Interpreting gene expression changes in single cells and during ageing using transcriptional regulatory networks

Inaugural-Dissertation

zur
Erlangung des Doktorgrades
der Mathematisch-Naturwissenschaftlichen Fakultät
der Universität zu Köln

vorgelegt von

Ana Carolina Leote

geboren in Lisboa

Köln, 2023

Berichterstatter:

Prof. Dr. Andreas Beyer

Dr. Peter Tessarz

Tag der mündlichen Prüfung: 23/11/2022

Table of Contents

ACKNOWLEDGEMENTS VI

ERKLÄRUNG ZUR DISSERTATION VII

ABSTRACT VIII

1. GENERAL INTRODUCTION 1

1.1 Regulation of gene expression 1
1.1.1 Transcription 1
1.1.2 Transcriptional regulation of gene expression 1
1.1.3 Non-transcriptional regulation of gene expression 2

1.2 Inference of gene regulatory relationships 3
1.2.1 Inference of gene regulatory relationships based on transcriptome data 3
1.2.2 Integrative inference of gene regulatory relationships 4

1.3 Single-cell RNA-sequencing 5
1.3.1 Overview of single-cell RNA-sequencing methodologies 5
1.3.2 Computational challenges of single-cell RNA-sequencing data analysis 6

1.4 Gene regulation changes during ageing 7
1.4.1 Cell-to-cell variability 7
1.4.2 Transcriptomic coordination 8
1.4.3 Potential mechanisms of gene regulation changes during ageing 8

2. AIMS OF THE PROJECT 9

3. REGULATORY NETWORK-BASED IMPUTATION OF DROPOUTS IN SINGLE-CELL RNA SEQUENCING DATA 10

3.1 Introduction 10

3.2 Results 13
3.2.1 Imputing dropouts using a transcriptional regulatory network 13
3.2.2 Transcriptional regulatory network information improves scRNA-seq dropout imputation 14
3.2.3 Gene features determine the best performing imputation method 17
3.2.4 Network-based imputation aids downstream analyses 19
Cell trajectory inference 19
Data visualization 19
Cluster marker detection 20
Determining transcriptional regulators 21

3.3 Methods 24

Pre-processing of cancer cell line data for transcriptional regulatory network inference 24
Network inference based on stability selection 24
Network validation using TCGA and GTEx data 24
Single-cell test data processing 25
Dropout imputation 25
Masking procedures 26
Imputation performance analysis 26
Top performing methods 27
Cell trajectory inference 27
Dimensionality reduction 27
Marker detection 27
GO term enrichment and transcription factor analyses 28
Determination of the optimal imputation method per gene 28
The ADImpute R package 28

3.4 Discussion **30**

3.5 Contributions **32**

3.6 Supporting Information **33**

4. AGE-RELATED CHANGES IN GENE EXPRESSION REGULATION ACROSS HUMAN TISSUES 50

4.1 Introduction **50**

4.2 Results **51**
4.2.1 Capture of gene-gene relationships from transcriptome data 51
4.2.2 Age-related changes in gene-gene relationships across tissues 52
4.2.3 Regulatory relationship changes with age 56

4.3 Methods **61**

Network processing **61**

Genotype-Tissue Expression (GTEx) data preprocessing **61**
Data download 61
Data preprocessing 61
Subsetting 61
Expression profile reconstruction using the regulatory model 62
Identification of poorly predicted genes 62
Gene filtering based on average predictability across age groups 62
Background distribution of predictability slopes 63
Network propagation of predictability slopes 63
Gene Set Enrichment Analysis 63
Quantification of age-related expression and variance changes 63

4.4 Discussion **64**

4.5 Contributions **66**

4.6 Supporting Information **67**

5. GENERAL DISCUSSION 75

5.1 Similarity of gene regulatory relationships across tissues and cell types **75**
  5.1.1 Limitations 75
  5.1.2 Future work 76

5.2 Application of gene regulatory relationships to single-cell RNA-sequencing data processing **76**
  5.2.1 Limitations 77
  5.2.2 Future prospects 77

5.3 Gene expression regulation changes during ageing **77**
  5.3.1 Limitations 77
  5.3.2 Future prospects 78

5.4 Concluding remarks **78**

BIBLIOGRAPHY 79

> “Science is not only a discipline of reason, but also one of romance and passion.”
>
> - Stephen Hawking (1942-2018)

Acknowledgements

My first word of thanks goes to my supervisor Prof. Dr. Andreas Beyer, whose constant availability and scientific curiosity made this effort lighter. I am grateful for the many scientific discussions we had over pizza and coffee and for the feedback on analyses given at 10pm. I am also extremely thankful for the time and thought invested in advising me professionally beyond the everyday science. The thoughtfulness and scientific rigor I witnessed have been and will remain a source of inspiration throughout my scientific career.

I also express my sincere gratitude to Dr. Peter Tessarz for his helpful comments on the progress of this project as a member of my Thesis Advisory Committee, as well as casual scientific interactions that I greatly appreciated. I thank Dr. Adam Antebi as well, for his input on this project as a member of my Thesis Advisory Committee. An additional word of appreciation goes to Dr. Matthias Weith, Dr. Jan Grossbach, Francisco Lopes and Taoyu Mei, who contributed to the projects described in this dissertation with ideas, interpretations and concerns.

I am thankful to the CGA coordinators Jenny Ostermann, Dr. Daniela Morick, Dr. Julia Zielinski, Saskia Wilming and Dr. Doris Birker for their enormous help with the logistic aspects of moving from a foreign country and radically simplifying many of the interactions with the University. I also thank the faculty members of the CGA for sharing their expertise and providing valuable scientific advice within the program.

It is not only the science, but also the people, that bring colour and character to the years I've dedicated to the projects described in this dissertation. In that respect, I was fortunate to be surrounded by the supportive and helpful people that have been part of the Beyer group through these years. In all their diversity, they've greatly helped me feel integrated and inspired me in countless personal and professional aspects. A special word of gratitude goes to Jan Grossbach, Luise Nagel, Oliver Hahn and Tim Padvitski, for their thoughtfulness, support, friendship, and captivating conversations, both about scientific and personal aspects. I've been also fortunate to be part of the Cologne Graduate School for Ageing Research (CGA) and surrounded by excellent and supportive fellow doctoral students who have been a constant source of inspiration. A special mention goes to Hafiza Alirzayeva, Paulo Lopes and Stephanie Fernandes for their warm friendship and support.

Finally, my most heartfelt gratitude goes to my family and partner. Their unconditional support and love are the strongest foundations for the personal and professional growth I've experienced during my doctoral studies.

"No man is an island,
Entire of itself.
Each is a piece of the continent,
A part of the main.
If a clod be washed away by the sea,
Europe is the less.
As well as if a promontory were.
As well as if a manor of thine own
Or of thine friend's were.
Each man's death diminishes me,
For I am involved in mankind.
Therefore, send not to know
For whom the bell tolls,
It tolls for thee."

John Donne (1572-1631)

Erklärung zur Dissertation

Hiermit versichere ich an Eides statt, dass ich die vorliegende Dissertation selbstständig und ohne die Benutzung anderer als der angegebenen Hilfsmittel und Literatur angefertigt habe. Alle Stellen, die wörtlich oder sinngemäß aus veröffentlichten und nicht veröffentlichten Werken dem Wortlaut oder dem Sinn nach entnommen wurden, sind als solche kenntlich gemacht. Ich versichere an Eides statt, dass diese Dissertation noch keiner anderen Fakultät oder Universität zur Prüfung vorgelegen hat; dass sie - abgesehen von unten angegebenen Teilpublikationen und eingebundenen Artikeln und Manuskripten - noch nicht veröffentlicht worden ist sowie, dass ich eine Veröffentlichung der Dissertation vor Abschluss der Promotion nicht ohne Genehmigung des Promotionsausschusses vornehmen werde. Die Bestimmungen dieser Ordnung sind mir bekannt. Darüber hinaus erkläre ich hiermit, dass ich die Ordnung zur Sicherung guter wissenschaftlicher Praxis und zum Umgang mit wissenschaftlichem Fehlverhalten der Universität zu Köln gelesen und sie bei der Durchführung der Dissertation zugrundeliegenden Arbeiten und der schriftlich verfassten Dissertation beachtet habe und verpflichte mich hiermit, die dort genannten Vorgaben bei allen wissenschaftlichen Tätigkeiten zu beachten und umzusetzen. Ich versichere, dass die eingereichte elektronische Fassung der eingereichten Druckfassung vollständig entspricht.

Teilpublikationen:

Leote AC, Wu X, Beyer A (2022) Regulatory network-based imputation of dropouts in single-cell RNA sequencing data. PLOS Computational Biology 18(2): e1009849.
https://doi.org/10.1371/journal.pcbi.1009849

Leote AC, Lopes F, Beyer A (2023) Age-related changes in gene expression regulation across human tissues (Manuscript in preparation)

Ana Carolina Leote

(Ana Carolina Leote)

30/11/2023, South San Francisco, CA, USA

Abstract

Regulation of gene expression allows for cells to execute highly specialized functions and orchestrate a variety of responses to stimuli. Impaired coordination of gene expression is linked to several diseases – including ageing – highlighting the relevance of comprehensively understanding gene expression coordination. However, current models of gene expression coordination are often limited to specific biological conditions or focused on individual cellular processes.

Here, we report a transcriptional regulatory model, derived from more than 1000 expression datasets, able to capture tissue- and cell-type-specific gene-gene relationships, as well as global (cross-tissue) relationships.

We take advantage of the wide applicability of this model to tackle two distinct biological problems. The first concerns the computational estimation of missing values in single-cell RNA-sequencing (scRNA-seq) data. scRNA-seq methods are typically unable to quantify the expression levels of all genes in a cell (dropout events). Several methods have been proposed for the estimation of dropout expression (dropout imputation), with no clear method outperforming others across datasets and downstream analyses. We propose a new method that makes use of the transcriptional regulatory model to estimate the expression of dropouts and show it outperforms published state-of-the-art methods, especially for lowly expressed genes, including cell-type-specific transcriptional regulators. We observed gene- and dataset-dependent performance of the methods we tested, leading us to implement an R package, ADImpute, that automatically determines the best imputation method for each gene in a dataset. This work represents a paradigm shift by demonstrating that there is no single best imputation method. Instead, we propose that imputation should maximally exploit external information and be adapted to gene-specific features, such as expression level and expression variation across cells.

The second problem addressed in this work was the impact of ageing on gene expression coordination. We use existing RNA-seq data of human tissues at different ages to investigate the impact of ageing on the gene-gene relationships captured in the transcriptional regulatory model. We observed age-related changes towards both a strengthening and a loosening of gene-gene relationships with age, mostly impacting genes with mitochondrial functions and cell cycle regulation. We detected age-related changes in relationships between genes involved in the same functional module, as well as genes in distinct functional modules, highlighting the impact of ageing on the coordination of cellular processes. This work demonstrates the importance of zooming out of the effect of ageing on individual genes or cellular processes and investigating how their crosstalk is affected at a systems level.

Put together, the work presented here shows that transcriptional gene-gene relationships can be ‘learned’ from a rather limited set of example datasets and subsequently applied to a wide range of cell- and tissue types, where pathological breakdown of gene-gene relationships can be investigated.

1. General Introduction

1.1 Regulation of gene expression

Although DNA content is practically the same across all cells in an organism, different cell types can execute different functions. This is possible due to an extensive regulation of the several processes involved in the synthesis of a functional product from the information encoded in the respective DNA region. This chapter describes such regulatory processes in eukaryotes with a focus on the first step of genetic information flow in the cell: transcription.

1.1.1 Transcription

Transcription consists of the synthesis of an RNA molecule (transcript) from a gene, i.e., a DNA region coding either for the mRNA encoding a polypeptide chain or for a functional RNA molecule. This process is executed by the RNA polymerase enzyme in three stages. Firstly, during initiation, the RNA polymerase binds a DNA region upstream of the gene body, the promoter, and unwinds the double helix structure of the DNA. Secondly, during elongation, the RNA polymerase uses one of the DNA strands as a template for the synthesis of the nascent RNA, one ribonucleotide at a time. Lastly, termination involves the release of the polymerase and the nascent RNA from the DNA template. Depending on the type of gene being transcribed (and thus RNA polymerase being used), termination can occur once a termination sequence in the DNA is encountered or through more complex mechanisms related to 3’ processing and not yet completely understood [1]. As it is transcribed, the nascent RNA undergoes processing through 3’ capping, 5’ polyadenylation and splicing. The resulting messenger RNA (mRNA) is then often translated into a protein [2].

1.1.2 Transcriptional regulation of gene expression

Transcription, while effected by RNA polymerase, is under intricate regulation. Much of this regulation is mediated by transcription factors (TFs), proteins with the ability to bind specific DNA sequences (motifs). TF binding to motifs within the promoter enables transcription, but motifs can also be found in distal regions. For instance, TF binding to enhancer regions can lead to the formation of DNA loops that bring the TF physically close to the promoter, with a regulatory effect on transcription (by stabilizing RNA polymerase binding or other mechanisms) [3].

Regulatory relationships between transcription factors and their targets are remarkably complex. Firstly, a single TF usually regulates the transcription of more than one gene. Secondly, TF activity is itself regulated as well, for instance through post-translational modifications in response to stimuli, which can change TF subcellular localization, stability and protein-protein interactions [4]. Thirdly, TFs are known to collaborate in multiple ways, either indirectly or through protein-protein interactions. Examples of this behaviour include the formation of TF dimers with a higher binding affinity to DNA, or a cooperative impact on local chromatin structure that facilitates exposure of DNA for binding [5].

In addition to the activity of transcription factors, epigenetic modifications, such as DNA methylation and histone modifications, also contribute to transcriptional regulation, due to their impact on DNA accessibility to TFs and the transcriptional machinery. DNA methylation, that is, the methylation of cytosines – mostly located 5' to a guanine (CpGs) – can have different impacts on expression depending on the position relative to the gene body [6]. For instance, promoter methylation mostly leads to decreased expression of the affected genes [7], [8]. However, methylation at the gene body has been reported to lead to both repression [9] and activation of gene expression [10]–[13].

On the other hand, histone proteins can also be modified, with impact on the conformation of chromatin. Chromatin is a 3D DNA-protein complex, whose repeating unit is the nucleosome, an octamer of histone proteins with DNA tightly coiled around it. Nucleosomes are further coiled into higher order structures of varying accessibility. Both this higher order structure and the organization into nucleosomes pose a physical barrier for the transcription factor machinery and TFs to contact with the DNA and carry out transcription. Histones possess an exposed 'tail' of residues which can be suffer post-translational modifications, such as acetylation, methylation or phosphorylation, with diverse consequences on chromatin structure. For instance, acetylation of histone H4 on lysine 16 (H4K16ac) leads to more lose chromatin [14], which in turn leads to increased accessibility and transcription. In addition, histone modifications also contribute to recruit or disturb the binding of specific proteins to chromatin [15], [16].

DNA methylation and chromatin conformation have also been shown to be linked. For instance, hypermethylated regions that are transcriptionally inactive also show histone modifications leading to a more compact chromatin conformation, and DNA methyltransferases are known to interact with histone modifying enzymes [17].

In summary, transcription factor activity, chromatin conformation and DNA methylation play a role in regulating gene expression. The impact of each of these factors on gene expression is not independent, greatly increasing the complexity of the gene regulatory landscape.

1.1.3 Non-transcriptional regulation of gene expression

The mechanisms of gene expression regulation covered so far focused on the regulation of the process of transcription. However, gene expression can also be regulated at post-transcriptional, translational or post-translational stages through diverse mechanisms. This is well exemplified by the inclusion of intronic sequences in mRNAs through alternative splicing, often leading to the presence of premature termination codons in the mRNA and its subsequent degradation [18]. Another example is the regulatory role of microRNAs (miRNAs), small non-coding RNAs with regulatory roles mostly mediated by their binding to the 3' untranslated region (UTR) of mRNAs to induce their degradation or repress their translation [19].

In this section, mechanisms of gene expression regulation were briefly covered, with a focus on transcriptional regulation. While many regulatory mechanisms are still not well understood, it is evident that final gene products (e.g., an active protein) are the result of a complex interplay between diverse fine-tuned mechanisms.

1.2 Inference of gene regulatory relationships

Due to the complexity of gene expression regulation, highlighted in Chapter 1, systems approaches are required to model regulatory relationships on a genome-wide scale. In this chapter, different data and methodologies used for the inference of gene regulatory relationships are presented and compared, with an emphasis on regularized linear regression applied to transcriptome data.

1.2.1 Inference of gene regulatory relationships based on transcriptome data

Due to the increasing accessibility of transcriptome profiling data, mainly from RNA-sequencing (RNA-seq), approaches to infer gene regulatory relationships often rely on this source of data alone. The underlying assumption of these approaches is that similar profiles of expression are indicative of a similar underlying mechanism of regulation.

Gene co-expression networks [20] are the most common of such approaches and are frequently applied to predict gene functions or find putative pathway members. Such methods are based on metrics of similarity, computed between the expression profile of each possible gene pair. Although here we focus on the use of transcriptome data, this approach can also be applied to other molecular layers, such as protein abundances [21] or metabolic data. Pearson correlation and Mutual Information (MI) are common choices of similarity metric for gene co-expression network inference. Multiple of the tools readily available for network inference are based in MI, including the well-known ARACNE algorithm [22]. While MI provides the advantage of detecting non-linear relationships between the expression profiles, comparison of co-expression measures concluded it provides little advantage over correlation [23], [24], while having the disadvantage of requiring larger sample sizes and being computationally more challenging.

However, the analysis of pairwise similarity presents limitations. Firstly, gene expression is regulated by a combination of multiple factors, but gene-gene similarity is limited to independent analysis of gene pairs. Secondly, metrics of similarity cannot be used for quantitative predictions.

The use of linear regression approaches can overcome these limitations by modelling quantitatively the impact of multiple genes (predictors) on the expression of another (target) gene. In this case, it is assumed that the contribution of each predictor is independent, and that the expression profile of the target gene is a weighted linear combination of the expression profiles of the predictor genes. Due to the vast search space for possible predictor genes (the entire transcriptome), this approach is usually combined with feature selection. Feature selection can be based on external information (for instance, by limiting potential predictors to known transcription factors) or in a data-driven way (for instance, using regularization approaches). Regularized regression adds additional constraints to the least squares problem in linear regression. In particular, Lasso regression adds a penalty on the sum of the absolute values of the linear regression coefficients, setting most to zero [25]. The obtained linear model is thus very sparse and contains only the most predictive explanatory variables. Thus, regularized regression can be used for feature selection, and has been employed in the learning of gene regulatory relationships [26], [27].

Additionally, the incorporation of stability selection [28], consisting of several runs of the linear model adjustment, each with a random subset of the training data and the explanatory variables, has been shown to improve performance [29]. This improved performance can be explained by the fact that this procedure results in the selection of robust predictors that are informative through variations in the training data.

The models described above have the limitation that they do not reflect direct regulatory relationships between genes. While computational approaches have been proposed to remove indirect relationships between genes, this is more commonly achieved by the addition of other molecular layers of information beyond transcript abundance.

1.2.2 Integrative inference of gene regulatory relationships

The role of the epigenetic landscape in gene expression regulation, highlighted in section 1.1.2, can be leveraged to improve gene regulatory relationship inference.

Transcription factor binding events can be captured through chromatin immunoprecipitation, followed by sequencing (ChIP-seq) [30]. Briefly, proteins transiently bound to the DNA are fixed to it through a crosslinking step. This is followed by the fragmentation of chromatin and immunoprecipitation, where an antibody against the protein of interest (a given TF) is used to isolate its target TF and the DNA fragments it is bound to. The resulting DNA fragments then undergo next generation sequencing. This procedure makes it possible to identify transcription factor binding events in the studied conditions. While co-expression can be observed for genes that are not under the same regulator (transcription factor), the observation of binding of that TF to binding sites upstream of the target genes allows to infer direct regulation [31].

Together with other epigenomic profiling methods [32], ChIP-seq data have been used to inform on gene regulation in combination with transcriptome data [33]. One disadvantage of this approach is that a large fraction of the binding events detected (30-40%) is context-dependent and thus cannot be extrapolated to other cell types or conditions [34].

In this section, different approaches to model gene regulatory relationships have been presented and compared. While this is not a comprehensive review of existing literature, it provides an overview of the most common approaches.

1.3 Single-cell RNA-sequencing

Transcriptome profiling is a commonly employed technique to understand cellular and tissue activity at a molecular level. On one hand, it offers a dynamic picture of activity when compared to the static measurements of genomic data. On the other hand, RNA-profiling technologies allow for a more comprehensive picture of the molecular composition of the cells and tissues at hand than current state-of-the-art protein expression profiling. In this section, an overview of different methodologies for profiling the transcriptome of individual cells is provided, along with a discussion of common challenges in the computational analysis of the resulting data.

1.3.1 Overview of single-cell RNA-sequencing methodologies

Several methods have been proposed for profiling the transcriptome of individual cells ([35]–[38], among many others, reviewed in [39]–[41]), each with their individual strengths and preferred applications. Despite their differences, all these methods follow common steps. Firstly, cells must be isolated from their tissue of origin, and then lysed to release their RNA content. This is followed by RNA capture, usually by poly(A)+ selection, which targets mRNA, or ribosomal RNA depletion, to remove the rRNA fraction (over 80% of the RNA molecules in a cell) [42]. This step ensures that sequencing efforts are focused on the RNA species of interest. RNA capture efficiency is consistently low across different methods – around 10-20% – which leads to the loss of many of the RNA molecules originally present in the cell. Upon capture, RNA molecules are reverse-transcribed into cDNA, a more stable molecule. cDNA fragments are then amplified via PCR or *in vitro* transcription and tagmented into a cDNA library ready for sequencing.

Different approaches exist for the first step of isolation of individual cells. The most popular scRNA-seq methods are often based on flow-activated cell sorting (FACS) or microdroplet-based microfluidics. FACS-based approaches allow for the isolation of cells of interest using fluorescently labelled antibodies against cell-surface markers of interest. Droplet-based approaches rely on dispersing aqueous droplets, containing individual cells and the necessary reagents for cell lysis, reverse transcription and amplification, in an oil phase within a microfluidics device [43], [44]. Whereas droplet-based approaches have the disadvantage of requiring a dedicated platform, they allow for a higher cell throughput than FACS-based approaches. Additionally, the use of microdroplets may lead to issues in downstream analysis, related to the presence of more than one cell per droplet, or contamination from ambient RNA.

Another main difference between scRNA-seq methods, with impact in the distribution of the resulting data, is the inclusion of Unique Molecular Identifiers (UMIs). UMIs consist of short random sequences of nucleotides which are ligated to the cDNA prior amplification and allow to track, throughout all downstream analyses, the identity of the original RNA molecule [36], [45]. Upon sequencing, the different fragments generated by amplification can be collapsed back into the original molecule, avoiding excessive technical noise originating from the amplification step.

The use of UMIs is commonly paired with 3' end sequencing of the cDNA fragments, as to capture the UMI sequence. This comes at the expense of full-length coverage, which allows for the distinction between different transcript isoforms, or identification of RNA editing events. Notably, the recently developed Smart-seq3 protocol combines UMI use with full-length sequencing [46].

1.3.2 Computational challenges of single-cell RNA-sequencing data analysis

Several challenges in the analysis of scRNA-sequencing data depend on the methodology used to generate them [47]. Library preparation steps can have a downstream consequence on the analysis steps, such as the use of droplets to encapsulate individual cells *versus* other isolation techniques, the use of UMIs, and sequencing of the 3' end *versus* full-length of the cDNA fragment.

The use of UMIs can avoid amplification bias resulting from PCR amplification of small quantities of RNA. Discussions in the field [48]–[50] have resulted in the consensus that UMI data does not suffer from zero inflation (i.e., an excessive amount of zeros compared to the negative binomial model), whereas data without UMIs is best modelled by zero-inflated negative binomial distributions. In fact, such distributions have been used by several methods for the statistical analysis of scRNA-seq data ([51], [52], among others) and are one approach to handle the excessive amount of dropouts. Another approach is to computationally estimate the true expression levels of genes that drop out. This is complicated by the fact that some genes are not expressed in the cell (biological zeros), while others are expressed but not captured (technical zeros). Dropout imputation is introduced and discussed in greater detail in Chapter 3.

Other challenges in the analysis of scRNA-seq data are common throughout different sequencing methodologies. This is the case for biological sources of variation between cells of the same cell type. Among these sources are differences in the cell cycle stage of individual cells, or transcriptional bursts, corresponding to a stochastic activation and inactivation of transcription.

Existing work presented in this section highlights how the advent of scRNA-sequencing motivated the development of a growing body of computational tools to tackle several of the challenges associated to these data. In Chapter 3, original work adding to this community effort is presented.

1.4 Gene regulation changes during ageing

Ageing is the primary risk factor for multiple diseases, making it an attractive biomedical target [53]. Given the ease with which transcriptomes can currently be profiled, a wide range of research has been focused on the impact of ageing on the transcriptome of individual tissues and cells. An emerging observation from these studies is that ageing-related changes in the transcriptome seem to be tissue-specific [54]. However, global trends have also been uncovered, including a downregulation of genes encoding for mitochondrial proteins and protein synthesis machinery, a dysregulation of immune system genes, reduction in growth factor signalling, constitutive responses to stress and DNA damage and a dysregulation of gene expression and mRNA processing [55]–[58]. In addition to average expression level changes with age, an age-related increase in expression heterogeneity has also been repeatedly reported. This is commonly referred to as ‘transcriptional noise’ and encompasses different concepts explored by different groups, which are explored in this section.

1.4.1 Cell-to-cell variability

The first type of ‘noise’ corresponds to expression variability between cells, either at the level of individual genes or the whole transcriptome. Examples of noise quantification per gene can be found in earlier research work, dating back to 2006. Bahar and colleagues first reported an increase in cell-to-cell variation of the expression levels of a panel of genes, quantified by RT-PCR from individual cardiomyocytes isolated from young (6 months) and old (27 months) mice [59]. The considered genes served diverse functions: seven housekeeping genes, three heart-specific genes, two protease-encoding genes and three mitochondrial genes. The authors reported an increase in cell-to-cell variation of expression levels of all nuclear genes tested, measured as variance between cells isolated from mice of each age group. Later, Warren and colleagues built upon this work by quantifying expression levels of six medium-to-high abundance genes with qRT-PCR in four hematopoietic cell types, purified with flow cytometry from young and old mice. The authors quantified the cell-to-cell variability of each gene with different metrics, such as the coefficient of variation (CV), the inter-quartile range (IQR) and the geometric standard deviation. When comparing cell-to-cell variability of the chosen genes in these cell types, the authors found no difference with age. Furthermore, when quantifying the coupling of the expression levels of correlated ribosomal genes Rpl5 and Rpl19, the authors found no age-related de-coupling of expression [60].

With the advent of high-throughput sequencing technologies at the single cell level, it became possible to investigate cell-to-cell variability in expression for the whole transcriptome rather than only a small panel of genes. Enge and colleagues were the first to take advantage of these data for the interrogation of cell-to-cell expression variability in 2017. Among different approaches used by the authors to quantify noise is one based on the similarity (Euclidean distance or Pearson correlation) of the whole transcriptome of each individual cell compared to the mean of the cell type. This similarity is computed based on the mRNA molecules (biological noise) and ERCC spike-ins (technical noise) to compute a ratio of biological to technical noise. The authors applied this approach to scRNA-seq data of pancreata from eight human donors of ages 1 month to 54 years and reported increased noise levels in the old adult group compared to children and young adults [61]. Angelidis and colleagues later applied a similar approach to scRNA-seq data from lungs of young (3 months) and old (24 months) mice, reporting an increase in transcriptional noise with age in most lung cell types [62]. More recently, Marti and colleagues developed a parametric approach to separate technical and biological noise in scRNA-seq data and reported cases of both increase (in genes involved in antigen presenting) and decrease (for genes involved in oxidative phosphorylation) in transcriptional noise with age [63].

1.4.2 Transcriptomic coordination

A conceptually different definition of 'noise' is focused on the coordination of gene expression across the transcriptome. This has been addressed by Southworth and colleagues by comparing co-expression networks derived from young (16 month old) and old (24 month old) mice. The authors observed a trend towards correlation decline with age impacting modules of genes involved in ribosome biogenesis, transcriptional regulation and mitochondrial functions. The authors combined this co-expression analysis with a computational identification of targets of the transcription factor NK-KappaB, to show concerted downregulation of these targets [64]. More recently, Levy and colleagues made use of publicly available scRNA-seq data to interrogate changes in coordination of gene expression with age. To this end, the authors developed a transcriptome-wide metric of gene expression coordination, agnostic to gene-gene regulatory relationships and based on the average dependency levels between random gene subsets. This work also revealed a decrease in transcriptional coordination in aged cells across different organisms and cell types.

