Adapted from Goodfellow, Bengio, et al. (2016) and Eraslan, Simon, et al. (2019).

real datasets using various downstream tasks by utilizing additional sources of prior information. In our evaluations (Section 4.2.2), we observed that DCA enhances biological discovery in these tasks. Our Python implementation, which is provided as a Python package and as a command line tool, is publicly available on GitHub (`https://github.com/theislab/dca`) and in Scanpy (Wolf et al. 2018) as an external module (`https://scanpy.readthedocs.io/en/latest/external/index.html#imputation`).

4.2.1 Methods

Distributions for overdispersed count data

Negative binomial (NB) and zero-inflated negative binomial (ZINB) distributions are commonly used for modeling overdispersed count data (Perumean-Chaney et al. 2013). NB is typically parameterized by the mean (μ) and dispersion (θ) parameters, whereas ZINB is a mixture model consisting of an NB component representing the count process and a point mass at zero, which accounts for excess zeros in the count data. π and $1-\pi$ parameters are used as mixture weights for the point mass and NB components, respectively (see Section 2.1.1). In the context of scRNA-seq, the NB component represents the process that generates the counts and the point mass represents the dropout process which inflates the expected amount of zeros in the count data, hence zero-inflation. Therefore, the π parameter can be interpreted as the probability that a dropout event occurs. Likelihood functions of NB and ZINB distributions whose logarithmized forms are used as loss functions in our approach are given below:

$$\begin{aligned} \mathcal{L}_{\mathrm{NB}}(x;\mu,\theta) &= \frac{\Gamma(x+\theta)}{\Gamma(\theta)\Gamma(x+1)}\left(\frac{\theta}{\theta+\mu}\right)^{\theta}\left(\frac{\mu}{\theta+\mu}\right)^{x} \\ \mathcal{L}_{\mathrm{ZINB}}(x;\pi,\mu,\theta) &= \pi\delta_0(x) + (1-\pi)\mathcal{L}_{\mathrm{NB}}(x;\mu,\theta) \end{aligned} \tag{4.1}$$

where $\Gamma()$ and $\delta_0()$ refer to the gamma function, an extension of the factorial function to complex numbers, and the Dirac delta function, representing the point mass at zero, respectively.

Architecture and loss functions

Autoencoders consist of encoder and decoder components with a lower-dimensional hidden layer between the encoder and the decoder called "the bottleneck", representing the latent space representations of data points (see Section 2.1.3). In the traditional autoencoders, the decoder has a single output layer that produces the predictions for the mean parameter of the normal distribution when the mean squared error is used as a loss function. Because NB and ZINB distributions require multiple parameters, DCA infers mean (μ) and dispersion (θ) (additionally the dropout parameter, π, for ZINB) using multiple output layers of the decoder.

The activation functions of the output layers (e.g. inverse link functions) are determined by taking the constraints of each parameter into account, e.g. exponential for mean and dispersion, to ensure non-negativity and sigmoid for the dropout parameter. Note that inferring dispersion and dropout parameters through the output layers implicitly conditions these parameters on the latent (bottleneck) variables and hence cell type and state. Alternatively, they can be defined as independent parameters (i.e. not conditioned on the data) and/or a scalar parameter shared across all genes.

The formulation of the architecture in the matrix form is as follows:

$$\begin{aligned}
\mathbf{E} &= \mathrm{ReLU}(\bar{\mathbf{X}}\mathbf{W}_E) \\
\mathbf{B} &= \mathrm{ReLU}(\mathbf{E}\mathbf{W}_B) \\
\mathbf{D} &= \mathrm{ReLU}(\mathbf{B}\mathbf{W}_D) \\
\bar{\mathbf{M}} &= \exp(\mathbf{D}\mathbf{W}_\mu) \\
\Pi &= \mathrm{sigmoid}(\mathbf{D}\mathbf{W}_\pi) \\
\Theta &= \exp(\mathbf{D}\mathbf{W}_\theta)
\end{aligned} \tag{4.2}$$

Here encoder, bottleneck and decoder are represented as $\mathbf{E}$, $\mathbf{B}$ and $\mathbf{D}$ variables, respectively. Encoder and decoder hidden layers contain 64 neurons in the standard DCA architecture, whereas the bottleneck layer has 32 neurons. Rectified linear unit (ReLU) is used as the activation function in these layers. Mean, dispersion and dropout parameters, $\bar{\mathbf{M}}$, Π and Θ, are defined as output layers. The input $\bar{\mathbf{X}}$ denotes total count normalized, log-transformed and scaled (via z-score) raw counts:

$$\bar{\mathbf{X}} = \mathrm{zscore}(\log(\mathrm{diag}(s_i)^{-1}\mathbf{X} + 1)) \tag{4.3}$$

where $\mathbf{X}$ represents raw counts and s_i is the total number of counts for cell i (e.g. $s_i = \sum_j x_{ij}$).

NB loss function of DCA, which is optimized via stochastic gradient descent (SGD), can be written as:

$$\underset{\mathbf{M},\Theta}{\mathrm{argmin}} - \log \mathcal{L}_{\mathrm{NB}}(\mathbf{X}; \mathrm{M}, \Theta) \tag{4.4}$$

$$= \underset{\mathbf{M},\Theta}{\mathrm{argmin}} \sum_i \sum_j - \log \mathcal{L}_{\mathrm{NB}}(x_{ij}; \mu_{ij}, \theta_{ij}) \tag{4.5}$$

where $\mathcal{L}_{\mathrm{NB}}$ refers to the likelihood function given in Equation 4.1, i and j are cell and gene indices. Note that the loss functions uses $\mathbf{M}$ instead of $\bar{\mathbf{M}}$. This aims to keep the mean predictions independent of total count bias (e.g. due to differing library sizes) by scaling the

predicted means ($\bar{\mathbf{M}}$) back to the original count scale using the total counts (s_i):

$$\mathbf{M} = \mathrm{diag}(s_i)\bar{\mathbf{M}} \tag{4.6}$$

Similarly, ZINB loss function is defined as follows:

$$\begin{aligned} &\underset{\Pi,\mathbf{M},\Theta}{\operatorname{argmin}} - \log \mathcal{L}_{\mathrm{ZINB}}(\mathbf{X}; \Pi, \mathrm{M}, \Theta) + \lambda ||\Pi||_{\mathrm{F}}^2 \\ =&\underset{\Pi,\mathbf{M},\Theta}{\operatorname{argmin}} \sum_i \sum_j - \log \mathcal{L}_{\mathrm{ZINB}}(x_{ij}; \mu_{ij}, \theta_{ij}, \pi_{ij}) + \lambda \pi_{ij}^2 \end{aligned} \tag{4.7}$$

where the tunable λ hyperparameter controls the strength of the ridge prior over the dropout probabilities Π. Especially for the lowly expressed genes, dropout probability may approach to 1.0, which prevents the weights driving the NB parameters from being updated. The ridge prior aims to avoid that by shrinking the dropout parameters. Moreover, DCA provides a hyperparameter search implementation to facilitate the optimization of the λ hyperparameter.

Training

We used the RMSProp variant[3] of the SGD with a learning rate of 0.001 for the optimization of DCA parameters. A learning rate scheduling scheme is performed where the learning rate was scaled by 0.1, if the validation error does not improve for 20 epochs. Early stopping was employed to speed up the training process and avoid overtraining, where training is stopped if the validation loss does not improve for 25 epochs. Gradients are clipped to 5.0 in order to stabilize the training and the batch size of 32 is used in the training of all datasets.

Denoising

The mean parameter of the NB component before the total count scaling ($\bar{\mathbf{M}}$ in Equation 4.2 and 4.6) represents the "denoised" and total count normalized version of the data. Therefore, it is the primary outcome of DCA, which is used in downstream applications. From a NB/ZINB GLM regression perspective, DCA can be intuitively interpreted as a two-step process where 1) representations of cells via some latent features are inferred by the encoder and decoder (denoted as $\mathbf{D}$ in Equation 4.2 which is the layer before the output layer) and 2) gene expression is regressed on these "new" features via NB/ZINB regression. Note that this is only an intuitive explanation and the autoencoder framework allows joint training of these two steps.

[3]See the notes of lecture 6 from the online course "Neural Networks for Machine Learning" by Geoffrey Hinton for the details on RMSProp.

Implementation

DCA is implemented as a stand-alone command line tool and as a Python 3 package using the deep learning frameworks Keras and Tensorflow, which support training on CPUs and GPUs. hyperopt (Bergstra et al. 2015) and kopt (`https://github.com/Avsecz/kopt`) Python packages are used for the optimization of the hyperparameter including the number of layers, the number of neurons and the ridge prior weight (λ) for the ZINB loss. For hyperparameter optimization, we trained DCA with different configurations for 100 epochs using the Tree-structured Parzen Estimator (TPE) (Bergstra et al. 2015) and picked the model with the lowest validation loss.

Zero inflation tests

To test whether zero inflation significantly exists in single-cell datasets, we fit NB and ZINB distributions to the selected clusters in four real scRNA-seq datasets (three UMI- and one read-based protocol) and one simulated dataset. For the NB model, first, the dispersion parameter (θ) is estimated using the relationship between mean (μ) and variance (σ^2) i.e. $\sigma^2 = \mu + \theta\mu^2$. For the ZINB fits, the zero-inflation parameter is inferred as an affine function of the observed mean jointly with the dispersion parameter using numerical optimization where we compared the empirical and the predicted dropout rates using binary cross entropy (BCE). Finally, we performed likelihood ratio tests between NB and ZINB fits using BCE to calculate p-values and test whether the zero inflation is significant.

Simulations of scRNA-seq data

To simulate realistic single-cell datasets, we used the Splatter package (Zappia et al. 2017) in R. `splatSimulate()` function was used with `groupCells=2`, `nGenes=200`, `dropout.present=TRUE`, `dropout.shape=-1`, `dropout.mid=5` parameters to simulate a dataset with two clusters and with `groupCells=6`, `nGenes=200`, `dropout.present=TRUE`, `dropout.shape=-1`, `dropout.mid=1` to simulate a dataset with six clusters. With these parameters, 63% and 35% of the matrix entries were set to zero, for two- and six-group datasets, respectively. Note that the dropout noise was conditioned on the mean expression such that the dropout likelihood was higher in lowly expressed genes.

68k peripheral blood mononuclear cell analysis

Gene expression count matrix and the cell type annotations of 68k peripheral blood mononuclear cell (PBMC) dataset (Zheng et al. 2017) were downloaded from the GitHub repository of 10X Genomics at `http://www.github.com/10XGenomics/single-cell-3prime-paper`. We collapsed the granular cell states/subtypes of CD4+ and CD8+ T cells (e.g. CD4+/CD25+

Reg. T, CD4+/CD45RO+ memory T) due to the high overlap within these cell types. tSNE coordinates of the cells were reproduced using the code available at the same GitHub repository. We used an architecture with two neurons in the bottleneck layer for visualization purposes and 16 neurons in the two additional hidden layers. We subsetted the dataset to top 1000 highly variable genes for this analysis using `sc.pp.filter_genes_dispersion()` function in Scanpy. `sklearn.metrics.silhouette_score()` function from the scikit-learn Python package was used to calculate the cell type separation on 2D representations.

Pseudotime and correlation analysis of blood differentiation

Gene expression count matrix and the cell type annotations of Paul et al. (2015) hematopoietic stem cell (HSC) differentiation data (containing 2730 cells and 3451 informative genes) were obtained via "`sc.datasets.paul15()`" Scanpy function. Diffusion map and pseudotime (DPT)were computed with "`sc.tl.dpt(adata, n_branchings=1)`" after constructing a k-nearest neighbors graph on logarithmized and normalized counts. Pseudotime estimates of the cells in the MEP and GMP branches of the differentiation manifold are scaled between $[0, 1]$ and $[0, -1]$ respectively to facilitate the interpretation and visualization. Pearson correlation coefficients were calculated with the "`corrcoef`" function from the `numpy` Python package.

CITE-seq cord blood mononuclear cells analysis

UMI and antibody-derived tag (ADT) counts of cord blood mononuclear cells profiled by the CITE-seq protocol (Stoeckius et al. 2017) were downloaded from the Gene Expression Omnibus (GEO) via the accession number GSE100866. Data is preprocessed and annotated as in the Seurat multimodal analysis vignette (`https://satijalab.org/seurat/archive/v2.4/multimodal_vignette.html`). Mouse cells, unknown cells and megakaryocytes were removed which yielded a total of 7617 cells. Centered log ratio (CLR)-transformed ADT counts provided by the authors were used for the protein expression. We used the top 5000 highly variable genes for denoising the RNA counts with DCA. Co-expression of three known protein markers (CD3, CD11c and CD56) and the corresponding mRNA markers (CD3E, ITGAX, and NCAM1) was compared using Pearson correlation (`corrcoef()` function in NumPy Python package) across all cells.

Scalability analysis with 1.3 million cells

10X Genomics scRNA-seq dataset containing 1.3 million mouse brain cells was downloaded from the dataset webpage of 10X Genomics available at `https://support.10xgenomics.com/single-cell-gene-expression/datasets/1.3.0/1M_neurons`. After removing cells and genes without expression, we subsetted genes to top thousand highly variable genes using

the `"filter_genes_dispersion"` function of Scanpy with `n_top_genes=1000` argument. The rows of the data matrix were downsampled to 100, 1,000, 2,000, 5,000, 10,000 and 100,000 cells to compare different denoising/imputation approaches using datasets with various sizes. Next, we denoised each matrix with the five methods as well as the GPU version of DCA and measured the runtime.

Definitive endoderm differentiation analysis

Endoderm differentiation single-cell gene expression data from Chu et al. (2016) were subsetted to the human embryonic stem cells (H1) using the provided annotation and the 1000 most highly variable genes. The dataset which is based on read counts was then compared to the datasets with UMI-counts in the zero-inflation analysis.

DCA applications

We used the default DCA command line arguments (which implies `-type zinb` i.e. ZINB loss) for the simulated datasets. NB loss function was used in the analyses of cord blood mononuclear cell CITE-seq dataset, 68k peripheral blood mononuclear cell dataset, and the Paul et al. (2015) blood cell differentiation dataset via the `-type nb` command line argument.

Code availability

Code for reproducing the figures in this chapter and the DCA tutorial are available at `https://github.com/theislab/dca`.

4.2.2 Results

Count-based noise model is necessary for denoising simulated scRNA-seq data

We first evaluated the performance of DCA on gene expression recovery and clustering tasks using simulated datasets where the ground truth expression and clusters, which represent broad cell classes, are known. Using Splatter (Zappia et al. 2017), we simulated two datasets consisting of two and six cell groups both with 200 genes and 2000 cells (see Section 4.2.1 for simulation details). We observed that adding substantial dropout noise to the simulated expression data with two groups obscured the cluster structure (Figure 4.4A-B). We were able to recover the cluster structure after denoising the data using DCA (Figure 4.4A-B). Unlike DCA with ZINB loss function, which explicitly takes the dropout noise into account, a regular autoencoder trained with a mean squared error (MSE) loss function that was applied to the logarithmized counts failed to recover the cell groups (Figure 4.4A-B). This indicates that the count characteristics and dropout noise in scRNA-seq data necessitate custom loss functions

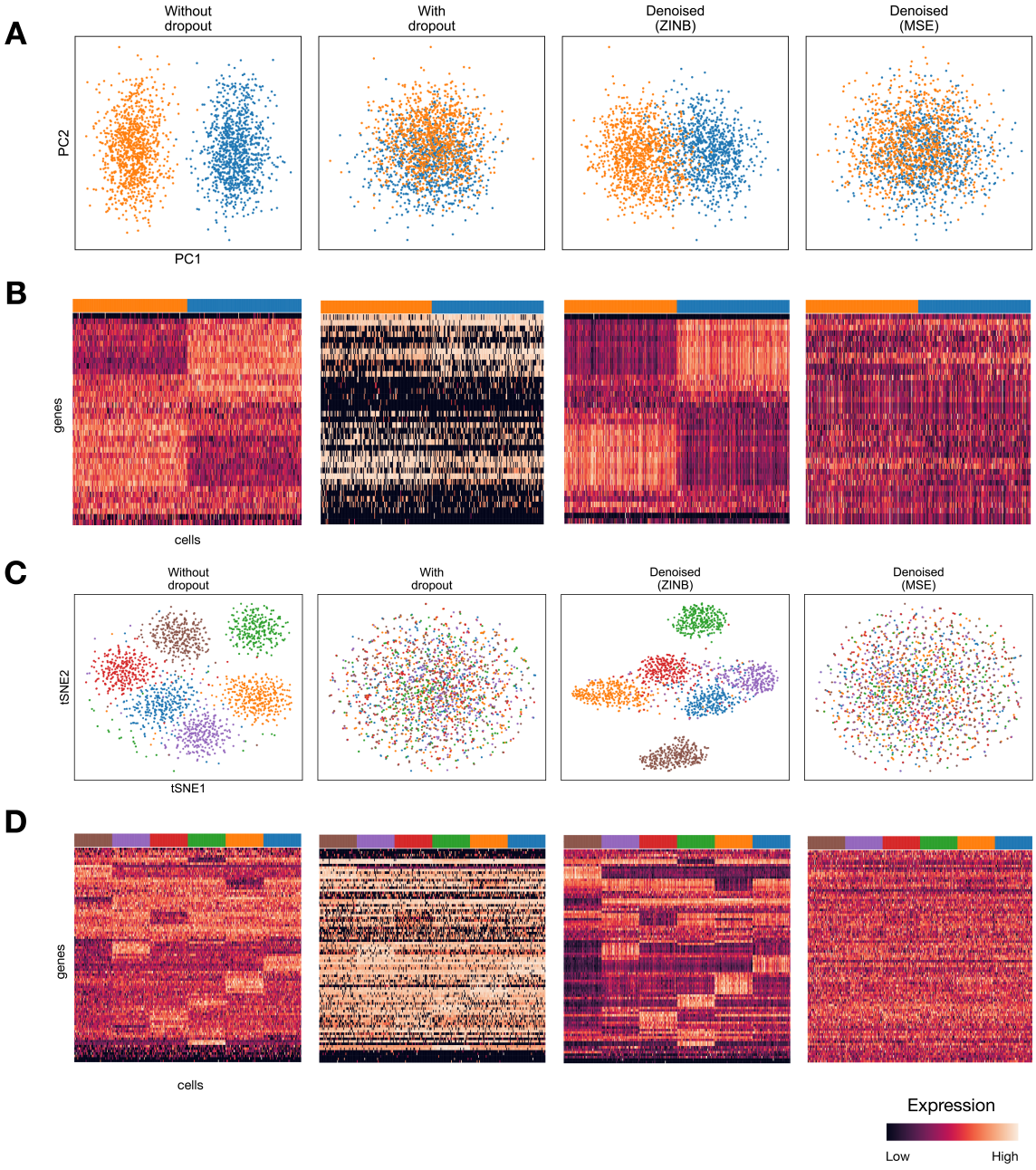


Figure 4.4: **A)** PCA representation of simulated scRNA-seq data with two groups. From left to right: Data without dropout noise, with dropout noise, denoised with DCA and denoised with an autoencoder with MSE loss function. **B)** Gene expression of two-group simulation data visualized with heatmaps. **C)** tSNE representation of simulated data with six cell types. **D)** Gene expression of six-group simulation data visualized with heatmaps. Cells are colored by the ground truth cell groups in panels A and C. Adapted from Eraslan, Simon, et al. (2019) (CC-BY-4.0).

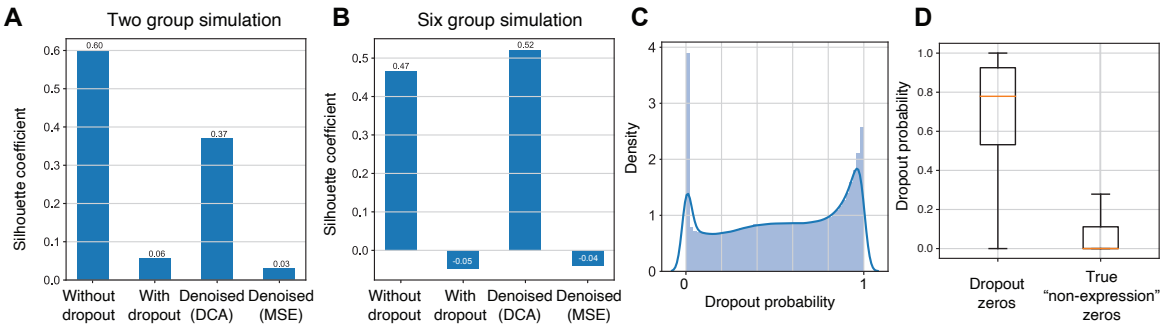


Figure 4.5: Consistency of cell groups in terms of Silhouette coefficient for four representations of simulated datasets in PCA **(A)** and tSNE space **(B)**. **C)** The distribution of dropout probabilities inferred by DCA via the π output layer. **D)** Dropout probability distributions shown separately as boxplots for the entries that were subject to dropout noise in the simulation (i.e. dropout zeros) versus the zeros representing no expression (i.e. true "non-expression" zeros). Adapted from Eraslan, Simon, et al. (2019) (CC-BY-4.0).

based on the likelihood of count distributions. We achieved similar results in the simulated dataset with six groups (Figure 4.4C-D).

We next quantitatively compared the clustering performances of DCA and the MSE autoencoder using the Silhouette coefficient, which showed that denoising with DCA highly improved the cluster structure in both two-group and six-group simulation datasets compared to the MSE loss (Figure 4.5A-B). The simulations also allowed us to investigate whether DCA is able to distinguish ground truth "dropout zeros", which are the entries that were set to zero by the dropout noise, and the "true zeros" which represent zeros due to non-expression. We examined the distribution of dropout probabilities (Figure 4.5C) which are the outputs of "dropout" layer (i.e. π parameter, Figure 4.3A). We observed that dropout zeros were assigned much higher dropout probabilities (median: 0.79) compared to the true zeros (median: 0.0, Figure 4.5D).

We further compared the true expression values (i.e. counts before adding dropout noise) with the denoised values to quantify the performance of gene expression recovery. We observed that the gene expression predicted by DCA highly correlates with the true expression (Pearson r=0.978, Figure 4.6A) and outperformed the MSE autoencoder (r=0.779). As expected, the MSE autoencoder performed better in the high expression regime (e.g. $x > 1000$, r=0.8) since the error in this regime simply contributes more to the overall loss. However, the predictions of the MSE autoencoder showed different trends in the medium to low expression regimes. Lowly expressed genes ($x < 50$) were highly underestimated (r=0.615) and the values in the medium expression regime ($50 < x < 1000$) were both over- and underestimated (r=0.66). In order to investigate this trend further, we correlated the true values of the dropout zeros with their denoised predictions. While DCA showed a correlation comparable to its performance on all counts (r=0.95, Figure 4.6B), MSE performance dropped to 0.649. Moreover, no overestimation

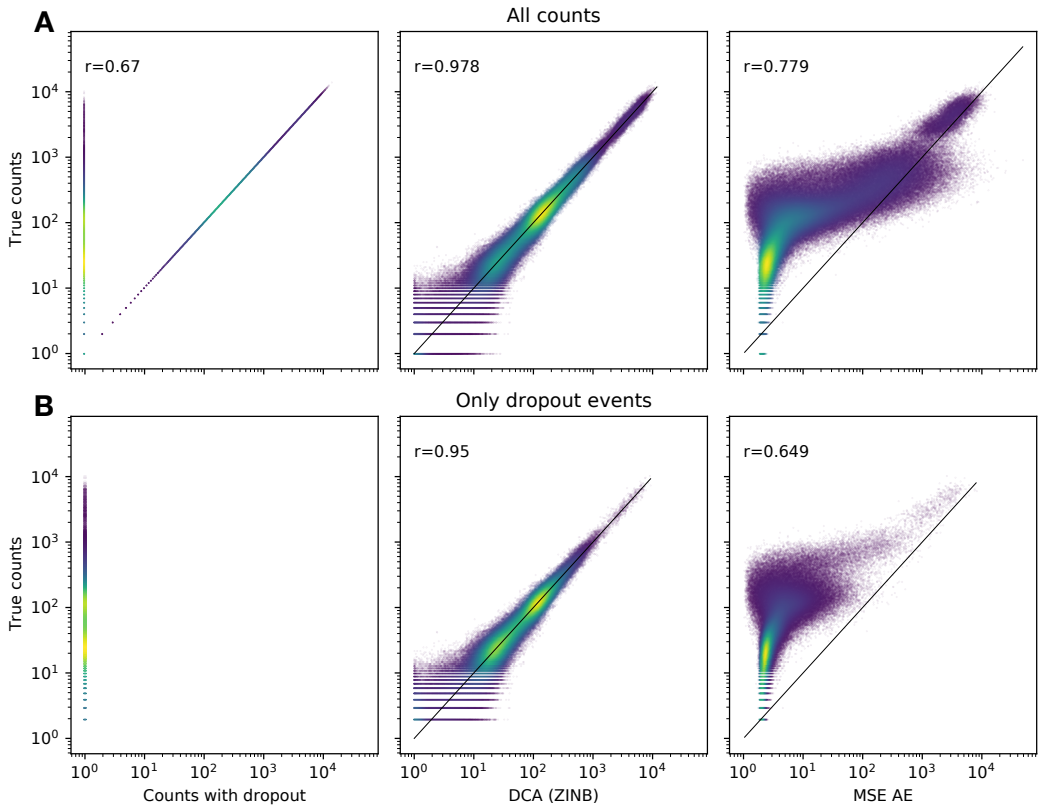
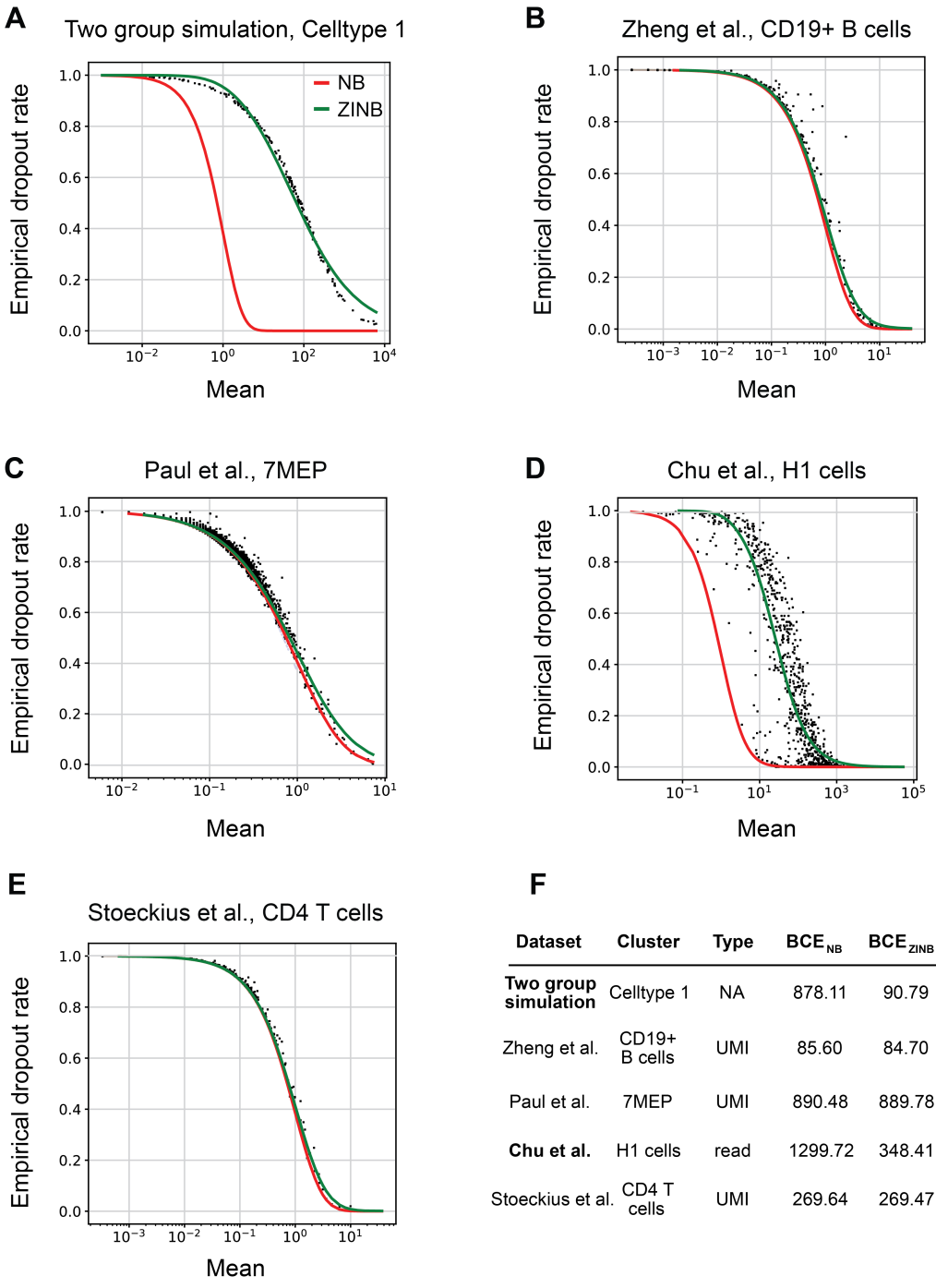


Figure 4.6: Correlation between the ground truth expression and the expression denoised with DCA and MSE autoencoder. All counts **(A)** and the non-zero counts that were set to zero due to the dropout noise **(B)** are shown separately. Pearson correlation coefficients are shown in the upper-left corner of the panels where applicable. The color of the dots represents density.

by the MSE autoencoder was observed in the medium expression regime.

UMI count datasets are not zero-inflated

To select the appropriate count distribution-based noise model implemented in DCA (i.e. NB and ZINB), it is important to determine whether the zero-inflation trend is present in a given dataset. To explore this and provide a simple test that may guide the users in the selection of the noise model, we implemented a zero-inflation test as a part of DCA and applied it to the simulated and real single-cell datasets. In this test, first NB and ZINB distributions are fitted to the relationship between the mean expression and the empirical dropout rates (i.e. fraction of observed zeros) using the expression values of cells in a selected cluster. Next, the goodness of these two fits are compared using the likelihood ratio test (see Section 4.2.1 for details). Expectedly, the ZINB fit showed significantly better likelihood compared to the NB model in the simulated two-group dataset (Figure 4.7A,F). Among the four real datasets analyzed, the ZINB model fit was significantly better only in the Chu et al. (2016)



F

Dataset	Cluster	Type	BCE_{NB}	BCE_{ZINB}	p-value
Two group simulation	Celltype 1	NA	878.11	90.79	< 1e-16
Zheng et al.	CD19+ B cells	UMI	85.60	84.70	0.405
Paul et al.	7MEP	UMI	890.48	889.78	0.498
Chu et al.	H1 cells	read	1299.72	348.41	< 1e-16
Stoeckius et al.	CD4 T cells	UMI	269.64	269.47	0.844

Figure 4.7: **(A-E)** Mean expression and empirical dropout rate (i.e. fraction of zeros) are shown for each gene in selected clusters of five datasets. Red and green curves represent negative binomial (NB, red) and zero-inflated negative binomial (ZINB, green) fits. Each dot is a gene. **F)** Selected clusters, type of the datasets, negative log-likelihood estimates of NB and ZINB fits and the significance of the difference between the goodness of fits of two distributions are given. BCE: binary cross entropy (i.e. negative log-likelihood of Bernoulli). Adapted from Eraslan, Simon, et al. (2019) (CC-BY-4.0).

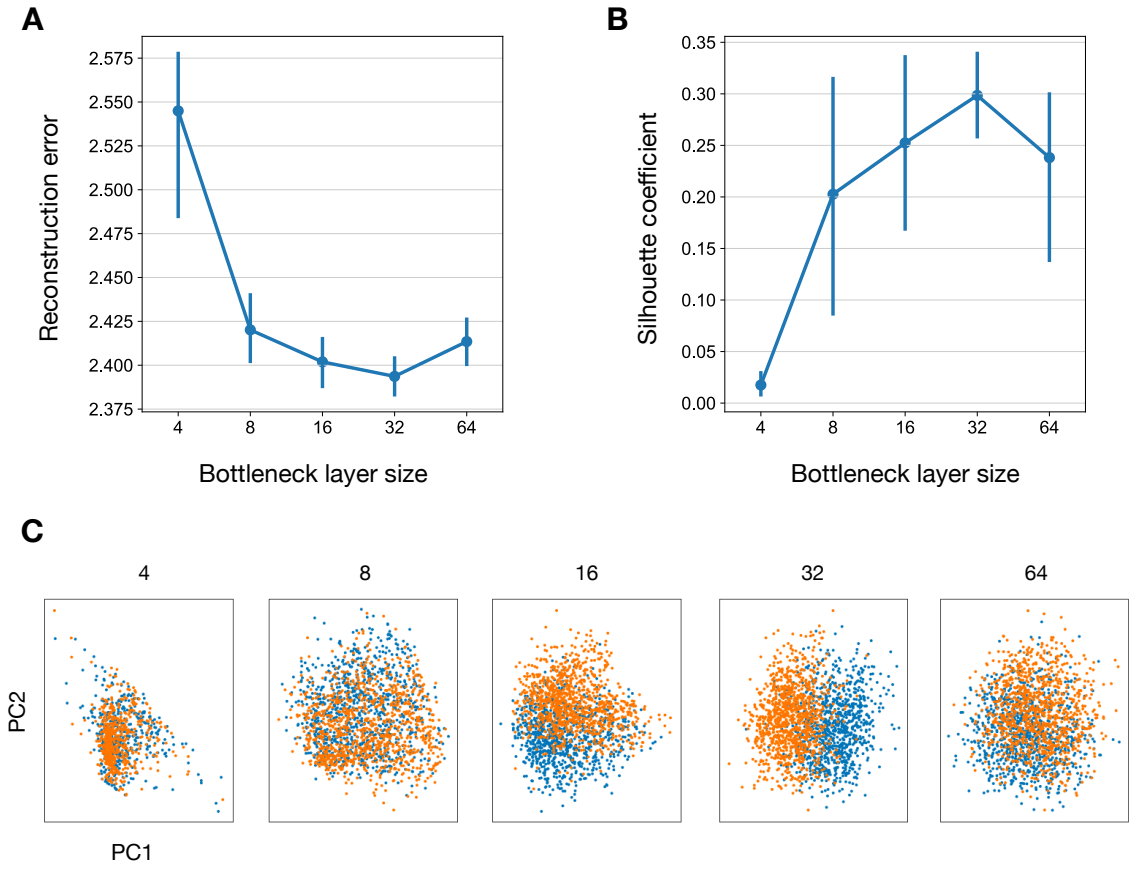


Figure 4.8: The relationship between the size of the DCA bottleneck layer and the reconstruction loss **(A)** as well as the separation of cell groups in terms of Silhouette coefficients **(B)**. The error bars represent standard error across five denoising runs. **C)** PCA representations of cells denoised using DCA with five different bottleneck layer sizes. Colors represent ground truth groups of simulated cells. Adapted from Eraslan, Simon, et al. (2019) (CC-BY-4.0).

definitive endoderm differentiation dataset which is the only read-based protocol analyzed (Figure 4.7B-F). Concordant with the observation reported by Wenan Chen et al. (2018), we conclude that UMI data does not exhibit any sign of zero-inflation, whereas the read-based protocols are more likely to have a zero inflation trend.

Downstream effects of hyperparameter selection

Neural networks, including autoencoders, typically require a set of hyperparameters (e.g. number of layers, size of the layers, regularization and dropout coefficients) to be specified by the user before the training. Similar to the choice of noise model, we implemented a hyperparameter search procedure in DCA, which can compare different hyperparameters via random or grid search. This can be used to evaluate the effects of hyperparameter selection on

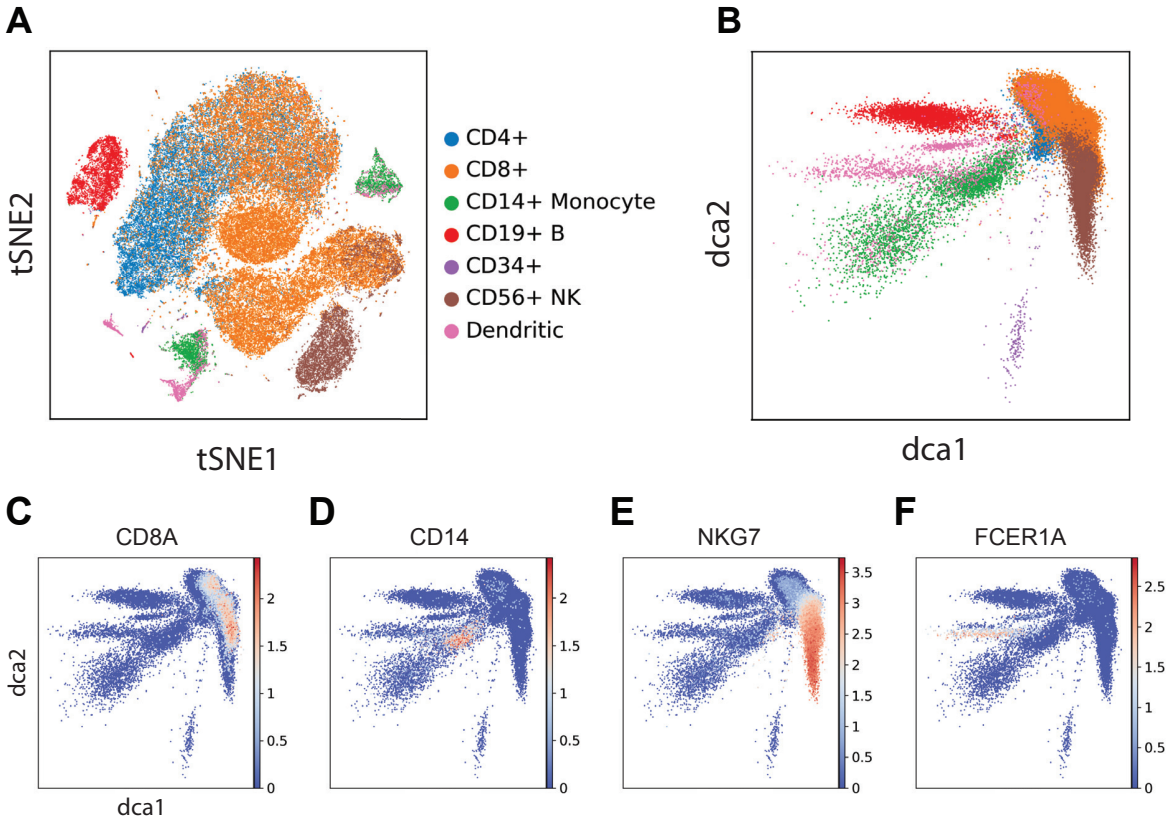


Figure 4.9: **A)** tSNE representation of 68k PBMCs reproduced from Zheng et al. (2017). **B)** DCA latent space visualization of the same dataset where the 2D bottleneck layer activations are used for visualization. Cells are colored by the cell types obtained from Zheng et al. (2017) where CD4 and CD8 subtypes are merged into coarser groups in panel A and B. **(C-F)** DCA representations of cells colored by the log-transformed expression of the CD8A, CD14, NKG7 and FCER1A genes which are the known cell type markers of CD8+ T cells, CD14+ monocytes, CD56+ natural killer cells and dendritic cells, respectively. Adapted from Eraslan, Simon, et al. (2019) (CC-BY-4.0).

the denoising performance of DCA and to guide the users. Using the grid search approach, we denoised the two-group simulated scRNA-seq data with varying bottleneck layer sizes (i.e. 4, 8, 16, 32 and 64) and evaluated the denoising performance using the reconstruction error (Figure 4.8A) and separation of ground truth cell groups via the Silhouette coefficient (Figure 4.8B). We ran each configuration five times to calculate the standard errors. We obtained the optimal values for both the reconstruction error and the Silhouette coefficient with the bottleneck size of 32 neurons. We further visualized the denoised cells with PCA (Figure 4.8C), which showed good cluster separation for the architecture with 32 bottleneck neurons (Silhouette: 0.3).