1.4.3 Potential mechanisms of gene regulation changes during ageing

Following reports of age-related increase in transcriptional noise, explored in the previous sections, significant effort has been made towards understanding the underlying mechanisms behind this observation.

One possibility explored in the literature is that the observed transcriptional noise stems from accumulated DNA damage in the affected genes. This possibility has been addressed by Enge and colleagues, who explored the link between mutational load and transcriptional noise using scRNA-seq data from the human pancreas. In this study, the authors found no link between mutation accumulation and transcriptional noise [61]. This is in contrast with the results reported by Levy and colleagues, showing a link between loss of transcriptional coordination and the accumulation of an age-related mutational signature [65].

Another possibility is that transcriptional noise originates from an age-related increase in epigenetic deregulation. Changes in epigenetic features such as DNA methylation, chromatin compactness and histone modifications have been reported with age (reviewed in [66]–[68]), with impact on gene regulation. Cheung and colleagues have addressed this topic by determining chromatin modifications at the single cell level using mass spectrometry. The authors report increased variability of chromatin modifications in older adults compared to young, and use a twin cohort to show that the observed age-related chromatin changes are mostly driven by non-heritable factors [69].

The literature mentioned in this section provides evidence for an age-related disruption of gene regulation. This can have multiple downstream consequences, such as the impairment of cell-type-specific responses [70] and loss of cellular identity [61], which compromise the function of individual cells and, as a consequence, of tissues and organisms.

2. Aims of the project

As described in Chapter 1, regulation of gene expression is fundamental for cellular integrity and function, and has been shown to be affected during ageing.

In this work, we make use of a transcriptional regulatory network to i) investigate the similarity of gene-gene relationships across tissues and cell types, ii) improve the estimation of uncaptured gene expression levels (dropout imputation) in scRNA-seq, and iii) explore gene expression regulation changes taking place during human ageing using bulk RNA-seq.

3. Regulatory network-based imputation of dropouts in single-cell RNA sequencing data

3.1 Introduction

Single-cell RNA sequencing (scRNA-seq) has become a routine method, revolutionizing our understanding of biological processes as diverse as tumor evolution, embryonic development, and ageing. However, current technologies still suffer from the problem that large numbers of genes remain undetected in single cells, although they actually are expressed (dropout events). Although dropouts are enriched among lowly expressed genes, relatively highly expressed genes can be affected as well. Of course, the dropout rate is also dependent on the sampling depth, i.e. the number of reads or transcript molecules (determined with unique molecular identifiers, UMIs) quantified in a given cell. Imputing dropouts is necessary for fully resolving the molecular state of the given cell at the time of the measurement. In particular, genes with regulatory functions - e.g. transcription factors, kinases, regulatory non-coding RNAs (ncRNAs) - are typically lowly expressed and hence particularly prone to be missed in scRNA-seq experiments. This poses problems for the interpretation of the experiments if one aims at understanding the regulatory processes responsible for the transcriptional makeup of the given cell.

The task of correctly imputing dropouts is further complicated by the fact that not all undetected genes are undetected for the same reasons. Some genes are originally expressed in the cell but fail to be detected due to incomplete RNA capturing. These are commonly referred to as technical dropouts. However, some genes are originally not expressed in the cell, and thus not detected (biological zeros). Biological zeros carry information about cell types and states, and incorrect estimation of non-zero expression in these cases may confound cellular profiles [71], [72]. Thus, computational methods for dropout imputation face two distinct challenges: on the one hand, to correctly call technical dropouts and, on the other hand, estimate their expression level. If not done carefully, dropout imputation can introduce false positive results in downstream analyses and amplify confounding signals such as batch effects [73].

Most dropout imputation methods are based on the underlying (explicit or implicit) assumption that detected and undetected genes are subject to the same regulatory processes, and hence detected genes can serve as a kind of 'fingerprint' of the state at which the cell was at the time of lysis. Several popular methods are based on some type of grouping (clustering) of cells based on the similarity of their expression patterns. Missing values are then imputed as a (weighted) average across those similar cells where the respective gene was detected [74]–[77]. For example, the Markov Affinity-based Graph Imputation of Cells (MAGIC) algorithm [74] creates a network of cells by linking cells with similar gene expression signatures. Missing values are subsequently imputed by computing an average over linked cells, where cells get weighted based on how similar or dissimilar their expression signatures are compared to the target cell. DrImpute [76], scImpute [75] and k-nearest neighbor smoothing (kNN-smoothing) [78] have further developed this notion and have been shown to outperform MAGIC in recent comparisons[73]. These methods rest on two important assumptions: (1) the global expression pattern of a cell (i.e. across the subset of detected genes) is predictive for all genes; (2) the (weighted) average of co-clustering (i.e. similar) cells is a good estimator of the missing value. The first assumption is violated if the expression of a dropout gene is driven by only a small subset of genes and hence the global expression pattern does not accurately reflect the state of the relevant sub-network.

Any global similarity measure of the whole transcriptome will be dominated by the majority of genes [73]. The second assumption is violated if the data is scarce, i.e. when either only few similar cells were measured or if the particular gene was detected in only a small subset of cells. In that case the average is computed across a relatively small number of observations and hence unstable.

Methods like Single-cell Analysis Via Expression Recovery (SAVER) [79], or Sparse Gene Graph of Smooth Signals (G2S3) [80, p. 3], employ a different strategy that can overcome some of these limitations. Instead of using the whole transcriptome of a cell to predict the expression level of a given gene, these methods learn gene-gene relationships from the dataset and use only the specific subset of genes that are expected to be predictive for the particular gene at hand. For example, SAVER learns gene-gene relationships using a penalized regression model, whereas G2S3 optimizes a sparse gene graph. However, if the scRNA-seq dataset at hand is sparse, the usefulness of the gene-gene relationships learnt from that dataset can be limited.

Single Cell RNA-Seq imputAtion constrained By BuLk RNAsEq data(SCRABBLE) [81] is different compared to all of the other methods mentioned above, because it can use bulk sequencing data to assist in the imputation. SCRABBLE combines a de-noising step with a moderated imputation moving the sample means towards the observed (bulk-derived) mean expression values. Limitations of this approach are that first, a matching bulk RNA sequencing data set needs to be available and second that the method only uses external data to adapt the distribution of the single cell data, but does not use it to inform gene-gene relationships.

Here, we compare published approaches that are representative for current state-of-the-art methods to two fundamentally different approaches. The first is a very simple baseline method that we use as a reference approach: we estimate missing values as the average of the expression level of the given gene across all cells in the dataset where the respective gene was detected. Initially intended to serve just as a reference for minimal expected performance, this sample-wide averaging approach turned out to perform surprisingly well and in many instances even better than state-of-the-art methods. The simple explanation is that estimating the average using all cells is a much more robust estimator of the true mean than using only a small set of similar cells, especially when the gene was detected in only few cells and/or if the gene's expression does not vary much across cells.

The second new approach avoids using a global similarity measure comparing entire transcriptomes. Instead, similar to SAVER or G2S3 it rests on the notion that genes are part of regulatory networks and only a small set of correlated or functionally associated genes should be used to predict the state of undetected genes. However, unlike other methods, we propose to use transcriptional regulatory networks trained on independent (bulk seq) data to rigorously quantify the transcriptional relationships between genes. Missing values are then imputed using the expression states of linked genes in the transcriptional regulatory network and exploiting the known quantitative relationships between genes. This approach allows imputing missing states of genes even in cases where the respective gene was not detected in any cell or in only extremely few cells. This second new approach rests on the assumption that the network describes the true regulatory relationships in the cells at hand with sufficient accuracy. Here, we show that this is indeed the case and that combining the two new approaches with published state-of-the-art methods drastically improves the imputation of scRNA-seq dropouts. Importantly, the performance of an imputation method is dependent on the 'character' of a gene (e.g. its expression level or the variability of expression between cells). Hence, we implemented an R-package (Adaptive Dropout Imputer, or ADImpute) that determines the best imputation method for each gene through a cross-validation approach.

3.2 Results

3.2.1 Imputing dropouts using a transcriptional regulatory network

In order to understand whether the inclusion of external gene regulatory information allows for more accurate scRNA-seq dropout imputation, we derived a regulatory network from bulk gene expression data in 1,376 cancer cell lines with known karyotypes. While the expression levels of genes in this data will be cell type-specific, the relationships between genes (e.g. concerted up-regulation of a transcription factor and its targets) are frequently conserved across cell types, allowing us to pool the data together to learn a generic gene regulatory network. For this purpose, we modelled the change (compared to average across all samples) of each gene as a function of its own copy number state and changes in predictive genes:

$$y_i = \alpha_i \cdot c_i + \sum_{j \neq i} \alpha_{ij} \cdot y_j + \varepsilon_i, \qquad (3.1)$$

where y_i is the expression deviation (log fold change) of gene i from the global average, c_i is the known (measured) copy number state of gene i, α the vector of regression coefficients, $y_j$the observed change in expression of gene j and ε_i the i.i.d. error of the model. To estimate a set of predictive genes j, we made use of Least Absolute Shrinkage and Selection Operator (LASSO) regression [25], which penalizes the L1 norm of the regression coefficients to determine a sparse solution. LASSO was combined with stability selection [82] to further restrict the set of predictive genes to stable variables and to control the false discovery rate (Methods). This approach ensures that the algorithm only selects gene-gene relationships that are invariant across most or all training data. Thus, interactions that would be specific to a single cell type will be excluded from the model. Using the training data, models were fit for 24,641 genes, including 3,696 non-coding genes. The copy number state was only used during the training of the model, since copy number alterations are frequent in cancer and can influence the expression of affected genes. If copy number states are known, they can of course also be used during the dropout imputation phase. Using cell line data for the model training has the advantage that the within-sample heterogeneity is much smaller than in tissue-based samples [27]. However, in order to evaluate the general applicability of the model across a wide range of conditions, we validated its predictive power on a diverse set of tissue-based bulk-seq expression datasets from the The Cancer Genome Atlas (4,548 samples from 13 different cohorts; see Methods and Supplementary Figure 3.1) and the Genotype-Tissue Expression (17,382 samples from 30 different healthy tissues; see Methods and Supplementary Figure 3.2).

Such a model allows us to estimate the expression of a gene that is not quantified in a given cell based on the expression of its predictors in the same cell. Here, the difficulty lies in the fact that imputed dropout genes might themselves be predictors for other dropout genes, i.e. the imputed expression of one gene might depend on the imputed expression of another gene. In order to derive the imputation scheme based on the model from equation (3.1), we revert to an algebraic expression of the problem,

$$Y = AY, \qquad (3.2)$$

where A is the adjacency matrix of the transcriptional network, with its entries α_{ij}being fitted using the regression approach described above, and $Y$is the vector of gene expression deviations from the mean across all cells in a given cell. In the current implementation we assume no copy number changes and hence, we exclude the $c_i$term from equation (3.1). Like in equation (3.1), we omit the intercept since we are predicting the deviation from the mean. Subsequently, imputed values are re-centered using those means to shift imputed values back to the original scale (see Methods). Further note that we drop the error term ε from equation (3.1), because this is now a prediction task (and not a regression). Here, we exclusively aim to predict dropout values, and (unlike SAVER) our goal is not to improve measured gene expression values. Hence, measured values remain unchanged. It is therefore convenient to further split Yinto two sub-vectors $Y^m$and Y^n, representing the measured and non-measured expression levels, respectively. Likewise, A is reduced to the rows corresponding to non-measured expression levels and split into A^m (dimensionality $|n| \times |m|$) and A^n (dimensionality $|n| \times |n|$), accounting for the contributions of measured and non-measured genes, respectively. The imputation problem is then reduced to:

$$Y^n = A^n Y^n + A^m Y^m \qquad (3.3)$$

As Y^m is known (measured) and will not be updated by our imputation procedure, the last term can be condensed in a fixed contribution, $F = A^m Y^m$, accounting for measured predictors:

$$Y^n = A^n Y^n + F \qquad (3.4)$$

The solution Y^n for this problem is given by:

$$Y^n = (I - A^n)^{-1} F \qquad (3.5)$$

The matrix $(I - A^n)$ may not be invertible, or if it is invertible, the inverse may be unstable. Therefore, we computed the pseudoinverse $(I - A^n)^+$ using the Moore-Penrose inversion. Computing this pseudoinverse for every cell is a computationally expensive operation. Thus, we implemented an additional algorithm finding a solution in an iterative manner (Methods). Although this iterative second approach is not guaranteed to converge, it did work well in practice (see Supplementary Figure 3.3, Methods). While our R-package implements both approaches, subsequent results are based on the iterative procedure.

3.2.2 Transcriptional regulatory network information improves scRNA-seq dropout imputation

To assess the performance of our network-based imputation method and compare it to that of previously published methods, we considered eight different single-cell RNA-sequencing datasets [83]–[87], covering a wide range of sequencing techniques (Smart-seq versions 1, 2 and 3 and droplet-based method 10X) and biological contexts (healthy tissue, cancer, stem cell differentiation and Human Embryonic Kidney - HEK - cells). A summary of the dataset characteristics, including number of cells and average number of quantified genes per cell, is provided in Supplementary Table 3.1. It was important to include a range of different healthy cell types in the evaluation, because the transcriptional regulatory network was trained on cancer cell line data. Thus, by including data from non-cancerous tissues, we could evaluate possible restrictions induced by the model training data.

In order to quantify the performance of both proposed and previously published imputation methods, we randomly set a fraction of the quantified values in the test data to zero according to two different schemes (Methods) and stored the original values for later comparison with the imputed values. Imputation was then performed on the masked dataset using our network-based approach, DrImpute [76], kNN-smoothing [78], SAVER [79], scImpute [75] and SCRABBLE [81].

Those methods were chosen since they were shown to be among the top-performing state-of-the-art dropout imputation methods [88], [89]. For masked entries imputed by all tested methods, the quality of imputation was assessed for each gene, using two approaches: computing the correlation between observed and imputed values for each gene, and computing the Mean Squared Error (MSE) of imputation (Methods).

According to the correlation quality measure our network-based approach (called ‘Network’) mostly outperformed other methods (Figure 3.1), especially in UMI-based datasets. We quantified the percentage of genes in the transcriptome of each dataset best imputed by each method (highest correlation) and verified that in six out of our eight test datasets, the network-based approach resulted in the highest performance for most genes (Table 3.1). Additionally, we observed that Network was less affected by low average expression levels compared to all other imputation methods (Figures 3.1 and S3.4, expression quartiles Q1 and Q2). This was expected since our network-based approach relies on information external to the dataset for dropout imputation, while other methods require sufficient observations of a gene to learn its expression characteristics from the single cell data itself. This result is in line with, for instance, a previous observation that scImpute is sensitive to missing information about genes across cells [88]. SCRABBLE is able to incorporate the average expression in matched bulk RNA-seq data to aid imputation, effectively taking advantage of external information like Network. Although this information is not available for most scRNA-seq datasets, it was available for the human embryonic stem cell (hESC) dataset, prompting us to use it in an additional SCRABBLE test (Figures 3.1, S3.4, and S3.5). We observed that SCRABBLE’s performance was not improved when incorporating this additional information.

As the network-based approach uses information regarding other genes contained in the same cell, we hypothesized its accuracy might be more affected by increasingly sparse information per cell when compared to other methods. However, the relative performance differences between methods were largely invariant to the number of missing genes per cell (Supplementary Figure 3.7), suggesting that other methods also suffer from the scarcity of information for cells with low sampling efficiency. The sensitivity of the imputation to the proportions of missing values was dependent on a multitude of factors, including the quality of the data, the specific gene(s) expressed in the cell type(s) at hand, the number of missing values, variability of expression/cell type heterogeneity within the study and possibly many others. Further, the network-based method performed well over a range of different cell types and showed decreased performance upon randomization of the transcriptional network (Methods, Supplementary Figure 3.8). Thus, the diversity of cell lines used in the training data seemed to capture a large fraction of all possible regulatory relationships in the human transcriptome.

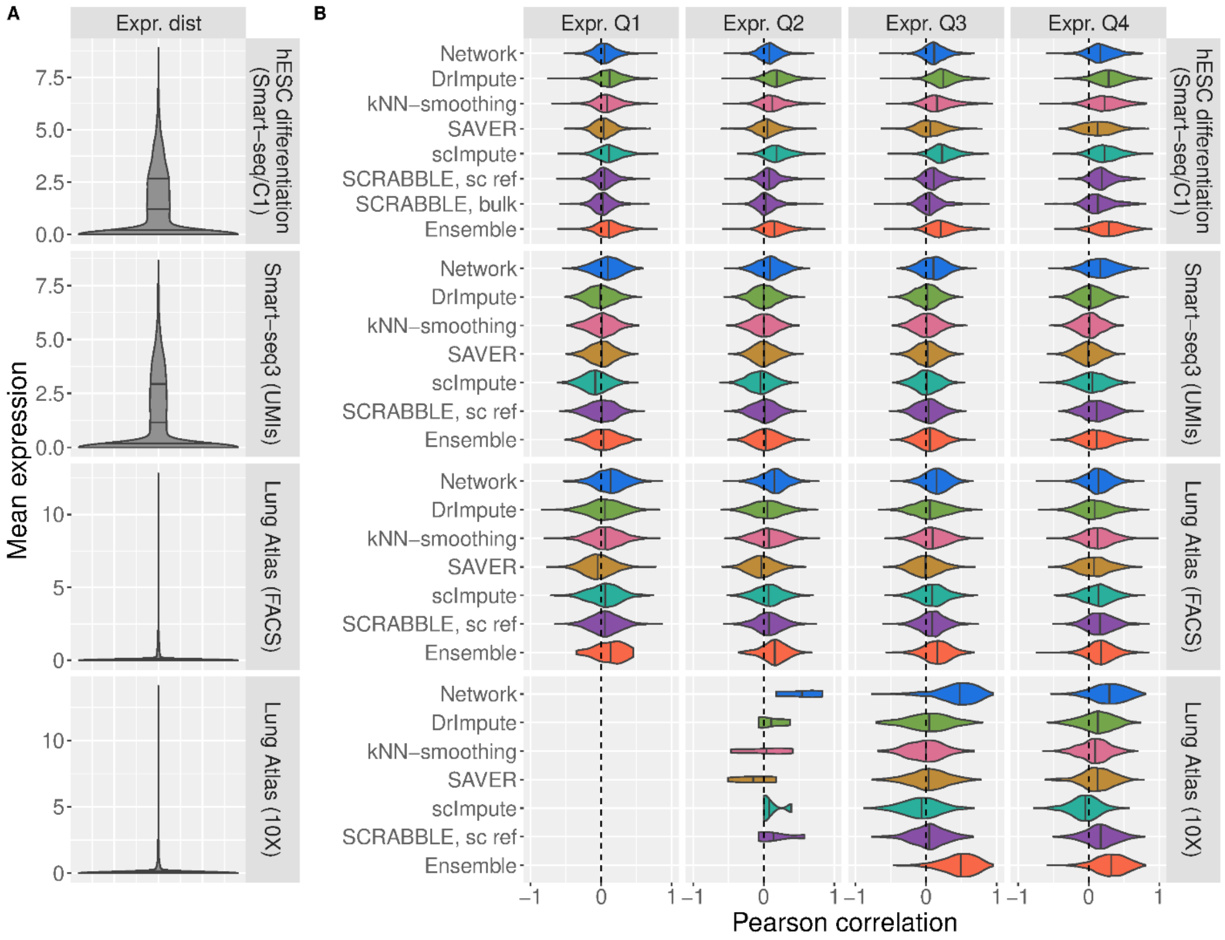


Figure 3.1: Imputation performance of Network (blue), DrImpute (green), kNN-smoothing (pink), SAVER (yellow), scImpute (turquoise), SCRABBLE (purple) and Ensemble (orange). A) Distribution of average expression levels (per gene) in each dataset. Quartiles are represented by vertical lines. **B)** Pearson correlation coefficient, for each gene, between the imputation by the specified method and the original values before masking. Only values that could be imputed by all methods (non-zero imputation) were considered. Correlations per gene across cells were computed for all genes for which at least 10 imputation values were available for analysis. Expression quartiles were determined for each dataset separately, on the masked data.

We additionally evaluated imputation performance using the mean squared error (MSE) between original and imputed values. In this analysis, we included a ‘Baseline’ method, which imputes dropouts with the average expression value of the gene across all cells, as a reference. Using the MSE as an alternative metric of performance, we also observed that Network outperformed previously published methods across all tested datasets and expression quartiles (Supplementary Figure 3.6). However, the performance of the Baseline method (Supplementary Figure 3.6, grey violin), which does not account for any expression variation between cells, was surprising to us. While SAVER showed a poor performance in comparison to all other methods (also described in [88]), it should be noted that this method aims to estimate the true value for all genes, not only for the dropout genes. This results in a change of both dropout and quantified values, which explains the good correlation performance but relatively high MSE.

Table 3.1: Percentage of genes best imputed by each method (highest Pearson correlation coefficient) in the seven test datasets restricted to values that could be imputed by all individual methods.

	Network	**DrImpute**	**kNN-smoothing**	**SCRABBLE, bulk**	**SCRABBLE, sc ref**	**scImpute**	**SAVER**
hESC differentiation	14.0%	**31.0%**	13.2%	5.9%	7.5%	22.0%	6.3%
hESC time course	16.6%	**29.0%**	22.7%	6.7%	9.2%	10.4%	5.5%
Oligodendroglioma	**29.3%**	18.4%	-	-	15.0%	20.1%	17.1%
Lung Atlas (10X)	**58.4%**	11.7%	6.0%	-	14.4%	2.3%	7.2%
Lung Atlas (FACS)	**31.5%**	13.5%	14.7%	-	17.8%	15.1%	7.4%
Renal Cell Carcinoma	**63.2%**	5.3%	3.1%	-	12.2%	2.0%	14.2%
Smart-seq3 (reads)	**26.8%**	11.3%	12.1%	-	18.1%	19.4%	12.3%
Smart-seq3 (UMIs)	**32.3%**	12.8%	12.4%	-	18.2%	11.5%	12.9%

Taken together, these results indicate that Network often – but not always - leads to more accurate imputations than state-of-the-art imputation methods (Supplementary Figure 3.6), while preserving variation across cells as captured by the correlation analysis (Figures 3.1 and S3.4). However, our analysis also uncovered that the advantage of using specific methods varies between datasets and expression quartiles, suggesting that there is no universally best performing method that outperforms the others in all cases. This motivated us to develop an ensemble approach, where we determine in a cross-validation scheme the best imputation method for each gene in the dataset at hand. We tested its performance (Figures 3.1, S3.4, and S3.6) and observed that the ensemble method tends to approach the performance of the best performing method.

3.2.3 Gene features determine the best performing imputation method

To better understand what gene features drive the performance differences between methods, we characterized the genes best imputed by each of the methods. We determined, for each gene in each test dataset, the method resulting in the highest correlation between imputed and original values (Table 3.1). Methods with performance close to the best (correlation difference of 0.1, see Methods) were considered to be “top performers”. We chose this approach because some methods apply similar strategies for dropout imputation and thus are expected to perform best for the same set of genes. Our analysis allows to capture these similarities between methods. The genes for which a given method was among the “top performers” were compared against a background including all genes for which all methods were able to perform imputations (Figures 3.2 and S3.10). As expected, methods relying on the similarity of cellular transcriptomes (scImpute, DrImpute and kNN-smoothing) performed best for more frequently detected genes (lower percentage of NAs per gene across cells). Conversely, the Network (and, to some extent, SAVER and SCRABBLE) were among the top performers especially on genes with many missing values. SAVER and Network are model-based methods, not relying on the comparison of entire transcriptomes between cells. SCRABBLE and Network are methods using external information for the imputation. Based on the above results we conclude that both aspects (model-based imputation and using external information) are advantageous for the imputation of rarely detected genes. We also determined the top performing methods for each gene based on the MSE instead of correlations. We observed that, as expected, the Baseline method performed best for genes with low expression levels and low variance (Supplementary Figure S3.11).

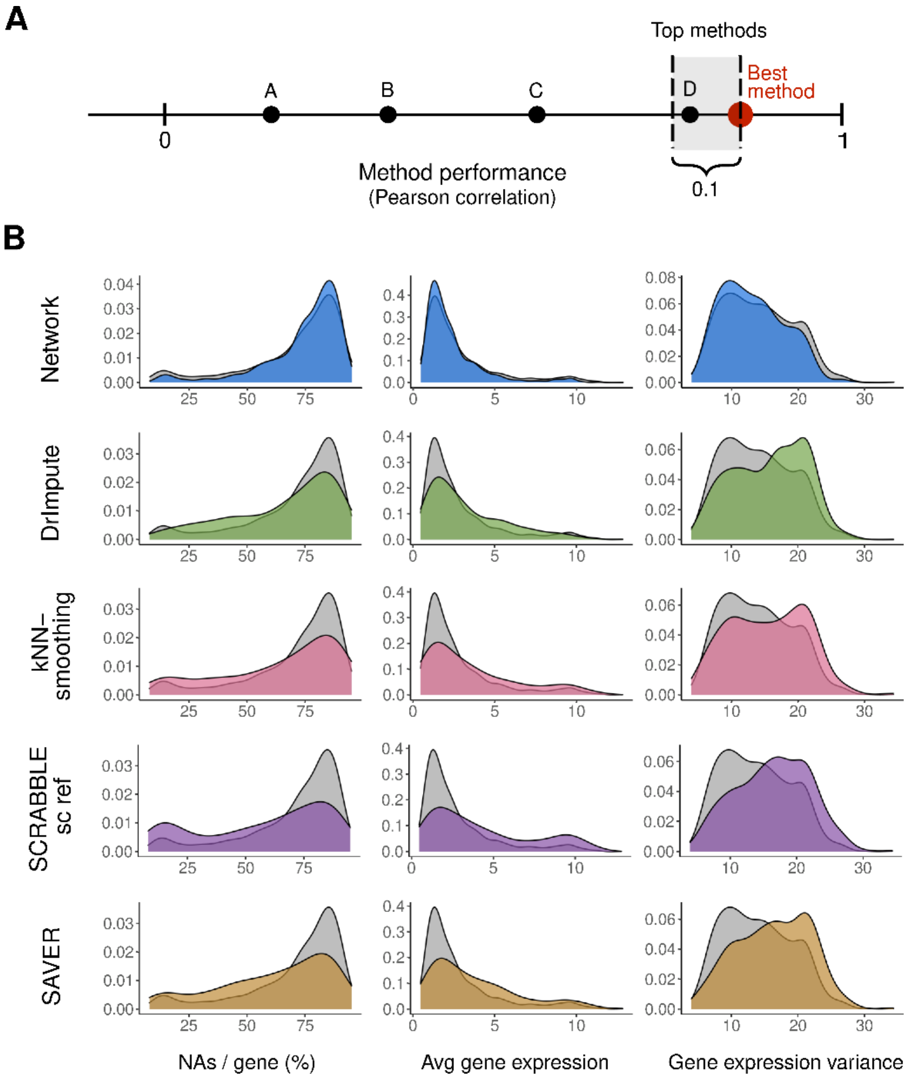


Figure 3.2: Characterization of the genes best predicted by Network (blue), DrImpute (green), kNN-smoothing (pink), SCRABBLE (bulk and single cell data as reference; purple) and SAVER (yellow) in the Lung Atlas 10X dataset. A) Determination of the top best performing methods for each gene. Methods performing best and close to the best method (correlation not smaller than 0.1 – best method) were selected as top performers. Genes for which all methods were top performers were included in the background but not in the foreground. **B)** Distribution of missing values per gene, average expression levels and variance of the genes for which a given method is one of the top performers, compared against all tested genes (background). Average gene expression is shown as $\log_2$-transformed normalized expression. Too few genes were best predicted by scImpute, so no distributions were drawn for this method.

3.2.4 Network-based imputation aids downstream analyses

Cell trajectory inference

scRNA-seq data is particularly suitable for the study of dynamic processes such as development or differentiation, due to the high numbers of individual cells sequenced and differences in progression along the dynamic process of choice between them [90]. Here, we make use of a time course hESC differentiation dataset [84] containing six distinct timepoints to infer cell trajectories through the course of differentiation, in order to assess the impact of dropout imputation on the inferred trajectories. To this end, we computed cell trajectories before and after imputation with slingshot [91], a method well evaluated in an independent work [89], and compared them to the known timepoint labels of the dataset (Figure 3.3). Imputation with any of the tested methods led to a better agreement between timepoint labels and inferred pseudotime, highlighting the usefulness of dropout imputation for downstream analyses such as trajectory inference. Additionally, Baseline and bulk-based SCRABBLE imputations showed the worst performance among the compared methods. Baseline's poor performance was expected, as signals across the whole dataset were averaged for dropout imputation, which dilutes the progressive changes across the course of differentiation. However, it was surprising to us that using Baseline was still better than not performing any dropout imputation. A possible explanation might be that leaving technical zeros in the data introduces additional noise thereby complicating the correct positioning of cells on the pseudo time trajectory. Bulk-based SCRABBLE's poor performance may be explained by the small number of samples per time point available in the bulk reference (n = 2 or 3). An average across such few samples is unstable, and it remains to be seen whether a more reliable bulk reference results in better performance for SCRABBLE. Finally, our results support the use of an Ensemble method where the best performing method is picked for each gene via a cross-validation approach, as the performance of this Ensemble method was practically indistinguishable from the best performing method.