Denoising process accounts for the cell population structure in real data

We next sought to examine whether DCA is able to capture cellular heterogeneity, e.g. broad cell classes, which typically constitutes a major source of variation in complex single-cell datasets (Zheng et al. 2017). Observing cell type-specific structure in the latent space suggests that the noise model parameters are inferred conditionally on cell types and therefore, the denoising procedure takes the cell population structure into account. We tested this hypothesis by training DCA with only two bottleneck neurons on 68,579 peripheral blood mononuclear cells (Zheng et al. 2017) (Figure 4.9A, Silhouette: -0.01). Visualizing the bottleneck neuron activations on 2D yielded cell type-specific variation in the data (Figure 4.9B, Silhouette: 0.07). Moreover, we visualized log-normalized expression of the markers of four major cell populations (CD8+ T cells, CD14+ monocytes, NK cells and dendritic cells) on the same DCA latent space (Figure 4.9C-F). Our results indicate that DCA latent space captures cell type-specificity and therefore, the denoising process accounts for the the cellular heterogeneity in the data.

DCA captures the cell differentiation process

Clustering, an unsupervised technique commonly used in single-cell genomics, relies on the discretization of cellular states that are expected to align with disparate cell types such as T cells and dendritic cells. However, there are also cases such as cell differentiation in which the transcriptome landscape can be modeled as a continuum where the phenotype is inferred as a continuous variable e.g. pseudotime (Haghverdi, Büttner, et al. 2016). Similarly to the cell population analysis given previously, we investigated whether DCA is able to capture continuous phenotypes e.g. differentiation process in the latent space. We trained DCA with two bottleneck neurons on the blood differentiation dataset by Paul et al. (2015) to test this hypothesis. The visualization of the latent space revealed that the major branches of the differentiation trajectory i.e. megakaryocyte–erythroid progenitors (MEP) and granulocyte-macrophage progenitors (GMP) are captured by DCA (Figure 4.10A). Moreover, we examined whether this representation can be used directly for the diffusion pseudotime (DPT) inference and whether the inferred pseudotemporal ordering of cells in DCA latent space (Figure 4.10B) correlates with the pseudotime order of cells in gene expression space. The comparison of these two DPT results showed a very high correlation (Figure 4.10C, Pearson correlation: 0.95). This indicates that DCA is able to capture biologically meaningful information (e.g. pseudotime) in its latent space when applied to continuous phenotypes such as cell differentiation processes and the latent space representation can be further used in downstream analysis.

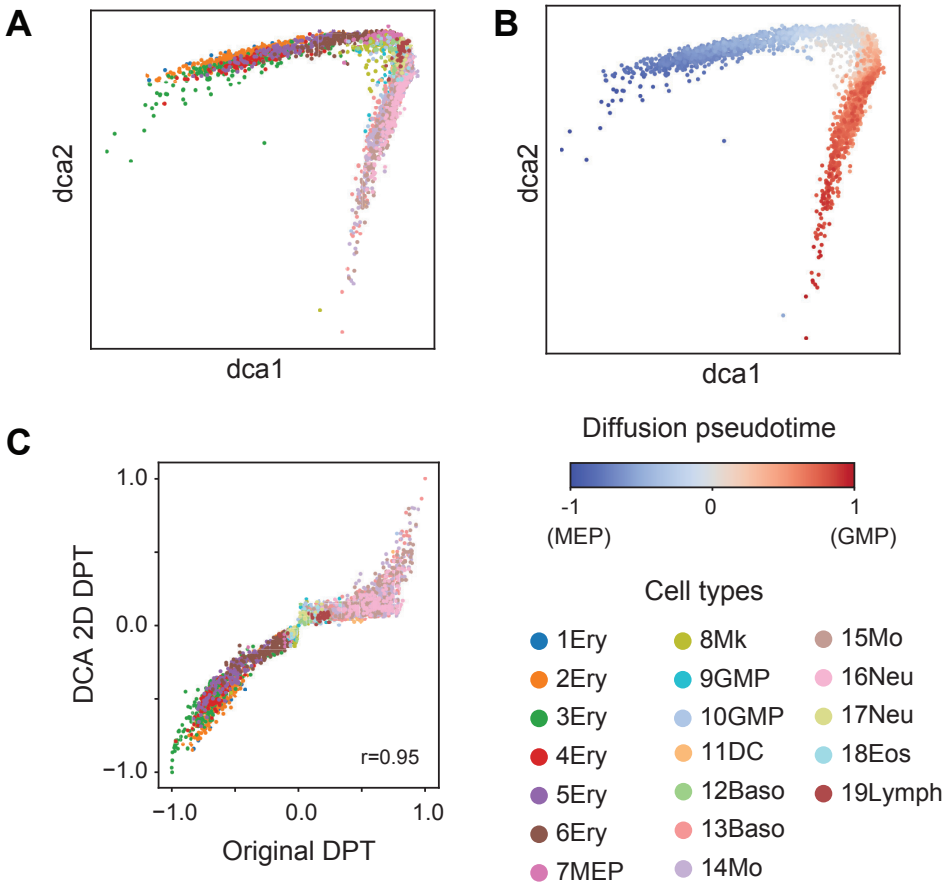


Figure 4.10: DCA latent space representations of differentiating blood cells from Paul et al. (2015) colored by annotated cell clusters **(A)** and diffusion pseudotime (DPT) **(B)**. **C)** Comparison of the DPT calculated on gene expression (x-axis) against the DPT on 2D DCA latent space (y-axis). Pearson correlation coefficient is shown in the lower right corner. Cell type annotations are obtained from Paul et al. (2015). **Abbreviations**: Ery (erythrocytes), Mk (megakaryocytes), DC (dendritic cells), Baso (basophils), Mo (monocytes), Neu (neutrophils), Eos (eosinophils), Lymph (lymphoid cells), MEP (Megakaryocyte/erythrocyte progenitor), GMP (Granulocyte/macrophage progenitor). Adapted from Eraslan, Simon, et al. (2019) (CC-BY-4.0).

DCA improves the correlation structure of key regulatory genes

Downstream analyses in single-cell genomics, such as finding differentially expressed genes between cell types or experimental conditions, are typically performed at a single gene level, meaning that the test statistics and/or association p-values are calculated separately for each gene. To gain a higher level and functional understanding of the findings, such as determining relevant pathways and processes, pre-defined gene sets (e.g. Gene Ontology) are used to query these results in the form of enrichment tests. As an alternative approach, downstream analyses can be conducted at the level of gene modules by inferring data-driven modules and interrogating the effects of e.g. experimental conditions on these modules. This can potentially

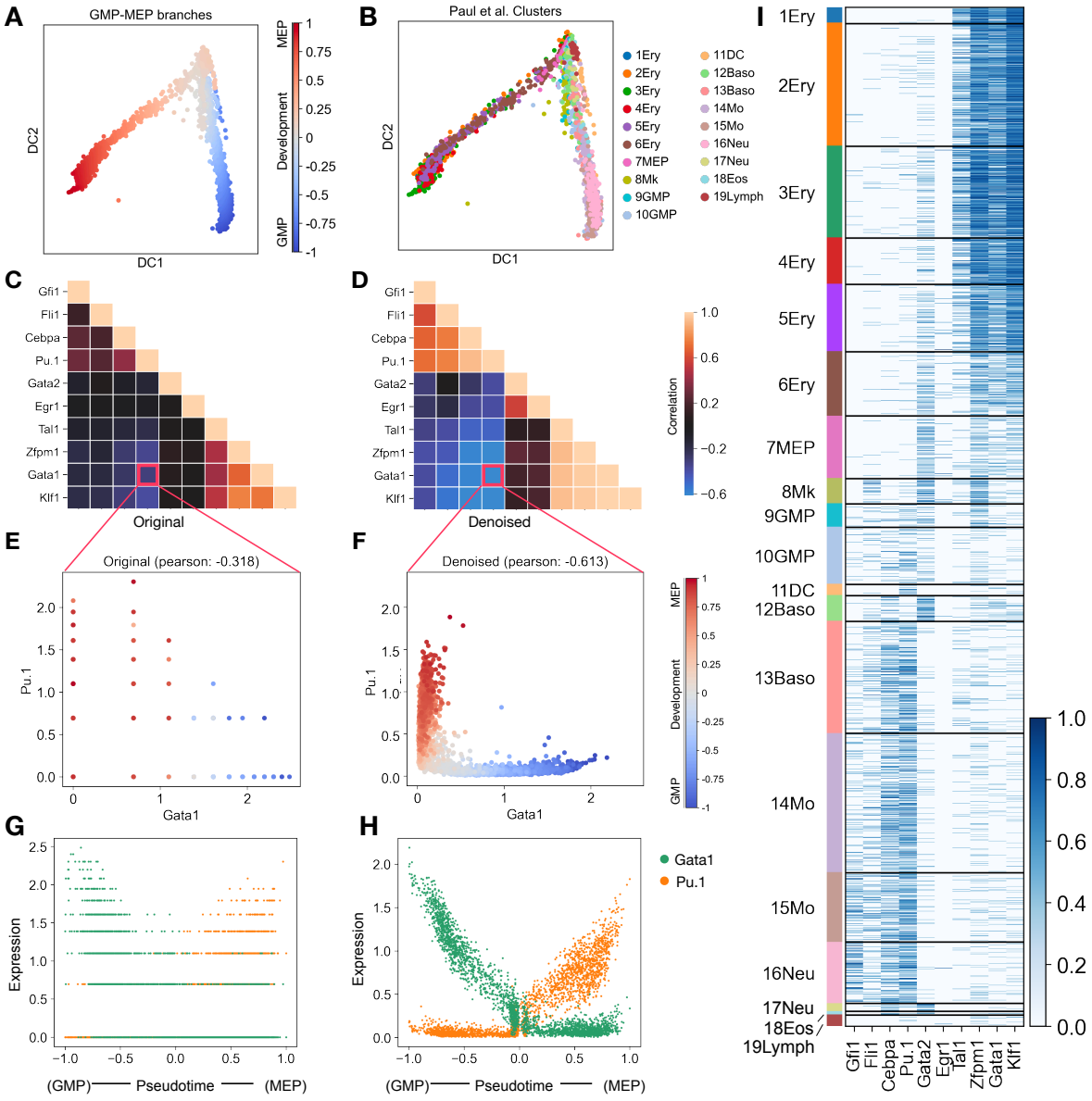


Figure 4.11: Diffusion map representations of two major branches of blood cell differentiation (GMP and MEP) colored by diffusion pseudotime **(A)** and annotated cell clusters **(B)**. Heatmaps of Pearson correlation coefficients for well-known blood regulators (Krumsiek et al. 2011) before **(C)** and after **(D)** denoising. Correlation of Pu.1 - Gata1 transcription factors are highlighted in the heatmaps. Anti-correlation patterns of *Gata1* and *Pu.1* before **(E)** and after **(F)** denoising. Cells are colored by pseudotime. Gene expression levels of *Gata1* and *Pu.1* before **(G)** and after denoising **(H)** visualized along the inferred DPT trajectory. **(I)** Heatmap showing the expression of regulatory genes (min-max scaled) for each cell. **Abbreviations**: Ery (erythrocytes), Mk (megakaryocytes), DC (dendritic cells), Baso (basophils), Mo (monocytes), Neu (neutrophils), Eos (eosinophils), Lymph (lymphoid cells), MEP (Megakaryocyte/erythrocyte progenitor), GMP (Granulocyte/macrophage progenitor). Adapted from Eraslan, Simon, et al. (2019) (CC-BY-4.0).

enrich our functional interpretation by allowing access to modules that are not covered by pre-defined gene sets.

Inferring data-driven gene modules requires robust and accurate estimation of gene-gene relationships (e.g. via correlation or predictive methods (Aibar et al. 2017)), which is hampered by noise in scRNA-seq datasets. To demonstrate this phenomenon and test whether denoising can improve the structure of gene modules and facilitate interpretation of gene-gene relationships, we analyzed the relationships between ten key regulatory genes in the Paul et al. (2015) blood development dataset. As previously described, this study presents the transcriptome landscape of blood cell differentiation through MEP and GMP branches (Figure 4.11A-B). Using well-studied transcription factors with key roles in the differentiation process (Krumsiek et al. 2011), we examined the effects of denoising on the correlation structure. Denoising with DCA improved the correlation structure where the anticorrelation between two factors that are known to inhibit each other (Orkin and Zon 2008), *Gata1* and *Pu.1,* increased (Pearson r: -0.318 and -0.613 without and with denoising, respectively, Figure 4.11C-F). Visualization of the expression of these genes along the differentiation trajectory shows that the mutually exclusive expression pattern becomes more visible and granular after denoising (Figure 4.11G-H). Moreover, the correlation of genes within the two gene modules that are active in the MEP and GMP branches (MEP module: *Gfi1, Fli1, Cebpa, Pu.1*; GMP module: *Tal1, Zfpm1, Gata1, Klf1*) also increased after denoising (Figure 4.11C-D). Although the correlation of genes within and between the modules is clearly visible when the gene expression is visualized in a heatmap(Figure 4.11I), this structure is weakly present in raw data due to the noise (Figure 4.11C). These results demonstrate that DCA enhances the correlation structure, which may facilitate the discovery of gene modules and hence improve the results of downstream analyses that rely on gene module inference.

Comparison with other denoising methods using bulk transcriptomics as ground truth

Comparisons of DCA with other denoising methods were performed by Lukas M. Simon and Maria Mircea in our publication (Eraslan, Simon, et al. 2019). First, we simulated mean-dependent single-cell noise and added this noise to the bulk RNA-seq time-course data of C. elegans development (Francesconi and Lehner 2013) similarly to the analysis in Dijk et al. (2018). We observed that the data denoised with DCA showed the highest correlation with the original bulk data compared to the other denoising methods MAGIC (Dijk et al. 2018), SAVER (Huang et al. 2018) and scImpute (W. V. Li and J. J. Li 2018). Second, we performed differential expression (DE) analysis on the definitive endoderm cells (DEC) bulk and single-cell data from Chu et al. (2016) and reported that DCA denoising improved the concordance between the bulk and single-cell DE results. Please see Figure 4 and Figure

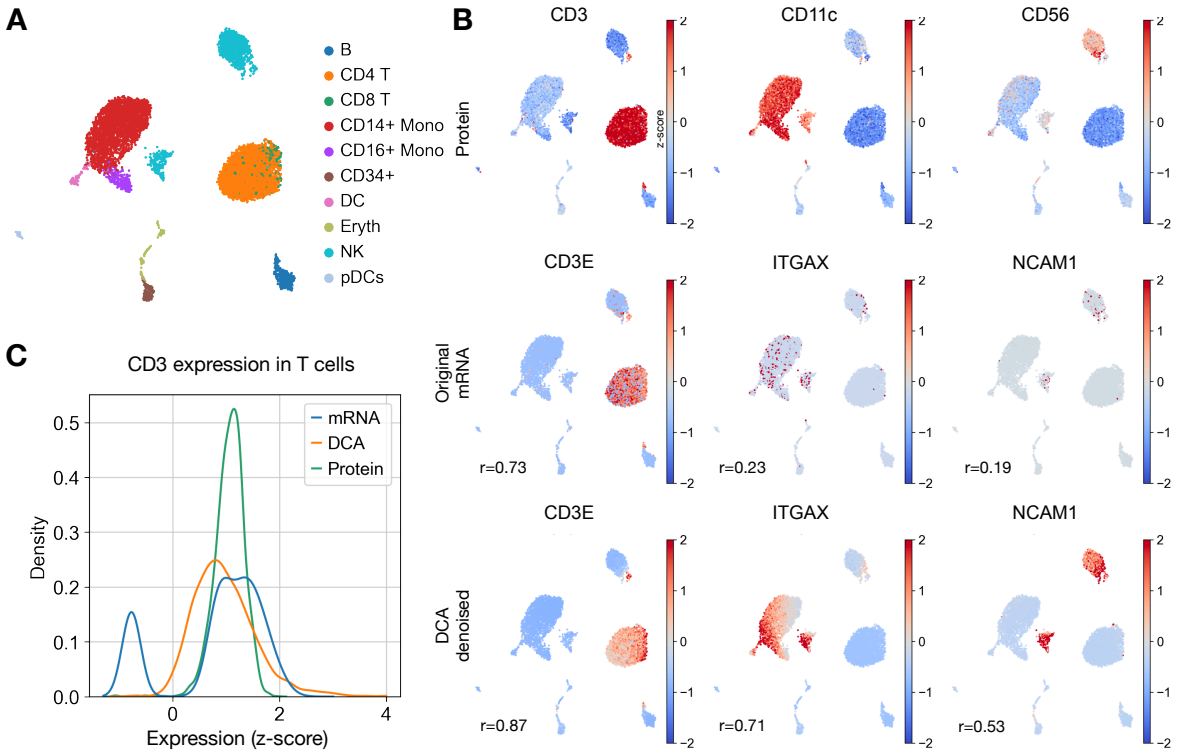


Figure 4.12: **A)** UMAP representation of cord blood mononuclear cells from Stoeckius et al. (2017). Cells are colored by the cell types. **B)** UMAP view of the same cells colored by the normalized expression of CD3, CD11c and CD56 proteins and corresponding mRNAs (CD3E, ITGAX and NCAM1). Rows show protein expression, RNA expression without denoising and RNA expression with denoising, respectively. Columns represent CD3, CD11c, and CD56 proteins and corresponding RNAs. Pearson correlation coefficients between protein and mRNA expression are given in the lower-left corner of each panel. **C)** Distribution of CD3 protein expression (green), mRNA expression (blue) and denoised mRNA expression (orange) in T cells are shown.

5 as well as the methods sections in Eraslan, Simon, et al. (2019) for more details of these applications.

Denoising increases protein and RNA co-expression

Multimodal single-cell profiling methods enable concomitant readouts such as protein and mRNA expression at cellular resolution. In addition to enriching the representation of cells, such approaches can also be used to obtain better estimates of the expression levels of genes with low mRNA expression, which are more difficult to detect with scRNA-seq due to dropout. Here, we sought to evaluate whether denoising with DCA accurately recovers expression by comparing DCA output with measured protein expression. For this analysis, we used CITE-seq, a single-cell profiling method that provides the full transcriptome readout as well as the expression of selected proteins where the authors profiled over 8,000 human cord blood mononuclear cells with a panel of 13 antibodies and identified major immune cell types

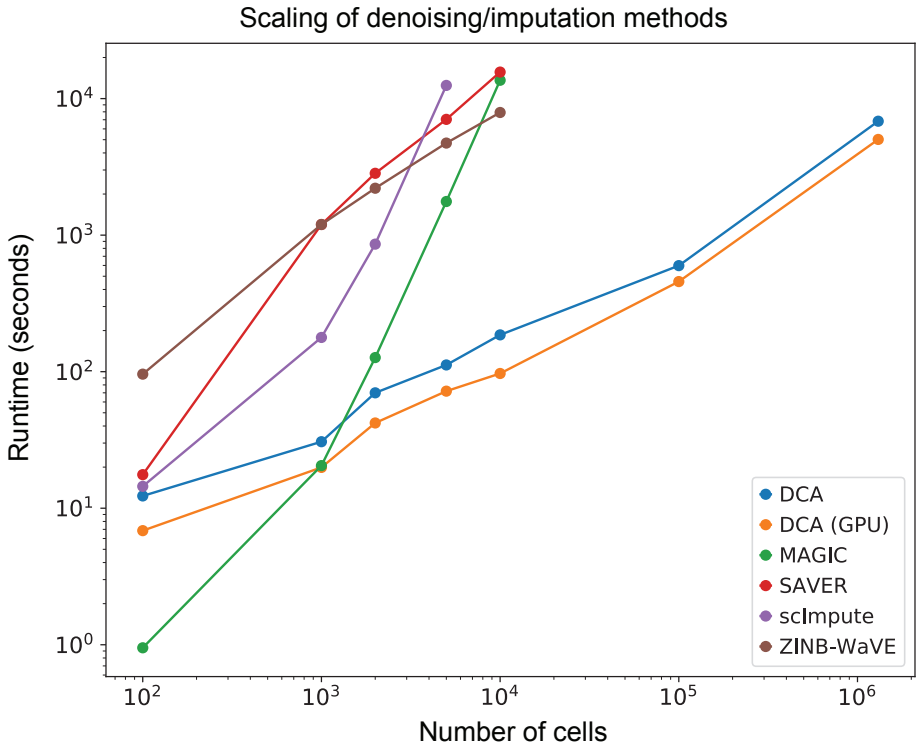


Figure 4.13: Runtimes for denoising of 1.3 million mouse brain cells as well as its random subsets of various sizes ranging from 100 to 100,000 (10X Genomics 2017). The colors denote different denoising/imputation methods. DCA(GPU) represents the DCA method run on the GPU. Adapted from Eraslan, Simon, et al. (2019) (CC-BY-4.0).

(Stoeckius et al. 2017) (Figure 4.12A). We used DCA with NB noise model to denoise the mRNA count data. After denoising, the expression of CD3, CD11c and CD56 proteins, which are known to be expressed, showed a higher correlation with mRNA expression of corresponding genes (Figure 4.12B). Although CD3 protein is expressed in 99.9% of annotated T cells, mRNA expression was detected in only 80% of T cells in the original data. After denoising, the fraction of T cells with CD3E expression increased to 99.9% (Figure 4.12C).

DCA is able to denoise massive datasets

Given that the number of profiled cells in each study in single-cell genomics is rapidly increasing, it is critical to develop performant scRNA-seq methods and implementations. Therefore we evaluated the scalability and runtime performance of five denoising/imputation methods (Dijk et al. 2018; Huang et al. 2018; W. V. Li and J. J. Li 2018; Risso et al. 2018) by applying these methods to the largest publicly available scRNA-seq dataset, which consists of 1.3 million mouse brain cells made available by 10X Genomics (`https://support.10xgenomics.com/single-cell-gene-expression/datasets/1.3.0/1M_neurons`). In addition to the full

dataset, we used the subsampled versions of it consisting of 100, 1,000, 2,000, 5,000, 10,000 and 100,000 cells. All count matrices were first subsetted to the top 1000 highly variable genes. Next, we denoised these matrices and measured the runtime. As expected, the runtime of DCA scaled linearly with the number of denoised cells. While the other methods did not even scale beyond 10,000 cells, DCA outperformed most methods and showed considerable performance advantage over other tested methods, especially in the regime of relatively large datasets.

4.2.3 Conclusion

The technical variation in scRNA-seq data is still one of the major challenges in single-cell data analysis. It has been shown in the recent studies that accounting for the sources of technical variation may potentially enhance downstream analysis (Brennecke et al. 2013; Buettner et al. 2015; Vallejos, Marioni, et al. 2015; B. Ding et al. 2015). Here we presented an unsupervised and scalable denoising approach based on neural networks, which is designed specifically for the sparse and count structure of scRNA-seq datasets. We further showed that the removal of technical variation via denoising improves various downstream analyses, including clustering, protein-mRNA expression concordance, gene module identification and pseudotime analysis. Moreover, we demonstrated that DCA is able to scale up to over a million cells.

Besides the simulations where the ground truth is known, denoising process is difficult to evaluate and/or validate. Using prior biological knowledge and additional data modalities, we here described a number of ways that can be used for the systematic evaluation of denoising methods in the future.

Notably, determining when denoising may improve downstream analysis in single-cell genomics can be difficult. For example, we observed that denoising resulted in increased gene-gene correlations, which is expected due to the design of the autoencoders (or any decomposition method in general). While increased correlations facilitated the detection of desired regulatory structure in our applications, it might also lead to overimputation, especially when the hyperparameter selection is performed poorly. To mitigate hyperparameter-related overimputation issues, we implemented a variety of regularization techniques such as L1 and L2 regularization as well as dropout, where neuron outputs are randomly dropped out, in order to avoid overfitting. This is especially important when training on datasets with small sample size. Moreover, DCA offers systematic hyperparameter search implementations to guide users about the right set of hyperparameters.

DCA is available as a standalone command line tool and as a Python package at `https://github.com/theislab/dca` and is able to work in harmony with established single-cell frameworks such as SCANPY (Wolf et al. 2018).

Chapter 5

Summary and outlook

Computational biology provides tools and techniques to characterize biological systems by processing, abstracting and interpreting experimental data. Like Søren Kierkegaard's famous quote, *"Life can only be understood backwards, but it must be lived forwards."* modeling allows us to go from the data, which is recorded as an imprint of a biological process through measurements, back to biology itself. Relatively recently, machine learning has become an essential tool in the modeling toolbox of computational biology by substantially contributing to this characterization with two major aspects: first, extrapolating beyond already known by enabling *in silico* experiments and second, providing biologically useful abstractions by unearthing relevant patterns while dimming irrelevant factors in the data. The first aspect refers to supervised learning, where we use trained machine learning models to make predictions in previously unseen settings. For example, training a model for predicting the event of a transcription factor binding given an unseen sequence can be considered a plain prediction task from a standard machine learning point-of-view. In contrast, in biology, this translates into an *in silico* ChIP-seq experiment without directly measuring the binding event. The second aspect refers to unsupervised learning tasks, such as visualization, feature selection and denoising, where we seek representations of the biological entities of interest that highlight their biologically meaningful, and more importantly, relevant features.

In this thesis, we focused on developing novel methods that leverage machine learning techniques, particularly neural networks, to improve the characterization of clinical phenotypes such as MS and MDD and molecular phenotypes such as gene expression of single cells. This goal can be summarized more specifically with the following research questions: 1) Can we generate functional hypotheses on clinical phenotypes by combining the genotype data from individuals with the variant effect predictions (VEPs) produced by machine learning models? 2) Can we recover the gene expression signal of individuals cells corrupted by the measurement process to obtain a representation that is more faithful to the underlying biology and improve downstream analysis? Now let us look at how the two aspects of machine learning mentioned

above are linked to these two research questions.

5.1 Towards functional hypotheses in GWAS

With the increasing availability of genotype datasets such as biobanks (Bycroft et al. 2018; Njølstad et al. 2019) and large GWAS cohorts (M. Liu et al. 2019; Jansen et al. 2019) together with the growing diversity of phenotypes (Buniello et al. 2019), human genetics is making tremendous progress towards the goal of pinpointing the genetic architecture of complex traits and diseases. Although playing a pivotal role in this progress, the primary result of GWAS is limited to a set of loci with genetic variation that is significantly linked to the phenotype of interest. Therefore functional interpretation and prioritization of resulting loci and variants bear great importance.

Various variant prioritization and annotation tools for investigating both coding and non-coding variants were proposed in the literature. The functions that these tools perform can be grouped into three categories. The first category is the investigation of the consequences of variants on the coding sequence, such as amino acid changes, events like stop gained, missense, stop lost, frameshift, and how the protein structure and function are altered. The second category is the integration of a multitude of resources such as gene and transcript annotations with intron and exon information and overlapping these with the given variants. The last category includes the examination of the regulatory effects of variants based on the positional overlap of given variants with comprehensive epigenetic and regulatory elements such as promoters, enhancers, and TF binding sites. The review of these tools is given in section 3.1 and published as a part of the Bioinformatics in Psychiatric Genetics chapter of the book titled Psychiatric Genetics: A Primer for Clinical and Basic Scientists (Schulze and F. McMahon 2018).

The majority of significant risk loci identified via GWAS harbor non-coding variants with unknown regulatory effects (Visscher et al. 2017) which makes the functional interpretation a particularly challenging task. In order to address this challenge, we developed two approaches, Misina and DeepWAS, which are given in section 3.2 and 3.3, respectively.

In our work, Misina (section 3.2), we focused on the link between GWAS risk variants and microRNAs because microRNAs are a key component of the regulome and they are reported to play an essential role in disease mechanisms through miRNA-mediated dysregulation (Chin et al. 2008; Esteller 2011). In this approach, we developed a simple variant prioritization tool for identifying variants that cause dysregulation of the microRNA target genes by disrupting or enhancing the miRNA binding sites either directly or via LD proxies. We further integrated experimental data of miRNA expression as well as the expression of the target genes. Using Misina, we investigated miRNA-mediated effects of risk variants associated with Alzheimer's

disease. One risk variant, rs6859, reside where miRNA hsa-miR199a-5p binds the target gene PVRL2 (NECTIN2) and potentially alter the binding affinity and hence lead to dysregulation of PVRL2. We also further characterized the relationship between the SNP, miRNA and the target gene with more experimental evidence. Note that it has been reported that the miRNA-mediated dysregulation of PVRL2 by rs6859 variant is a potential risk factor for Alzheimer's disease (X. Zhou et al. 2019). In summary, our approach is a novel way to interrogate the functional role of risk variants by integrating prior knowledge.

An alternative method for interpreting GWAS variants is to complement these results with the results from the experiments where the other effects of genetic variation are studied, such as the variant-gene associations obtained in eQTL studies in a post hoc manner. For example, a locus that is strongly associated with obesity is located in the first intron of the FTO gene (Frayling et al. 2007). Later, it has been shown that the obesity-associated variant is actually an eQTL variant regulating the expression of the transcription factor IRX3, which is megabases away from the FTO gene, through the enhancers located in this intronic region (Smemo et al. 2014). It has been further validated that deletion of Irx3 gene in developing adipocytes in mice protects against obesity (Smemo et al. 2014). Therefore, having experimental data that might suggest the function of the variant can greatly improve our understanding of the phenotype. Another valuable experimental approach for functional validation is the binding QTLs (bQTLs), where the genetic variations that alter the binding of transcription factors are identified via pooled ChIP-seq experiments (Tehranchi et al. 2016). Although such experiments elegantly highlight the potential regulatory role of variants, they are expensive and scale to only a few transcription factors.

As an alternative to the experimental validation, J. Zhou and Troyanskaya (2015) proposed a radically different approach, DeepSEA, for variant effect predictions using the idea of *in silico* perturbations, exploiting the pattern recognition capabilities of deep neural sequence models. DeepSEA relies on a classifier that predicts the binary binding event of multiple transcription factors (and other chromatin features like DNase hypersensitivity and histone modifications) from a given 1kb sequence. Trained using publicly available ChIP-seq and DNase-seq datasets, this classifier is then used for estimating the effects of variation in the input sequence. Compared to the experimental validation like binding QTLs, DeepSEA variant effect predictions serve as *in silico* perturbation experiments for quantifying the regulatory effect of a given variant.

In our work, DeepWAS (section 3.3, published in Arloth and Eraslan et al., 2020), we proposed a new approach where putative regulatory variants are jointly tested for genotype-phenotype associations rather than post hoc functional analyses conducted after GWAS. First, the DeepWAS workflow predicts regulatory effects of all measured variants and their LD proxies using the *in silico* experiments performed by the pre-trained DeepSEA model (J. Zhou

and Troyanskaya 2015), which yields potential binding-QTLs, histone-QTLs and DNase-QTLs similar to their experimental equivalents (Tehranchi et al. 2016). Second, we group the variants that potentially act through the same regulatory mechanism (e.g. variants altering the binding of MafK transcription factor in cell line K562). Last, we test genotype-phenotype associations within each group using multivariate Lasso models with stability selection (Meinshausen and Bühlmann 2010), allowing us to control for error rates of false discoveries. We applied DeepWAS to three phenotypes, multiple sclerosis (MS), major depressive disorder (MDD) and height, to uncover variants with potential regulatory effects to the phenotypes using the genotype data from relatively small cohorts (MS: 15,283, MDD: 3,514 and height: 5,866 individuals). We identified several variants (MS: 53, MDD: 61, height: 43) along with the transcription factors that might be affected by these variants, as well as the cell lines or tissues where the variants might be acting.

In the comparisons of our results with the results of the identical phenotypes from much larger GWASes (115k–807k individuals), we found that there are both novel risk variants that are either untested or sub-threshold in GWAS and the variants that are common. Because the DeepWAS approach prioritizes variants with likely regulatory effects, even if their effect sizes are relatively small, we expected to find that many DeepWAS hits were sub-threshold in GWAS. We further highlighted four major findings involving non-coding variants that collectively contribute to the pathology of MS and MDD by affecting different mechanisms.

Our first key finding was a group of variants that potentially affect the binding of a family of TFs called small Maf proteins (sMafs). Variants rs62420820, rs12768537 and rs137969 likely alter the binding of not only MafF and MafK TFs, which are sMafs, but also Bach1 and NF-E2, which are known to form heterodimers with sMafs (Katsuoka and Yamamoto 2016). We hypothesize that these variants modulate the regulation of the targets of either sMaf homodimers (e.g. MafF-MafF or MafF-MafK) as well as the heterodimers (e.g. MafF-Bach1 and MafK-NF-E2). Notably, sMafs are known to play a critical role in CNS (Katsuoka, Motohashi, et al. 2003). Furthermore, GRAP2, an MS susceptibility gene identified by Berge et al. (2019), is an eQTL gene for rs137969 (Võsa et al. 2018), which further supports our hypothesis. Although MafG is not a part of the DeepSEA predictive model, and hence DeepWAS, we hypothesize that Maf family TFs likely play an important role in MS since the variants identified by DeepWAS are potentially affecting the binding of various components of TF protein complexes involving Mafs.

Our second highlight, rs1985372, is an intronic variant located in CLEC16A which is a significant locus previously identified in the largest MS GWAS (International Multiple Sclerosis Genetics Consortium 2019). This DeepWAS hit potentially alters the binding of multiple TFs, including GABP, GATA-1, GATA-2, p300, STAT1, STAT2, STAT5A, and TBLR1. rs1985372 is also significantly associated with CLEC16A expression in various tissues by GTEx eQTL data

(see the GTEx portal `https://gtexportal.org/home/snp/rs1985372`). DeepWAS suggests a testable hypothesis that rs1985372 plays a role in MS pathology by altering the binding of multiple TFs and causing dysregulation of CLEC16A.

Third, we focused on rs175714, an intergenic DeepWAS hit for MS with potentially severe regulatory effects on 29 chromatin features, including 18 TFs, chromatin accessibility, and 10 histone marks in total of 116 cell lines. MAZ, which is among the affected TFs, is a significant locus in the cohort-matched GWAS. Another MS DeepWAS hit rs11000015, which also impacts the binding of the MAZ TF together with rs175714, is an eQTL for the PSAP gene in GTEx data. PSAP gene is a critical component of the sphingolipid pathways which were previously linked to MS (O'Brien and Kishimoto 1991). Furthermore, PSAP was previously associated with MS through a genome-wide expression study (Kemppinen et al. 2011). Therefore, DeepWAS can also nominate key regulators of diseases by identifying variants with severe regulatory effects as well as other variants exacerbating the regulatory effects by targeting the same TFs.

Our last highlight was an intergenic SNP, rs7839671, in the analysis of MDD. This variant potentially affects the binding of a key TF, MEF2C, which was already reported to be an important risk gene in the PGC GWAS for MDD (Howard et al. 2019). Furthermore, MEF2C is a member in the MEF2 TF family which was previously identified as the master regulator of developmental metaplasticity (S. X. Chen et al. 2012) and also linked to activity-dependent dendritic spine growth and suppression of memory growth (Barbosa et al. 2008).

In silico experiments provide a powerful means to enrich the characterization of variants when they are leveraged in post hoc functional analysis in GWAS. In our approach, we showed that moving these predictions to the center of genotype-phenotype associations, we not only identify sub-threshold variants that likely play an important role in complex diseases through regulatory effects but also generate hypotheses regarding how potential dysregulations might occur, which addresses our research question on generating functional hypotheses on complex traits and diseases.

5.2 Enhancing scRNA-seq characterization with unsupervised machine learning

We leveraged supervised machine learning models that can accurately predict how sequence variation can affect molecular events like TF binding or chromatin accessibility for improving variant prioritization. Unlike this task that requires ground truth knowledge, unsupervised machine learning focuses on enhancing the data representation for facilitating data exploration and characterization. For example, t-distributed stochastic neighbor embedding (tSNE) (Maaten and G. Hinton 2008) is an unsupervised machine learning method for embedding

high-dimensional data into two-dimensional space while preserving the local structure of the data so that the points that are close to each other in the high-dimensional space are embedded close to each other in the two-dimensional space. In many cases, this representation is preferable over the high-dimensional representation since it enables exploring both the oddities and the regularities of the data visually. This massive transformation of the data is done with the hope that the biologically relevant patterns are magnified while irrelevant sources of variation in the data are hidden.

In our next method, DCA (chapter 4, published in Eraslan, Simon, et al. 2019), we utilized an unsupervised machine learning technique, namely autoencoders, to address one of the fundamental challenges in single-cell RNA-seq (scRNA-seq), which is the technical variation introduced by substantial measurement noise. A major component of this noise arises from the amplification of low amount of the input RNA and low RNA capture rate, and leads to the failure of detection of expressed genes, a phenomenon called dropout, which results in "false zeros" in addition to the true zeros of "non-expression". Overall, the measurement noise might affect the biological interpretation of the data and must be accounted for (Brennecke et al. 2013; Buettner et al. 2015; Vallejos, Marioni, et al. 2015; B. Ding et al. 2015).

Following the brief introduction of two denoising/imputation methods, MAGIC (Dijk et al. 2018) and SAVER (Huang et al. 2018), in section 4.1, we introduced a scalable and robust machine learning method that is tailored to the structure of the scRNA-seq data for recovering the gene expression signal, improving various downstream analyses and enhancing the characterization of the biological sample of interest in section 4.2.

First, we discussed the appropriate noise models for the scRNA-seq data analysis and how distribution assumptions might affect the quality of the downstream analysis, which are highly debated topics in the field of single-cell genomics (Grün et al. 2014). We fitted negative binomial (NB) and zero-inflated negative binomial (ZINB) models to one simulated and four real datasets (one read-based (Chu et al. 2016) and three UMI-based (Stoeckius et al. 2017; Zheng et al. 2017; Paul et al. 2015)) and compared the models using likelihood ratio test. In agreement with previously reported results (Wenan Chen et al. 2018), we concluded that, unlike the read-based count data, counts produced by the UMI-based scRNA-seq protocols do not show zero inflation. The negative binomial noise model is preferable over the zero-inflated negative binomial for UMI-based technologies.

Second, using the simulated scRNA-seq datasets, we demonstrated how severe dropout noise can weaken the cluster structure in the data and obscure the cell type identities. Moreover, we showed that the NB loss function, which takes the count structure of the data into account, can recover the cluster structure impaired by the substantial noise better than the loss function with normal distribution assumptions.

Third, we examined whether our method can capture the population structure in real

datasets and can perform denoising in a cell type-specific manner since the datasets generated from tissues typically exhibit high biological complexity and cellular heterogeneity. Using 68,579 peripheral blood mononuclear cells (Zheng et al. 2017), we restricted the latent space of DCA to only two dimensions and visualized the latent variables our model extracted from the data in order to uncover the determinants of the denoising process. This two-dimensional view of the data overlapped well with the known cell population structure of the data. Since the latent variables represent the compressed view and the source of information used for denoising, the denoising process was highly cell type-specific. We also investigated the ability of DCA to capture continuous phenotypes, for example, the differentiation trajectory of myeloid progenitors. When applied to this dataset, the data manifold captured by DCA highly correlated with the pseudotime estimated by the diffusion pseudotime (DPT) method (Haghverdi, Büttner, et al. 2016) (Pearson's rho: 0.95), indicating that DCA captures biologically meaningful features.

Fourth, we utilized the data produced by Stoeckius et al. (2017) using the CITE-seq protocol, which enables the measurements of both mRNA and protein expressions from the same cells in order to quantify the mRNA-protein correlations and used this as a metric to evaluate the denoising performance. We showed that DCA increased the concordance between mRNA and protein expression.

Fifth, we examined the effects of denoising on the correlation structure of well-studied key regulators of the blood development using the myeloid progenitor differentiation scRNA-seq data from Paul et al. (2015). Denoised data exhibited higher pairwise correlations for the genes that are part of the same module such as Cebpa-Pu.1 and Gata1-Tal1, whereas the genes that are part of mutually exclusive gene programs such as Pu.1-Gata1 (Nerlov et al. 2000) showed higher anticorrelation (correlation without and with denoising were -0.318 and -0.439, respectively). Although these examples might seem anecdotal, correlation-based inference of regulatory networks and gene modules is a common task in single-cell genomics (Aibar et al. 2017; Smillie et al. 2019). Therefore inferring the gene modules without recovering the correlation structure that is lost due to the measurement noise might lead to underestimated gene-gene relationships and disconnected network structure, which might further propagate to the findings built on these networks.