Data visualization

Another popular application of scRNA-seq is the identification of discrete sub-populations of cells in a sample in order to, for example, identify new cell types. The clustering of cells and the visual 2D representation of single-cell data is affected by the choice of the dropout imputation method[13]. Therefore, we assessed the impact of dropout imputation on data visualization using Uniform Manifold Approximation and Projection (UMAP) [92] on the hESC data before and after imputation by all methods. The snapshot hESC dataset was particularly suitable in this case, because it was of high quality and it consisted of six well-annotated distinct cell types. This analysis confirmed that the choice of the imputation method impacts on the grouping/clustering of cells (Supplementary Figure 3.12). Application of other dimensionality reduction techniques (t-distributed stochastic neighbor embedding, t-SNE, and Zero-Inflated Negative Binomial-based Wanted Variation Extraction ZINB-WaVE) showed varying results depending on the chosen method, suggesting that visual clustering upon dimensionality reduction is an inconclusive criterion for evaluating dropout imputation (Supplementary Figure 3.13).

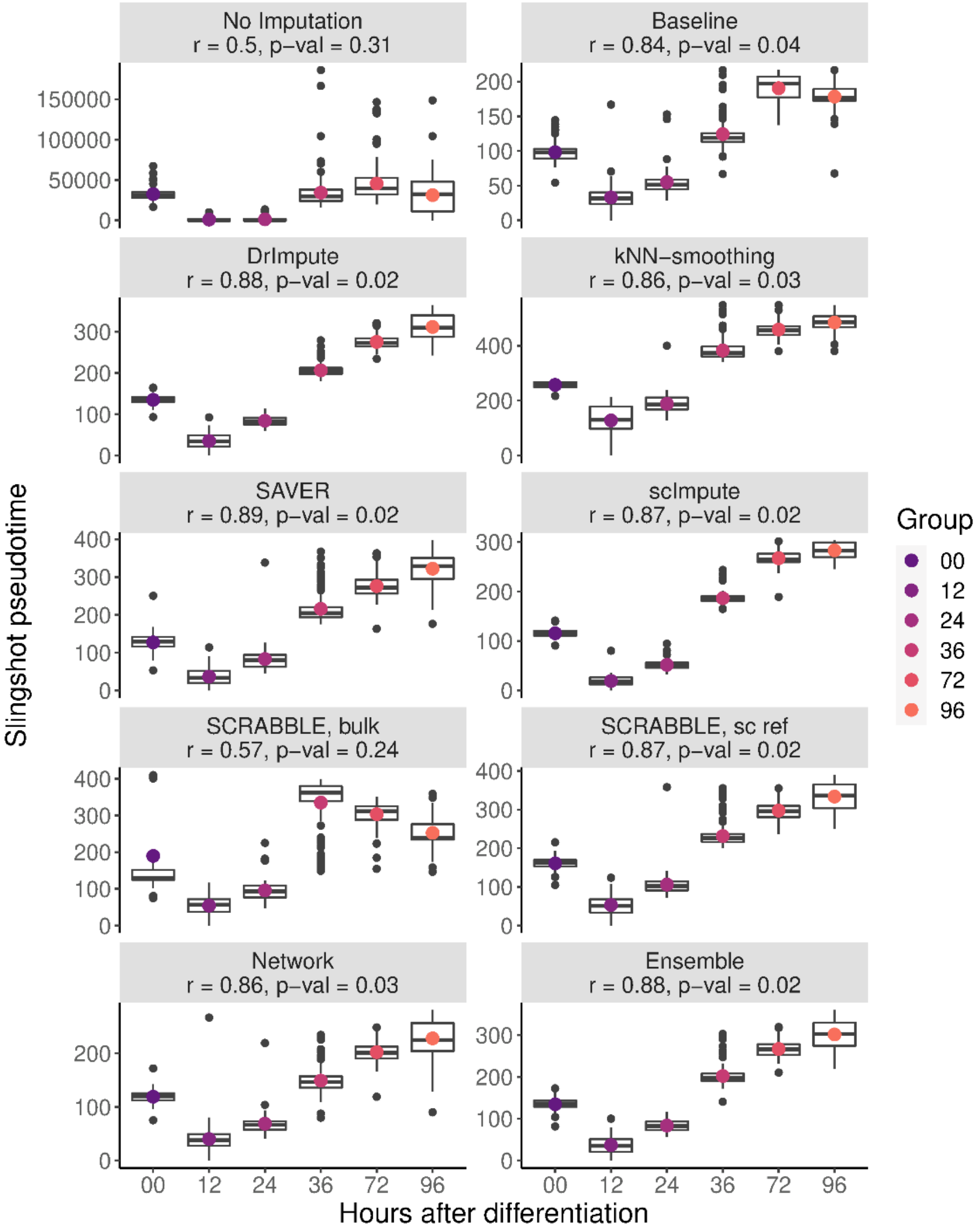


Figure 3.3: Comparison of impact of different imputation methods on cell trajectory inference for time course hESC differentiation data. Cells were projected onto the trajectories determined with slingshot to compute a pseudotime of differentiation. The pseudotime assigned to each individual cell (y-axis) is then compared to the time point label of the sample (x-axis). Colored points represent the mean pseudotime per known time point. In the title, the Pearson's r between pseudotime and time point labels (and the respective p-value) are shown. Note that slingshot always fails to correctly position the sample at time point 0, suggesting an artifact in the original data.

Cluster marker detection

We next asked to what extent the detection of cluster markers would be affected by the choice of the imputation method. Thus, we applied Seurat [93] to the hESC differentiation dataset, which was composed of a well-defined set of distinct cell types, before and after imputation. We then defined genes that were significantly differentially expressed between one cluster and all the others as cluster markers (Methods). We observed a considerable overlap between markers detected before and after applying the tested imputation methods (Figure 3.4.A, horizontal dashed line), suggesting a common core of detected cluster markers across methods. Additionally, the numbers of significant markers detected after Network and Baseline imputations were lower than for other imputation methods (Figure 3.4.A). Imputation with kNN-smoothing, scImpute and, to a smaller extent, with DrImpute, led to the highest number of significant markers (Figure 3.4.A). We hypothesized that many of these marker genes may result from artefactual clustering of cells. In order to test that notion we first determined all Gene Ontology (GO) biological process terms that were enriched in the respective cell clusters without any dropout imputation. We termed them 'high confidence GO terms' since they are independent of the choice of the imputation method. It turned out that kNN-smoothing, scImpute and DrImpute had the weakest enrichments in high confidence GO biological process terms (Figure 3.4.B and 3.4.C; Methods; Supplementary Table 3.4), suggesting that the extra markers found upon applying scImpute and DrImpute contained many false positives, which diluted biological signals. Conversely, Network and SCRABBLE led to the strongest enrichments in high confidence GO biological process terms (Figure 3.4.B and 3.4.C).

Determining transcriptional regulators

Genes with regulatory functions are particularly important for understanding and explaining the transcriptional state of a cell. However, since genes with regulatory functions are often lowly expressed [94], they are frequently subject to dropouts. Since our analysis had shown that the network-based approach is especially helpful for lowly expressed genes (Figure 3.2), we hypothesized that the imputation of transcript levels of regulatory genes would be particularly improved. In order to test this hypothesis, we further characterized those cluster markers that were exclusively detected using the network-based method. Indeed, we observed regulatory genes to be enriched among those markers (Figure 3.5.A). The transcription factor ETS Homologous Factor (EHF) was the second most significant trophoblast-specific among these markers exclusively detected upon network-based imputation. EHF is a known epithelium-specific transcription factor that has been described to control epithelial differentiation [95] and to be expressed in trophoblasts (TB) [96], even though at very low levels (EHF expression was found among the first quintile of bulk TB RNA-seq data from the same authors). While EHF transcripts were not well captured in TB single-cell RNA-seq data (only detected in 39 out of 775 TB cells), a trophoblast-specific expression pattern was recovered after network-based imputation (Figure 3.5.B, upper panel), but not with any of the other tested imputation methods (Supplementary Figure 3.14). Similarly, Odd-Skipped Related Transcription Factor 1 (OSR1) has been described as a relevant fibroblast-specific transcription factor [97] which failed to be detected without imputation. Imputing with Network lead to the strongest fibroblast-specific expression pattern of OSR1 (Figures 3.5.B, lower panel, and S3.14). Interestingly, TWIST2 and PRRX1, described by Tomaru *et al.* [97] to interact with OSR1, also showed fibroblast-specific expression (Supplementary Figure 3.15). Taken together, these results suggest that imputation based on transcriptional regulatory networks can recover the expression levels of relevant, lowly expressed regulators affected by dropouts.

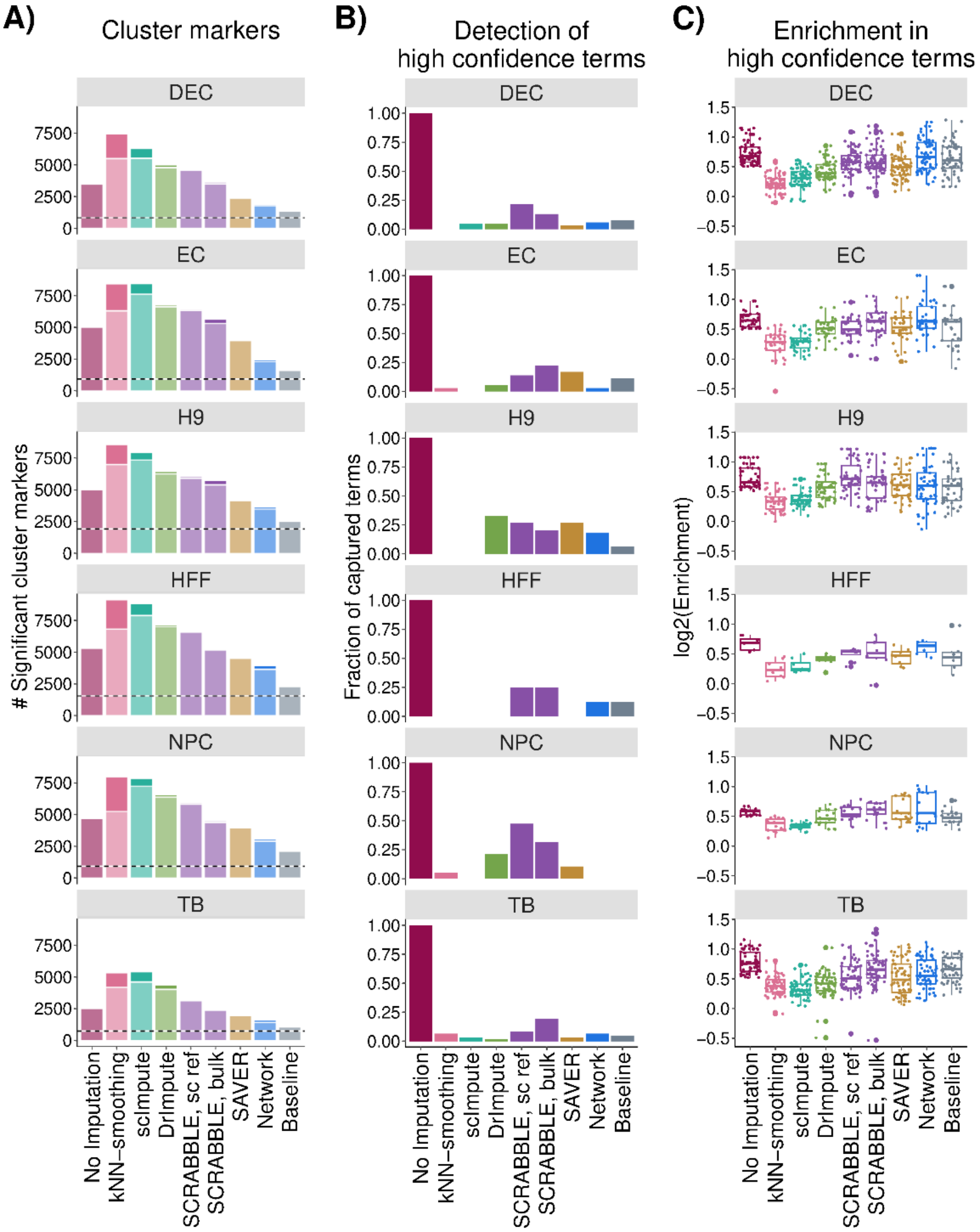


Figure 3.4: Detection of cell type-specific markers before and after imputation. A) Number of significant (FDR < 0.05, $|\log_2 FC| > 0.25$) cell type markers detected with no dropout imputation and using the tested imputation methods. Horizontal dashed lines correspond to the number of markers detected irrespectively of imputation. The fraction of the bar in a darker shade corresponds to the number of markers detected exclusively when using a given imputation approach. **B)** and **C)** fraction of captured high confidence terms, defined as significantly enriched (p.value < 0.001 and $\log_2$Enrichment > 0.5, Methods) GO biological process terms among the cluster markers detected without imputation. **B)** Sensitivity: fraction of high confidence terms detected as significantly enriched (p.value < 0.001 and $\log_2$Enrichment > 0.5) among the cluster markers detected with each imputation method. **C)** $\log_2$-enrichment of all high confidence terms among the cluster markers detected with each imputation method. DEC: definitive endoderm cells; EC: endothelial cells; H9: undifferentiated human embryonic stem cells; HFF: human foreskin fibroblasts; NPC: neural progenitor cells; TB: trophoblast-like cells.

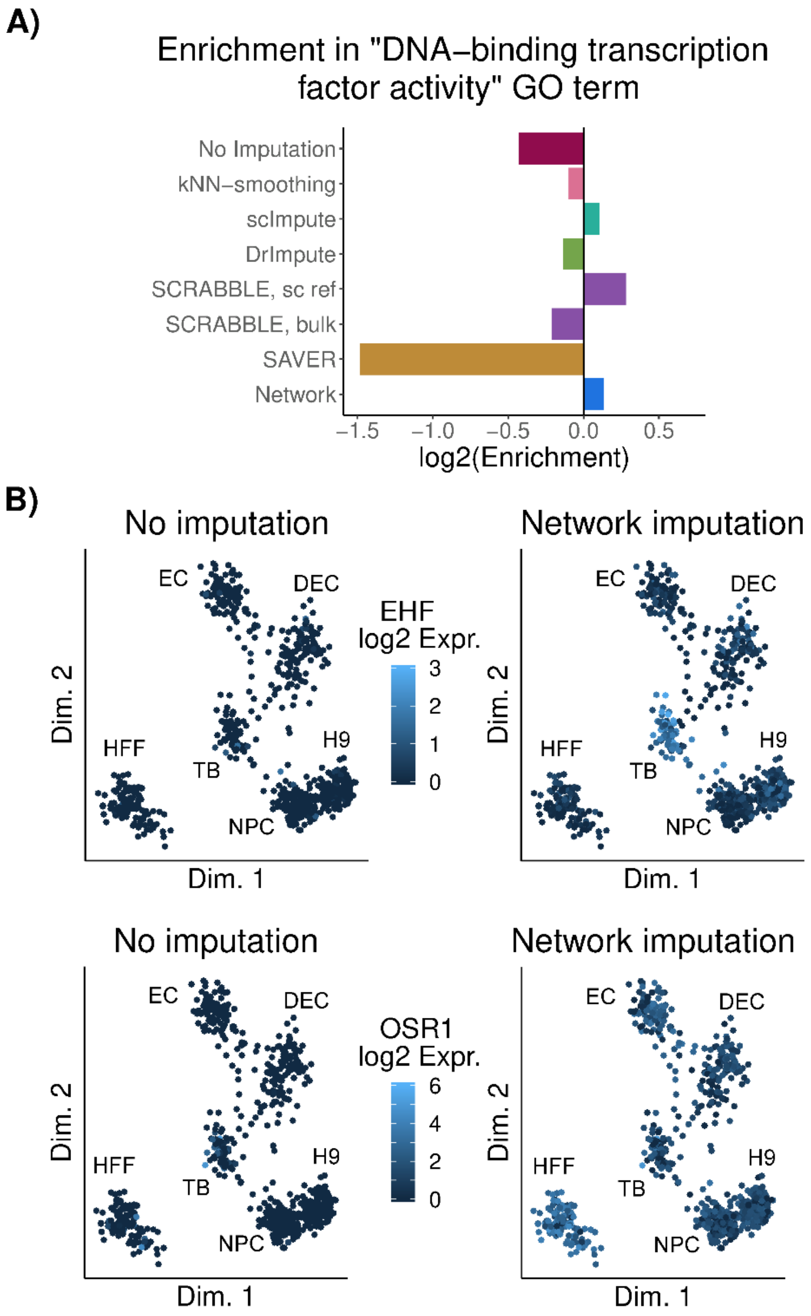


Figure 3.5: Detection of cell type-specific transcription factors is improved upon network-based imputation. A) Enrichment score in GO term “DNA-binding transcription factor activity” among the genes uniquely detected after each imputation approach. **B)** Projection of cells onto a low dimension representation of the data before imputation, using ZINB-WaVE [98]. Color represents normalized expression levels of EHF (top) and OSR1 (bottom) before and after Network-based imputation. DEC: definitive endoderm cells; EC: endothelial cells; H9: undifferentiated human embryonic stem cells; HFF: human foreskin fibroblasts; NPC: neural progenitor cells; TB: trophoblast-like cells.

3.3 Methods

Pre-processing of cancer cell line data for transcriptional regulatory network inference

Entrez IDs and corresponding gene symbols were retrieved from the NCBI (https://www.ncbi.nlm.nih.gov/gene/?term=human%5Borgn%5D). Genome annotation was obtained from Ensembl (*Biomart*). Finally, genes of biotype in protein coding, ncRNA, snoRNA, scRNA, snRNA were used for network inference. For CCLE [99], 768 cell lines that were used in Seifert et al. [27] were used. Raw CEL files were downloaded from https://portals.broadinstitute.org/ccle/ and processed using the R package RMA in combination with a BrainArray design file (HGU133Plus2_Hs_ENTREZG_21.0.0). Final expression values were in $\log_2$ scale. Expression levels and CNV data set from RNA-seq were downloaded from Klijn et al. [100]. Before combining, each dataset is $\log_2$ transformed and scaled to (0,1) for all genes in each sample using R function scale. Then datasets were merged and the function ComBat from the sva R package [101] was used to remove batch effect of the data source. The final combined data set contains 24641 genes in 1443 cell lines. Finally, expression levels of genes were subtracted by the average expression level across all cell lines of the corresponding gene.

Network inference based on stability selection

The network inference problem can be solved by inferring independent gene-specific sub-networks. We used the linear regression model from equation (3.1) to model the change in a target gene as dependent on the combination of the gene-specific CNA and changes in all other genes. Here the intercept is not included because the data is assumed to be centered. We used LASSO with stability selection [82] to find optimal model parameters α_{ij}.

The R package *stabs* was employed to implement stability selection and the *glmnet* package was used to fit the generalized linear model. Two parameters regarding error bounds were set with the cutoff value being 0.6 and the per-family error rate being 0.05. A set of stable variables were defined by LASSO in combination with stability selection. Then coefficients of the selected variables were estimated by fitting generalized linear models using the R function *glm*.

Network validation using TCGA and GTEx data

Gene expression and gene copy number data of 14 different tumor cohorts (4548 tumor patients in total) from TCGA collected in a previous study [27] were used for validation. We examined the predictive power of our inferred networks on each TCGA cohort by predicting the expression level of each gene for each tumor using the corresponding copy number and gene expression data.

Additionally, in order to validate the applicability of the learnt network to healthy tissues, we further leveraged gene expression data from the Genotype-Tissue Expression (GTEx) Project. Read counts were downloaded from the portal website (version 8), normalized using the R package *DESeq2* and centered gene-wise across tissues.

For each TCGA cohort or GTEx tissue, the expression levels of each gene were predicted using the network and expression quantification of the interacting genes in the same sample. The predicted value was then compared to the observed value, present in the original dataset. The quality of prediction for each TCGA cohort or GTEx tissue was quantified as either the correlation between predicted and observed expression of a gene across all samples or the MSE of prediction of a gene across all samples. A strong positive correlation or low MSE for a gene suggests high predictive power by the network on the respective gene.

Single-cell test data processing

Human embryonic stem cell differentiation data [84] were downloaded from the Gene Expression Omnibus (GEO, accession number GSE75748) in the format of expected counts. The downloaded data were converted to RPM (reads per million). Renal cell carcinoma data [87], in the format of normalized UMI counts, and corresponding metadata, were download via the Single Cell Portal. Data was reduced to cells from patient P915. Cells with library sizes more than 3 median absolute deviations above the median were removed as potential doublets. 2000 cells were randomly selected for further analysis and underwent reversion of the log-transformation. Human embryonic kidney (HEK) cell read and UMI data, sequenced with Smart-seq3 [83, p. 3], were downloaded from ArrayExpress (accession code E-MTAB-8735). Ensembl IDs were converted to gene symbol and data was normalized for library size (RPM). Oligodendroglioma data [85] were downloaded from GEO (accession number GSE70630) as $\log_2$(TPM/10+1) and converted back to TPM. Lung Atlas 10X and Smart-seq2 data [86], together with corresponding metadata, were downloaded from Synapse (ID syn21041850). For both sequencing methods, data was restricted to cells from the lung of patient 1. Potential doublets (cells with library sizes above the median) were removed from the 10X data. For this dataset, library sizes above the median were considered potential doublets due to the bi-modal distribution of library sizes below the usual threshold of median+(3*MAD). After doublet removal, 2000 cells were randomly selected for RPM normalization and further analysis for both sequencing methods.

Dropout imputation

Version 0.0.9 of scImpute [75] was used for dropout imputation, in “TPM” mode for the oligodendroglioma dataset and “count” mode for all other datasets, without specifying cell type labels. The parameters were left as default, except for drop_thre = 0.3 (upon artificial masking), as the default of 0.5 resulted in no imputations performed. Cell cluster number (Kcluster) was left at the default value of 2 for imputation of all datasets except for the hESC differentiation datasets (snapshot and timecourse), where it was set to 6 in order to match the number of cell clusters identified by the authors [84], and the Smart-seq3 dataset, where it was set to 1 because only one cell type was sequenced. SAVER 1.1.1 was used with size.factors = 1. SCRABBLE 0.0.1 was run with the parameters suggested by the authors and using by default the average gene expression across cells as the bulk reference. In the case of the hESC differentiation dataset, bulk data from the same study was available, and thus was also used as reference. kNN-smoothing [78] (python implementation v. 2.1) was run on the data before library size normalization, as this method performs a different correction. Since the Oligodendroglioma data was retrieved in TPM format, kNN-smoothing was not included in the method comparison for this dataset. For all other imputation methods, the data was log2-transformed with a pseudocount of 1. DrImpute 1.0 was run using the default parameters. For dropout imputation by average expression (‘Baseline’), gene expression levels were log2-transformed with a pseudocount of 1 and the average expression of each gene across all cells, excluding zeros, was used for imputation. For network-based imputation, expression values were log2-transformed with a pseudocount of 1 and centered gene-wise across all cells. The original centers were stored for posterior re-conversion. Subsequently, cell-specific deviations of expression levels from those centers were predicted using either equation (3.5) or the following iterative procedure. During the iteration genes were first predicted using all measured predictors. Subsequently, genes with dropout predictors were re-predicted using the imputed values from the previous iteration. This was repeated for at most 50 iterations. The obtained values were added to the gene-wise centers. We note that the values after imputation cannot be interpreted as TPMs/RPMs, as the sum of the expression levels per sample is no longer guaranteed to be the same across samples. However, one could still perform a new normalization by total signal (sum over all genes) to overcome this issue.

Masking procedures

In order to compare the imputation error of the tested methods, we randomly masked (set to zero) some of the values for each gene, using two different approaches.

The first approach consisted of setting a fraction of the quantified, uniformly sampled values to zero for each gene (Figure 3.1, S3.4, and S3.5) - 35% for the hESC differentiation and Smart-seq3 datasets, 10% for the oligodendroglioma and Lung Atlas (FACS-sorted) datasets and 8% for the Renal Cell Carcinoma and Lung Atlas 10X datasets. In case of Supplementary Figure 3.6 30% of the cells (not genes) were sampled. This unbiased masking scheme is in agreement with previous work [102]. The differing percentages of masked values per gene in each dataset result in a comparable sparsity of the data after masking.

As an alternative masking procedure that represents more closely a downsampling process, we modelled for each gene its probability to be an observed zero in the following way: the fraction of cells where each gene was not captured (zero in the original data) was modelled as a function of its average expression across cells (Supplementary Figure 3.8). For this, a cubic spline was used, with knots at each 10% quantile of the average expression levels, excluding the 0% and 100% quantiles. A cubic spline was chosen so that it could properly fit to both UMI-based and non UMI-based datasets. With this model, a ‘dropout probability’ p was computed for each gene from its mean expression. The masking procedure then consisted of, for each entry, sampling a Bernoulli distribution with probability of success 1-p, where 0 corresponds to a mask (the entry is set to 0) and 1 to leaving the data as it is. Thus, each entry in the data matrix may be masked with a probability p, which is gene-specific and based on the observed dropout rates in the dataset at hand.

We observed the same relative performance of the imputation methods under this alternative masking scheme (Supplementary Figure 3.9), and for this reason present the results obtained with the first masking approach.

Imputation performance analysis

Imputation was performed with each of the four tested methods separately and the imputed masked entries were then compared to the original ones. For genes where at least 10 values were imputed (non-zero after imputation, zero after masking) by all methods, the Pearson correlation between original and imputed values across cells was computed for each gene individually.

Additionally, we used the mean of the squared imputation error across all imputations for a given gene:

$$\frac{\sum(original-imputed)^2}{number\ of\ imputations} \qquad (3.1)$$

In order to avoid higher errors for more highly expressed genes, we split the genes into expression quartiles when reporting the imputation error.

Top performing methods

For each gene, the best performing method (highest correlation / lowest MSE) was computed. For the correlation-based analysis, methods with a correlation difference to the maximum no bigger than 0.1 were considered to be top performing. If all methods for a given gene were within this range (all top performers), the gene was not included in the foreground of Figures 3.2 and S3.10. For the MSE-based analysis, the maximum MSE difference to the best performer was computed for each gene individually, since the MSE range can be quite different from gene to gene. The maximum MSE difference was determined as 5% of the MSE range, in order to make it comparable to the threshold used in the correlation-based analysis and thus, methods in the top 5% of the gene-specific MSE range were considered to be top performing.

Cell trajectory inference

Trajectory inference was performed on the original and imputed hESC time course dataset, using the R package *slingshot* (v. 1.8.0). Dimensionality reduction was done via PCA without scaling, as advised by the authors. The first 2 Principal Components were used, as retaining more did not change the relative performance of the methods. Since specifying the start and/or end of clusters sometimes resulted in branched trajectories, cluster labels were not given as input in order to avoid branching (branching was not expected in this dataset). The Pearson correlation between the obtained pseudotime and the known timepoint labels of the dataset was used for evaluation.

Dimensionality reduction

Dimensionality reduction on the original hESC differentiation data (Figure 3.5.B) was performed using ZINBWaVe, t-SNE and UMAP (Supplementary Figures 3.12 and 3.13). H1 and TB cells in Batch 3 were removed to avoid confounding batch effects and dimensionality reduction was performed for the remaining cells. UMAP was performed on the first 5 principal components obtained from the top 1000 most variable genes in the hESC differentiation data (normalized, $\log_2$-transformed) before and after imputation (Supplementary Figure 3.12) using the *Seurat* R package [93], version 4.0.4, with parameters *umap.method* = *"umap-learn"* and *metric* = *"correlation"*. ZINB-WaVE, implemented in the R package *zinbwave* [98], was used to extract 2 latent variables from the information contained in the top 1000 genes with highest variance across cells. Batch information and the default intercepts were included in the ZINB-WaVE model, using *epsilon* = 1000 (Supplementary Figure 3.13). K-means clustering (k = 6) on the 2 latent variables strongly matched the annotated cell type labels (0.977 accuracy), confirming the reliability of this approach. t-SNE was performed on the normalized and $\log_2$-transformed data using the *Rtsne* R package with default settings (Supplementary Figure 3.13).

Marker detection

Cluster-specific markers were detected from the $\log_2$-transformed normalized data using Seurat version 4.0.4. Detection rate was regressed out using the *ScaleData* function with *vars.to.regress*

= *nGene*. Markers were detected with the *FindAllMarkers* function, using MAST [103] test and setting *logfc.threshold* and *min.pct* to 0, and *min.cells.gene* to 1.