Finally, we compared the runtime performance of our approach with four other denoising/imputation methods, namely MAGIC (Dijk et al. 2018), SAVER (Huang et al. 2018), scImpute (W. V. Li and J. J. Li 2018) and ZINB-WaVE (Risso et al. 2018). In this analysis, we used the downsampled versions of the largest single-cell RNA-seq data, which has 1.3M mouse brain cells (10X Genomics 2017), in addition to the full dataset. The runtime of DCA scaled linearly with the number of cells, while other methods did not scale beyond 10,000 cells. We did not encounter any issues when we denoised the entire dataset with 1.3M cells using

DCA.

5.3 Outlook

In this thesis, we utilized the DeepSEA model as the basis for the predictions of molecular modalities such as TF binding and chromatin accessibility. Recently, alternative approaches have been proposed and substantially expanded the repertoire of biological sequence modeling in different directions. For example, ExPecto (J. Zhou, C. L. Theesfeld, et al. 2018) incorporated gene expression into the variant prioritization and estimated the effects of sequence variation on gene expression. This model can be viewed as *in silico* eQTL experiments. Basenji (Kelley, Reshef, et al. 2018) applied dilated convolutions to biological sequences to predict histone modifications, chromatin accessibility and gene expression from 131kb-long DNA sequences to identify the promoters and distal regulatory elements in the genome. BPNet (Avsec et al. 2019) is another example which predicts TF binding events at base-resolution using ChIP-nexus profiles (Q. He et al. 2015) for a given DNA sequence, instead of making binary binding predictions (i.e. binding/not binding). We expect that these next-generation predictive sequence models will be widely used in the future in two forms. First, as a discovery tool, predictive models are able to extrapolate beyond already characterized settings to those where measurements are not available. For example, molecular patterns of gene expression, TF binding and chromatin accessibility in tissues and cell types where the measured data is scarce can be predicted, which might give us the ability to rapidly hypothesize about the underlying biology. Second, such models can be used in approaches like DeepWAS where sequence models play a central role in investigating and interpreting genotype–phenotype associations.

In parallel to the developments in sequence-based predictive models and variant prioritization approaches in computational genomics, new machine learning techniques for sequence models, language models and word embedding methods have been proposed in the fields of natural language processing and deep learning (Vaswani et al. 2017; Peters et al. 2018; Devlin et al. 2018). We envision that these new techniques will find applications in our field and enable better *in silico* experiments which in turn will contribute to our understanding of various complex phenotypes, biological mechanisms and regulatory codes of the genome.

Machine learning is becoming an indispensable tool in single-cell genomics. In Eraslan, Simon, et al. (2019), we proposed one of the first deep learning-based representation learning applications in single-cell which found many applications (Schiller et al. 2019; Y. Deng et al. 2019; Arisdakessian et al. 2019; Jingshu Wang et al. 2019; Hafemeister and Satija 2019). Together with scVI (Lopez et al. 2018), our work also showed that the downstream tasks can be performed well in the latent space for the first time which was followed up by other studies. For example, Lotfollahi et al. (2019) proposed an *in silico* perturbation approach where latent

space vector arithmetics were employed to predict single-cell perturbation responses. This method also serves as a discovery tool for exploring molecular modalities beyond known cell types, studies and species. With the emergence of highly multiplexed combinatorial single-cell drug screens and perturbation experiments, we expect more machine learning approaches will be developed to predict the effects of drugs and genetic perturbations in the future. These experiments will not only enhance our understanding of causal effects in biology and discern correlation and causation, but also give rise to methodological improvements in the domain of causal inference and indirectly help other fields.

Moreover, predictive models are exploited for addressing challenges in single-cell immunobiology. Fischer et al. (2019) used deep learning to predict the antigen specificity of single T cells using multimodal datasets where coupled transcriptome, T-cell receptor (TCR) sequence and surface proteins of single-cells are available. Several unsupervised approaches that aim to improve data representation and to facilitate biological discoveries in single-cell genomics are proposed (Amodio et al. 2018; Cho et al. 2018; Jiarui Ding, Condon, et al. 2018; Lopez et al. 2018). We expect machine learning to be used for a wider range of tasks, both supervised and unsupervised, in single-cell genomics in the future.

Today not only the volume and but also the diversity and the complexity of biological datasets are increasing as new experimental techniques are introduced. For example, high-throughput spatial protocols (Xiao Wang et al. 2018; Eng et al. 2019), single-cell perturbation experiments with different types of readouts (Dixit et al. 2016; Datlinger et al. 2017; Mimitou et al. 2019), simultaneous measurements of multiple modalities in single cells (J. Cao, Cusanovich, et al. 2018; Reyes et al. 2019; L. Liu et al. 2019; S. J. Clark et al. 2018) are becoming a standard practice. As these techniques mature and the datasets grow in quality and quantity, we will reach a better coverage of genomic, epigenomic and genetic landscapes, which in turn will improve the quality and complexity of the training datasets for future machine learning methods. Combined with the increasing computational power of new hardware and the predictive power of new scalable models, such training datasets have the potential to revolutionize not only molecular biology but also personalized medicine and drug development.

Bibliography

1000 Genomes Project Consortium (2015). "A global reference for human genetic variation". In: *Nature* 526.7571, pp. 68–74.

10X Genomics (2017). *1.3 Million Brain Cells from E18 Mice*. `https://support.10xgenomics.com/single-cell-gene-expression/datasets/1.3.0/1M_neurons`. Accessed: 2017-11-21.

Abadi, Martín, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, et al. (2015). *TensorFlow: Large-scale machine learning on heterogeneous systems*.

Adzhubei, Ivan A, Steffen Schmidt, Leonid Peshkin, Vasily E Ramensky, Anna Gerasimova, Peer Bork, Alexey S Kondrashov, and Shamil R Sunyaev (Apr. 2010). "A method and server for predicting damaging missense mutations". en. In: *Nat. Methods* 7.4, pp. 248–249.

Agarwal, Vikram, George W Bell, Jin-Wu Nam, and David P Bartel (2015). "Predicting effective microRNA target sites in mammalian mRNAs". In: *Elife* 4, e05005.

Aibar, Sara, Carmen Bravo González-Blas, Thomas Moerman, Hana Imrichova, Gert Hulselmans, Florian Rambow, Jean-Christophe Marine, Pierre Geurts, Jan Aerts, Joost van den Oord, et al. (2017). "SCENIC: single-cell regulatory network inference and clustering". In: *Nature methods* 14.11, pp. 1083–1086.

Alipanahi, Babak, Andrew Delong, Matthew T Weirauch, and Brendan J Frey (Aug. 2015). "Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning". en. In: *Nat. Biotechnol.* 33.8, pp. 831–838.

Allen, Hana Lango, Karol Estrada, Guillaume Lettre, Sonja I Berndt, Michael N Weedon, Fernando Rivadeneira, Cristen J Willer, Anne U Jackson, Sailaja Vedantam, Soumya Raychaudhuri, et al. (2010). "Hundreds of variants clustered in genomic loci and biological pathways affect human height". In: *Nature* 467.7317, p. 832.

Alpaydin, Ethem (2009). *Introduction to machine learning*. MIT press.

Altshuler, David, Peter Donnelly, International HapMap Consortium, et al. (2005). "A haplotype map of the human genome". In: *Nature* 437.7063, p. 1299.

Amodio, Matthew, David Van Dijk, Krishnan Srinivasan, William S Chen, Hussein Mohsen, Kevin R Moon, Allison Campbell, Yujiao Zhao, Xiaomei Wang, Manjunatha Venkataswamy,

et al. (2018). “Exploring Single-Cell Data with Deep Multitasking Neural Networks”. In: *bioRxiv*.

Andersson, Robin, Claudia Gebhard, Irene Miguel-Escalada, Ilka Hoof, Jette Bornholdt, Mette Boyd, Yun Chen, Xiaobei Zhao, Christian Schmidl, Takahiro Suzuki, Evgenia Ntini, Erik Arner, Eivind Valen, Kang Li, Lucia Schwarzfischer, Dagmar Glatz, Johanna Raithel, Berit Lilje, Nicolas Rapin, Frederik Otzen Bagger, Mette Jørgensen, Peter Refsing Andersen, Nicolas Bertin, Owen Rackham, A Maxwell Burroughs, J Kenneth Baillie, Yuri Ishizu, Yuri Shimizu, Erina Furuhata, Shiori Maeda, Yutaka Negishi, Christopher J Mungall, Terrence F Meehan, Timo Lassmann, Masayoshi Itoh, Hideya Kawaji, Naoto Kondo, Jun Kawai, Andreas Lennartsson, Carsten O Daub, Peter Heutink, David A Hume, Torben Heick Jensen, Harukazu Suzuki, Yoshihide Hayashizaki, Ferenc Müller, Alistair R R Forrest, Piero Carninci, Michael Rehli, and Albin Sandelin (Mar. 2014). “An atlas of active enhancers across human cell types and tissues”. en. In: *Nature* 507.7493, p. 455.

Andlauer, Till FM, Dorothea Buck, Gisela Antony, Antonios Bayas, Lukas Bechmann, Achim Berthele, Andrew Chan, Christiane Gasperi, Ralf Gold, Christiane Graetz, et al. (2016). “Novel multiple sclerosis susceptibility loci implicated in epigenetic regulation”. In: *Science advances* 2.6, e1501678.

Arisdakessian, Cédric, Olivier Poirion, Breck Yunits, Xun Zhu, and Lana X Garmire (2019). “DeepImpute: an accurate, fast, and scalable deep neural network method to impute single-cell RNA-seq data”. In: *Genome biology* 20.1, pp. 1–14.

Arloth, Janine, Gökcen Eraslan, Till FM Andlauer, Jade Martins, Stella Iurato, Brigitte Kühnel, Melanie Waldenberger, Josef Frank, Ralf Gold, Bernhard Hemmer, et al. (2020). “DeepWAS: Multivariate genotype-phenotype associations by directly integrating regulatory information using deep learning”. In: *PLoS computational biology* 16.2, e1007616.

Arnold, Matthias, Johannes Raffler, Arne Pfeufer, Karsten Suhre, and Gabi Kastenmüller (2014). “SNiPA: an interactive, genetic variant-centered annotation browser”. In: *Bioinformatics*, btu779.

Aune, Dagfinn, Ana Rita Vieira, Doris Sau Man Chan, Deborah A Navarro Rosenblatt, Rui Vieira, Darren C Greenwood, Janet E Cade, Victoria J Burley, and Teresa Norat (2012). “Height and pancreatic cancer risk: a systematic review and meta-analysis of cohort studies”. In: *Cancer Causes & Control* 23.8, pp. 1213–1222.

Avsec, Žiga, Melanie Weilert, Avanti Shrikumar, Amr Alexandari, Sabrina Krueger, Khyati Dalal, Robin Fropf, Charles McAnany, Julien Gagneur, Anshul Kundaje, et al. (2019). “Deep learning at base-resolution reveals motif syntax of the cis-regulatory code”. In: *BioRxiv*, p. 737981.

Azizi, Elham, Sandhya Prabhakaran, Ambrose Carr, and Dana Pe'er (2017). "Bayesian Inference for Single-cell Clustering and Imputing". In: *Genomics and Computational Biology* 3.1, p. 46.

Bach, Karsten, Sara Pensa, Marta Grzelak, James Hadfield, David J Adams, John C Marioni, and Walid T Khaled (2017). "Differentiation dynamics of mammary epithelial cells revealed by single-cell RNA sequencing". In: *Nature communications* 8.1, p. 2128.

Bacher, Rhonda, Li-Fang Chu, Ning Leng, Audrey P Gasch, James A Thomson, Ron M Stewart, Michael Newton, and Christina Kendziorski (2017). "SCnorm: robust normalization of single-cell RNA-seq data". In: *Nature methods* 14.6, p. 584.

Balakrishnan, Ilango, Xiaodong Yang, Joseph Brown, Aravind Ramakrishnan, Beverly Torok-Storb, Peter Kabos, Jay R Hesselberth, and Manoj M Pillai (2014). "Genome-wide analysis of miRNA-mRNA interactions in marrow stromal cells". In: *Stem cells* 32.3, pp. 662–673.

Baranzini, Sergio E and Jorge R Oksenberg (2017). "The genetics of multiple sclerosis: from 0 to 200 in 50 years". In: *Trends in Genetics*.

Barbosa, Ana C, Mi-Sung Kim, Mert Ertunc, Megumi Adachi, Erika D Nelson, John McAnally, James A Richardson, Ege T Kavalali, Lisa M Monteggia, Rhonda Bassel-Duby, and Eric N Olson (July 2008). "MEF2C, a transcription factor that facilitates learning and memory by negative regulation of synapse numbers and function". en. In: *Proc. Natl. Acad. Sci. U. S. A.* 105.27, pp. 9391–9396.

Bartel, David P (2009). "MicroRNAs: target recognition and regulatory functions". In: *Cell* 136.2, pp. 215–233.

Berge, Tone, Anna Eriksson, Ina Skaara Brorson, Einar August Høgestøl, Pål Berg-Hansen, Anne Døskeland, Olav Mjaavatten, Steffan Daniel Bos, Hanne F Harbo, and Frode Berven (2019). "Quantitative proteomic analyses of CD4+ and CD8+ T cells reveal differentially expressed proteins in multiple sclerosis patients and healthy controls". In: *Clinical proteomics* 16.1, p. 19.

Bergen, V, M Lange, S Peidli, FA Wolf, and FJ Theis (2019). "Generalizing RNA velocity to transient cell states through dynamical modeling. bioRxiv 820936". In:

Bergstra, James, Brent Komer, Chris Eliasmith, Dan Yamins, and David D Cox (2015). "Hyperopt: a Python library for model selection and hyperparameter optimization". In: *Comput. Sci. Discov.* 8.1, p. 014008.

Bersanelli, Matteo, Ettore Mosca, Daniel Remondini, Enrico Giampieri, Claudia Sala, Gastone Castellani, and Luciano Milanesi (2016). "Methods for the integration of multi-omics data: mathematical aspects". In: *BMC bioinformatics* 17.2, S15.

Betel, Doron, Anjali Koppal, Phaedra Agius, Chris Sander, and Christina Leslie (2010). "Comprehensive modeling of microRNA targets predicts functional non-conserved and non-canonical sites". In: *Genome biology* 11.8, R90.

Bishop, Christopher M (1995). *Neural networks for pattern recognition*. Oxford university press.

Box, George EP (1976). "Science and statistics". In: *Journal of the American Statistical Association* 71.356, pp. 791–799.

Boyle, Alan P, Eurie L Hong, Manoj Hariharan, Yong Cheng, Marc A Schaub, Maya Kasowski, Konrad J Karczewski, Julie Park, Benjamin C Hitz, Shuai Weng, J Michael Cherry, and Michael Snyder (Sept. 2012). "Annotation of functional variation in personal genomes using RegulomeDB". en. In: *Genome Res.* 22.9, pp. 1790–1797.

Brennecke, Philip, Simon Anders, Jong Kyoung Kim, Aleksandra A Kołodziejczyk, Xiuwei Zhang, Valentina Proserpio, Bianka Baying, Vladimir Benes, Sarah A Teichmann, John C Marioni, and Marcus G Heisler (Nov. 2013). "Accounting for technical noise in single-cell RNA-seq experiments". en. In: *Nat. Methods* 10.11, pp. 1093–1095.

Buades, Antoni, Bartomeu Coll, and Jean-Michel Morel (2005). "A review of image denoising algorithms, with a new one". In: *Multiscale Modeling & Simulation* 4.2, pp. 490–530.

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". In: *Nature* 523.7561, p. 486.

Buettner, Florian, Kedar N Natarajan, F Paolo Casale, Valentina Proserpio, Antonio Scialdone, Fabian J Theis, Sarah A Teichmann, John C Marioni, and Oliver Stegle (Feb. 2015). "Computational analysis of cell-to-cell heterogeneity in single-cell RNA-sequencing data reveals hidden subpopulations of cells". en. In: *Nat. Biotechnol.* 33.2, pp. 155–160.

Buniello, Annalisa, Jacqueline A L MacArthur, Maria Cerezo, Laura W Harris, James Hayhurst, Cinzia Malangone, Aoife McMahon, Joannella Morales, Edward Mountjoy, Elliot Sollis, et al. (2019). "The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019". In: *Nucleic acids research* 47.D1, pp. D1005–D1012.

Bycroft, Clare, Colin Freeman, Desislava Petkova, Gavin Band, Lloyd T Elliott, Kevin Sharp, Allan Motyer, Damjan Vukcevic, Olivier Delaneau, Jared O'Connell, et al. (2018). "The UK Biobank resource with deep phenotyping and genomic data". In: *Nature* 562.7726, pp. 203–209.

Calo, Eliezer and Joanna Wysocka (2013). "Modification of enhancer chromatin: what, how, and why?" In: *Molecular cell* 49.5, pp. 825–837.

Cao, Junyue, Darren A Cusanovich, Vijay Ramani, Delasa Aghamirzaie, Hannah A Pliner, Andrew J Hill, Riza M Daza, Jose L McFaline-Figueroa, Jonathan S Packer, Lena Christiansen, et al. (2018). "Joint profiling of chromatin accessibility and gene expression in thousands of single cells". In: *Science* 361.6409, pp. 1380–1385.

Cao, Junyue, Jonathan S Packer, Vijay Ramani, Darren A Cusanovich, Chau Huynh, Riza Daza, Xiaojie Qiu, Choli Lee, Scott N Furlan, Frank J Steemers, et al. (2017). "Comprehensive single-cell transcriptional profiling of a multicellular organism". In: *Science* 357.6352, pp. 661–667.

Carvalho, Benilton S, Thomas A Louis, and Rafael A Irizarry (2009). "Quantifying uncertainty in genotype calls". In: *Bioinformatics* 26.2, pp. 242–249.

Casper, Jonathan, Ann S Zweig, Chris Villarreal, Cath Tyner, Matthew L Speir, Kate R Rosenbloom, Brian J Raney, Christopher M Lee, Brian T Lee, Donna Karolchik, et al. (2018). "The UCSC genome browser database: 2018 update". In: *Nucleic acids research* 46.D1, pp. D762–D769.

Chang, Winston, Joe Cheng, JJ Allaire, Yihui Xie, and Jonathan McPherson (2015). *shiny: Web Application Framework for R*. R package version 0.12.2. URL: `http://CRAN.R-project.org/package=shiny`.

Chen, Geng, Baitang Ning, and Tieliu Shi (2019). "Single-cell RNA-seq technologies and related computational data analysis". In: *Frontiers in genetics* 10, p. 317.

Chen, Simon Xuan, Angus Cherry, Parisa Karimi Tari, Kaspar Podgorski, Yue Kay Kali Kwong, and Kurt Haas (Sept. 2012). "The transcription factor MEF2 directs developmental visually driven functional and structural metaplasticity". en. In: *Cell* 151.1, pp. 41–55.

Chen, Wenan, Yan Li, John Easton, David Finkelstein, Gang Wu, and Xiang Chen (May 2018). "UMI-count modeling and differential expression analysis for single-cell RNA sequencing". en. In: *Genome Biol.* 19.1, p. 70.

Chi, Sung Wook, Julie B Zang, Aldo Mele, and Robert B Darnell (2009). "Argonaute HITS-CLIP decodes microRNA–mRNA interaction maps". In: *Nature* 460.7254, pp. 479–486.

Chin, Lena J, Elena Ratner, Shuguang Leng, Rihong Zhai, Sunitha Nallur, et al. (2008). "A SNP in a let-7 microRNA complementary site in the KRAS 3' untranslated region increases non–small cell lung cancer risk". In: *Cancer research* 68.20, pp. 8535–8540.

Cho, Hyunghoon, Bonnie Berger, and Jian Peng (2018). "Generalizable and scalable visualization of single-cell data using neural networks". In: *Cell systems* 7.2, pp. 185–191.

Chu, Li-Fang, Ning Leng, Jue Zhang, Zhonggang Hou, Daniel Mamott, David T Vereide, Jeea Choi, Christina Kendziorski, Ron Stewart, and James A Thomson (Aug. 2016). "Single-cell RNA-seq reveals novel regulators of human embryonic stem cell differentiation to definitive endoderm". en. In: *Genome Biol.* 17.1, p. 173.

Clark, Stephen J, Ricard Argelaguet, Chantriolnt-Andreas Kapourani, Thomas M Stubbs, Heather J Lee, Celia Alda-Catalinas, Felix Krueger, Guido Sanguinetti, Gavin Kelsey, John C Marioni, et al. (2018). "scNMT-seq enables joint profiling of chromatin accessibility DNA methylation and transcription in single cells". In: *Nature communications* 9.1, pp. 1–9.

Coetzee, Simon G, Gerhard A Coetzee, and Dennis J Hazelett (2015). "motifbreakR: an R/Bioconductor package for predicting variant effects at transcription factor binding sites". In: *Bioinformatics* 31.23, pp. 3847–3849.

Coetzee, Simon G, Suhn K Rhie, Benjamin P Berman, Gerhard A Coetzee, and Houtan Noushmehr (Oct. 2012). "FunciSNP: an R/bioconductor tool integrating functional non-coding data sets with genetic association studies to identify candidate regulatory SNPs". en. In: *Nucleic Acids Res.* 40.18, e139.

Datlinger, Paul, André F Rendeiro, Christian Schmidl, Thomas Krausgruber, Peter Traxler, Johanna Klughammer, Linda C Schuster, Amelie Kuchler, Donat Alpar, and Christoph Bock (2017). "Pooled CRISPR screening with single-cell transcriptome readout". In: *Nature methods* 14.3, pp. 297–301.

Delay, Charlotte, Wim Mandemakers, and Sébastien S Hébert (2012). "MicroRNAs in Alzheimer's disease". In: *Neurobiology of disease* 46.2, pp. 285–290.

Deng, Jia, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei (2009). "Imagenet: A large-scale hierarchical image database". In: *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*. Ieee, pp. 248–255.

Deng, Yue, Feng Bao, Qionghai Dai, Lani F Wu, and Steven J Altschuler (2019). "Scalable analysis of cell-type composition from single-cell transcriptomics using deep recurrent learning". In: *Nature methods* 16.4, pp. 311–314.

Deplancke, Bart, Daniel Alpern, and Vincent Gardeux (2016). "The genetics of transcription factor DNA binding variation". In: *Cell* 166.3, pp. 538–554.

Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova (2018). "Bert: Pre-training of deep bidirectional transformers for language understanding". In: *arXiv preprint arXiv:1810.04805*.

Dijk, David van, Roshan Sharma, Juozas Nainys, Kristina Yim, Pooja Kathail, Ambrose J Carr, Cassandra Burdziak, Kevin R Moon, Christine L Chaffer, Diwakar Pattabiraman, Brian Bierie, Linas Mazutis, Guy Wolf, Smita Krishnaswamy, and Dana Pe'er (July 2018). "Recovering Gene Interactions from Single-Cell Data Using Data Diffusion". en. In: *Cell* 174.3, 716–729.e27.

Ding, Bo, Lina Zheng, Yun Zhu, Nan Li, Haiyang Jia, Rizi Ai, Andre Wildberg, and Wei Wang (July 2015). "Normalization and noise reduction for single cell RNA-seq experiments". en. In: *Bioinformatics* 31.13, pp. 2225–2227.

Ding, Jiarui, Xian Adiconis, Sean K Simmons, Monika S Kowalczyk, Cynthia C Hession, Nemanja D Marjanovic, Travis K Hughes, Marc H Wadsworth, Tyler Burks, Lan T Nguyen, et al. (2019). "Systematic comparative analysis of single cell RNA-sequencing methods". In: *BioRxiv*, p. 632216.

Ding, Jiarui, Anne Condon, and Sohrab P Shah (2018). "Interpretable dimensionality reduction of single cell transcriptome data with deep generative models". In: *Nature communications* 9.1, p. 2002.

Ding, Jun, Bruce J Aronow, Naftali Kaminski, Joseph Kitzmiller, Jeffrey A Whitsett, and Ziv Bar-Joseph (2018). "Reconstructing differentiation networks and their regulation from time series single-cell expression data". In: *Genome research*.

Dixit, Atray, Oren Parnas, Biyu Li, Jenny Chen, Charles P Fulco, Livnat Jerby-Arnon, Nemanja D Marjanovic, Danielle Dionne, Tyler Burks, Raktima Raychowdhury, et al. (2016). "Perturb-Seq: dissecting molecular circuits with scalable single-cell RNA profiling of pooled genetic screens". In: *Cell* 167.7, pp. 1853–1866.

Djebali, Sarah, Carrie A Davis, Angelika Merkel, Alex Dobin, Timo Lassmann, Ali Mortazavi, Andrea Tanzer, Julien Lagarde, Wei Lin, Felix Schlesinger, et al. (2012). "Landscape of transcription in human cells". In: *Nature* 489.7414, p. 101.

Dorn, Gerald W, Scot J Matkovich, William H Eschenbacher, and Yan Zhang (2012). "A Human 3' miR-499 Mutation Alters Cardiac mRNA Targeting and FunctionNovelty and Significance". In: *Circulation research* 110.7, pp. 958–967.

Edwards, Stacey L, Jonathan Beesley, Juliet D French, and Alison M Dunning (Nov. 2013). "Beyond GWASs: illuminating the dark road from association to function". en. In: *Am. J. Hum. Genet.* 93.5, pp. 779–797.

Eicher, John D, Christa Landowski, Brian Stackhouse, Arielle Sloan, Wenjie Chen, et al. (2015). "GRASP v2.0: an update on the Genome-Wide Repository of Associations between SNPs and phenotypes". In: *Nucleic acids research* 43.D1, pp. D799–D804.

ENCODE Project Consortium (2007). "Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project". In: *Nature* 447.7146, p. 799.

ENCODE Project Consortium (2012). "An integrated encyclopedia of DNA elements in the human genome". In: *Nature* 489.7414, pp. 57–74.

Eng, Chee-Huat Linus, Michael Lawson, Qian Zhu, Ruben Dries, Noushin Koulena, Yodai Takei, Jina Yun, Christopher Cronin, Christoph Karp, Guo-Cheng Yuan, et al. (2019). "Transcriptome-scale super-resolved imaging in tissues by RNA seqFISH+". In: *Nature* 568.7751, pp. 235–239.

Eraslan, Gökcen, Žiga Avsec, Julien Gagneur, and Fabian J Theis (2019). "Deep learning: new computational modelling techniques for genomics". In: *Nature Reviews Genetics* 20.7, pp. 389–403.

Eraslan, Gökcen, Lukas M Simon, Maria Mircea, Nikola S Mueller, and Fabian J Theis (2019). "Single-cell RNA-seq denoising using a deep count autoencoder". In: *Nature communications* 10.1, pp. 1–14.

Ernst, Jason and Manolis Kellis (2012). "ChromHMM: automating chromatin-state discovery and characterization". In: *Nature methods* 9.3, p. 215.

Eser, Umut and L Stirling Churchman (2016). "FIDDLE: An integrative deep learning framework for functional genomic data inference". In: *Biorxiv*, p. 081380.

Esteller, Manel (2011). "Non-coding RNAs in human disease". In: *Nature Reviews Genetics* 12.12, pp. 861–874.

Femminella, Grazia D, Nicola Ferrara, and Giuseppe Rengo (2015). "The emerging role of microRNAs in Alzheimer's disease". In: *Frontiers in physiology* 6, p. 40.

Fiers, Mark WEJ, Liesbeth Minnoye, Sara Aibar, Carmen Bravo González-Blas, Zeynep Kalender Atak, and Stein Aerts (2018). "Mapping gene regulatory networks from single-cell omics data". In: *Briefings in functional genomics* 17.4, pp. 246–254.

Finak, Greg, Andrew McDavid, Masanao Yajima, Jingyuan Deng, Vivian Gersuk, Alex K Shalek, Chloe K Slichter, Hannah W Miller, M Juliana McElrath, Martin Prlic, et al. (2015). "MAST: a flexible statistical framework for assessing transcriptional changes and characterizing heterogeneity in single-cell RNA sequencing data". In: *Genome biology* 16.1, p. 278.

Fischer, David S, Yihan Wu, Benjamin Schubert, and Fabian J Theis (2019). "Predicting antigen-specificity of single T-cells based on TCR CDR3 regions". In: *bioRxiv*.

Fleming, Stephen Jordan, John C Marioni, and Mehrtash Babadi (2019). "CellBender remove-background: a deep generative model for unsupervised removal of background noise from scRNA-seq datasets". In: *bioRxiv*, p. 791699.

Francesconi, Mirko and Ben Lehner (2013). "The effects of genetic variation on gene expression dynamics during development". In: *Nature* 505.7482, pp. 208–211.

Frayling, Timothy M, Nicholas J Timpson, Michael N Weedon, Eleftheria Zeggini, Rachel M Freathy, Cecilia M Lindgren, John RB Perry, Katherine S Elliott, Hana Lango, Nigel W Rayner, et al. (2007). "A common variant in the FTO gene is associated with body mass index and predisposes to childhood and adult obesity". In: *Science* 316.5826, pp. 889–894.

Friedberg, Richard M (1958). "A learning machine: Part I". In: *IBM Journal of Research and Development* 2.1, pp. 2–13.

Friedman, Jerome, Trevor Hastie, Robert Tibshirani, et al. (2001). *The elements of statistical learning*. Vol. 1. 10. Springer series in statistics New York.

Fu, Yao, Zhu Liu, Shaoke Lou, Jason Bedford, Xinmeng Jasmine Mu, Kevin Y Yip, Ekta Khurana, and Mark Gerstein (2014). "FunSeq2: a framework for prioritizing noncoding regulatory variants in cancer". In: *Genome biology* 15.10, p. 480.

Furey, Terrence S (2012). "ChIP–seq and beyond: new and improved methodologies to detect and characterize protein–DNA interactions". In: *Nature Reviews Genetics* 13.12, pp. 840–852.

Gelernter, Herbert (1959). "Realization of a geometry theorem proving machine." In: *IFIP congress*, pp. 273–281.

Gierahn, Todd M, Marc H Wadsworth II, Travis K Hughes, Bryan D Bryson, Andrew Butler, Rahul Satija, Sarah Fortune, J Christopher Love, and Alex K Shalek (2017). "Seq-Well: portable, low-cost RNA sequencing of single cells at high throughput". In: *Nature methods* 14.4, p. 395.

Girshick, Ross, Jeff Donahue, Trevor Darrell, and Jitendra Malik (2014). "Rich feature hierarchies for accurate object detection and semantic segmentation". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 580–587.

Gladka, Monika M, Bas Molenaar, Hesther de Ruiter, Stefan van der Elst, Hoyee Tsui, Danielle Versteeg, Grègory P A Lacraz, Manon M H Huibers, Alexander van Oudenaarden, and Eva van Rooij (Jan. 2018). "Single-Cell Sequencing of the Healthy and Diseased Heart Reveals Ckap4 as a New Modulator of Fibroblasts Activation". en. In: *Circulation*.

Goodfellow, Ian, Yoshua Bengio, and Aaron Courville (2016). *Deep Learning*. `http://www.deeplearningbook.org`. MIT Press.

Goodfellow, Ian, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio (2014). "Generative adversarial nets". In: *Advances in neural information processing systems*, pp. 2672–2680.

Grün, Dominic, Lennart Kester, and Alexander Van Oudenaarden (2014). "Validation of noise models for single-cell transcriptomics". In: *Nature methods* 11.6, pp. 637–640.

GTEx Consortium (2017). "Genetic effects on gene expression across human tissues". In: *Nature* 550.7675, pp. 204–213.

Gunderson, Kevin L, Frank J Steemers, Grace Lee, Leo G Mendoza, and Mark S Chee (2005). "A genome-wide scalable SNP genotyping assay using microarray technology". In: *Nature genetics* 37.5, p. 549.

Guttman, Mitchell, Ido Amit, Manuel Garber, Courtney French, Michael F Lin, David Feldser, Maite Huarte, Or Zuk, Bryce W Carey, John P Cassady, et al. (2009). "Chromatin signature reveals over a thousand highly conserved large non-coding RNAs in mammals". In: *Nature* 458.7235, p. 223.

Hafemeister, Christoph and Rahul Satija (2019). "Normalization and variance stabilization of single-cell RNA-seq data using regularized negative binomial regression". In: *Genome biology* 20.1, pp. 1–15.

Haghverdi, Laleh, Florian Buettner, and Fabian J Theis (2015). "Diffusion maps for high-dimensional single-cell analysis of differentiation data". In: *Bioinformatics* 31.18, pp. 2989–2998.

Haghverdi, Laleh, Maren Büttner, F Alexander Wolf, Florian Buettner, and Fabian J Theis (Oct. 2016). "Diffusion pseudotime robustly reconstructs lineage branching". en. In: *Nat. Methods* 13.10, pp. 845–848.

Han, Xiaoping, Renying Wang, Yincong Zhou, Lijiang Fei, Huiyu Sun, Shujing Lai, Assieh Saadatpour, Ziming Zhou, Haide Chen, Fang Ye, et al. (2018). "Mapping the mouse cell atlas by microwell-seq". In: *Cell* 172.5, pp. 1091–1107.

Hasin, Yehudit, Marcus Seldin, and Aldons Lusis (2017). "Multi-omics approaches to disease". In: *Genome biology* 18.1, p. 83.

Hauberg, Mads Engel, Marie Hebsgaard Holm-Nielsen, Manuel Mattheisen, Anne Louise Askou, Jakob Grove, Anders Dupont Børglum, and Thomas Juhl Corydon (Sept. 2016). "Schizophrenia risk variants affecting microRNA function and site-specific regulation of NT5C2 by miR-206". en. In: *Eur. Neuropsychopharmacol.* 26.9, pp. 1522–1526.

He, Kaiming, Xiangyu Zhang, Shaoqing Ren, and Jian Sun (2015). "Delving deep into rectifiers: Surpassing human-level performance on imagenet classification". In: *Proceedings of the IEEE international conference on computer vision*, pp. 1026–1034.

He, Qiye, Jeff Johnston, and Julia Zeitlinger (2015). "ChIP-nexus enables improved detection of in vivo transcription factor binding footprints". In: *Nature biotechnology* 33.4, pp. 395–401.

Heaton, Haynes, Arthur M Talman, Andrew Knights, Maria Imaz, Richard Durbin, Martin Hemberg, and Mara Lawniczak (2019). "souporcell: Robust clustering of single cell RNAseq by genotype and ambient RNA inference without reference genotypes". In: *BioRxiv*, p. 699637.

Herring, Charles A, Amrita Banerjee, Eliot T McKinley, Alan J Simmons, Jie Ping, Joseph T Roland, Jeffrey L Franklin, Qi Liu, Michael J Gerdes, Robert J Coffey, and Ken S Lau (Jan. 2018). "Unsupervised Trajectory Analysis of Single-Cell RNA-Seq and Imaging Data Reveals Alternative Tuft Cell Origins in the Gut". en. In: *Cell Syst* 6.1, 37–51.e9.

Hinton, Geoffrey E and Ruslan R Salakhutdinov (2006). "Reducing the dimensionality of data with neural networks". In: *science* 313.5786, pp. 504–507.

Hofner, Benjamin, Luigi Boccuto, and Markus Göker (May 2015). "Controlling false discoveries in high-dimensional situations: boosting with stability selection". en. In: *BMC Bioinformatics* 16, p. 144.

Hoheisel, Jörg D (2006). "Microarray technology: beyond transcript profiling and genotype analysis". In: *Nature reviews genetics* 7.3, p. 200.

Hooke, Robert (1961). *Micrographia*. Allestry.

Hou, Lin and Hongyu Zhao (Dec. 2013). "A review of post-GWAS prioritization approaches". en. In: *Front. Genet.* 4, p. 280.

Hou, Wenpin, Zhicheng Ji, Hongkai Ji, and Stephanie C Hicks (2020). "A Systematic Evaluation of Single-cell RNA-sequencing Imputation Methods". In: *bioRxiv*.

Howard, David M, Mark J Adams, Toni-Kim Clarke, Jonathan D Hafferty, Jude Gibson, Masoud Shirali, Jonathan RI Coleman, Saskia P Hagenaars, Joey Ward, Eleanor M Wigmore, et al. (2019). "Genome-wide meta-analysis of depression identifies 102 independent variants and highlights the importance of the prefrontal brain regions". In: *Nature neuroscience* 22.3, pp. 343–352.

Hsu, Fan, W James Kent, Hiram Clawson, Robert M Kuhn, Mark Diekhans, and David Haussler (2006). "The UCSC known genes". In: *Bioinformatics* 22.9, pp. 1036–1046.

Huang, Mo, Jingshu Wang, Eduardo Torre, Hannah Dueck, Sydney Shaffer, Roberto Bonasio, John I Murray, Arjun Raj, Mingyao Li, and Nancy R Zhang (July 2018). "SAVER: gene expression recovery for single-cell RNA sequencing". en. In: *Nat. Methods* 15.7, pp. 539–542.

Hubbard, Tim, Daniel Barker, Ewan Birney, Graham Cameron, Yuan Chen, L Clark, Tony Cox, J Cuff, Val Curwen, Thomas Down, et al. (2002). "The Ensembl genome database project". In: *Nucleic acids research* 30.1, pp. 38–41.

Huber, Wolfgang, Vincent J Carey, Robert Gentleman, Simon Anders, Marc Carlson, et al. (2015). "Orchestrating high-throughput genomic analysis with Bioconductor". In: *Nature methods* 12.2, pp. 115–121.

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

Iacono, Giovanni, Ramon Massoni-Badosa, and Holger Heyn (2019). "Single-cell transcriptomics unveils gene regulatory network plasticity". In: *Genome biology* 20.1, pp. 1–20.

International HapMap Consortium (2003). "The International HapMap project". In: *Nature* 426.6968, p. 789.

International Human Genome Sequencing Consortium (2001). "Initial sequencing and analysis of the human genome". In: *nature* 409.6822, p. 860.

International Multiple Sclerosis Genetics Consortium (2019). "Multiple sclerosis genomic map implicates peripheral immune cells and microglia in susceptibility". In: *Science* 365.6460, eaav7188.

Ioffe, Sergey and Christian Szegedy (2015). "Batch normalization: Accelerating deep network training by reducing internal covariate shift". In: *arXiv preprint arXiv:1502.03167*.

Jaitin, Diego Adhemar, Ephraim Kenigsberg, Hadas Keren-Shaul, Naama Elefant, Franziska Paul, Irina Zaretsky, Alexander Mildner, Nadav Cohen, Steffen Jung, Amos Tanay, et al. (2014). "Massively parallel single-cell RNA-seq for marker-free decomposition of tissues into cell types". In: *Science* 343.6172, pp. 776–779.

Jansen, Philip R, Kyoko Watanabe, Sven Stringer, Nathan Skene, Julien Bryois, Anke R Hammerschlag, Christiaan A de Leeuw, Jeroen S Benjamins, Ana B Muñoz-Manchado,

Mats Nagel, et al. (2019). "Genome-wide analysis of insomnia in 1,331,010 individuals identifies new risk loci and functional pathways". In: *Nature genetics* 51.3, pp. 394–403.

Kanehisa, Minoru and Susumu Goto (2000). "KEGG: kyoto encyclopedia of genes and genomes". In: *Nucleic acids research* 28.1, pp. 27–30.

Katsuoka, Fumiki, Hozumi Motohashi, Yuna Tamagawa, Shigeo Kure, Kazuhiko Igarashi, James Douglas Engel, and Masayuki Yamamoto (2003). "Small Maf compound mutants display central nervous system neuronal degeneration, aberrant transcription, and Bach protein mislocalization coincident with myoclonus and abnormal startle response". In: *Molecular and cellular biology* 23.4, pp. 1163–1174.

Katsuoka, Fumiki and Masayuki Yamamoto (2016). "Small Maf proteins (MafF, MafG, MafK): history, structure and function". In: *Gene* 586.2, pp. 197–205.

Kelley, David R, Yakir A Reshef, Maxwell Bileschi, David Belanger, Cory Y McLean, and Jasper Snoek (2018). "Sequential regulatory activity prediction across chromosomes with convolutional neural networks". In: *Genome research* 28.5, pp. 739–750.