GO term enrichment and transcription factor analyses

All GO term enrichment analyses were performed with the *topGO* R package [104]. Enrichment in GO biological process terms among cluster-specific markers (Figures 3.4.B and 3.4.C) was performed for each cell cluster and (no) imputation method separately, using as foreground the set of significant cluster markers detected by Seurat, with FDR < 0.05 and |logFC| > 0.25, and as background all genes in the Seurat result (both significant and non-significant). The *classic* algorithm was used, in combination with Fisher test, and $\log_2$ enrichment was quantified as the $\log_2$ of the ratio between the number of significant and expected genes in each term. Significantly enriched (p-value < 0.001 and $\log_2$ enrichment > 0.5) GO biological process terms within each set of cluster markers, as detected in the original data (no masking, no imputation), were defined as "high confidence" terms.

For regulatory GO molecular function term enrichment analyses (Figure 3.5.A), significant (FDR < 0.05 and |logFC| > 0.25) markers uniquely detected without / with each imputation method were combined across all clusters and tested for enrichment in the term "DNA-binding transcription factor activity" against the background of all genes obtained as the result of Seurat (both significant and non-significant). The *classic* algorithm was used, in combination with Fisher test, and $\log_2$ enrichment was quantified as the $\log_2$ of the ratio between the number of significant and expected genes in each term.

To identify transcription factors (TFs) among cluster markers exclusively detected using the network-based method, a curated TF list was downloaded from http://tfcheckpoint.org/ .

Determination of the optimal imputation method per gene

In order to determine the best performing imputation method for each gene, 70% of the cells in each dataset were used as training data, where a percentage of the expression values were masked, as previously described. In the Smart-seq3 dataset, where only around 100 cells were available, the training was done in 98% of the dataset The remaining cells were used for testing. After masking, each of the tested imputation methods was applied to the training data and the imputed values of masked entries were then compared to the measured values. The Pearson correlation coefficient was computed for each gene with at least 10 imputed (with a non-zero value) masked entries. For each gene, the method leading to the highest correlation coefficient was chosen as optimal. When no decision could be done, the Baseline method was used as a default.

The ADImpute R package

The ADImpute R package is composed of two main functions, *EvaluateMethods* and *Impute*. *EvaluateMethods* determines, for each gene, the method resulting in the best imputation performance, in a cross-validation procedure. *Impute* performs dropout imputation according to the choice of method provided by the user. Currently supported methods are scImpute, DrImpute, SAVER, SCRABBLE, the Baseline and Network methods described in this manuscript and an Ensemble method, which takes the results from *EvaluateMethods* to select the imputation results from the gene-specific best method. Additionally, the user can choose to estimate the probability that each dropout value is a true zero, according to the approach used by scImpute, and leave the values unimputed if their probability of being a true zero falls above a user-defined threshold.

3.4 Discussion

This work has led to the following key findings: (i) a model-based approach using external data is particularly powerful for the imputation of rarely detected genes; (ii) based on the MSE criterion the expression of surprisingly many genes is best predicted by simply using their average expression across cells ('Baseline'); (iii) not all genes are equally well predicted by a single imputation approach; instead, one should adapt the imputation method to the specific gene in a given dataset. In addition, our work confirmed earlier findings, such as the artifactual clustering resulting from some imputation methods.

The consideration of external gene co-expression information for the dropout imputation substantially improved the performance in many cases, especially for lowly expressed genes. Since genes with regulatory functions are often lowly expressed [94], imputation of those genes might be critical for explaining expression variation between cells. Of note, cell type-specific regulatory genes were successfully imputed using information from our global gene co-expression network (Figure 3.5). This observation, together with the demonstrated predictive capacity of our network across cancer and healthy human data from a wide range of tissues, highlights the transferability of the gene-gene relationships learnt by our network. scImpute, kNN-smoothing and DrImpute elevated the number of cluster markers found in downstream DE analysis (Figure 3.4), but these additional markers seemed to dilute the true biological signal in the data. A similar behaviour has been described elsewhere for scImpute [73]. Similarly to our network-based approach, SAVER makes use of gene-gene relationships for imputation. However, SAVER learns these relationships in the dataset at hand, while Network makes use of externally trained relationships. The fact that Network outperforms SAVER in our comparisons suggests that the scRNA-seq data at hand may often be too sparse for relationships to be adequately learnt.

A potential limitation of our approach is that a transcriptional network derived from bulk-seq data may not fully capture gene-gene relationships that are detectable from single cell data. For example, gene regulatory relationships that are specific to a small sub-population of cells in a bulk tissue may not be correctly captured, because the signal would be too weak. A second example would be genes regulated during the cell cycle. Bulk tissue is usually not synchronized, i.e. it consists of a mix of cells at different cell cycle stages, which may prevent the detection of those relationships. To some extent these limitations were alleviated by using cell line data rather than actual tissue data for training the network. Of course, the network that we used here is still imperfect. However, despite that imperfection it demonstrated the power of our approach. Using it was clearly advantageous over not using it in most dropout imputation tests, and we showed its predictive power across independent datasets from cancer, healthy human tissue and cell lines.

A surprising finding of our analysis is the fact that the sample-wide average expression ('Baseline') performs well for the imputation of many genes when using the MSE as a performance measure (Supplementary Table 3.3). Originally, we developed Baseline only as a benchmark of a minimalistic approach in order to compare the performance gain of more sophisticated approaches against this one. Much to our surprise, many genes could not be imputed better by any of the competing approaches. As expected, genes whose expression levels were best imputed by Baseline were characterized by lower variance across cells and by remaining undetected in relatively many cells (Supplementary Figure 3.11). A potential problem of methods based on co-clustering cells is that the number of observations per cluster can get very small, which makes the estimation of the true mean more unstable. Thus, averaging across all cells is preferred when the gene was detected in only few cells and/or if the gene's expression does not vary much across cells. Further, our findings imply that cell-to-cell expression variation of many genes is negligible or at least within the limits of technical measurement noise. Obviously, Baseline does not support the identification of differentially expressed genes between different groups of cells. However, it may help reducing artifacts resulting from the clustering of cells with technical zeros (Figure 3.3).

The third -- and maybe most important -- conclusion is that the best performing imputation method is gene- and dataset-dependent. That is, there is no single best performing method. If the number of observations is high (many cells with detected expression) and if the expression quantification is sufficiently good, scImpute and DrImpute outperformed other methods. Importantly, the technical quality of the quantification depends on the read counts, which in turn depends on sampling efficiency, gene expression, transcript length and mappability – i.e. multiple factors beyond expression. If however, gene expression is low and/or too imprecise, scImpute and DrImpute were outcompeted by other methods, since expression data across neighboring cells is not informative in this case. This finding led us to conclude that a combination of imputation methods would be optimal. Hence, we developed an R-package that determines 'on the fly' for each gene the best performing imputation method by masking observed values (i.e. *via* cross validation). This approach has the benefit that it self-adapts to the specificities of the dataset at hand. For example, the network-based approach might perform well in cell types where the assumptions of the co-expression model are fulfilled, whereas it might fail (for the same gene) in other cell types, where these assumptions are not met. Hence, the optimal imputation approach is gene- and dataset-dependent. An adaptive method selection better handles such situations. Another benefit of this approach is that the cross validation imputation performance can be used as a quantitative guide on how 'imputable' a given gene is in a specific scRNA-seq dataset. We have therefore implemented and tested this approach (see Figures 3.1, S3.4, and S3.5). The resulting R-package (called ADImpute) is open to the inclusion of future methods and user-provided gene networks, includes scImpute's estimation of dropout probability and it can be downloaded from https://bioconductor.org/packages/release/bioc/html/ADImpute.html. While, by default, ADImpute makes use of the transcriptional regulatory network described here, the user can provide any desired gene network e.g. for a different species, similarly to [105]. Our network has been trained on a large set of human cancer cell lines and thus is expected to capture much of the possible regulatory interactions of interest to many users; however, ADImpute facilitates the use of custom-made more specialized networks.

We believe that this work presents a paradigm shift in the sense that we should no longer search for the single best imputation approach. Rather, the task for the future will be to find the best method for a particular combination of gene and experimental condition.

3.5 Contributions

The work described in this chapter is available in the following peer-reviewed publication:

Leote AC, Wu X, Beyer A (2022) Regulatory network-based imputation of dropouts in single-cell RNA sequencing data. PLOS Computational Biology 18(2): e1009849. https://doi.org/10.1371/journal.pcbi.1009849

X.W. trained the regulatory model on transcriptomic and copy number data in cancer cell lines. I implemented the dropout imputation approach, compared the performance to that of previously published methods and developed the ADImpute R package. I wrote the manuscript together with A.B.

3.6 Supporting Information

Supplementary Table 3.1: Characteristics of the test datasets.

Authors	Tissue	Sequencing method	# cells	Avg. quantified genes / cell
Hagemann-Jensen *et al.*	Human Embryonic Kidney (HEK) cells	Smart-seq3	117	10,743 (read counts)
Hagemann-Jensen *et al.*	Human Embryonic Kidney (HEK) cells	Smart-seq3	117	9,539 (UMI counts)
Chu *et al.*	hESC differentiation	Smart-seq/C1	1,018 (snapshot)	9,617 (snapshot)
Chu *et al.*	hESC differentiation	Smart-seq/C1	758 (time course)	8,695 (time course)
Tirosh *et al.*	Oligodendroglioma, 6 donors	Smart-seq2	4,347	5,169
Travaglini *et al.*	Healthy lung, 3 donors' lung and blood	10X Smart-seq2	7,524 (10X, donor 1 lung) 3,235 (Smart-seq2, donor 1 lung)	3,003 (Smart-seq2, donor 1 lung)
Travaglini *et al.*	Healthy lung, 3 donors' lung and blood	10X Smart-seq2	7,524 (10X, donor 1 lung) 3,235 (Smart-seq2, donor 1 lung)	2,105 (10X, donor 1 lung)
Bi *et al.*	Clear cell renal cell carcinoma, 8 donors	10X	6,541 (donor P915)	1,473 (donor P915)

Supplementary Table 3.2: Spearman rank correlation between original values and imputation results after masking. Correlation was computed between the vector of original entries and imputations common to all methods.

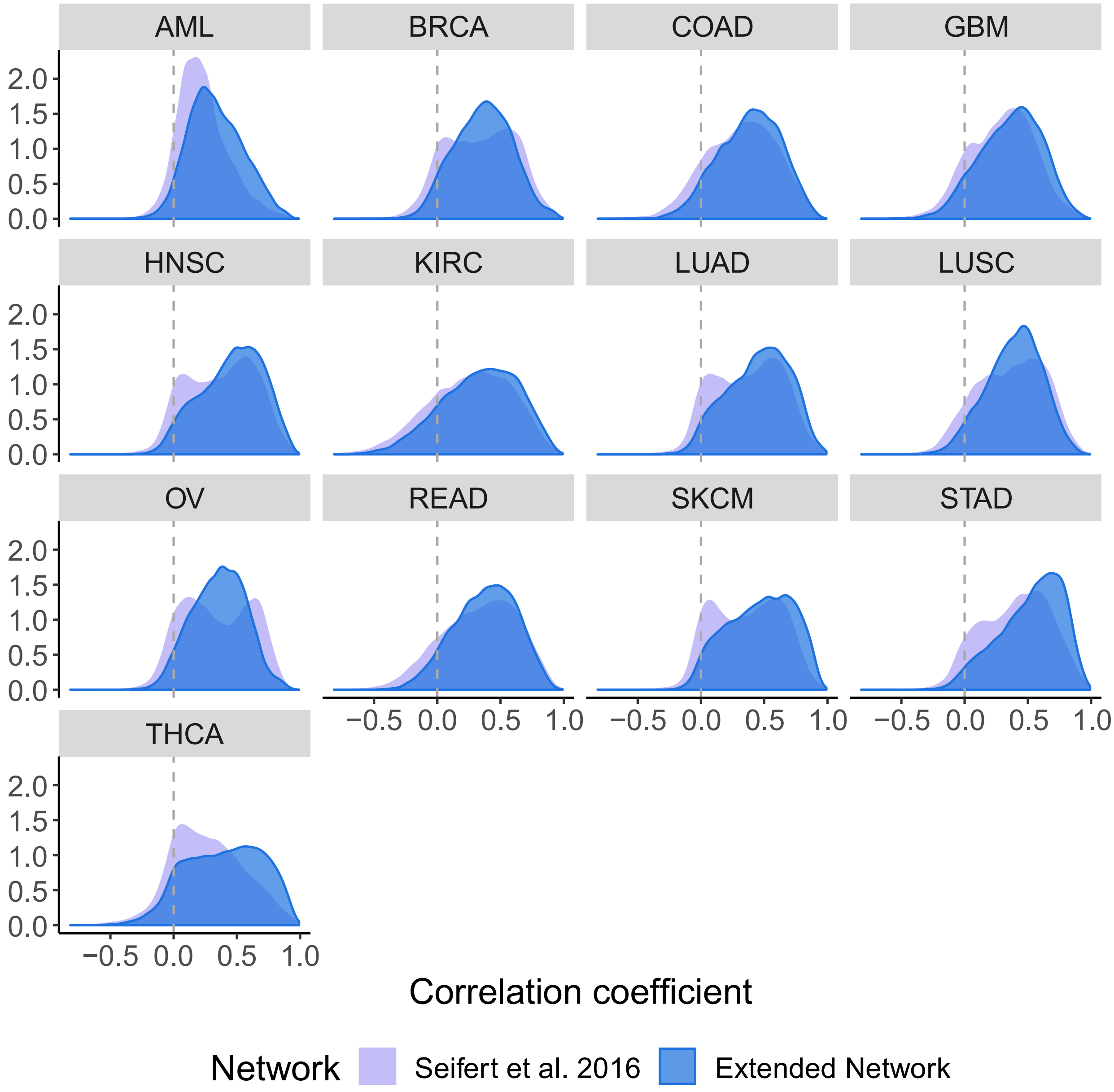
	Baseline	**DrImpute**	**kNN-smoothing**	**SAVER**	**scImpute**	**SCRABBLE, bulk**	**SCRABBLE, sc ref**	**Network**
Smart-seq3 (reads)	0.79	0.71	0.77	0.75	0.78	-	0.72	0.79
Smart-seq3 (UMIs)	0.77	0.72	0.77	0.75	0.74	-	0.70	0.77
hESC differentiation	0.65	0.64	0.66	0.51	0.63	0.39	0.52	0.65
hESC timecourse	0.68	0.66	0.69	0.53	0.56	0.48	0.54	0.67
Oligodendroglioma	0.65	0.41	-	0.29	0.43	-	0.45	0.64
Lung Atlas (10X)	0.70	0.64	0.57	0.51	0.50	-	0.50	0.73
Lung Atlas (FACS)	0.52	0.41	0.49	0.34	0.39	-	0.41	0.53
ccRCC	0.68	0.63	0.54	0.60	0.54	-	0.48	0.73

Supplementary Table 3.3: Percentage of masked dropouts imputed by each method in the tested datasets.

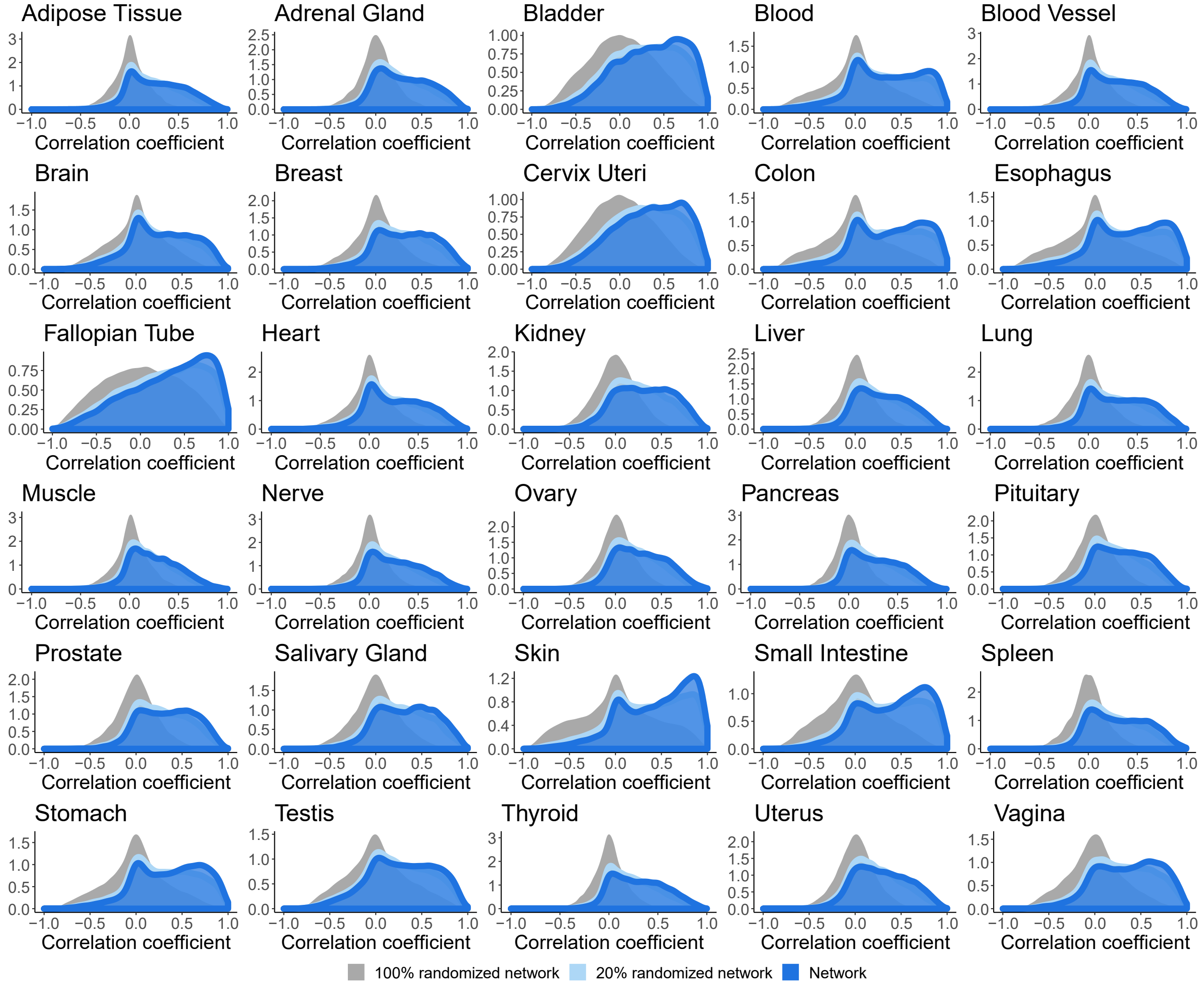
	Baseline	DrImpute	KNN smoothing	SAVER	scImpute	SCRABBLE, bulk	SCRABBLE, sc ref	Network
Smart-seq3 (reads)	81.8%	79.9%	79.2%	81.8%	80.7%	-	82.2%	75.7%
Smart-seq3 (UMIs)	80.6%	78.4%	80.5%	80.6%	72.8%	-	81.6%	77.7%
hESC differentiation	84.1%	83.7%	83.5%	84.1%	78.9%	63.5%	37.8%	87.7%
hESC timecourse	83.9%	83.4%	83.2%	83.9%	74.4%	68.5%	41.8%	87.5%
Oligodendroglioma	84.8%	84.1%	-	84.8%	80.6%	-	24.9%	85.2%
Lung Atlas (10X)	51.1%	49.3%	20.9%	51.1%	20.0%	-	18.8%	68.6%
Lung Atlas (FACS)	65.6%	64.8%	62.7%	65.6%	62.3%	-	15.6%	71.2%
ccRCC	59.7%	58.4%	27.1%	59.7%	29.1%	-	19.5%	73.7%

Supplementary Table 3.4: Percentage of genes best imputed by each method (lowest MSE) in the tested datasets restricted to values that could be imputed by all methods.

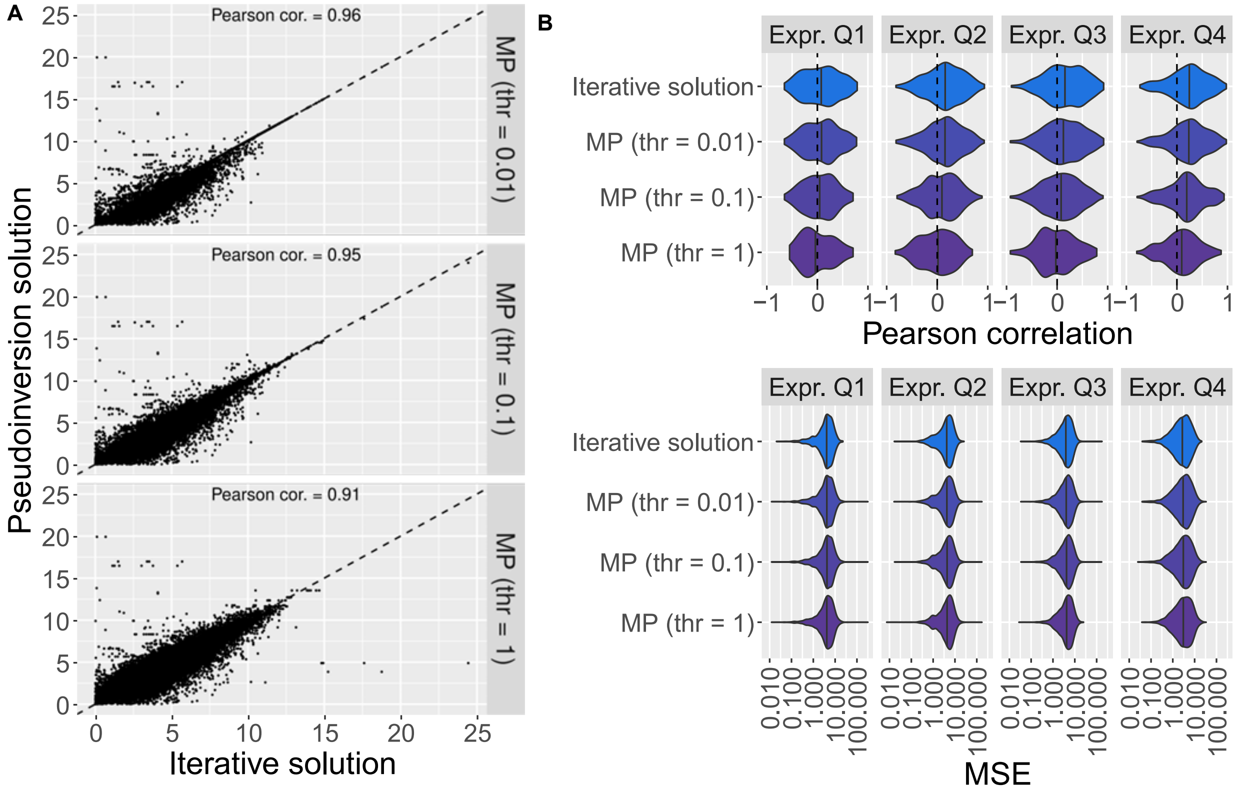
	Baseline	DrImpute	kNN smoothing	SAVER	scImpute	SCRABBLE, bulk	SCRABBLE, sc ref	Network
Smart-seq3 (reads)	**45.2%**	0%	1.0%	0%	9.6%	-	1.7%	42.5%
Smart-seq3 (UMIs)	**48.8%**	0%	3.5%	0%	0%	-	0.3%	48.8%
hESC differentiation	31.1%	0.5%	0.3%	0.1%	21.6%	5.2%	6.2%	**34.9%**
hESC timecourse	41.8%	0.3%	1.3%	0%	0%	4.1%	6.9%	**45.6%**
Oligodendroglioma	**47.8%**	0.2%	-	0%	4.6%	-	1.9%	45.5%
Lung Atlas (10X)	12.1%	0%	13.8%	0%	0%	-	0.1%	**74.0%**
Lung Atlas (FACS)	35.6%	0%	5.5%	0%	2.5%	-	6.9%	**49.4%**
ccRCC	11.3%	0.2%	4.7%	0%	0%	-	0.3%	**83.6%**



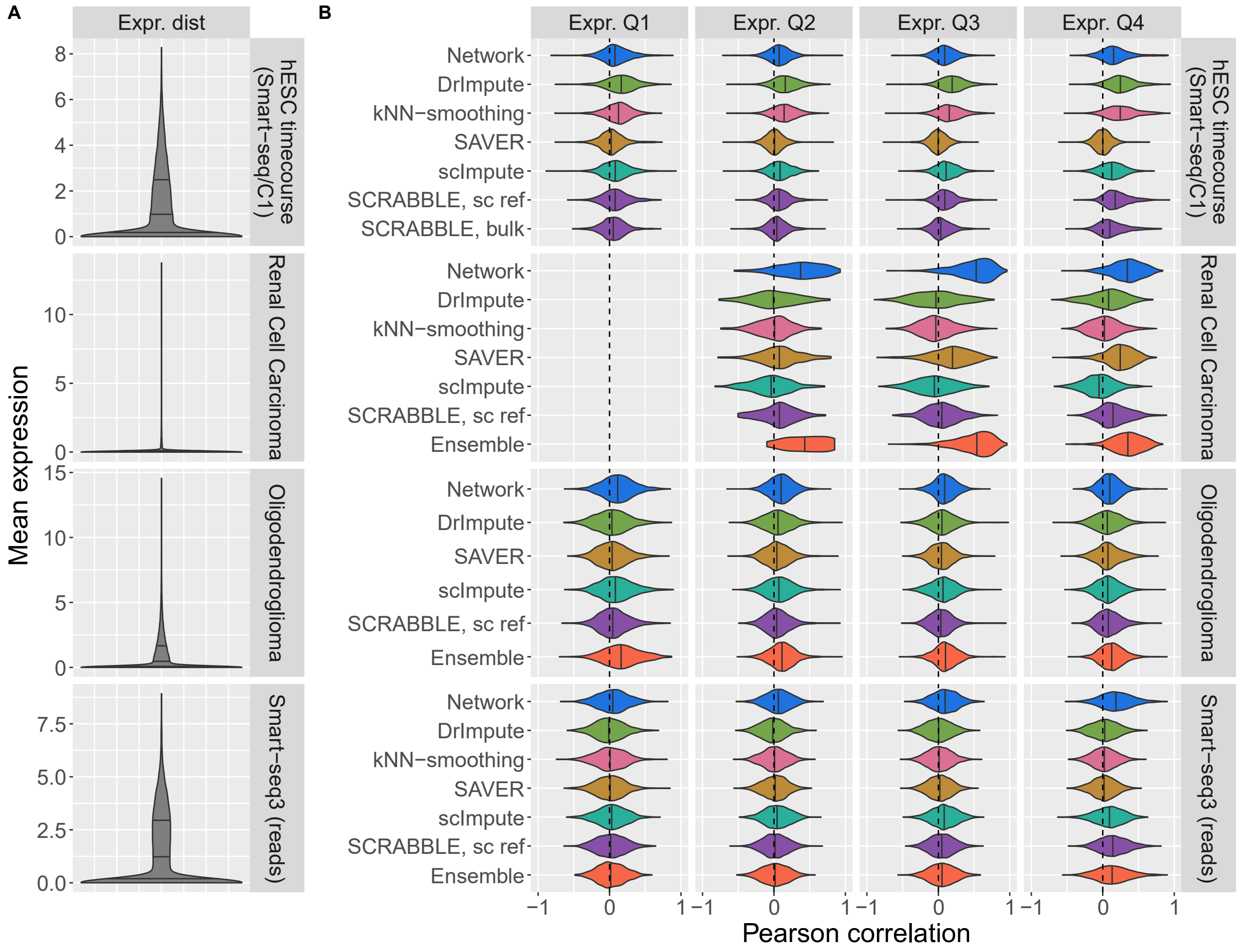
Supplementary Figure 3.1: Correlation between network predictions (using the model from Seifert et al. [27] **(light purple) and the improved network described here (blue)) and measured gene expression in diverse TCGA datasets.** For each gene its expression was predicted in a given tumor sample using the measured expression values of all detected predictors in the model. Subsequently, observed and predicted values were correlated across all samples from one cohort. The plots show the distributions of Pearson's correlation scores across all genes common between the network model and the respective TCGA dataset. Although there is variation with respect to how well genes in different tumor entities can be predicted, the distributions are always strongly skewed in favour of positive correlations. This trend is enhanced with the new model presented here. AML - Acute Myeloid Leukemia; BRCA - Breast Invasive Carcinoma; COAD - Colon Adenocarcinoma; GBM - Glioblastoma Multiforme; HNSC - Head and Neck Squamous Cell Carcinoma; KIRC - Kidney Renal Clear Cell Carcinoma; LUAD - Lung Adenocarcinoma; LUSC - Lung Squamous Cell Carcinoma; OV - Ovarian Serous Cystadenocarcinoma; READ - Rectum Adenocarcinoma; SKCM - Skin Cutaneous Melanoma; STAD - Stomach Adenocarcinoma; THCA - Thyroid Carcinoma.