Kelley, David R, Jasper Snoek, and John L Rinn (July 2016). "Basset: learning the regulatory code of the accessible genome with deep convolutional neural networks". en. In: *Genome Res.* 26.7, pp. 990–999.

Kemppinen, AK, Jaakko Kaprio, Aarno Palotie, and Janna Saarela (2011). "Systematic review of genome-wide expression studies in multiple sclerosis". In: *BMJ open* 1.1.

Keren-Shaul, Hadas, Amit Spinrad, Assaf Weiner, Orit Matcovitch-Natan, Raz Dvir-Szternfeld, Tyler K Ulland, Eyal David, Kuti Baruch, David Lara-Astaiso, Beata Toth, Shalev Itzkovitz, Marco Colonna, Michal Schwartz, and Ido Amit (2017). "A Unique Microglia Type Associated with Restricting Development of Alzheimer's Disease". In: *Cell* 169.7, 1276–1290.e17.

Kharchenko, Peter V, Lev Silberstein, and David T Scadden (July 2014). "Bayesian approach to single-cell differential expression analysis". en. In: *Nat. Methods* 11.7, pp. 740–742.

Kingma, Diederik P and Jimmy Ba (2014). "Adam: A method for stochastic optimization". In: *arXiv preprint arXiv:1412.6980*.

Kingma, Diederik P and Max Welling (2013). "Auto-encoding variational bayes". In: *arXiv preprint arXiv:1312.6114*.

Kircher, Martin, Daniela M Witten, Preti Jain, Brian J O'Roak, Gregory M Cooper, and Jay Shendure (Mar. 2014). "A general framework for estimating the relative pathogenicity of human genetic variants". en. In: *Nat. Genet.* 46.3, pp. 310–315.

Kitano, Hiroaki (2002). "Computational systems biology". In: *Nature* 420.6912, p. 206.

Klein, Allon M, Linas Mazutis, Ilke Akartuna, Naren Tallapragada, Adrian Veres, Victor Li, Leonid Peshkin, David A Weitz, and Marc W Kirschner (2015). "Droplet Barcoding for Single-Cell Transcriptomics Applied to Embryonic Stem Cells". In: *Cell* 161.5, pp. 1187–1201.

Kohl, Peter, Edmund J Crampin, TA Quinn, and Denis Noble (2010). "Systems biology: an approach". In: *Clinical Pharmacology & Therapeutics* 88.1, pp. 25–33.

Krizhevsky, Alex, Ilya Sutskever, and Geoffrey E Hinton (2012). "Imagenet classification with deep convolutional neural networks". In: *Advances in neural information processing systems*, pp. 1097–1105.

Krumsiek, Jan, Carsten Marr, Timm Schroeder, and Fabian J Theis (Aug. 2011). "Hierarchical differentiation of myeloid progenitors is encoded in the transcription factor network". en. In: *PLoS One* 6.8, e22649.

Kundaje, Anshul, Wouter Meuleman, Jason Ernst, Misha Bilenky, Angela Yen, Alireza Heravi-Moussavi, Pouya Kheradpour, Zhizhuo Zhang, Jianrong Wang, Michael J Ziller, et al. (2015). "Integrative analysis of 111 reference human epigenomes". In: *Nature* 518.7539, p. 317.

La Manno, Gioele, Ruslan Soldatov, Amit Zeisel, Emelie Braun, Hannah Hochgerner, Viktor Petukhov, Katja Lidschreiber, Maria E Kastriti, Peter Lönnerberg, Alessandro Furlan, et al. (2018). "RNA velocity of single cells". In: *Nature* 560.7719, pp. 494–498.

Lähnemann, David, Johannes Köster, Ewa Szczurek, Davis J McCarthy, Stephanie C Hicks, Mark D Robinson, Catalina A Vallejos, Kieran R Campbell, Niko Beerenwinkel, Ahmed Mahfouz, et al. (2020). "Eleven grand challenges in single-cell data science". In: *Genome biology* 21.1, pp. 1–35.

Lake, Blue B, Song Chen, Brandon C Sos, Jean Fan, Gwendolyn E Kaeser, Yun C Yung, Thu E Duong, Derek Gao, Jerold Chun, Peter V Kharchenko, et al. (2018). "Integrative single-cell analysis of transcriptional and epigenetic states in the human adult brain". In: *Nature biotechnology* 36.1, p. 70.

Lawrence, Michael, Wolfgang Huber, Hervé Pages, Patrick Aboyoun, Marc Carlson, et al. (2013). "Software for computing and annotating genomic ranges". In: *PLoS Comput Biol* 9.8, e1003118.

LeCun, Yann, Yoshua Bengio, and Geoffrey Hinton (May 2015). "Deep learning". en. In: *Nature* 521.7553, pp. 436–444.

Lee, James J, Robbee Wedow, Aysu Okbay, Edward Kong, Omeed Maghzian, Meghan Zacher, Tuan Anh Nguyen-Viet, Peter Bowers, Julia Sidorenko, Richard Karlsson Linnér, et al. (2018). "Gene discovery and polygenic prediction from a genome-wide association study of educational attainment in 1.1 million individuals". In: *Nature genetics*, p. 1.

Lee, Phil H, Colm O'Dushlaine, Brett Thomas, and Shaun M Purcell (July 2012). "INRICH: interval-based enrichment analysis for genome-wide association studies". en. In: *Bioinformatics* 28.13, pp. 1797–1799.

Lek, Monkol, Konrad J Karczewski, Eric V Minikel, Kaitlin E Samocha, Eric Banks, Timothy Fennell, Anne H O'Donnell-Luria, James S Ware, Andrew J Hill, Beryl B Cummings,

et al. (2016). “Analysis of protein-coding genetic variation in 60,706 humans”. In: *Nature* 536.7616, p. 285.

Leslie, Richard, Christopher J O’Donnell, and Andrew D Johnson (2014). “GRASP: analysis of genotype–phenotype results from 1390 genome-wide association studies and corresponding open access database”. In: *Bioinformatics* 30.12, pp. i185–i194.

Lewis, Benjamin P, I-hung Shih, Matthew W Jones-Rhoades, David P Bartel, and Christopher B Burge (2003). “Prediction of mammalian microRNA targets”. In: *Cell* 115.7, pp. 787–798.

Li, Jin Billy, Erez Y Levanon, Jung-Ki Yoon, John Aach, Bin Xie, Emily LeProust, Kun Zhang, Yuan Gao, and George M Church (2009). “Genome-wide identification of human RNA editing sites by parallel DNA capturing and sequencing”. In: *Science* 324.5931, pp. 1210–1213.

Li, Jun-Hao, Shun Liu, Hui Zhou, Liang-Hu Qu, and Jian-Hua Yang (2013). “starBase v2. 0: decoding miRNA-ceRNA, miRNA-ncRNA and protein–RNA interaction networks from large-scale CLIP-Seq data”. In: *Nucleic acids research*, gkt1248.

Li, Qiyuan, Ji-Heui Seo, Barbara Stranger, Aaron McKenna, Itsik Pe’er, Thomas LaFramboise, Myles Brown, Svitlana Tyekucheva, and Matthew L Freedman (2013). “Integrative eQTL-based analyses reveal the biology of breast cancer risk loci”. In: *Cell* 152.3, pp. 633–641.

Li, Wei Vivian and Jingyi Jessica Li (Mar. 2018). “An accurate and robust imputation method scImpute for single-cell RNA-seq data”. en. In: *Nat. Commun.* 9.1, p. 997.

Libbrecht, Maxwell W and William Stafford Noble (2015). “Machine learning applications in genetics and genomics”. In: *Nature Reviews Genetics* 16.6, p. 321.

Lighthill, Sir James (1973). *Artificial Intelligence: A General Survey. Part I of Artificial Intelligence’: a paper symposium. London: Science Research Council.*

Liu, Jimmy Z, Allan F McRae, Dale R Nyholt, Sarah E Medland, Naomi R Wray, Kevin M Brown, AMFS Investigators, Nicholas K Hayward, Grant W Montgomery, Peter M Visscher, Nicholas G Martin, and Stuart Macgregor (July 2010). “A versatile gene-based test for genome-wide association studies”. en. In: *Am. J. Hum. Genet.* 87.1, pp. 139–145.

Liu, Longqi, Chuanyu Liu, Andrés Quintero, Liang Wu, Yue Yuan, Mingyue Wang, Mengnan Cheng, Lizhi Leng, Liqin Xu, Guoyi Dong, et al. (2019). “Deconvolution of single-cell multi-omics layers reveals regulatory heterogeneity”. In: *Nature communications* 10.1, pp. 1–10.

Liu, Mengzhen, Yu Jiang, Robbee Wedow, Yue Li, David M Brazel, Fang Chen, Gargi Datta, Jose Davila-Velderrain, Daniel McGuire, Chao Tian, et al. (2019). “Association studies of up to 1.2 million individuals yield new insights into the genetic etiology of tobacco and alcohol use”. In: *Nature genetics* 51.2, pp. 237–244.

Long, Jonathan, Evan Shelhamer, and Trevor Darrell (2015). "Fully convolutional networks for semantic segmentation". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3431–3440.

Lonsdale, John, Jeffrey Thomas, Mike Salvatore, Rebecca Phillips, Edmund Lo, et al. (2013). "The genotype-tissue expression (GTEx) project". In: *Nature genetics* 45.6, pp. 580–585.

Lopez, Romain, Jeffrey Regier, Michael B Cole, Michael I Jordan, and Nir Yosef (2018). "Deep generative modeling for single-cell transcriptomics". In: *Nature methods* 15.12, pp. 1053–1058.

Lotfollahi, Mohammad, F Alexander Wolf, and Fabian J Theis (2019). "scGen predicts single-cell perturbation responses". In: *Nature methods* 16.8, pp. 715–721.

Ludwig, Nicole, Petra Leidinger, Kurt Becker, Christina Backes, Tobias Fehlmann, Christian Pallasch, Steffi Rheinheimer, Benjamin Meder, Cord Stähler, Eckart Meese, et al. (2016). "Distribution of miRNA expression across human tissues". In: *Nucleic acids research* 44.8, pp. 3865–3877.

Luecken, Malte D and Fabian J Theis (2019). "Current best practices in single-cell RNA-seq analysis: a tutorial". In: *Molecular systems biology* 15.6, e8746.

Lun, Aaron TL, Karsten Bach, and John C Marioni (2016). "Pooling across cells to normalize single-cell RNA sequencing data with many zero counts". In: *Genome biology* 17.1, p. 75.

Lun, Aaron TL, Samantha Riesenfeld, Tallulah Andrews, Tomas Gomes, John C Marioni, et al. (2019). "EmptyDrops: distinguishing cells from empty droplets in droplet-based single-cell RNA sequencing data". In: *Genome biology* 20.1, pp. 1–9.

Lv, Ke, Yingjun Guo, Yiliang Zhang, Kaiyu Wang, Yin Jia, and Shuhan Sun (2008). "Allele-specific targeting of hsa-miR-657 to human IGF2R creates a potential mechanism underlying the association of ACAA-insertion/deletion polymorphism with type 2 diabetes". In: *Biochemical and biophysical research communications* 374.1, pp. 101–105.

Maaten, Laurens van der and Geoffrey Hinton (2008). "Visualizing data using t-SNE". In: *Journal of machine learning research* 9.Nov, pp. 2579–2605.

Macosko, Evan Z, Anindita Basu, Rahul Satija, James Nemesh, Karthik Shekhar, Melissa Goldman, Itay Tirosh, Allison R Bialas, Nolan Kamitaki, Emily M Martersteck, John J Trombetta, David A Weitz, Joshua R Sanes, Alex K Shalek, Aviv Regev, and Steven A McCarroll (May 2015). "Highly Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets". en. In: *Cell* 161.5, pp. 1202–1214.

Marchini, Jonathan and Bryan Howie (2010). "Genotype imputation for genome-wide association studies". In: *Nature Reviews Genetics* 11.7, p. 499.

Markowetz, Florian (2017). "All biology is computational biology". In: *PLoS biology* 15.3, e2002050.

McCarthy, Davis J, Peter Humburg, Alexander Kanapin, Manuel A Rivas, Kyle Gaulton, Jean-Baptiste Cazier, Peter Donnelly, and asds (2014). "Choice of transcripts and software has a large effect on variant annotation". In: *Genome Med.* 6.3, p. 26.

McCulloch, Warren S and Walter Pitts (1943). "A logical calculus of the ideas immanent in nervous activity". In: *The bulletin of mathematical biophysics* 5.4, pp. 115–133.

McLaren, William, Laurent Gil, Sarah E Hunt, Harpreet Singh Riat, Graham R S Ritchie, Anja Thormann, Paul Flicek, and Fiona Cunningham (June 2016). "The Ensembl Variant Effect Predictor". en. In: *Genome Biol.* 17.1, p. 122.

Meinshausen, Nicolai and Peter Bühlmann (2010). "Stability selection". In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 72.4, pp. 417–473.

Mimitou, Eleni P, Anthony Cheng, Antonino Montalbano, Stephanie Hao, Marlon Stoeckius, Mateusz Legut, Timothy Roush, Alberto Herrera, Efthymia Papalexi, Zhengqing Ouyang, et al. (2019). "Multiplexed detection of proteins, transcriptomes, clonotypes and CRISPR perturbations in single cells". In: *Nature methods* 16.5, pp. 409–412.

Minsky, Marvin and Seymour Papert (1969). *Perceptrons: An Introduction to Computational Geometry.*

Moignard, Victoria, Steven Woodhouse, Laleh Haghverdi, Andrew J Lilly, Yosuke Tanaka, Adam C Wilkinson, Florian Buettner, Iain C Macaulay, Wajid Jawaid, Evangelia Diamanti, Shin-Ichi Nishikawa, Nir Piterman, Valerie Kouskoff, Fabian J Theis, Jasmin Fisher, and Berthold Göttgens (Mar. 2015). "Decoding the regulatory network of early blood development from single-cell gene expression measurements". en. In: *Nat. Biotechnol.* 33.3, pp. 269–276.

Muglia, P, F Tozzi, N W Galwey, C Francks, R Upmanyu, X Q Kong, A Antoniades, E Domenici, J Perry, S Rothen, C L Vandeleur, V Mooser, G Waeber, P Vollenweider, M Preisig, S Lucae, B Müller-Myhsok, F Holsboer, L T Middleton, and A D Roses (June 2010). "Genome-wide association study of recurrent major depressive disorder in two European case-control cohorts". en. In: *Mol. Psychiatry* 15.6, pp. 589–601.

Nerlov, Claus, Erich Querfurth, Holger Kulessa, and Thomas Graf (2000). "GATA-1 interacts with the myeloid PU.1 transcription factor and represses PU.1-dependent transcription". In: *Blood, The Journal of the American Society of Hematology* 95.8, pp. 2543–2551.

Ng, Pauline C and Steven Henikoff (July 2003). "SIFT: Predicting amino acid changes that affect protein function". en. In: *Nucleic Acids Res.* 31.13, pp. 3812–3814.

Njølstad, Pål Rasmus, Ole Andreas Andreassen, Søren Brunak, Anders D Børglum, Joakim Dillner, Tõnu Esko, Paul W Franks, Nelson Freimer, Leif Groop, Hakon Heimer, et al. (2019). "Roadmap for a precision-medicine initiative in the Nordic region". In: *Nature genetics* 51.6, pp. 924–930.

O'Brien, John S and Yasuo Kishimoto (1991). "Saposin proteins: structure, function, and role in human lysosomal storage disorders". In: *The FASEB Journal* 5.3, pp. 301–308.

Onos, Kristen D, Stacey J Sukoff Rizzo, Gareth R Howell, and Michael Sasner (2016). "Toward more predictive genetic mouse models of Alzheimer's disease". In: *Brain research bulletin* 122, pp. 1–11.

Oord, Aaron van den, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew Senior, and Koray Kavukcuoglu (2016). "Wavenet: A generative model for raw audio". In: *arXiv preprint arXiv:1609.03499*.

Orkin, Stuart H and Leonard I Zon (Feb. 2008). "Hematopoiesis: an evolving paradigm for stem cell biology". en. In: *Cell* 132.4, pp. 631–644.

Paila, Umadevi, Brad A Chapman, Rory Kirchner, and Aaron R Quinlan (July 2013). "GEMINI: integrative exploration of genetic variation and genome annotations". en. In: *PLoS Comput. Biol.* 9.7, e1003153.

Pan, Qun, Ofer Shai, Leo J Lee, Brendan J Frey, and Benjamin J Blencowe (2008). "Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing". In: *Nature genetics* 40.12, p. 1413.

Panwar, Bharat, Gilbert S Omenn, and Yuanfang Guan (2017). "miRmine: a database of human miRNA expression profiles". In: *Bioinformatics* 33.10, pp. 1554–1560.

Patsopoulos, Nikolaos A (July 2018). "Genetics of Multiple Sclerosis: An Overview and New Directions". en. In: *Cold Spring Harb. Perspect. Med.* 8.7.

Paul, Franziska, Ya'ara Arkin, Amir Giladi, Diego Adhemar Jaitin, Ephraim Kenigsberg, Hadas Keren-Shaul, Deborah Winter, David Lara-Astiaso, Meital Gury, Assaf Weiner, Eyal David, Nadav Cohen, Felicia Kathrine Bratt Lauridsen, Simon Haas, Andreas Schlitzer, Alexander Mildner, Florent Ginhoux, Steffen Jung, Andreas Trumpp, Bo Torben Porse, Amos Tanay, and Ido Amit (2015). "Transcriptional Heterogeneity and Lineage Commitment in Myeloid Progenitors". In: *Cell* 163.7, pp. 1663–1677.

Pers, Tune H, Juha M Karjalainen, Yingleong Chan, Harm-Jan Westra, Andrew R Wood, Jian Yang, Julian C Lui, Sailaja Vedantam, Stefan Gustafsson, Tonu Esko, Tim Frayling, Elizabeth K Speliotes, Genetic Investigation of ANthropometric Traits (GIANT) Consortium, Michael Boehnke, Soumya Raychaudhuri, Rudolf S N Fehrmann, Joel N Hirschhorn, and Lude Franke (Jan. 2015). "Biological interpretation of genome-wide association studies using predicted gene functions". en. In: *Nat. Commun.* 6, p. 5890.

Perumean-Chaney, Suzanne E, Charity Morgan, David McDowall, and Inmaculada Aban (2013). "Zero-inflated and overdispersed: what's one to do?" In: *Journal of Statistical Computation and Simulation* 83.9, pp. 1671–1683.

Peters, Matthew E, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer (2018). "Deep contextualized word representations". In: *arXiv preprint arXiv:1802.05365*.

Petrovski, Slavé, Quanli Wang, Erin L Heinzen, Andrew S Allen, and David B Goldstein (2013). "Genic intolerance to functional variation and the interpretation of personal genomes". In: *PLoS genetics* 9.8, e1003709.

Picelli, Simone, Åsa K Björklund, Omid R Faridani, Sven Sagasser, Gösta Winberg, and Rickard Sandberg (2013). "Smart-seq2 for sensitive full-length transcriptome profiling in single cells". In: *Nature methods* 10.11, p. 1096.

Pierson, Emma and Christopher Yau (2015). "ZIFA: Dimensionality reduction for zero-inflated single-cell gene expression analysis". In: *Genome biology* 16.1, p. 241.

Plass, Mireya, Jordi Solana, F Alexander Wolf, Salah Ayoub, Aristotelis Misios, Petar Glažar, Benedikt Obermayer, Fabian J Theis, Christine Kocks, and Nikolaus Rajewsky (2018). "Cell type atlas and lineage tree of a whole complex animal by single-cell transcriptomics". In: *Science* 360.6391.

Plaut, Elad (2018). "From principal subspaces to principal components with linear autoencoders". In: *arXiv preprint arXiv:1804.10253*.

Pruitt, Kim D, Tatiana Tatusova, and Donna R Maglott (2006). "NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins". In: *Nucleic acids research* 35.suppl_1, pp. D61–D65.

Purcell, Shaun, Benjamin Neale, Kathe Todd-Brown, Lori Thomas, Manuel AR Ferreira, David Bender, Julian Maller, Pamela Sklar, Paul IW De Bakker, Mark J Daly, et al. (2007). "PLINK: a tool set for whole-genome association and population-based linkage analyses". In: *The American Journal of Human Genetics* 81.3, pp. 559–575.