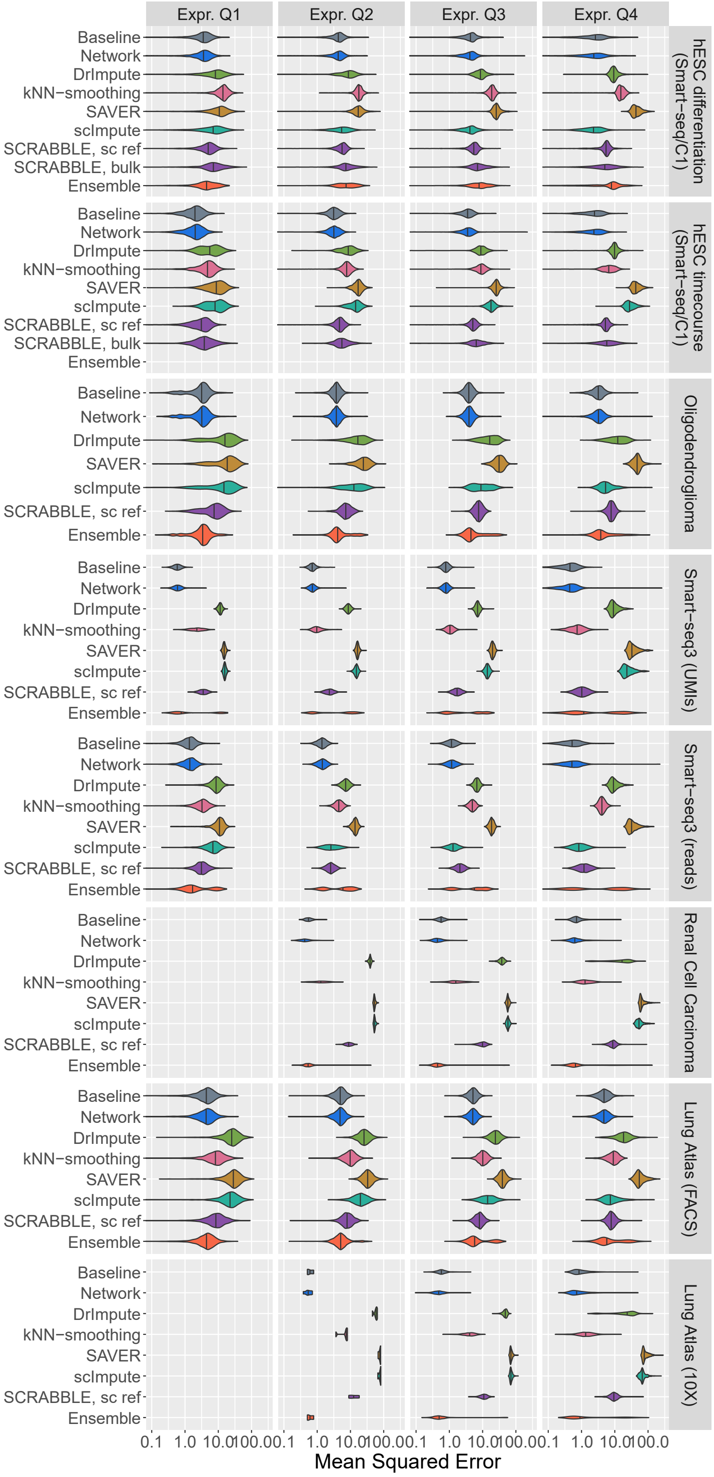
Supplementary Figure 3.2: Correlation between network predictions (using the network described here – blue - , a partially randomized network - light blue - and a fully randomized network – grey) and measured gene expression in diverse healthy tissues from the GTEx consortium. For each gene its expression was predicted in a given healthy tissue sample using the measured expression values of all detected predictors in the model. Subsequently, observed and predicted values were correlated across all samples from one tissue. The plots show the distributions of Pearson's correlation scores across all genes common between the network model and the GTEx data.



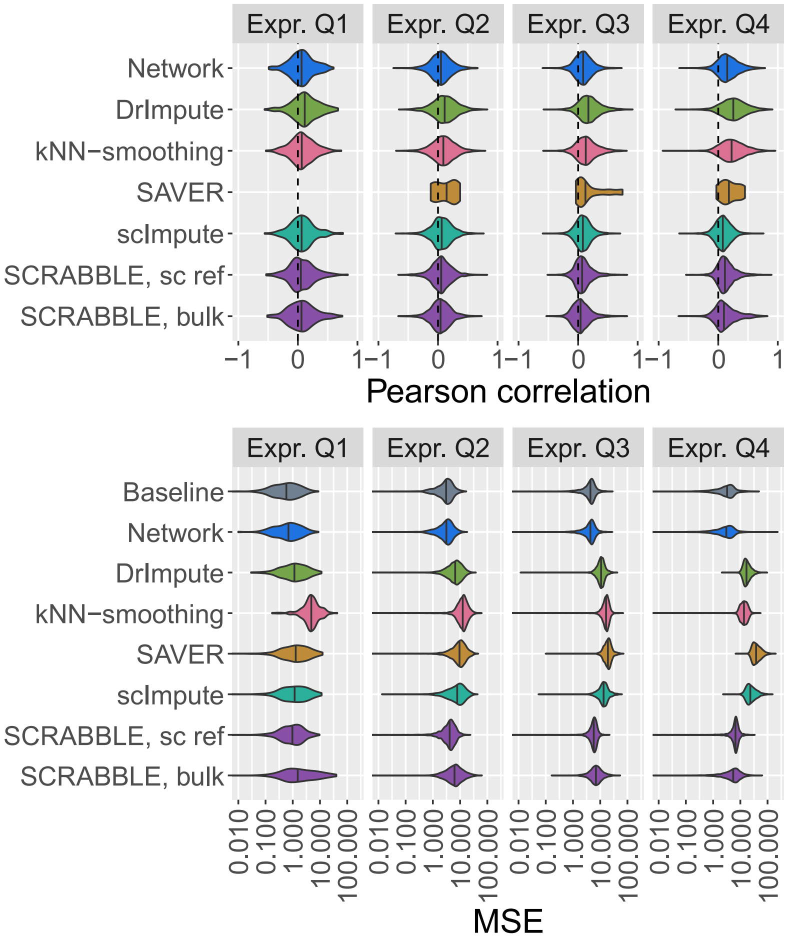
Supplementary Figure 3.3: Comparison between results of the network-based imputation using the iterative approach and the Moore-Penrose pseudoinversion (MP) with varying tolerance thresholds in a random subset of 20 cells from the hESC differentiation dataset. A) Correlation between the results of the iterative approach (x axis) and the Moore-Penrose pseudoinversion (y-axis), across the 20 random cells. B) Imputation performance per gene using the iterative approach and MP with different tolerance thresholds (Pearson correlation, top, and MSE, bottom), separated by expression quartile on the masked data. The higher the tolerance threshold, the fewer singular values are used for the pseudoinversion. Results were limited to imputations performed by all methods.



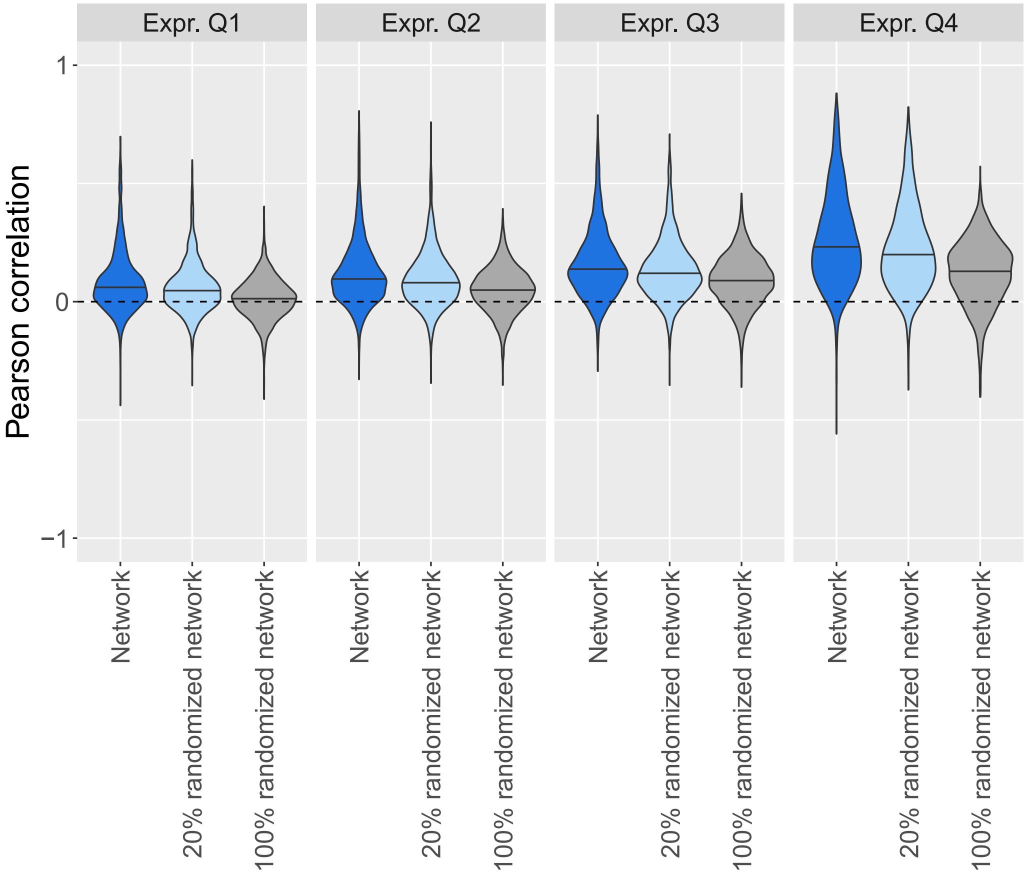
Supplementary Figure 3.4: Imputation performance of Network (blue), DrImpute (green), kNN-smoothing (pink), SAVER (yellow), scImpute (turquoise), SCRABBLE (with bulk or single-cell data as reference; purple) and Ensemble (orange), stratified by expression quartiles, on additional datasets. A) Distribution of average expression levels in each dataset. Quartiles are represented by vertical lines. **B)** Pearson correlation coefficient, for each gene, between the imputation by the specified method and the original values before masking. Only values that could be imputed by all methods were used for correlation computation. Expression quartiles are determined for each dataset separately, on the masked data.



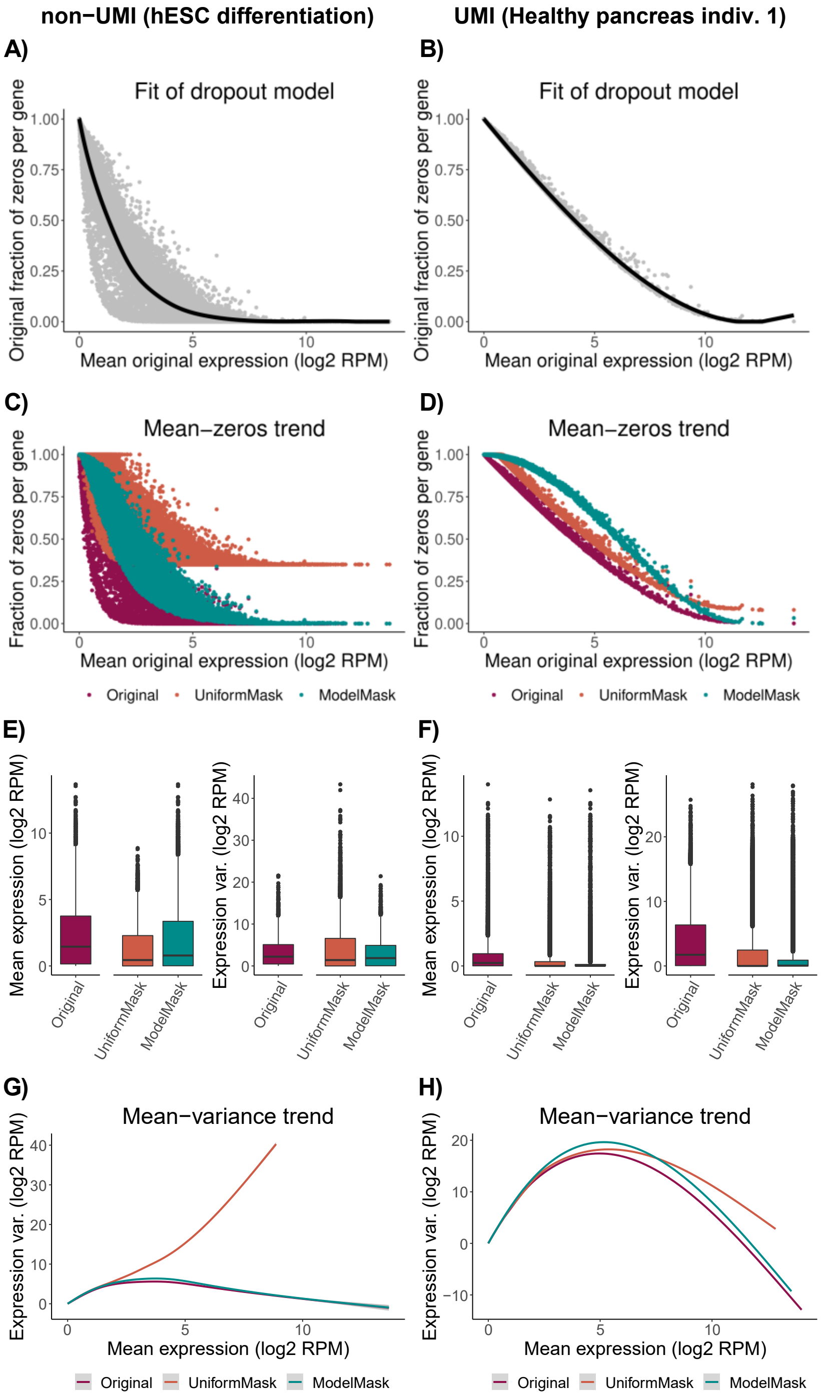
Supplementary Figure 3.5: Imputation error of Baseline (slate grey), Network (blue), DrImpute (green), kNN-smoothing (pink), SAVER (yellow), scImpute (turquoise), SCRABBLE (with bulk or single-cell data as reference; purple) and Ensemble (orange), stratified by expression quartiles. Only values that could be imputed by all methods were used for MSE computation. Expression quartiles are determined for each dataset separately, on the masked data. The x axis is presented log-transformed and was cropped at 0.1 to exclude the low-MSE tail from visualization and facilitate result comparison.



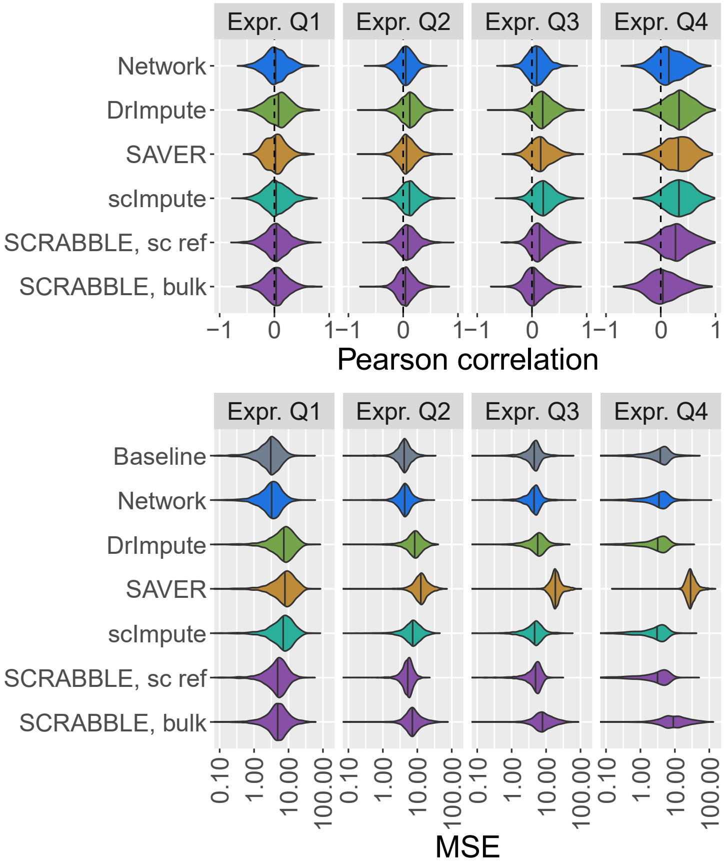
Supplementary Figure 3.6: Imputation performance of Baseline (slate grey), Network (blue), DrImpute (green), kNN-smoothing (pink), SAVER (yellow), scImpute (turquoise) and SCRABBLE (purple) methods in the hESC differentiation dataset, upon cellwise masking. Correlation coefficient between imputed and original values (top) and MSE of imputation (bottom), upon random masking of 30% of the quantified genes in each cell in the hESC differentiation dataset. The Baseline method computes the average expression of a gene across all cells and it is not using the information of any other genes. Hence, one would assume that its error should be independent of the number of missing genes per cell. This is however not the case due to a biased gene sampling: cells with few detected genes will preferentially report values for highly expressed genes, whereas cells with many detected genes will represent a less biased sample of the whole transcriptome. This is affecting the performance, which thus is slightly dependent on the number of missing genes per cell. Values were restricted to imputations performed by all tested methods.



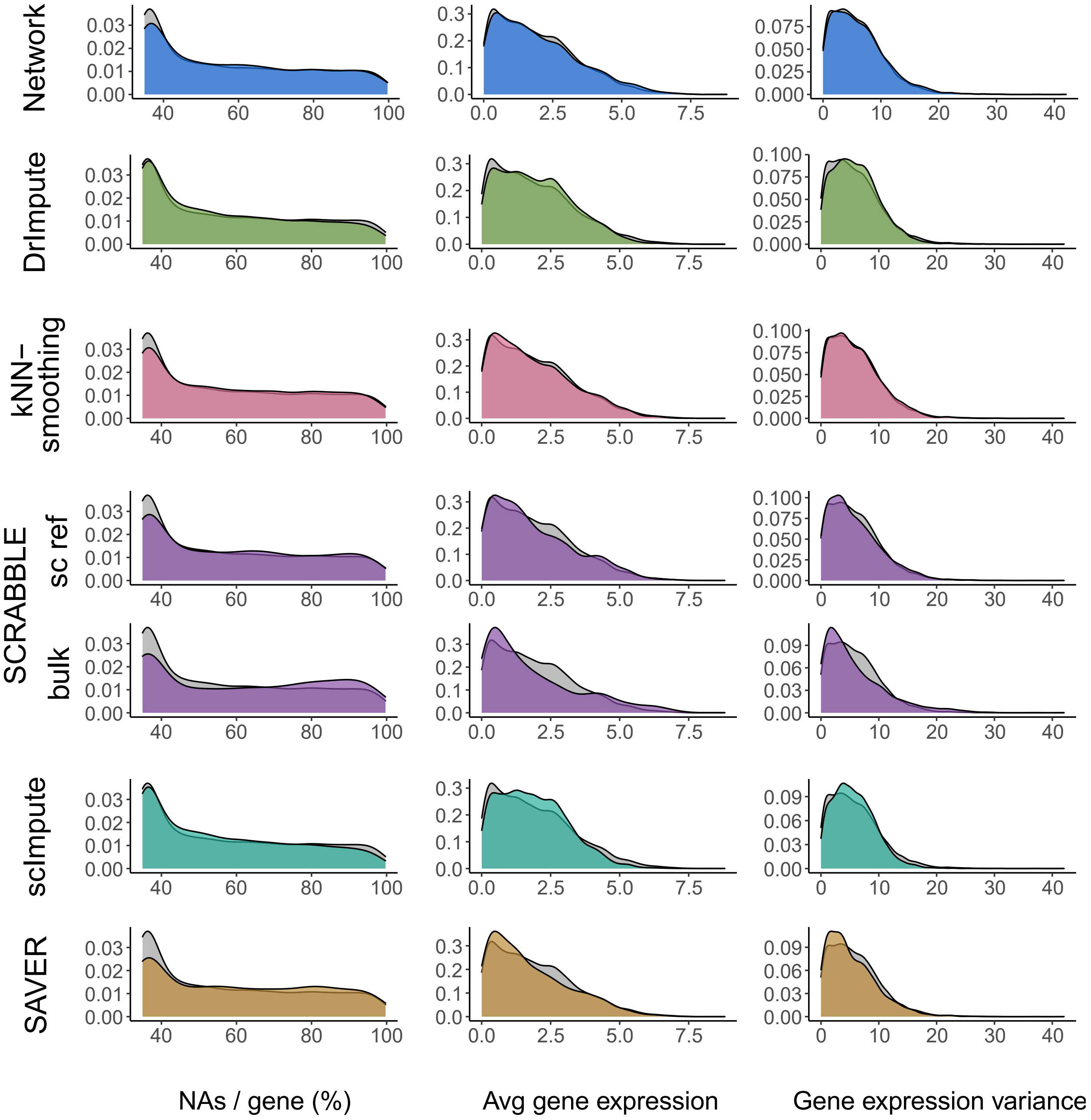
Supplementary Figure 3.7: Imputation performance of the Network method upon randomization in the hESC differentiation dataset. Dropout imputation was performed using the network described here (blue), a partially randomized network (light blue) and a fully randomized network (grey).



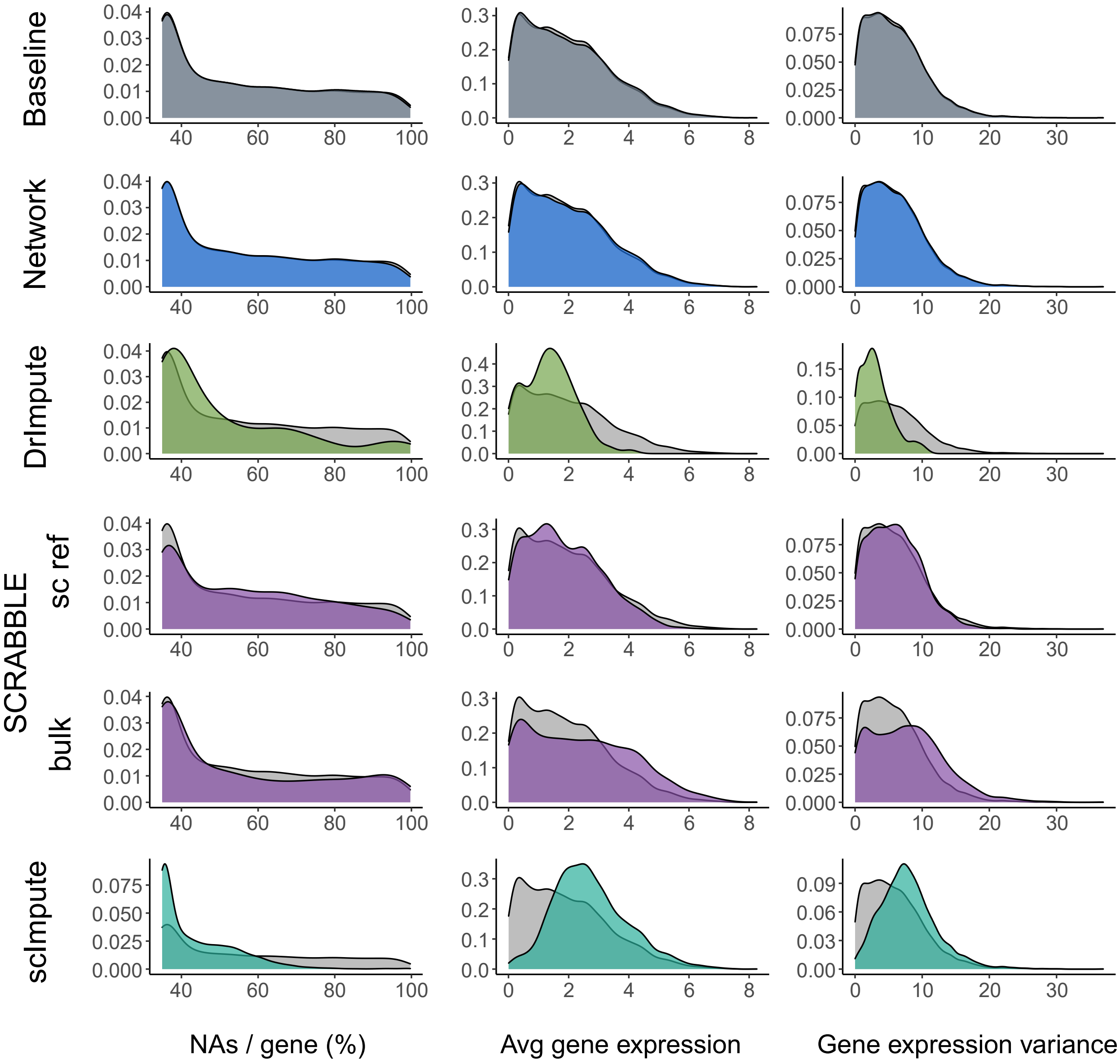
Supplementary Figure 3.8: Comparison of the two masking procedures employed in this work: a uniform masking procedure which sets to zero the same number of entries per gene (UniformMask) and a model-based procedure which sets entries to zero with a gene-specific probability obtained from the data. Comparisons were done on representative datasets of non-UMI data (**A)**, **C)**, **E)**, **G)**) and UMI data (**B)**, **D)**, **F)**, **H)**). **A)** and **B)** Fit of the spline model (Methods) to the original data. **C)** and **D)** Fraction of zeros in the data before (Original) and after (UniformMask, ModelMask) masking, compared to original average gene expression,. **E)** and **F)** Distribution of mean expression and expression variance before and after masking, compared to original average gene expression. **G)** and **H)** Mean-variance trend before and after masking.



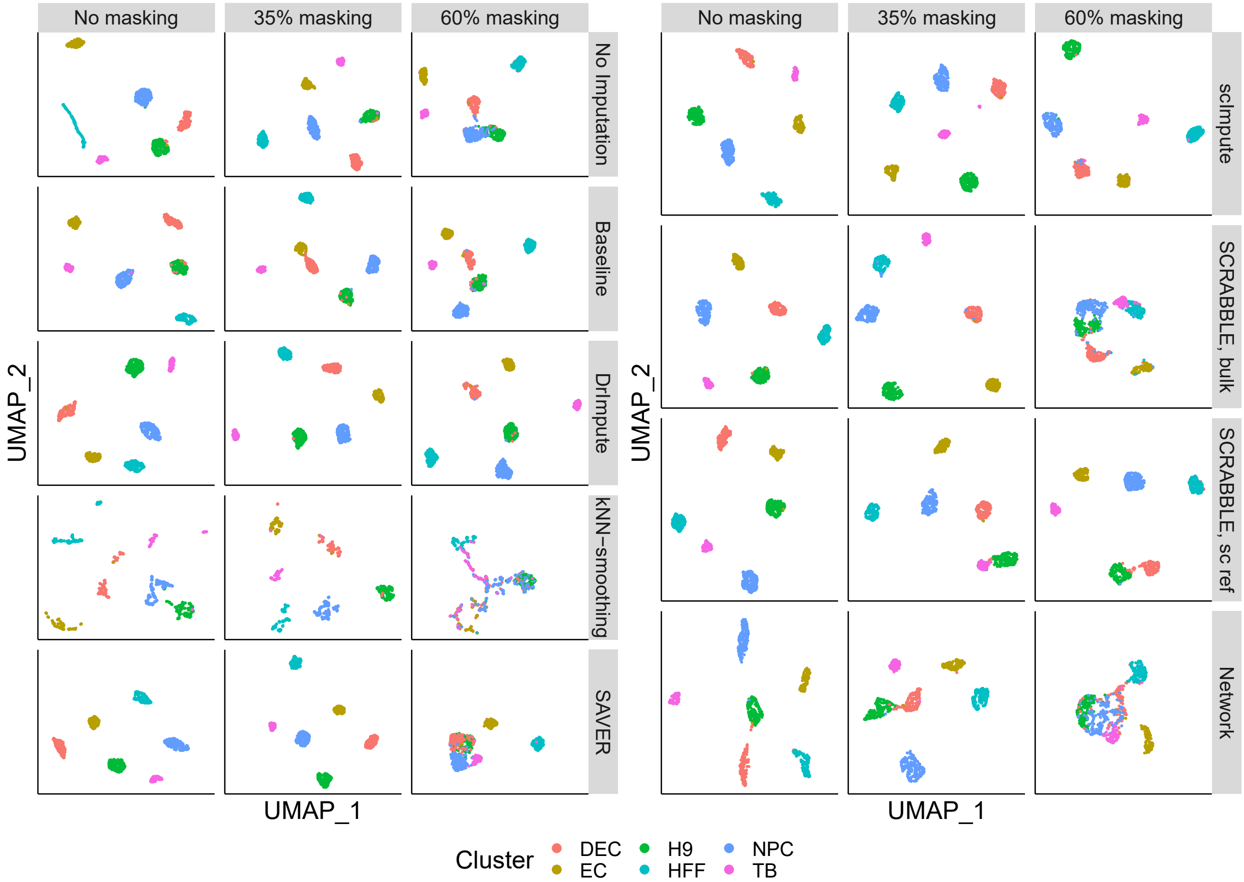
Supplementary Figure 3.9: Imputation performance of Baseline (slate grey), Network (blue), DrImpute (green), SAVER (yellow), scImpute (turquoise) and SCRABBLE (purple) methods in the hESC differentiation dataset, using a gene-specific masking procedure (Methods). Correlation coefficient between imputed and original values (top) and MSE of imputation (bottom), upon random masking of 30% of the quantified genes in each cell in the hESC differentiation dataset. Only values that could be imputed by all methods were used for performance analysis. Expression quartiles are determined on the masked data. The MSE axis is presented log-transformed and was cropped at 0.1 to exclude the low-MSE tail from visualization and facilitate result comparison. kNN-smoothing is not included in the comparison since no imputations could be performed by this method.



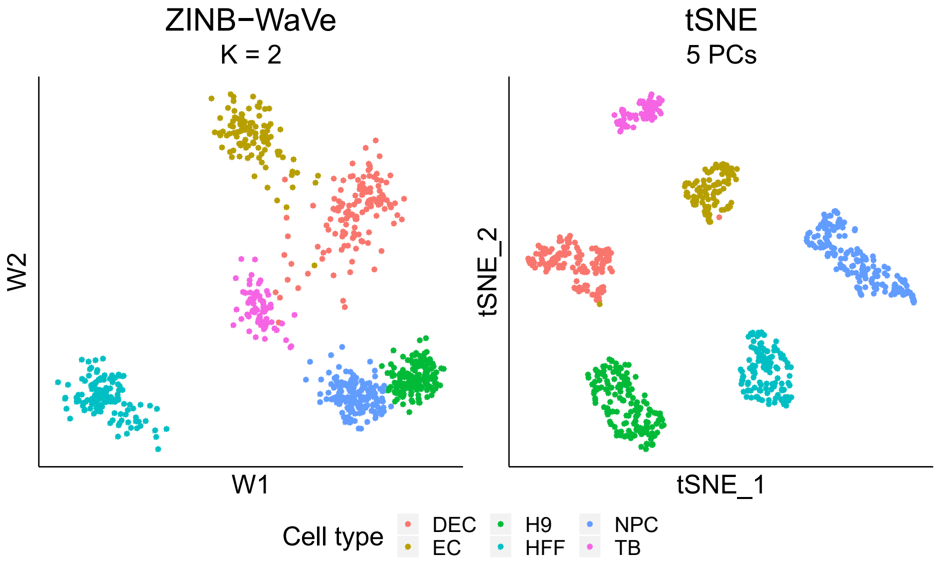
Supplementary Figure 3.10: Characterization of the genes best predicted by Network (blue), DrImpute (green), kNN-smoothing (pink), SCRABBLE (single cell and bulk data as reference; purple), scImpute (turquoise) and SAVER (yellow) in the hESC differentiation dataset. Distribution of missing values per gene, average expression levels and variance of the genes for which a given method is one of the top performers, compared against all tested genes (background). Average gene expression is shown as $\log_2$-transformed normalized expression. Methods with close performance to the best method (correlation not smaller than 0.1 – best method) are selected as top best performing. Genes for which all methods are best performers are included in the background but not in the foreground.



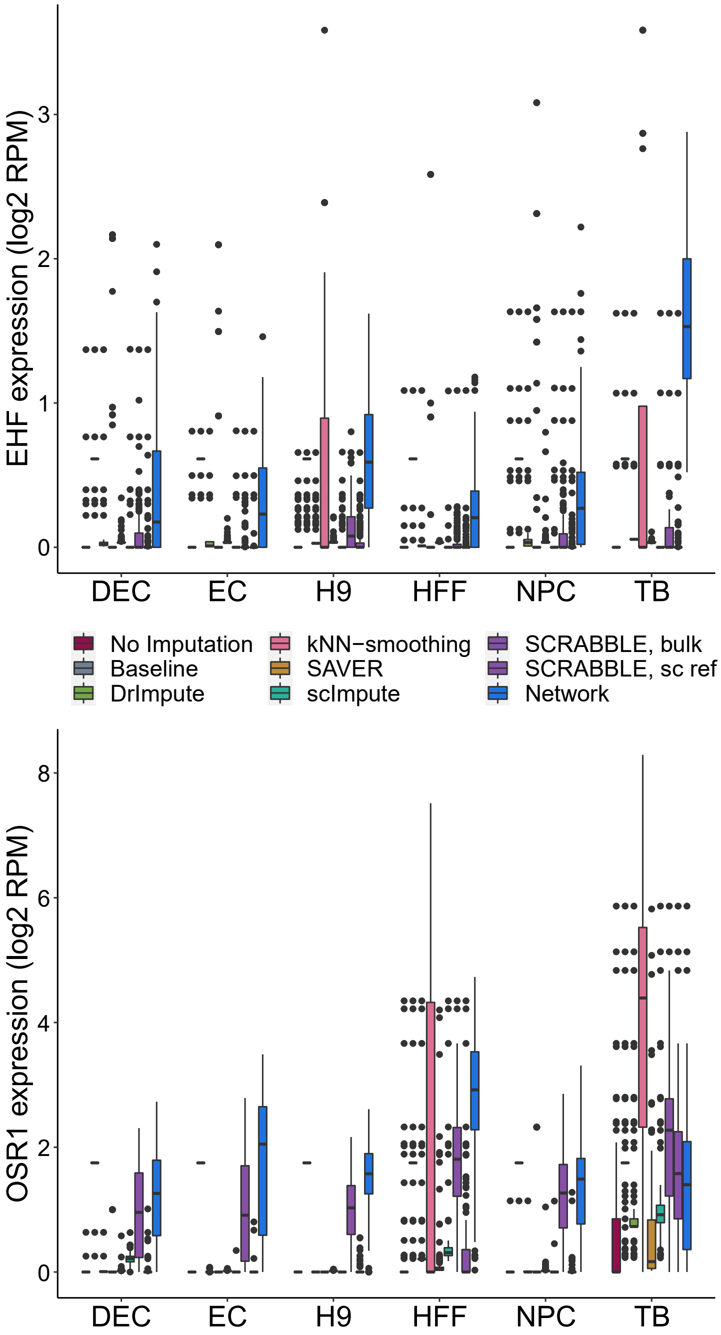
Supplementary Figure 3.11: Characterization of the genes best predicted by Network (blue), DrImpute (green), SCRABBLE (single cell and bulk data as reference; purple) and scImpute (turquoise) in the hESC differentiation dataset, using imputation error to quantify performance. Distribution of missing values per gene, average expression levels and variance of the genes for which a given method is one of the top performers, compared against all tested genes (background). Average gene expression is shown as $\log_2$-transformed normalized expression. Methods with close performance to the best method (MSE not higher than 1/20 of the MSE range for that given gene) are selected as top best performing. Genes for which all methods are best performers are included in the background but not in the foreground.



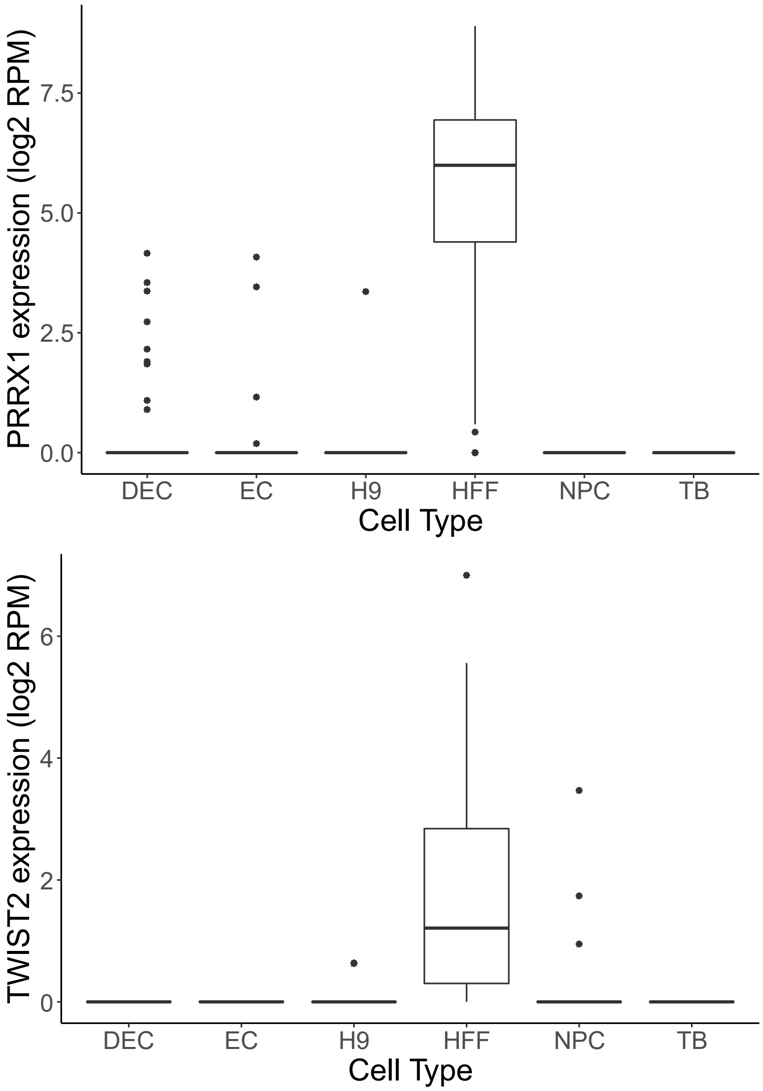
Supplementary Figure 3.12: Effect of imputation with Baseline, DrImpute, kNN-smoothing, SAVER, scImpute, SCRABBLE and Network methods on UMAP plots. Data was subject to: no masking (left column), relaxed masking (35% of quantified entries per gene were set to zero, middle column), stringent masking (60% of quantified entries set to zero, right column). The plot in the upper left reflects the clustering on the original, unchanged data. Imputation was performed for actually missing values in the original data (all columns) and on masked values (columns 2 & 3). Colors represent cell type label annotations from the original publication. DEC: definitive endoderm cells; EC: endothelial cells; H9: undifferentiated human embryonic stem cells; HFF: human foreskin fibroblasts; NPC: neural progenitor cells; TB: trophoblast-like cells.



Supplementary Figure 3.13: Dimensionality reduction results with different techniques: ZINBWaVe and t-SNE on the hESC dataset. Data was not subject to any masking. Colors represent cell type label annotations from the original publication. DEC: definitive endoderm cells; EC: endothelial cells; H9: undifferentiated human embryonic stem cells; HFF: human foreskin fibroblasts; NPC: neural progenitor cells; TB: trophoblast-like cells.



Supplementary Figure 3.14: EHF and OSR1 expression levels before and after dropout imputation, across cell types. DEC: definitive endoderm cells; EC: endothelial cells; H9: undifferentiated human embryonic stem cells; HFF: human foreskin fibroblasts; NPC: neural progenitor cells; TB: trophoblast-like cells.



Supplementary Figure 3.15: PRRX1 and TWIST2 expression levels before dropout imputation, across cell types. DEC: definitive endoderm cells; EC: endothelial cells; H9: undifferentiated human embryonic stem cells; HFF: human foreskin fibroblasts; NPC: neural progenitor cells; TB: trophoblast-like cells.

4. Age-related changes in gene expression regulation across human tissues

4.1 Introduction

Investigation of transcriptomic signatures of ageing has uncovered alterations in several processes, either as possible causes of cellular damage or as adaptive responses to age-related functional decline [106]. These include accumulation of DNA damage, loss of proteostasis, deregulated nutrient sensing, impaired mitochondrial function, accumulation of senescent cells, exhaustion of stem cell niches and altered intercellular communication. Although age-related changes within each of these processes have become better understood, many of the unanswered questions now revolve around the impact of ageing on the regulation of each of these processes individually, and in synchrony with other processes.

One of the standing questions in the field is how universal is age-related gene expression de-regulation, if it takes place at all. Early work has reported an age-related increase in the variability (noise) of expression levels of individual genes across cells in some cell types, but not others [59], [60]. Further work making use of single-cell RNA-sequencing data has led to the conclusion that increased cell-to-cell variation is not universally observed across all genes and cell-types [61]–[63].

Another open question is to what extent synchrony of gene expression becomes impaired with age. Since genes operate in functional modules [107], [108], expression coordination within and between gene modules is required to orchestrate a functional cellular response. For instance, impact of ageing on the coordination of T cell activation in response to stimuli has been addressed using single-cell RNA-sequencing data to show that T cell activation is impaired (weaker and more variable) in old mice compared to young [70]. Given the complexity of regulatory interactions between cellular processes, addressing the impact of ageing on within-and between-module coordination requires a systems-level approach. In this direction, Southworth and colleagues compared co-expression networks of young and aged mice [64] and found a tendency for a global decrease in pairwise gene correlations in old mice, impacting ribosome biogenesis, transcriptional regulation and mitochondrial functions, along with an increase in correlations between DNA-damage genes. More recently, Levy and colleagues developed a metric of overall transcriptomic coordination, which decreased with age in different organisms and cell types [65]. However, a comprehensive, cross-tissue analysis of the impact of ageing on the crosstalk between individual gene modules is still lacking.

Here, we describe an approach for the identification of alterations in gene-gene relationships with age in humans. First, we report a regulatory model, derived from more than one thousand expression data sets, that is capable of capturing tissue-specific and global (cross-tissue) gene-gene relationships. Motivated by this wide applicability, we use existing RNA-sequencing data of human tissues at different ages to investigate age-related changes in gene-gene relationships. We observe age-related changes in both directions – towards gain and loss in gene-gene relationship strength – and characterize differences, as well as similarities, between tissues. We observe the most widespread changes in genes involved in mitochondrial respiration and cell cycle regulation, and provide examples of age-related regulation changes in tissue-specific functions, as well as in the crosstalk between different cellular functions.

4.2 Results

4.2.1 Capture of gene-gene relationships from transcriptome data

In order to illustrate the modular nature of gene co-expression, we made use of RNA-seq data collected *postmortem* by the GTEx consortium [109] from 30 different human tissues, spanning 948 donors. We hand-picked 5 sets of genes, constructed based on GO term membership, and computed all pairwise Pearson correlations in two tissues with clear differences in cellular composition and function – Brain and Blood. In both tissues, we observed strong within-set correlations (Figure 4.1.B, diagonal blocks), in line with a modular organization of gene expression. Additionally, we observed between-set correlations (Figure 4.1.B, off-diagonal blocks), particularly between genes encoding for members of the mitochondrial respiratory chain and RNA polymerase II (Pol-II) core complex (in orange and light green, respectively), representative of regulatory crosstalk between functional modules. Notably, we observed differences in within- and between-module correlation between Brain and Blood.

Next, we aimed to investigate the extent to which this gene expression coordination within and between modules is conserved across tissues and cell types. To capture robust regulatory relationships that explain expression coordination across different tissues, we inferred gene-gene relationships using transcriptomic data collected from 1443 cancer cell lines [110], [111]. In order to derive these relationships from large transcriptomic datasets, we have devised a procedure that selects a small set of genes whose expression patterns are, as a linear combination, most predictive for the expression of a given target gene. These expression patterns correspond to the relative deviation, in each sample, with respect to the cross-sample (global) average. The network resulting from this procedure is much sparser than computing all pairwise correlations between all genes: only a very small fraction (0.04%) of gene pairs have non-zero relationships (ignoring directionality). These gene-gene relationships do not exclusively represent causal relationships between regulators and their targets. Instead they represent gene pairs that are robustly co-expressed across a wide range of human tissues and cell types [111].

Using the gene-gene relationships captured by our regulatory model, we reconstructed the expression pattern of each gene across samples of the same tissue (Figure 4.1.A, Methods). Thus, the reconstructed data consists of the expected (predicted) pattern for each gene given the expression pattern observed for its regulatory neighbours (i.e. directly connected genes in the network, corresponding to the most strongly co-regulated gene pairs). We observed that tissue-specific differences in the gene-gene correlation patterns observed in the original data (Figure 4.1.B) were largely preserved in the reconstructed data (Figure 4.1.C) – for instance, the strong anti-correlation between some of the extracellular matrix components in Brain, but not in Blood. This was true not only for within-module correlations, but also for correlations between different modules, as is the case between genes encoding for peptide antigen binding partners and other modules. This observation confirms that our network captures regulatory neighbourhoods that are conserved across cell types and tissues, and – at the same time – enables prediction of tissue-specific expression levels.

We note that the capacity of our model to correctly capture relevant regulatory neighbourhoods is dependent on the gene and dataset at hand [111]. To gain a better understanding of the limitations of our model, we identified genes whose observed expression profiles differ strongly from the predicted based on their regulatory neighbourhood (Methods). Further analyses revealed that most of the poorly predicted genes fell into one of two groups. The first group consisted of lowly expressed genes with few regulatory neighbours, which themselves were also lowly expressed (Supplementary Figure 4.1). Thus, these genes correspond to network regions that were essentially turned off in the respective tissue or cell type and therefore the observed expression variation reflected mostly background noise. The second group corresponded to highly expressed genes with extremely low variance and ubiquitous expression across tissues (Supplementary Figure 4.1), suggesting that these are housekeeping genes with essentially constant expression patterns. Since our model tries to explain expression variation across samples (i.e. relative differences in expression between different samples), a constant expression pattern results in virtually no relative differences between samples and is thus hard to model.

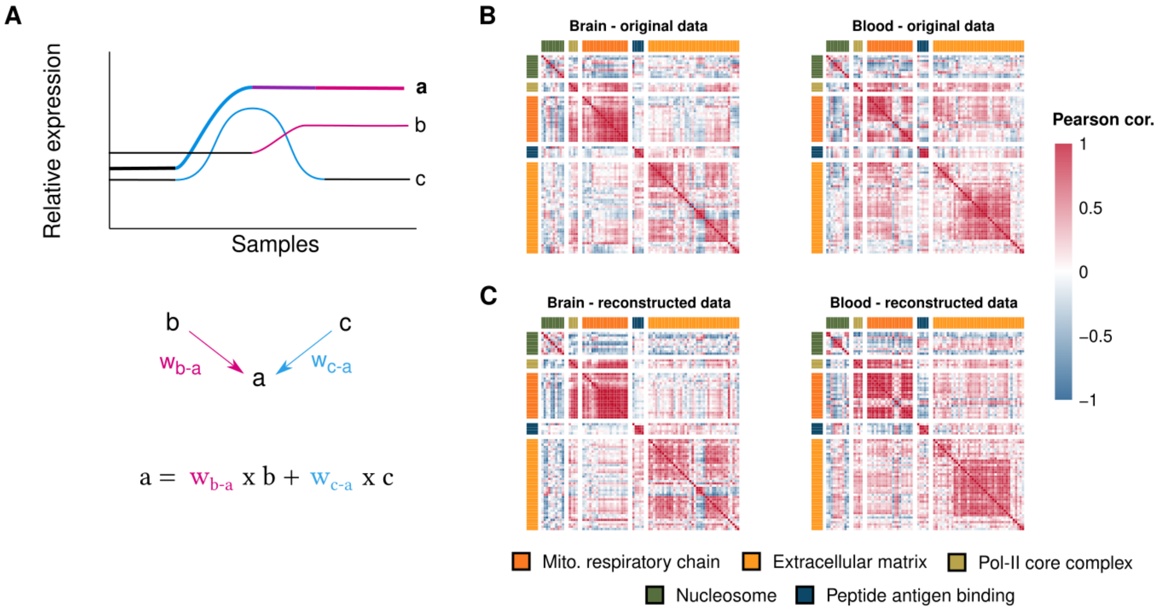


Figure 4.1: Representative gene-gene relationships captured by pairwise correlation and a gene regulatory network (GRN)-based approach. A) Methodological approach used to capture gene-gene relationships in our regulatory model. Through a combination of regularized linear regression (Lasso) and stability selection, we identify stable predictors for each gene in the transcriptome, i.e., genes whose expression pattern across the training data (cancer cell line transcriptomic data) is informative of the expression pattern of the target gene. A linear model is then fit to the explain the expression pattern of the target gene (a) based on the pattern of the stable predictors (b and c). The weights of this linear model can then be used in other datasets to reconstruct the expression pattern of the target gene a based on the expression pattern of the stable predictors b and c observed in those datasets. **B), C)** Pairwise Pearson correlation coefficients, separated by gene set, for selected cellular functions. Correlations were computed based on the original expression data (**B**) or using the reconstructed expression values based on model predictions from Panel A (**C**). For illustrative purposes, gene sets were restricted to the following functions: respiratory chain complex members I to IV (GO:0045271, GO:0005749, GO:0005750, GO:0005751, orange), components of collagen-containing extracellular matrix (GO:0062023, yellow), polymerase II core complex members (GO:0005665, light green), nucleosome members (GO:0000786, dark green) and peptide antigen binding partners (GO:0042605, blue).

4.2.2 Age-related changes in gene-gene relationships across tissues

We next took advantage of the wide applicability of our model to investigate age-related changes in expression regulation across multiple human tissues.

Our first step was to establish a metric of (de)regulation, based on the extent of agreement between the expression profile of a given gene in the data and the reconstructed profile according to its regulatory neighbourhood (i.e. comparing observed *versus* predicted expression levels). To quantify this agreement, we used Spearman's rho, computed between the expected and observed expression of a gene across samples. We refer to this metric as the "predictability" of a gene in each group of samples (Figure 4.2.A). Our choice of Spearman over Pearson correlation was based on its lower sensitivity to outliers, which avoids that a few individuals with outlying expression of the target gene dictate the overall predictability.

To capture age-related changes in predictability, we made use of the wide age range of the donors in the GTEx dataset. We restricted our analysis to the 20 tissues with highest sample number and split the data into 6 similarly sized age groups, defined as age decades (20-29 up to 70-79, Methods). Splitting the samples by decade resulted in 30 to 62 samples per tissue and age split (Supplementary Table 4.1). We then computed the predictability of each gene in every tissue-age split and restricted our analysis to genes showing sufficiently high average predictability across age groups (Methods). This resulted in the selection of 3291 to 5830 genes per tissue (Supplementary Table 4.3). We then regressed predictability of these genes as a function of age using a linear model ($Predictability \sim Age$). This procedure resulted in one slope for each gene in each tissue, quantifying the change of its predictability with age. To determine the statistical significance of the resulting predictability slopes, we compared the p-value distribution for the obtained regression slope with the null p-value distribution, obtained after repeatedly shuffling the age groups (Methods). We observed large differences in the predictability signatures of different tissues, with some tissues showing an enrichment in small p-values, but not others (Supplementary Figure 4.2). We focused our subsequent analysis on 8 tissues with substantially more genes showing an age-dependent predictability change than expected by chance, i.e. those tissues with an inflation of small p-values (Adipose – Visceral, Artery – Tibial, Blood, Brain, Breast – Mammary Tissue, Esophagus – Mucosa, Testis and Thyroid, Supplementary Table 4.1, Figure 4.1). When comparing the predictability slopes of individual genes across these tissues, we observed that many genes showed increasing or decreasing predictability changes consistently across multiple tissues (Figure 4.2.C). We selected for further analysis in each tissue the 100 genes with the most significant predictability changes with age (lowest regression p-values), independent of the direction of that change. Here we also observed considerable differences between tissues, with Blood, Thyroid and Adipose Tissue showing mostly decreases in predictability with age, and Testis, Breast, Brain and Esophagus showing mostly increases in predictability (Figure 4.2.D). To exclude that these differences between tissues are due to artefacts in our analysis, we repeated our analysis with similar sample numbers (n = 30) across tissues and age groups (Supplementary Figure 4.2, Supplementary Table 4.2), which removed differences between tissues due to different sample sizes. This resulted in statistical significance (Supplementary Figure 4.2.B) and slope directions (Supplementary Figure 4.2.D) consistent to those obtained with larger sample sizes. Finally, we also repeated our analysis excluding the age group 70-79, as the smaller number of samples in this group might have influenced the predictability values and skewed the regression slope (Supplementary Table 4.1). However, we once again observed consistent statistical significance (Supplementary Figure 4.3.B) and preference towards positive or negative slopes (Supplementary Figure 4.3.D) upon exclusion of this age group. Put together, these results suggest that predictability is affected by age in some tissues more than others, and that the direction of this effect varies between tissues.

We next asked which cellular functions were most affected by the observed age-related predictability changes. To identify network regions that are enriched for genes with age-associated predictability changes, we performed network propagation of the predictability slopes using a Random Walk with Restart algorithm [112] (Methods). We then performed Gene Set Enrichment Analysis [113] (GSEA) on the slopes after propagation, to identify gene sets with significant enrichment in age-related predictability changes in each tissue (Methods). We focused our analysis on gene sets (hallmarks, Figure 4.2.E or GO terms, Supplementary Figure 4.4) showing a significant enrichment in at least one tissue. Our analysis revealed that genes involved in oxidative phosphorylation were among the most affected by predictability changes across several tissues, showing a predictability decrease in all tissues apart from Brain and Testis (Figure 4.2.E, Supplementary Figure 4.4). Additionally, genes involved in the G2M checkpoint and E2F targets were identified as consistently enriched in predictability changes (mostly a decrease) across tissues. The G2M checkpoint leads to a cell cycle arrest at gap phase G2, before mitosis (M), in case of DNA damage, while E2F transcription factors regulate the expression of genes involved in the transition between the G1 and S (DNA synthesis) phases of the cell cycle. The identification of these two gene sets suggests an age-related predictability decrease, across several tissues, in genes involved in cell cycle regulation. On the other hand, targets of the oncogene MYC and genes involved in epithelial-mesenchymal transition (a phenomenon occurring during development, tissue repair and carcinogenesis [114]) showed different predictability trends in different tissues (Figure 4.2.E). In Esophagus and, to a smaller extent, Thyroid, Adipose Tissue and Blood, predictability of these genes decreased with age. We observed a trend in the opposite direction for Testis, Brain, Artery and Breast. Of note, gene sets with tissue-specific functions also showed age-related predictability changes in the respective tissue. Among these, we highlight the enrichment in predictability decrease among fatty acid metabolism genes in Adipose Tissue (Figure 4.2.E), the increased predictability of coagulation genes in Blood (Figure 4.2.E) and the decreased predictability of synaptic genes in Brain (Supplementary Figure 4.4).