R Core Team (2015). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria. URL: `https://www.R-project.org`.

Radford, Alec, Luke Metz, and Soumith Chintala (2015). "Unsupervised representation learning with deep convolutional generative adversarial networks". In: *arXiv preprint arXiv:1511.06434*.

Raychaudhuri, Soumya, Robert M Plenge, Elizabeth J Rossin, Aylwin C Y Ng, International Schizophrenia Consortium, Shaun M Purcell, Pamela Sklar, Edward M Scolnick, Ramnik J Xavier, David Altshuler, and Mark J Daly (June 2009). "Identifying relationships among genomic disease regions: predicting genes at pathogenic SNP associations and rare deletions". en. In: *PLoS Genet.* 5.6, e1000534.

Regev, Aviv, Sarah A Teichmann, Eric S Lander, Ido Amit, Christophe Benoist, Ewan Birney, Bernd Bodenmiller, Peter Campbell, Piero Carninci, Menna Clatworthy, et al. (2017). "Science forum: the human cell atlas". In: *Elife* 6, e27041.

Reuter, Jason A, Damek V Spacek, and Michael P Snyder (2015). "High-throughput sequencing technologies". In: *Molecular cell* 58.4, pp. 586–597.

Reyes, Miguel, Kianna Billman, Nir Hacohen, and Paul C Blainey (2019). "Simultaneous profiling of gene expression and chromatin accessibility in single cells". In: *Advanced biosystems* 3.11, p. 1900065.

Rietschel, Marcella, Manuel Mattheisen, Josef Frank, Jens Treutlein, Franziska Degenhardt, René Breuer, Michael Steffens, Daniela Mier, Christine Esslinger, Henrik Walter, et al. (2010). "Genome-wide association-, replication-, and neuroimaging study implicates HOMER1 in the etiology of major depression". In: *Biological psychiatry* 68.6, pp. 578–585.

Risso, Davide, Fanny Perraudeau, Svetlana Gribkova, Sandrine Dudoit, and Jean-Philippe Vert (Jan. 2018). "A general and flexible method for signal extraction from single-cell RNA-seq data". en. In: *Nat. Commun.* 9.1, p. 284.

Ritchie, Graham RS, Ian Dunham, Eleftheria Zeggini, and Paul Flicek (2014). "Functional annotation of noncoding sequence variants". In: *Nature methods* 11.3, p. 294.

Ronen, Jonathan and Altuna Akalin (Jan. 2018). "netSmooth: Network-smoothing based imputation for single cell RNA-seq". en. In: *F1000Res.* 7, p. 8.

Rosenberg, Alexander B, Charles M Roco, Richard A Muscat, Anna Kuchina, Paul Sample, Zizhen Yao, Lucas T Graybuck, David J Peeler, Sumit Mukherjee, Wei Chen, et al. (2018). "Single-cell profiling of the developing mouse brain and spinal cord with split-pool barcoding". In: *Science* 360.6385, pp. 176–182.

Rosenblatt, Frank (1958). "The perceptron: a probabilistic model for information storage and organization in the brain." In: *Psychological review* 65.6, p. 386.

Rosenblatt, Frank (1961). *Principles of neurodynamics. perceptrons and the theory of brain mechanisms.* Tech. rep. CORNELL AERONAUTICAL LAB INC BUFFALO NY.

Russell, Stuart J and Peter Norvig (2016). *Artificial intelligence: a modern approach.* Pearson Education Limited.

Samuel, Arthur L (1959). "Some studies in machine learning using the game of checkers". In: *IBM Journal of research and development* 3.3, pp. 210–229.

Schiller, Herbert B, Daniel T Montoro, Lukas M Simon, Emma L Rawlins, Kerstin B Meyer, Maximilian Strunz, Felipe A Vieira Braga, Wim Timens, Gerard H Koppelman, GR Scott Budinger, et al. (2019). "The Human Lung Cell Atlas: a high-resolution reference map of the human lung in health and disease". In: *American journal of respiratory cell and molecular biology* 61.1, pp. 31–41.

Schizophrenia Working Group of the Psychiatric Genomics Consortium (July 2014). "Biological insights from 108 schizophrenia-associated genetic loci". en. In: *Nature* 511.7510, pp. 421–427.

Schulze, Thomas and Francis McMahon (2018). *Psychiatric Genetics: A Primer for Clinical and Basic Scientists*. Oxford University Press.

Schwartzman, Omer and Amos Tanay (2015). "Single-cell epigenomics: techniques and emerging applications". In: *Nature Reviews Genetics* 16.12, p. 716.

Shabalin, Andrey A (2012). "Matrix eQTL: ultra fast eQTL analysis via large matrix operations". In: *Bioinformatics* 28.10, pp. 1353–1358.

Shah, Rajen D and Richard J Samworth (2012). "Variable selection with error control: another look at stability selection". In: *J. R. Stat. Soc. Series B Stat. Methodol.* 75.1, pp. 55–80.

Shao, Ling, Ruomei Yan, Xuelong Li, and Yan Liu (July 2014). "From heuristic optimization to dictionary learning: a review and comprehensive comparison of image denoising algorithms". en. In: *IEEE Trans Cybern* 44.7, pp. 1001–1013.

Sherry, Stephen T, M-H Ward, M Kholodov, J Baker, Lon Phan, et al. (2001). "dbSNP: the NCBI database of genetic variation". In: *Nucleic acids research* 29.1, pp. 308–311.

Shi, Shaohuai, Qiang Wang, Pengfei Xu, and Xiaowen Chu (2016). "Benchmarking state-of-the-art deep learning software tools". In: *Cloud Computing and Big Data (CCBD), 2016 7th International Conference on*. IEEE, pp. 99–104.

Siva, Nayanah (2008). *1000 Genomes Project*.

Slatkin, Montgomery (2008). "Linkage disequilibrium—understanding the evolutionary past and mapping the medical future". In: *Nature Reviews Genetics* 9.6, pp. 477–485.

Slowikowski, Kamil, Xinli Hu, and Soumya Raychaudhuri (2014). "SNPsea: an algorithm to identify cell types, tissues and pathways affected by risk loci". In: *Bioinformatics* 30.17, pp. 2496–2497.

Smemo, Scott, Juan J Tena, Kyoung-Han Kim, Eric R Gamazon, Noboru J Sakabe, Carlos Gómez-Marín, Ivy Aneas, Flavia L Credidio, Débora R Sobreira, Nora F Wasserman, et al. (2014). "Obesity-associated variants within FTO form long-range functional connections with IRX3". In: *Nature* 507.7492, pp. 371–375.

Smillie, Christopher S, Moshe Biton, Jose Ordovas-Montanes, Keri M Sullivan, Grace Burgin, Daniel B Graham, Rebecca H Herbst, Noga Rogel, Michal Slyper, Julia Waldman, et al. (2019). "Intra-and inter-cellular rewiring of the human colon during ulcerative colitis". In: *Cell* 178.3, pp. 714–730.

Soumillon, Magali, Davide Cacchiarelli, Stefan Semrau, Alexander van Oudenaarden, and Tarjei S Mikkelsen (2014). "Characterization of directed differentiation by high-throughput single-cell RNA-Seq". In: *BioRxiv*, p. 003236.

Stephens, Zachary D, Skylar Y Lee, Faraz Faghri, Roy H Campbell, Chengxiang Zhai, Miles J Efron, Ravishankar Iyer, Michael C Schatz, Saurabh Sinha, and Gene E Robinson (2015). "Big data: astronomical or genomical?" In: *PLoS biology* 13.7, e1002195.

Stephenson, William, Laura T Donlin, Andrew Butler, Cristina Rozo, Bernadette Bracken, Ali Rashidfarrokhi, Susan M Goodman, Lionel B Ivashkiv, Vivian P Bykerk, Dana E Orange, Robert B Darnell, Harold P Swerdlow, and Rahul Satija (Feb. 2018). "Single-cell RNA-seq of rheumatoid arthritis synovial tissue using low-cost microfluidic instrumentation". en. In: *Nat. Commun.* 9.1, p. 791.

Stevens, Hallam (2013). *Life out of sequence: a data-driven history of bioinformatics*. University of Chicago Press.

Stoeckius, Marlon, Christoph Hafemeister, William Stephenson, Brian Houck-Loomis, Pratip K Chattopadhyay, Harold Swerdlow, Rahul Satija, and Peter Smibert (Sept. 2017). "Simultaneous epitope and transcriptome measurement in single cells". en. In: *Nat. Methods* 14.9, pp. 865–868.

Svensson, Valentine, Eduardo da Veiga Beltrame, and Lior Pachter (2019). "A curated database reveals trends in single-cell transcriptomics". In: *BioRxiv*, p. 742304.

Tak, Yu Gyoung and Peggy J Farnham (Dec. 2015). "Making sense of GWAS: using epigenomics and genome engineering to understand the functional relevance of SNPs in non-coding regions of the human genome". en. In: *Epigenetics & Chromatin* 8, p. 57.

Tang, Fuchou, Catalin Barbacioru, Yangzhou Wang, Ellen Nordman, Clarence Lee, Nanlan Xu, Xiaohui Wang, John Bodeau, Brian B Tuch, Asim Siddiqui, et al. (2009). "mRNA-Seq whole-transcriptome analysis of a single cell". In: *Nature methods* 6.5, p. 377.

Tehranchi, Ashley K, Marsha Myrthil, Trevor Martin, Brian L Hie, David Golan, and Hunter B Fraser (Apr. 2016). "Pooled ChIP-Seq Links Variation in Transcription Factor Binding to Complex Disease Risk". en. In: *Cell* 165.3, pp. 730–741.

Thomas-Chollier, Morgane, Andrew Hufton, Matthias Heinig, Sean O'keeffe, Nassim El Masri, Helge G Roider, Thomas Manke, and Martin Vingron (2011). "Transcription factor binding predictions using TRAP for the analysis of ChIP-seq data and regulatory SNPs". In: *Nature protocols* 6.12, p. 1860.

Tibshirani, Robert (1996). "Regression shrinkage and selection via the lasso". In: *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 267–288.

Trapnell, Cole, Davide Cacchiarelli, Jonna Grimsby, Prapti Pokharel, Shuqiang Li, Michael Morse, Niall J Lennon, Kenneth J Livak, Tarjei S Mikkelsen, and John L Rinn (2014). "The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells". In: *Nature biotechnology* 32.4, p. 381.

Turing, Alan (1950). "Computing machinery and intelligence". In: *Mind* 59.236, p. 433.

Vallejos, Catalina A, John C Marioni, and Sylvia Richardson (June 2015). "BASiCS: Bayesian Analysis of Single-Cell Sequencing Data". en. In: *PLoS Comput. Biol.* 11.6, e1004333.

Vallejos, Catalina A, Davide Risso, Antonio Scialdone, Sandrine Dudoit, and John C Marioni (June 2017). "Normalizing single-cell RNA sequencing data: challenges and opportunities". en. In: *Nat. Methods* 14.6, pp. 565–571.

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin (2017). "Attention is all you need". In: *Advances in neural information processing systems*, pp. 5998–6008.

Velten, Lars, Simon F Haas, Simon Raffel, Sandra Blaszkiewicz, Saiful Islam, Bianca P Hennig, Christoph Hirche, Christoph Lutz, Eike C Buss, Daniel Nowak, et al. (2017). "Human haematopoietic stem cell lineage commitment is a continuous process". In: *Nature cell biology* 19.4, p. 271.

Venter, J Craig, Mark D Adams, Eugene W Myers, Peter W Li, Richard J Mural, Granger G Sutton, Hamilton O Smith, Mark Yandell, Cheryl A Evans, Robert A Holt, et al. (2001). "The sequence of the human genome". In: *science* 291.5507, pp. 1304–1351.

Vieth, Beate, Swati Parekh, Christoph Ziegenhain, Wolfgang Enard, and Ines Hellmann (2019). "A systematic evaluation of single cell RNA-seq analysis pipelines". In: *Nature communications* 10.1, pp. 1–11.

Vincent, Pascal, Hugo Larochelle, Yoshua Bengio, and Pierre-Antoine Manzagol (2008). "Extracting and composing robust features with denoising autoencoders". In: *Proceedings of the 25th international conference on Machine learning*. ACM, pp. 1096–1103.

Visscher, Peter M, Naomi R Wray, Qian Zhang, Pamela Sklar, Mark I McCarthy, Matthew A Brown, and Jian Yang (2017). "10 years of GWAS discovery: biology, function, and translation". In: *The American Journal of Human Genetics* 101.1, pp. 5–22.

Võsa, Urmo, Annique Claringbould, Harm-Jan Westra, Marc Jan Bonder, Patrick Deelen, Biao Zeng, Holger Kirsten, Ashis Saha, Roman Kreuzhuber, Silva Kasela, et al. (2018). "Unraveling the polygenic architecture of complex traits using blood eQTL metaanalysis". In: *BioRxiv*, p. 447367.

Vu, Trung Nghia, Quin F Wills, Krishna R Kalari, Nifang Niu, Liewei Wang, Mattias Rantalainen, and Yudi Pawitan (2016). "Beta-Poisson model for single-cell RNA-seq data analyses". In: *Bioinformatics* 32.14, pp. 2128–2135.

Wagner, Allon, Aviv Regev, and Nir Yosef (2016). "Revealing the vectors of cellular identity with single-cell genomics". In: *Nature biotechnology* 34.11, p. 1145.

Wang, Gaofeng, Joelle M van der Walt, Gregory Mayhew, Yi-Ju Li, Stephan Züchner, William K Scott, Eden R Martin, and Jeffery M Vance (2008). "Variation in the miRNA-433 binding site of FGF20 confers risk for Parkinson disease by overexpression of α-synuclein". In: *The American Journal of Human Genetics* 82.2, pp. 283–289.

Wang, Jingshu, Divyansh Agarwal, Mo Huang, Gang Hu, Zilu Zhou, Chengzhong Ye, and Nancy R Zhang (2019). "Data denoising with transfer learning in single-cell transcriptomics". In: *Nature methods* 16.9, pp. 875–878.

Wang, Kai, Mingyao Li, and Hakon Hakonarson (Sept. 2010). "ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data". en. In: *Nucleic Acids Res.* 38.16, e164.

Wang, Xiao, William E Allen, Matthew A Wright, Emily L Sylwestrak, Nikolay Samusik, Sam Vesuna, Kathryn Evans, Cindy Liu, Charu Ramakrishnan, Jia Liu, et al. (2018). "Three-dimensional intact-tissue sequencing of single-cell transcriptional states". In: *Science* 361.6400.

Wang, Xinchen, Nathan R Tucker, Gizem Rizki, Robert Mills, Peter Hl Krijger, Elzo de Wit, Vidya Subramanian, Eric Bartell, Xinh-Xinh Nguyen, Jiangchuan Ye, Jordan Leyton-Mange, Elena V Dolmatova, Pim van der Harst, Wouter de Laat, Patrick T Ellinor, Christopher Newton-Cheh, David J Milan, Manolis Kellis, and Laurie A Boyer (May 2016). "Discovery and validation of sub-threshold genome-wide association study loci using epigenomic signatures". en. In: *Elife* 5.

Wang, Xiuchao, He Ren, Tiansuo Zhao, Weidong Ma, Jie Dong, Shengjie Zhang, Wen Xin, Shengyu Yang, Li Jia, and Jihui Hao (2016). "Single nucleotide polymorphism in the microRNA-199a binding site of HIF1A gene is associated with pancreatic ductal adenocarcinoma risk and worse clinical outcomes". In: *Oncotarget* 7.12, p. 13717.

Ward, Lucas D and Manolis Kellis (Jan. 2016). "HaploReg v4: systematic mining of putative causal variants, cell types, regulators and target genes for human complex traits and disease". en. In: *Nucleic Acids Res.* 44.D1, pp. D877–81.

Weber, James L and Eugene W Myers (1997). "Human whole-genome shotgun sequencing". In: *Genome research* 7.5, pp. 401–409.

Wichmann, H-E, Christian Gieger, and Thomas Illig (2005). "KORA-gen-resource for population genetics, controls and a broad spectrum of disease phenotypes". In: *Das Gesundheitswesen* 67.S 01, pp. 26–30.

Wolf, F Alexander, Philipp Angerer, and Fabian J Theis (Feb. 2018). "SCANPY: large-scale single-cell gene expression data analysis". en. In: *Genome Biol.* 19.1, p. 15.

Wray, Naomi R, Stephan Ripke, Manuel Mattheisen, Maciej Trzaskowski, Enda M Byrne, Abdel Abdellaoui, Mark J Adams, Esben Agerbo, Tracy M Air, Till MF Andlauer, et al. (2018). "Genome-wide association analyses identify 44 risk variants and refine the genetic architecture of major depression". In: *Nature genetics* 50.5, p. 668.

Wu, Yonghui, Mike Schuster, Zhifeng Chen, Quoc V Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, et al. (2016). "Google's

neural machine translation system: Bridging the gap between human and machine translation". In: *arXiv preprint arXiv:1609.08144*.

Yu, Guangchuang, Li-Gen Wang, and Qing-Yu He (2015). "ChIPseeker: an R/Bioconductor package for ChIP peak annotation, comparison and visualization". In: *Bioinformatics* 31.14, pp. 2382–2383.

Zappia, Luke, Belinda Phipson, and Alicia Oshlack (Sept. 2017). "Splatter: simulation of single-cell RNA sequencing data". en. In: *Genome Biol.* 18.1, p. 174.

Zhao, Xu, Qing Ye, Kang Xu, Jinluo Cheng, Yanqin Gao, Qian Li, Juan Du, Hui Shi, and Ling Zhou (2013). "Single-nucleotide polymorphisms inside microRNA target sites influence the susceptibility to type 2 diabetes". In: *Journal of human genetics* 58.3, pp. 135–141.

Zheng, Grace X Y, Jessica M Terry, Phillip Belgrader, Paul Ryvkin, Zachary W Bent, Ryan Wilson, Solongo B Ziraldo, Tobias D Wheeler, Geoff P McDermott, Junjie Zhu, Mark T Gregory, Joe Shuga, Luz Montesclaros, Jason G Underwood, Donald A Masquelier, Stefanie Y Nishimura, Michael Schnall-Levin, Paul W Wyatt, Christopher M Hindson, Rajiv Bharadwaj, Alexander Wong, Kevin D Ness, Lan W Beppu, H Joachim Deeg, Christopher McFarland, Keith R Loeb, William J Valente, Nolan G Ericson, Emily A Stevens, Jerald P Radich, Tarjei S Mikkelsen, Benjamin J Hindson, and Jason H Bielas (Jan. 2017). "Massively parallel digital transcriptional profiling of single cells". en. In: *Nat. Commun.* 8, p. 14049.

Zhou, Jian, Christopher Park, Chandra Theesfeld, Yuan Yuan, Kirsty Sawicka, Jennifer Darnell, Claudia Scheckel, John Fak, Yoko Tajima, Robert Darnell, et al. (2018). "Whole-genome deep learning analysis reveals causal role of noncoding mutations in autism". In: *bioRxiv*, p. 319681.

Zhou, Jian, Chandra L Theesfeld, Kevin Yao, Kathleen M Chen, Aaron K Wong, and Olga G Troyanskaya (2018). "Deep learning sequence-based ab initio prediction of variant effects on expression and disease risk". In: *Nature genetics* 50.8, p. 1171.

Zhou, Jian and Olga G Troyanskaya (2015). "Predicting effects of noncoding variants with deep learning–based sequence model". In: *Nat. Methods* 12.10, pp. 931–934.

Zhou, Xiaopu, Yu Chen, Kin Y Mok, Timothy CY Kwok, Vincent CT Mok, Qihao Guo, Fanny C Ip, Yuewen Chen, Nandita Mullapudi, Paola Giusti-Rodríguez, et al. (2019). "Non-coding variability at the APOE locus contributes to the Alzheimer's risk". In: *Nature communications* 10.1, pp. 1–16.

Ziegenhain, Christoph, Beate Vieth, Swati Parekh, Björn Reinius, Amy Guillaumet-Adkins, Martha Smets, Heinrich Leonhardt, Holger Heyn, Ines Hellmann, and Wolfgang Enard (2017). "Comparative analysis of single-cell RNA sequencing methods". In: *Molecular cell* 65.4, pp. 631–643.