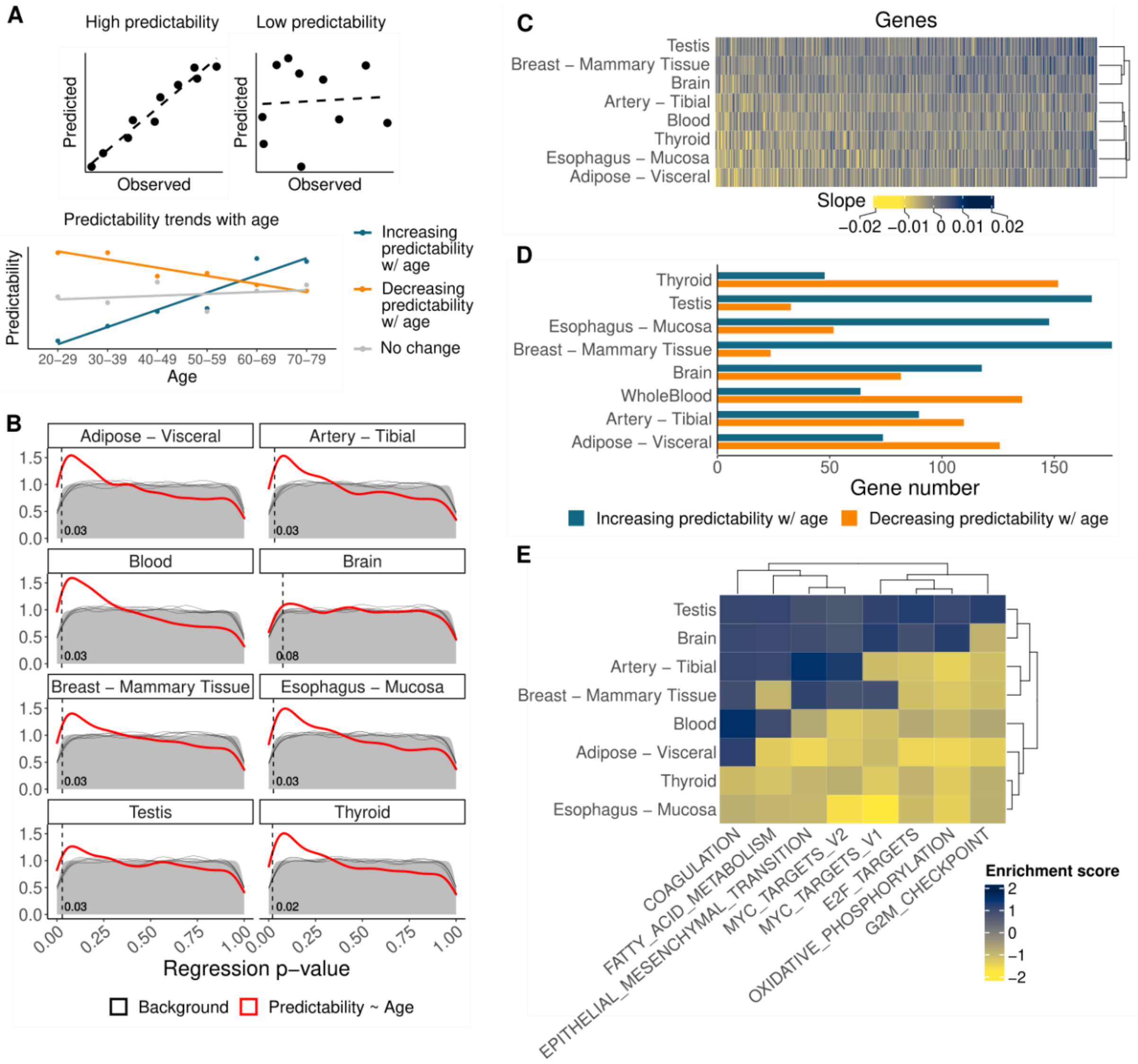


Figure 4.2: Transcriptome-wide predictability changes with age across tissues. A) Computational approach used to identify predictability changes with age. Predictability is quantified as the Spearman correlation between the observed expression patterns (in the original data) and the predicted expression patterns (in the reconstructed data), corresponding to the expected pattern given the expression of regulatory neighbours. A high correlation indicates that the expression pattern of a gene fits the regulatory relationships captured by the model (top left), while a low correlation indicates the opposite. Predictability is quantified for groups of samples at 6 different age groups: the decades spanning 20-29 to 70-79. For each gene, predictability is modelled as a linear function of age. **B)** Distribution of p-values for the regression of predictability values within each age group against the mean age of the age group. Red line: p-value distribution obtained from the real data. Grey background: average p-value distribution across 100 permutations of the age groups. Black lines: 5 individual permutations randomly picked from the background. The dashed vertical line indicates the highest p-value among the genes considered statistically significant in each tissue (orange and blue bars in D). The number of genes included in each tissue-specific analysis can be found in Supplementary Table 4.3. **C)** Heatmap of the predictability slopes across all 376 genes, independently of significance level, for which the regression analysis was performed in all 8 tissues. These 376 genes had a high average predictability in all 8 tissues (Supplementary Table 4.3). **D)** Number of genes with predictability increase (blue) and decrease (orange) among the top 100 most significant genes per tissue. **E)** Hallmark gene sets enriched in age-related gene-gene relationship changes, captured by Gene Set Enrichment Analysis. The heatmap shows all hallmarks with statistically significant (FDR < 0.1) enrichment in at least one tissue.

4.2.3 Regulatory relationship changes with age

Different factors can impact the predictability of a gene. One of such factors is the average expression level of the gene at hand: if average gene expression levels are low, the quantification becomes noisier (fewer reads per gene), resulting in lower predictability (Figure 4.3.A, “Low expression”). Another factor is gene expression variance, as a small range of expression values impairs correlation quantification and thus predictability (Figure 4.3.A, “Low variance”). Finally, low predictability can also result from changes in the structure of the regulatory model itself, i.e. if the regulatory relationships captured by our model are not met. This can be quantitatively captured as a low correlation between the gene at hand and its regulatory neighbourhood (Figure 4.3.A, “Correlation loss”).

First, we quantified average expression level and variance changes with age. We observed a clear trend towards variance increase with age across all tissues except of Artery (Figure 4.3.B, y-axis), in line with earlier reports of increased inter-individual transcriptomic variability with age [115], [116]. The direction of average expression changes with age varied across tissues, with Artery, Brain and Breast showing mostly expression decrease with age, Thyroid showing mostly expression increase, and the remaining tissues showing comparable expression changes in both directions (Figure 4.3.B, x-axis). We found the quantification of average expression and variance changes with age to be influenced by the 70-79 age group in some tissues (Supplementary Figure 4.5 *versus* Figure 4.3.B). For this reason, the process described below for the identification of genes with predictability, average expression, variance, or correlation changes is limited to genes showing the same trend when including and excluding the 70-79 age group (Methods).

To quantify the contribution of the factors in Figure 4.3.A (average expression, variance, and correlation to regulatory neighbourhood) to age-associated predictability changes, we compared age-related changes in each of the 3 factors with age-related predictability changes. We applied a cut-off on the absolute values of the expression fold changes or variance slopes with age (0.001), and then counted the number of predictability hits (either increase or decrease in predictability) that also showed age-related changes in expression levels (Figure 4.3.C, “Expression changes”) and in expression variance (Figure 4.3.C, “Variance changes”).

To capture correlation changes between the target gene and its regulatory neighbours, we regressed each of the correlation coefficients (one per neighbour) in each age group against the average age of the group ($Correlation \sim Age$). We applied a cut-off on the absolute values of the slopes (0.005) and took a weighted average of the slopes passing this cut-off (with weights corresponding to the weights of the respective neighbour in the regulatory model, Figure 4.1.A). We then compared the number of predictability hits (either increase or decrease in predictability) with the number of positive versus negative averaged correlation slopes (Figure 4.3.C, “Correlation changes”).

We observed relatively few genes whose age-associated expression changes were linked to predictability changes (Figure 4.3.C, “Expression change”), suggesting that age-related differences in expression levels do not explain the predictability changes captured by our approach. Genes with increased variance with age almost always increased their predictability and, conversely, genes with decreased variance mostly decreased their predictability with age (Figure 4.3.C, “Variance change”), suggesting that age-related changes in gene expression variance influence changes in predictability. Correlation changes were the most strongly linked to predictability changes (Figure 4.3.C, “Correlation change”) – in all tissues analysed, predictability hits showed stronger age-related changes in correlation with their regulatory neighbours than in average expression levels or variance. Furthermore, the direction of correlation change matched the direction of predictability change: genes with decreasing predictability showed decreasing correlation to their regulatory neighbours, and the opposite trend was observed for genes with increased predictability.

Put together, our analyses suggest that predictability changes captured by our regulatory model are rarely explained by age-related changes in expression levels. Instead, the relationships between genes (between targets and their predictors) seem to change with age for a large number of genes.

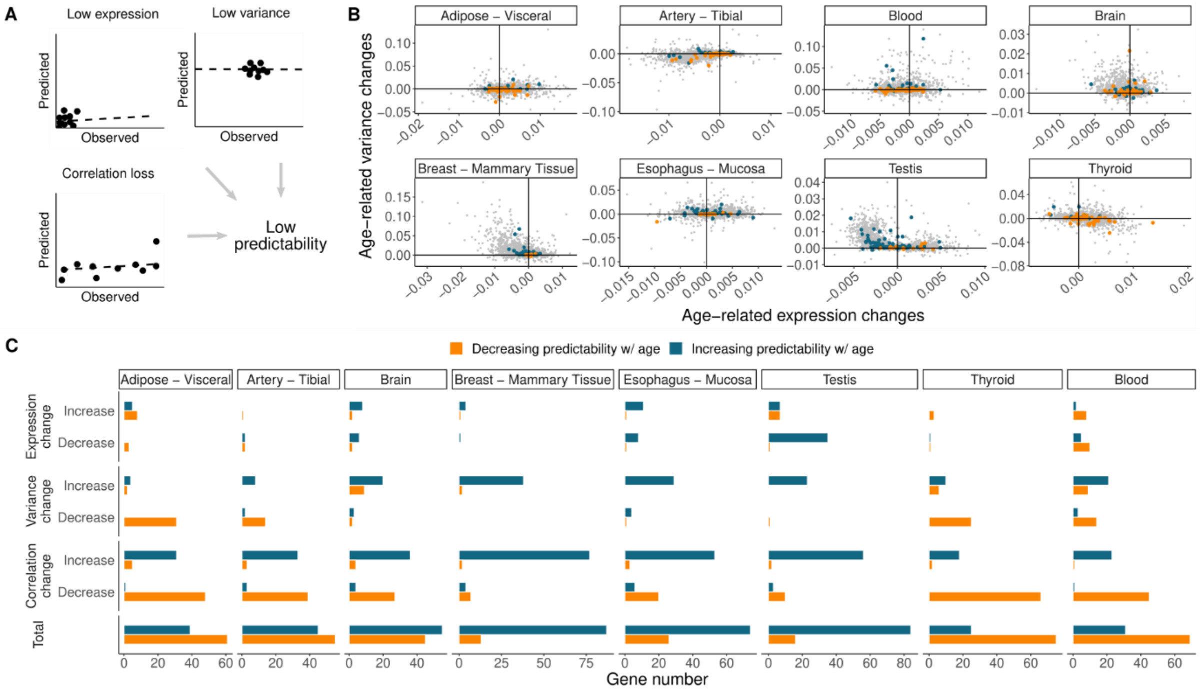


Figure 4.3: Factors underlying age-related predictability changes. A) Possible factors explaining observations of low predictability, and resulting comparisons of observed values (in the original data, x-axis) versus predicted values (data reconstructed with our regulatory model, y-axis). Top left: low average expression of the target gene results in noisy quantifications. Top right: low variance of the target gene results in almost constant expression values. Correlation approaches thus capture noisy fluctuations around the mean. Bottom left: Loss of correlation with the regulatory neighbourhood results in a failure of the model to correctly capture the expression pattern of the target gene. **B)** Expression slope across age (x-axis) against variance slope across age (y-axis) for background genes (small light grey points), top significant genes with predictability increase (large blue points) and decrease (large orange points) with age. **C)** Number of genes with predictability decrease (orange) or increase (blue) in each of the 3 scenarios tested: “Expression change”, “Variance change” or “Correlation change”. The total number of predictability hits is shown in “Total” as a reference.

To assess whether age-related predictability changes tend to impact individual genes or are spread throughout a wider regulatory neighbourhood, we focused on the subnetworks created by the 20 strongest (highest absolute value of age slopes) predictability hits in each tissue and their immediate regulatory neighbours. We observed both disconnected subnetworks, where one or very few genes showed strong predictability changes with age (exemplified in Supplementary Figure 4.6), and larger subnetworks connecting several genes with strong predictability changes (exemplified in Figure 4.4.A,F).

We then selected for further analysis two of the obtained subnetworks, showing distinct age-related behaviours and each captured in a different tissue.

The first subnetwork (Figure 4.4.A) is centred on mTOR signalling, which is a well known contributor to age-related phenotypes and involved in lifespan extension [117], [118]. This subnetwork included the predictability hit LAMTOR5, a gene encoding for a member of the Ragulator complex, located in the lysosomal membrane and involved in mTORC1 activation upon nutrient sensing. In nutrient rich conditions, mTORC1 is recruited to the lysosomal membrane [119], resulting in the promotion of cellular growth and proliferation, coupled to translation and biosynthesis [120, p. 1]. In our analysis, LAMTOR5 showed a loss of correlation with most of its regulatory neighbours in Artery - Tibial: cytochrome C oxidases COX14 and COX6B1, NADH dehydrogenase complex member NDUFB3, vacuolar ATPase cytosolic (V1) domain subunit ATP6V1F and oligosaccharyl transferase complex subunit OST4 (Figure 4.4.B). This loss of correlation took place in parallel with a decrease in predictability with age (Figure 4.4.C).

Interestingly, also in Artery - Tibial, we observed age-related changes in the predictability of another vacuolar ATPase (v-ATPase) subunit (in the transmembrane domain, Supplementary Figure 4.6), also linked to another Ragulator complex member, LAMTOR1. v-ATPases are located on the membrane of multiple organelles, including the lysosome, where they pump protons into the lumen and acidify it [121]. The observed relationships between Ragulator subunits (LAMTOR1 and LAMTOR5) and v-ATPase subunits in both the transmembrane and cytosolic domains (ATP6V1F, ATP6V0D1, ATP6V0C, ATP6V1H, ATP6V1D and ATP6V1A) are consistent with the knowledge that v-ATPases are required for mTORC1 activation by aminoacids [122]. Furthermore, v-ATPases have been described to physically interact with the Ragulator complex through LAMTOR1 [122], in line with the relationship observed between this gene and multiple of the v-ATPase subunits (Supplementary Figure 4.6).

In addition, we observed tight gene-gene relationships between Ragulator complex members and electron transport chain members (NDUFB3, COX14 and COX6B1 as direct neighbours of LAMTOR5; UQCRFS1, COX6A1, COX8A, COX17 and COX7A2 as indirect neighbours). These observations are in line with previous work showing that mTORC1 promotes the expression of mitochondrial regulators such as TFAM and electron transport chain components [123]. These gene-gene relationships were also altered with age, as exemplified by the age-related loss of correlation between LAMTOR5 and its direct electron transport chain neighbours NDUFB3, COX14 and COX6B1 (Figure 4.4.B).

WDR61 is another direct network neighbour of NADH dehydrogenase complex member NDUFB3, also showing an age-related predictability decrease (Figure 4.4.E). WDR61 is a subunit of the superkiller complex SKI, involved in co-translational mRNA surveillance [124]. We observed an age-related loss of correlation to its regulatory neighbours, including TBC1D7. TBC1D7 is a member of the TSC-TBC complex, a negative regulator of the mTORC1 signalling cascade [125]–[127], highlighting the crosstalk between the different cellular processes captured in this subnetwork.

The second subnetwork is composed of serum-specific proteins, most of which showed increased predictability with age in Blood (Figure 4.4.F). This subnetwork included apolipoproteins APOA1, APOA2, APOB and APOH, involved in lipid transport between tissues, members of the coagulation cascade F2 (thrombin), FGA (a subunit of fibrinogen) and F11 (coagulation factor XI), inter-alpha-trypsin inhibitor chains ITIH2 and AMBP, and serine protease HABP2, involved in hyaluronic acid binding. Inter-alpha-inhibitor activity is required for the formation of the hyaluronan coat surrounding some cell types [128], supporting the observed crosstalk between these functions.

We observed an age-related increase of ITIH2 predictability (Figure 4.4.H), in parallel with an increased correlation to its regulatory neighbours (Figure 4.4.G) and increased variance across individuals. Inter-alpha inhibitors suppress proinflammatory responses and have been shown to improve the outcome of ischemic stroke [129], which may point towards an adaptive response to an age-related increase in inflammation. In the neighbourhood of ITIH2, and also undergoing an increase in predictability with age is ALB, encoding for albumin (Figure 4.4.I,J), a major transporter in the plasma and the most abundant protein in human blood. Both ITIH2 and ALB show an age-related increase in the correlation to their regulatory neighbourhood, which includes several members of the coagulation cascade (F2, FGA and F11). These observations are in line with an increasing blood coagulation potential during healthy ageing [130].

Put together, our analyses have uncovered gene-gene relationships that seem to be affected by the aging process. The weakening of gene-gene relationships between mTOR regulators, v-ATPase subunits and electron transport chain members may suggest an age-related de-coupling of these processes, which should be coordinated to ensure appropriate response to cellular growth cues. On the other hand, the strengthening of gene-gene relationships between diverse serum proteins may reflect an age-related increase in the coordination of the coagulation cascade and an adaptive response to increased inflammation.

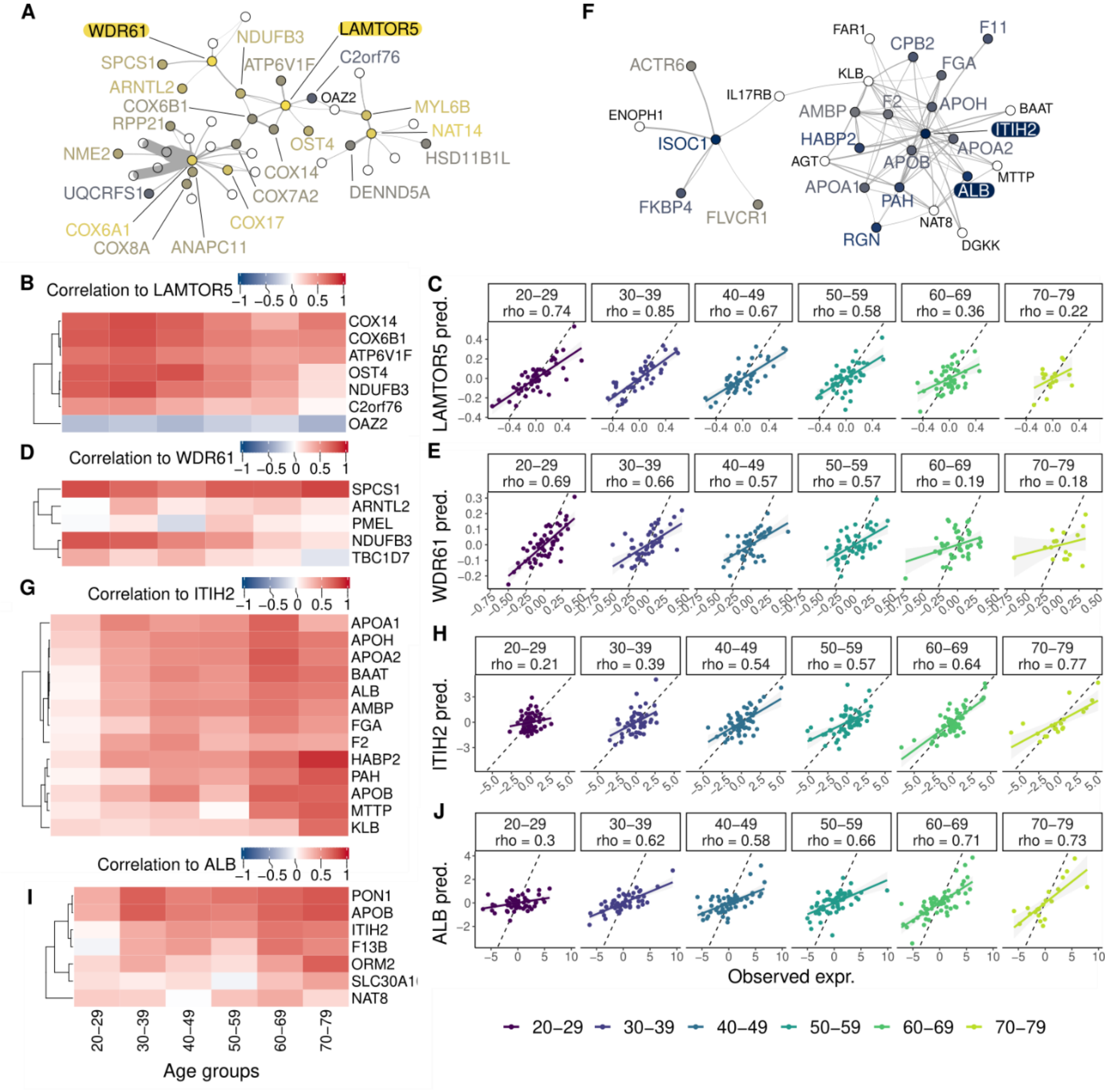


Figure 4.4: Regulatory relationships altered with age. A,F) Subnetwork of the neighbourhood of genes with age-related predictability changes. Nodes represent genes, colored by predictability slope in the respective tissue. Connections between nodes represent gene-gene relationships captured by the regulatory model. **A)** Neighbourhood of LAMTOR5 and WDR61, colored by predictability slope with age in Artery - Tibial. **F)** Neighbourhood of ITIH2 and ALB, colored by predicatability slope with age in Blood. **B), D), G), I)** Correlation between expression of genes with predictability changes (LAMTOR5 in B, WDR61 in D, ITIH2 in G and ALB in I) and regulatory neighbours across age groups. **C), E), H), J)** Comparison of expression of genes LAMTOR5 (C), WDR61 (E), ITIH2 (H) and ALB (J) in the original data (observed expression, x-axis) and reconstructed expression based on the regulatory neighbourhood (predicted expression, y-axis).

4.3 Methods

Analyses were conducted on R-4.0.3.

Gene regulatory network training

The gene regulatory network was trained as previously described [111]. A combination of regularized linear regression and stability selection was used to identify genes whose expression pattern is informative (predictive) of the expression pattern of a given gene. The expression pattern of each gene was then modelled as a linear function of the expression patterns of informative genes. This procedure was applied to 1,376 cancer cell lines with known karyotypes [131], [132]. The weights of this linear model, learnt in the cancer cell line data, capture the direction and strength of the relationship between the predictor and target genes.

Network processing

Firstly, residual edges were removed. This was done by computing, for each edge pair (i,j) (j,i), the ratio of between the lowest and the sum of the absolute edge weights. When this ratio was below 0.1, the weakest edge (lowest absolute value) was removed from the network. To avoid effects captured by the network which are not due to regulatory interactions between genes but rather local effects (e.g. both genes impacted by a copy number change in the training data), we removed predictors located in the same chromosome arm as the target. The largest connected component of the resulting network was then determined using the function *components()* from the *igraph* R package (v. 1.2.6), with default parameters. All subsequent analyses were performed on the largest connected component.

Genotype-Tissue Expression (GTEx) data preprocessing

Data download

Read counts were downloaded from the GTEx portal (version 8).

Data preprocessing

Genes were filtered based on biotype and average expression across the whole dataset. Biotypes were limited to protein coding, lincRNA, snRNA, miRNA and snoRNA. Genes with average read count below 100 across the whole dataset were excluded.

Filtered data were normalized using DESeq2 v.1.30.1. Sample size factors were estimated with the function estimateSizeFactors() and normalization was done with the *counts()* function, with *normalized = TRUE*. Normalized expression levels were log-transformed (base 2 with pseudocount of 1).

To regress out the effect of confounding variables on gene expression levels, expression levels across samples of each tissue (SMTSD) were modelled as a function of ischemic time (SMTSISCH), batch (SMGEBTCH), Hardy scale (DTHHRDY) and sex (SEX), using base R's *lm()* function. For Testis and Breast samples, sex was left out of the linear model. The residuals of the linear regression were used as batch-corrected data.

Subsetting

To assure the same number of samples in each age group, each age group was downsampled to the size of the smallest age group (determined excluding the 70-79 group). Samples from different regions of the brain were grouped into the same tissue by selecting, for each subregion in each age group, 5 samples. For groups of similar regions (cerebellar cortex and cerebellum, different regions of the cortex, and different regions of the basal ganglia), 3 samples from each of the subregions were picked instead of 5. The selection of the 3 or 5 samples per region and age group was done in a way that maximizes sampling across different donors. Within each age group, donors with the least available brain samples (across all regions) were prioritized and after each round of sample selection (one round per region), the selected donors acquired the lowest priority. The age group 70-79 was included provided that at least 10 samples were available.

To assure the same number of samples for each tissue (SMSTD) and in each age group (same-sized subset analyses, Supplementary Figure 4.2, Supplementary Tables 4.2 and 4.4), 30 samples were randomly selected per age group. Brain samples were handled in the same way as described above, with the exception that 3 samples were chosen per region and age group, except for regions in groups of similar regions, for which 2 samples were chosen per age group. The age group 70-79 was included provided that at least 10 samples were available.

Expression profile reconstruction using the regulatory model

The training procedure in the cancer cell line data resulted in a set of robust predictors and respective weights, for each gene. To reconstruct profiles based on the regulatory relationships captured by the model, the expression data were centered at 0 for each gene and the relative expression pattern of the predictor genes was multiplied by the respective weights from the regulatory model.

Identification of poorly predicted genes

Poorly predicted genes (Supplementary Figure 4.1) were identified based on the Spearman correlation between the observed expression levels (original data after normalization and batch correction) and the predicted expression levels (reconstructed data based on the regulatory model). Genes with average Spearman correlation coefficient across tissues below 0.2 were identified as poorly predicted.

Gene filtering based on average predictability across age groups

The analysis of age-related changes in predictability was restricted to genes with average predictability across age groups above tissue-specific thresholds. To determine the tissue-specific thresholds, gene-gene relationships in the regulatory model were first randomized. This was achieved by randomly choosing two predictor-target gene pairs and swapping the two targets and repeating this procedure to cover all gene pairs in the regulatory model. The randomized regulatory model is expected to hold no biological significance, while maintaining the topological properties of the original model. The randomized model was then used to reconstruct the expression profiles of tissue-specific subsets including donors of all ages. The Spearman correlation was computed between the reconstructed (predicted) and observed expression patterns of each gene in each tissue. For each tissue, the distribution of Spearman correlation coefficients across all genes was computed, and the right-side tail containing the top 5% values was determined. The value, for each tissue, corresponding to the definition of that 5% tail in the randomized model was then used as threshold for the average predictability of the original model across age groups. Threshold values varied between 0.39 (Thyroid) and 0.76 (Colon –

Transverse). Genes with average predictability across age groups below this threshold were excluded from the analysis in that tissue.

Background distribution of predictability slopes

Background distributions of predictability slopes were generated by repeated (100) random permutation of the age vector used as explanatory variable in the linear regression.

Network propagation of predictability slopes

Network propagation was done using the R package BioNetSmooth (https://github.com/beyergroup/BioNetSmooth) with *alpha = 0.2* and a user-defined network. The network defined by our regulatory model was made undirected, by summing the weights the weights of edge pairs between two nodes (edge $i \rightarrow j$ and edge $j \rightarrow i$), to capture the full strength of the relationship between gene pairs. The adjacency matrix of the resulting network was then row-normalized by dividing each row by the sum of all its entries (each node by its degree). This avoids that highly connected nodes dominate the represented topology [133].

Gene Set Enrichment Analysis

Gene Set Enrichment Analysis [113] was performed with the graphical interface of the GSEA software v4.2.3 for Linux in PreRanked mode. For each tissue, genes were ranked according to their age-related predictability slope. The Molecular Signature Database (MSigDB) [134] (v2022.1, updated August 2022) was used to retrieve Hallmark and Gene Ontology (including Biological Processes, Molecular Functions and Cellular Components) gene sets.

Quantification of age-related expression and variance changes

Age-related expression and variance changes were computed after preprocessing of the GTEx data (filtering, normalization and batch correction). For the quantification of expression level changes with age, we used the R package limma [135] to fit a linear model with age as explanatory variable, using the function *lm.fit()* followed by empirical Bayes moderation with the function *eBayes()*. For the quantification of variance changes with age, we computed the variance within each group and fit a linear model with age as explanatory variable, using R's base *lm()* function. In both analyses, age was represented numerically, as the average of the corresponding age group (e.g. 25 for the age group 20-29).

4.4 Discussion

Our analyses show that our regulatory model, trained on expression data from a wide collection of cancer cell lines, successfully captures gene-gene relationships that explain the expression pattern of individual genes based on the expression of their regulatory neighbourhood (Figure 4.1). Notably, our model was able to capture tissue-specific differences in relationships between genes involved in the same function, as well as genes involved in different functions, across 30 healthy human tissues [111]. At first, it may seem surprising that our model predicts tissue-specific gene co-expression patterns, because the model itself was not adapted to the tissue. The explanation for this finding is as follows: whereas the regulatory structure remains invariant across tissues, predictors (i.e. regulatory neighbours) have different expression levels in different tissues. As a consequence, predicted expression levels of their targets will also be tissue-specific.

Based on this regulatory model, we were able to quantify age-related changes in the relationship between individual genes and their regulatory neighbours. Our approach has the advantage of providing such a score for each gene, unlike global approaches based on the whole transcriptome. Additionally, by using robust gene-gene relationships trained on external data, our model captures the transcriptomic constraints common to all observed samples. Thus, our metric of predictability corresponds to the extent to which the gene-gene relationships of a given gene fit those captured in the training data. This is conceptually different from a differential co-expression analysis, where gene-gene relationships are compared in two conditions (old *versus* young), without any external information. It should be noted that our regulatory model reflects only a part of the complete space of gene-gene relationships that can take place. In particular, the captured relationships are those that are common between the GTEx dataset and the different cancer cell lines in the training dataset. This means that we cannot rule out the possibility that observations of low predictability come from a gain in correlation with different regulatory neighbours that were not captured by the training data, rather than a complete loss of regulatory control.

Using the regulatory relationships captured in our model to quantify changes in regulation (predictability), we observed substantial difference in the ageing patterns of different tissues. Not all tissues showed significant changes in predictability (Supplementary Figure 4.2) and, among those that did, the direction of the global predictability trends sometimes differed between tissues. It is important to note that changes in the signal strength between tissues may originate from differences in data quality or confounding factors that we could not fully correct for, despite our pre-processing efforts.

Predictability changes that were consistent across tissues mostly linked to mitochondrial functions and cell cycle regulation. Mitochondrial dysfunction is an established hallmark of ageing [106] and a decrease in ATP production has been reported with age [136], along with alterations in mitochondrial morphology and mitochondrial protein expression [137]. In parallel to age-related deregulation of mitochondrial functions, we also observed a deregulation of genes involved in the cell cycle (E2F targets and G2M checkpoint, Figure 4.2.E). We note that our regulatory model has been trained in cancer cell lines (proliferating, often with mutations in the tumour-suppressor p53 and oncogene MYC), which requires some caution when analysing results concerning genes involved in cell cycle regulation. Our analysis of age-related predictability changes only includes genes with high average predictability across age groups, which excludes genes whose regulatory relationships are exclusively valid in the context of cancer data but not healthy tissues.

Thus, the observed predictability gain of cell cycle genes cannot be attributed to a poor predictive performance of the model for such genes at young ages.

We observed that, for most tissues, the largest contribution to the observed predictability trends came from the gain or loss of correlation with regulatory neighbours in the tissue at hand (Figure 4.3.C). One exception to this was Testis, where genes with increased age-related predictability showed decreased expression. This would point towards a concerted (regulated) down-regulation of specific transcriptional programmes, mainly related to cell cycle (Figure 4.2.E, E2F targets and G2M checkpoint). These results can be explained by an exhaustion of the stem cell niche sustaining spermatogenesis [138]. In all other tissues, increased variance was linked to no predictability changes, or increased predictability, which would correspond to genes composing regulated responses, possibly as an adaptive response to the ageing process, with different levels of activity in different individuals.

An advantage of our approach is that it can capture gene-gene relationships both within and between different functional gene modules (Figure 4.1). We detected age-related changes in gene-gene relationships in both cases. The impact of ageing on the crosstalk between cellular processes was exemplified in Figure 4.4.A, for a subnetwork of genes centred on mTOR signalling. The variety of processes captured in this subnetwork is in line with the known mechanisms of mTORC1 activation and its role in the regulation of cellular growth, which requires the coordination of distinct cellular processes. In our subnetwork, the connection between mTORC1 activation and mitochondrial functions was particularly evident. The observed loss of correlation between all these interconnected cellular processes suggests that, with age, cells are no longer able to coordinate different processes to put in place an integrated, complex response such as cellular growth. This result is particularly interesting in the context of the role of mTOR signalling in ageing and longevity. Inhibition of mTOR with rapamycin is well known to extend lifespan across species [139], [140], and an increase in mTOR activation has been reported with ageing in some tissues [141]. Our results suggest a decoupling between the mechanistic activation of mTORC1 and some of its downstream target pathways that may be linked to an inability of the cell to sustain cellular growth.

Finally, we note that we use bulk RNA-seq data, which consists of a mixture of the transcriptome of the different cells present within the tissue. Thus, our approach cannot capture transcriptional noise resulting from random expression in individual cells. In our case, such noise would be averaged out across the different cells in the tissue. This is in line with the fact that we barely observe any genes with increased variance and decreased predictability with age (Figure 4.3.B). We hypothesize that the gene-gene regulation changes we find may in part be driven by age-related alterations in the cell type composition of the tissues, or by a rewiring of the relationships in one leading cell type within the tissue. Single-cell RNA-seq data from multiple human tissues is beginning to accumulate. Once this data covers a sufficiently large age range, we will be able to quantify age-related changes in predictability in a cell-type-specific manner and elucidate the contribution of different cell types to the results described here.

Our work highlights the importance of zooming out of the effect of ageing in individual genes or cellular processes and investigating how their crosstalk is affected at a systems level. Given the complex and multifactorial nature of ageing, we expect that approaches such as the one presented here will contribute to a better understanding of the global alterations in functional coordination that take place with age.

4.5 Contributions

F. Lopes preprocessed the GTEx data (filtering, normalization and batch correction). I performed all the analysis of pairwise correlation (Figure 4.1) and predictability (Figures 4.2 to 4.4) changes with age. I wrote the manuscript together with A. Beyer.

4.6 Supporting Information

Supplementary Table 4.1: Number of samples per tissue and age group for computation of predictability within age group (Figure 4.2).

Tissue	**20-29**	**30-39**	**40-49**	**50-59**	**60-69**	**70-79**
Adipose – Visceral	42	42	42	42	42	16
Artery - Tibial	54	54	54	54	54	18
Blood	62	62	62	62	62	21
Brain	46	42	49	49	49	48
Breast – Mammary Tissue	36	36	36	36	36	15
Esophagus - Mucosa	51	51	51	51	51	12
Testis	30	30	30	30	30	-
Thyroid	44	44	44	44	44	20

Supplementary Table 4.2: Number of samples per tissue and age group, after randomly selecting similar sample numbers per tissue and age group (Supplementary Figure 4.2).

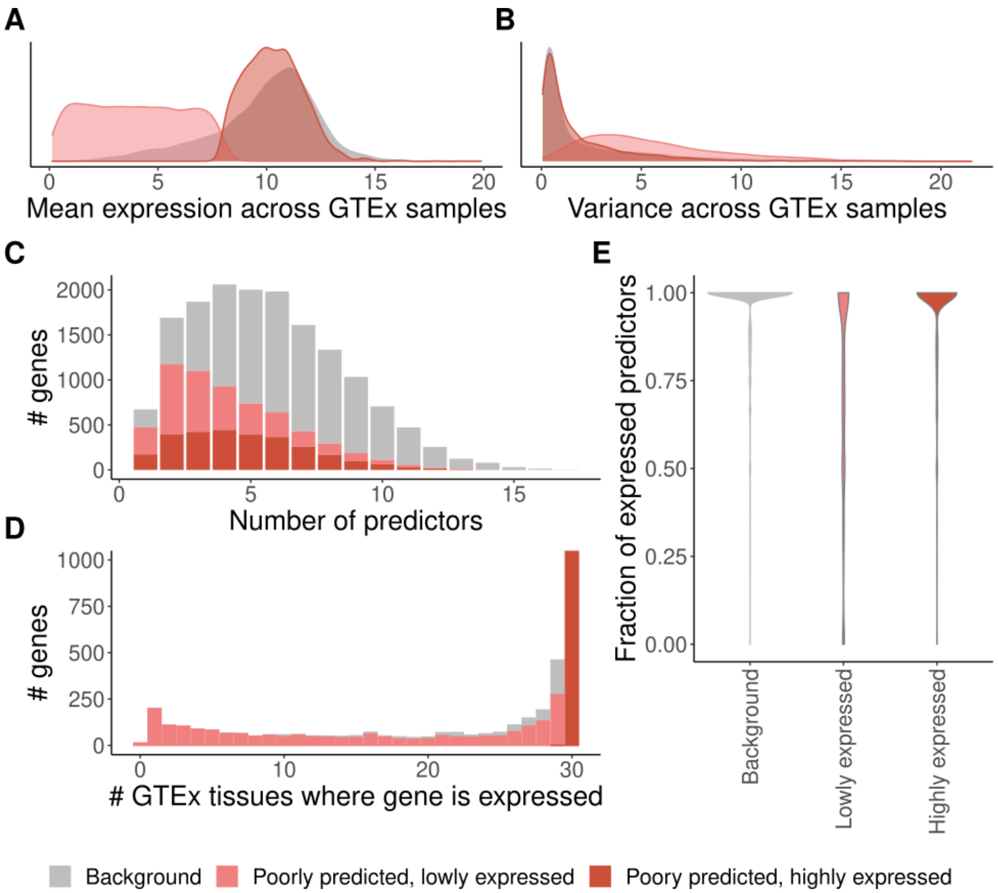
Tissue	**20-29**	**30-39**	**40-49**	**50-59**	**60-69**	**70-79**
Adipose - Subcutaneous	30	30	30	30	30	18
Adipose – Visceral	30	30	30	30	30	16
Artery - Tibial	30	30	30	30	30	18
Blood	30	30	30	30	30	21
Brain	30	30	30	30	30	48
Breast – Mammary Tissue	30	30	30	30	30	15
Colon - Transverse	30	30	30	30	30	-
Esophagus - Mucosa	30	30	30	30	30	12
Esophagus - Muscularis	30	30	30	30	30	-
Lung	30	30	30	30	30	17
Muscle - Skeletal	30	30	30	30	30	26
Nerve - Tibial	30	30	30	30	30	18
Skin – Not sun exposed	30	30	30	30	30	22
Skin – Sun exposed	30	30	30	30	30	26
Testis	30	30	30	30	30	-
Thyroid	30	30	30	30	30	20

Supplementary Table 4.3: Number of genes for which predictability trends with age were computed (Figure 4.2).

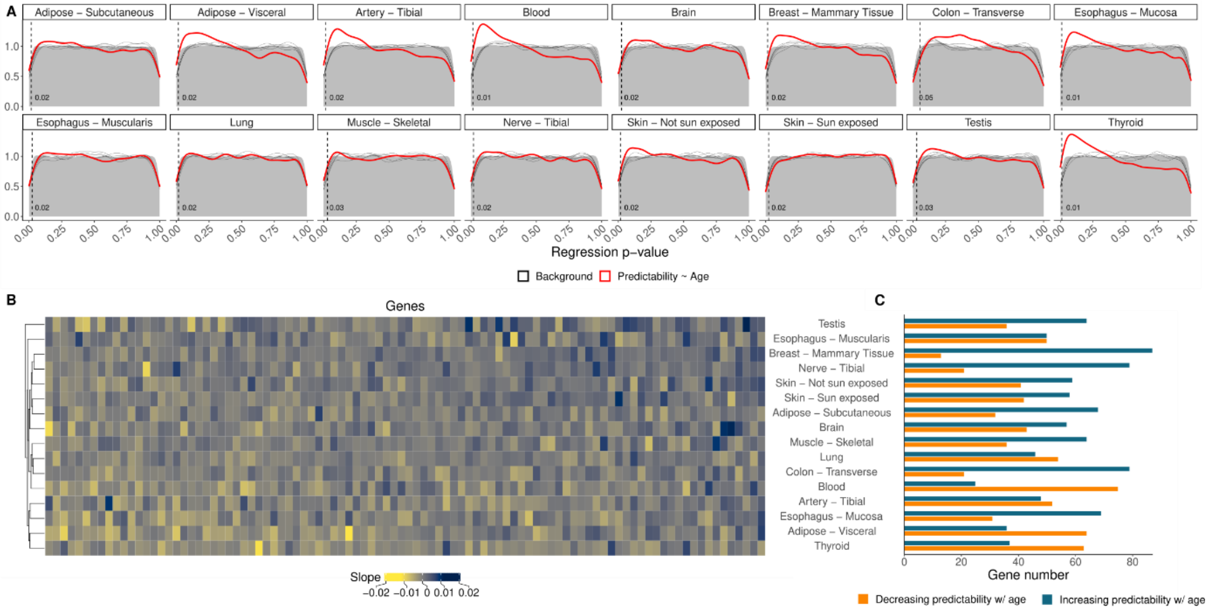
Tissue	**# genes**
Adipose – Visceral	4691
Artery - Tibial	3964
Blood	4768
Brain	3728
Breast – Mammary Tissue	4709
Esophagus - Mucosa	5830
Testis	3291
Thyroid	4984

Supplementary Table 4.4: Number of genes for which predictability trends with age were computed, after randomly selecting similar sample numbers per tissue and age group (Supplementary Figure 4.2).

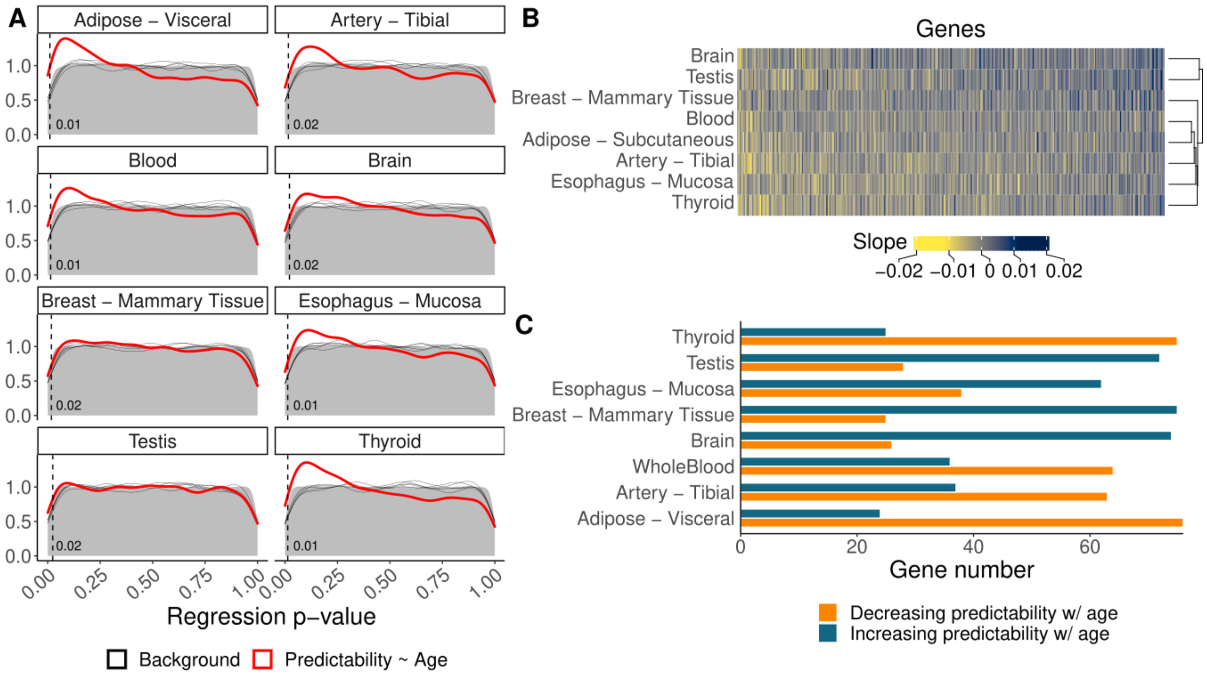
Tissue	**# genes**
Adipose - Subcutaneous	4569
Adipose – Visceral	4593
Artery - Tibial	4008
Blood	4641
Brain	3708
Breast – Mammary Tissue	4873
Colon - Transverse	2407
Esophagus - Mucosa	5890
Esophagus - Muscularis	4149
Lung	5120
Muscle - Skeletal	2726
Nerve - Tibial	3854
Skin – Not sun exposed	5038
Skin – Sun exposed	5498
Testis	3111
Thyroid	5179



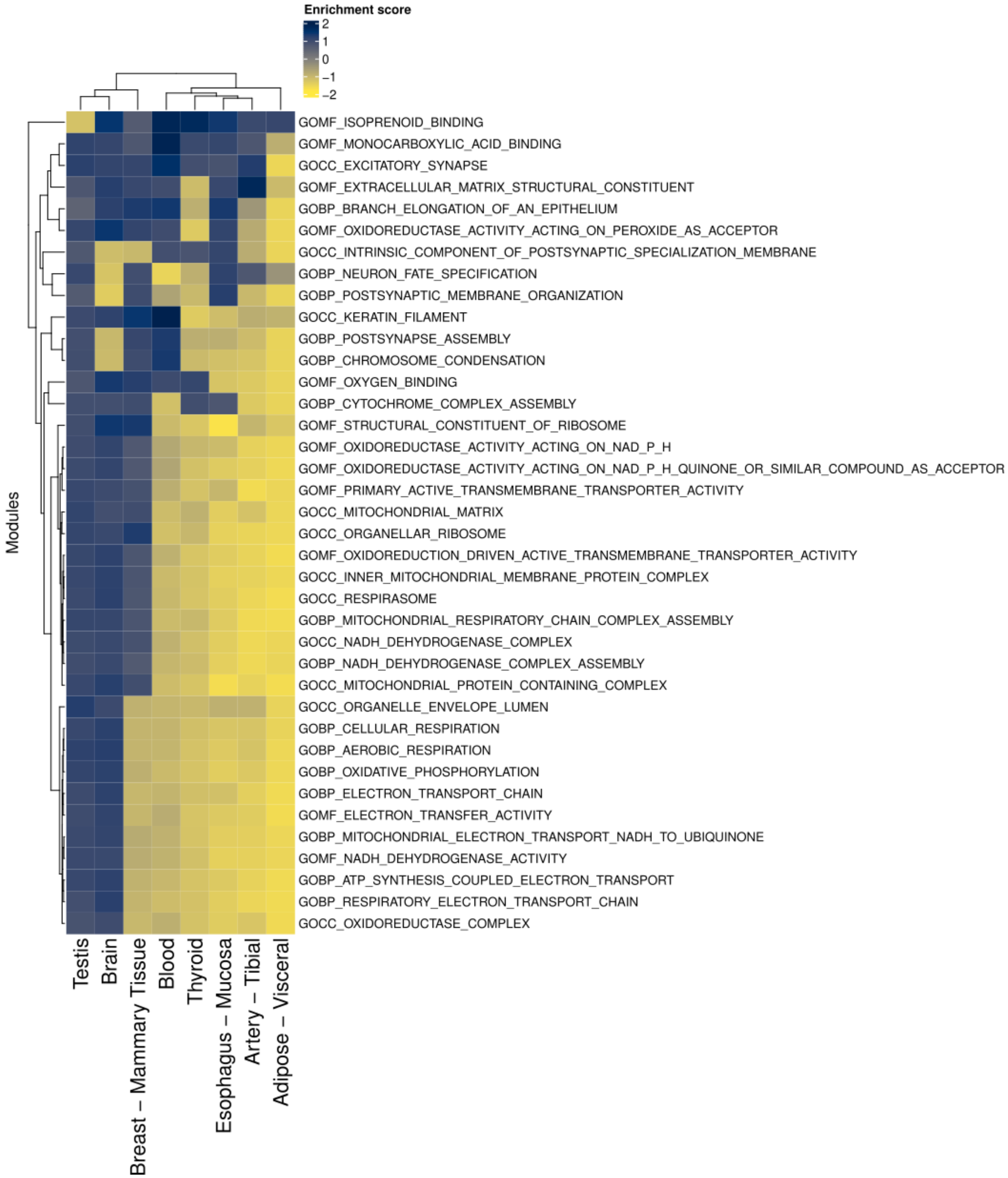
Supplementary Figure 4.1: Characterization of genes poorly predicted by the regulatory model. Poorly predicted genes were defined as having an average Spearman's rho across all 30 tissues < 0.2 and split into 2 groups according to their average expression levels across GTEx tissues: lowly expressed genes (light pink) and highly expressed genes (dark orange). **A)** Average expression levels across all GTEx samples. **B)** Expression variance across all GTEx samples. **C)** Histogram of number of predictors (genes with expression profiles informative for the target gene). **D)** Histogram of number of tissues where the target gene is expressed (average normalized expression within tissue > 2). **E)** Distribution of the fraction of predictors of a given target that are expressed (average normalized expression across tissues > 2). Histogram bars in C and D are stacked.



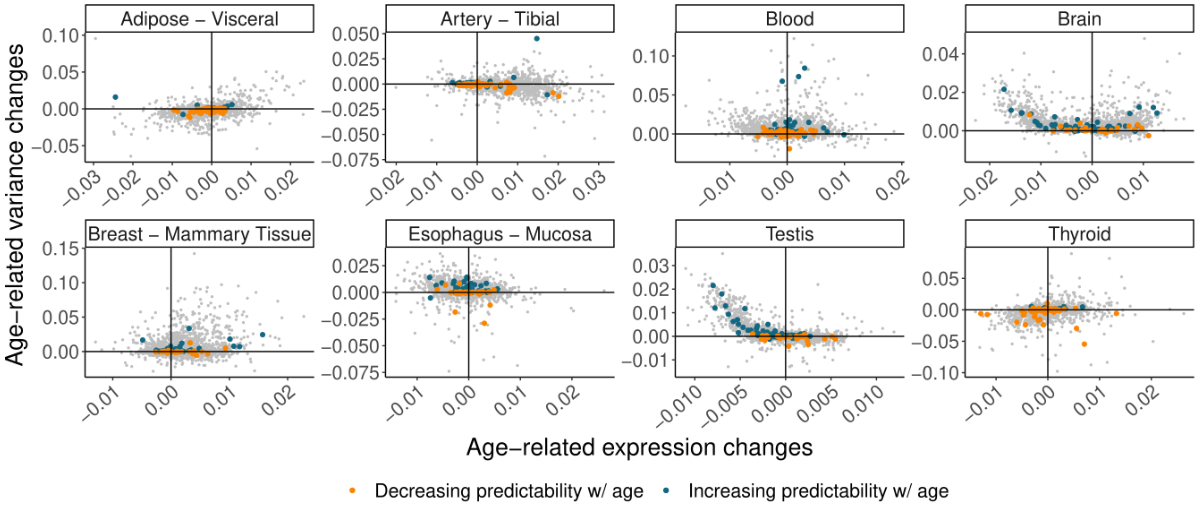
Supplementary Figure 4.2: Transcriptome-wide predictability changes with age: expanded for all 16 tissues. Across all tissues, age groups were downsampled to 30 samples. The age group 70-79 was included in the analysis, with as many samples as were available. For sample size details see Supplementary Table 4.2. **A)** Distribution of p-values for the regression of predictability values within each age group against the mean age of the age group. Red line: p-value distribution obtained from the real data. Grey background: average p-value distribution across 100 permutations of the age groups. Black lines: 5 individual permutations randomly picked from the background. The dashed vertical line indicates the highest p-value among the genes considered statistically significant in each tissue (orange and blue bars in C). The number of genes included in each tissue-specific analysis can be found in Supplementary Table 4.4. **B)** Heatmap of the predictability slopes across all genes, independently of significance level, for which the regression analysis was performed in all 16 tissues. These genes had a high average predictability in all 16 tissues (Supplementary Table 4.4). **C)** Number of genes with predictability increase (blue) and decrease (orange) among the top 100 most significant genes per tissue.



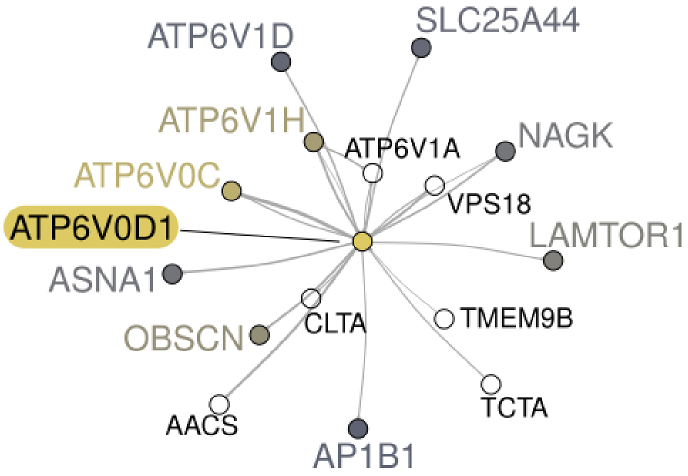
Supplementary Figure 4.3: Transcriptome-wide predictability changes with age: excluding the age group of 70-79. A) Distribution of p-values for the regression of predictability values within each age group against the mean age of the age group. Red line: p-value distribution obtained from the real data. Grey background: average p-value distribution across 100 permutations of the age groups. Black lines: 5 individual permutations randomly picked from the background. The dashed vertical line indicates the highest p-value among the genes considered statistically significant in each tissue (orange and blue bars in C). **B)** Heatmap of the predictability slopes across all genes, independently of significance level, for which the regression analysis was performed in all 8 tissues. These genes had a high average predictability in all 8 tissues. **C)** Number of genes with predictability increase (blue) and decrease (orange) among the top 100 most significant genes per tissue.



Supplementary Figure 4.4: GO terms enriched in age-related gene-gene relationship changes, captured by Gene Set Enrichment Analysis. The heatmap shows all GO terms with statistically significant (FDR < 0.1) enrichment in at least one tissue.



Supplementary Figure 4.5: Expression slope across age (x-axis) against variance slope across age (y-axis), excluding the 70-79 age group. Background genes are represented as small light grey points and top significant genes with age-related predictability changes as large colored points (predictability increase in blue and decrease in orange).



Supplementary Figure 4.6: Regulatory neighbourhood of ATP6V0D1. This gene encodes for a subunit of the transmembrane (V0) domain of vacuolar ATPase. Nodes represent genes, colored by predictability slope in Artery - Tibial. Connections between nodes represent gene-gene relationships captured by the regulatory model.

5. General Discussion

Modelling gene regulation is key to understanding cellular behaviours in response to internal and external stimuli. In the work presented here, we have made use of a global transcriptional regulatory model to estimate expression of uncaptured (dropout) genes in scRNA-seq data – Chapter 3 – and to investigate age-related changes in gene-gene relationships – Chapter 4.

5.1 Similarity of gene regulatory relationships across tissues and cell types

A common observation underlying the work described in this dissertation is that gene-gene relationships captured by our model are valid across different tissues and cell types (Supplementary Figures 3.1 and 3.2). This suggests that, while different cell types have distinct, specialized transcriptomes, gene-gene relationships are more global. Our results additionally show that these consistent relationships go beyond housekeeping genes and that our model correctly captures gene expression patterns of cell type markers (Figure 3.5) and tissue-specific regulatory subnetworks (Figure 4.4.F).

A vast portion of the efforts in transcriptional regulatory network inference so far have focused on training specialized regulatory networks that capture gene-gene relationships observed in the context of a given tissue or disease. However, some studies have emerged suggesting the existence of a 'core network', with regulatory relationships common across different cell types. Evidence in this direction can be found in a study based on scRNA-seq from brain cell types, showing that gene co-expression patterns are mostly preserved across the different cell types [142]. In this context, the work described in this dissertation exemplifies the potential of harnessing such 'core networks'.

On the other hand, recent work using co-expression networks trained for individual cell types in the fruit fly has found a small set of overlapping edges across cell-type-specific networks, mainly enriched in genes encoding for ribosomal proteins [143]. The disparity between these results and ours can be partially explained by the fact that we look for such a 'core network' by learning gene-gene relationships in different biological conditions (cancer cell lines) simultaneously. This facilitates the detection of relationships between genes whose variation is small within biological conditions, but larger between conditions, as can be expected for cell-type-specific genes.

Another conclusion that can be drawn from our work is that gene expression is modular (Figure 4.1). This is in line with observations of the topology of regulatory networks, inferred from transcriptome or metabolome data [20], [144]. The concept of modular gene expression has been explored in recent years using single-cell RNA-sequencing data: in this context, functional modules (referred to as 'regulons') are determined as sets of co-regulated genes under the regulatory influence of a transcription factor [108]. Of note, activity of such regulons in single cells has been shown to be sufficient to distinguish between different types [145], in line with the idea that cell type identity is determined based on the level to which such modules are active, rather than on the composition of the modules themselves.

5.1.1 Limitations

Since our global transcriptional regulatory model is based on gene expression levels, gene-gene relationships are not indicative of a direct causal regulatory relationship. For this reason, our model cannot be used to investigate the molecular mechanisms behind the loss or gain of specific gene-gene relationships.

Additionally, our model has been trained on transcriptome data at the gene level. As mentioned in section 1.1.3, alternative splicing plays a relevant role in the regulation of gene expression. This role of alternative splicing, as well as other regulatory mechanisms not detectable at the transcript level (e.g., protein activation by post-translational modifications), is not included in our global regulatory model.

Of further note, our model has been trained in cancer cell lines. While this provides advantages in terms of heterogeneity of expression levels compared to bulk RNA-seq data from human tissues, and robustness of expression quantification compared to scRNA-seq, it also results in limitations. Since these are proliferating cells, this can introduce a bias in the relationships captured for cell cycle genes. Additionally, mutations observed across multiple cancer cell lines (e.g., p53 mutations) may influence the structure of our model. However, due to the variety of cancer cell lines used to train our model, we expect this to be the case only for the subset of genes mutated across a wide range of the cancer cell lines.

5.1.2 Future work

As mentioned in the previous section, gene-gene relationships in our global transcriptional model do not reflect direct regulatory relationships. In order to add such information to the model, other sources of information would have to be integrated. Genomic perturbation data, making use of natural or engineered [146] genetic variation, could provide causal information by observing the effect of genomic perturbations of a given gene on the expression profile of others. DNA binding events of specific transcription factors captured through ChIP-seq are commonly used in the context of regulatory network inference, and another example source of causal information. Message-passing algorithms can be a tool for integrating data from distinct sources [147] and have been used to integrate co-expression, protein-protein interaction and TF-target information [148].

5.2 Application of gene regulatory relationships to single-cell RNA-sequencing data processing

The results described in Chapter 3 show the advantage of utilizing information external to the dataset at hand for the estimation of missing expression values in scRNA-seq data (dropout imputation). The use of external data avoids amplification of the signal within the dataset, which has been shown to introduce false positives in downstream analyses such as differential expression [73].

Given the limitations of the regulatory model discussed in the previous section, the ADImpute R package does not rely solely on network-based imputation. Instead, it can determine which imputation method, out of a preselected set, performs the best for each gene in the dataset at hand. For highly sparse genes, inclusion of external information is usually beneficial, and the method based on the regulatory model is chosen. However, if the regulatory relationships in the model do not correctly capture the expression patterns of the gene being imputed, other methods, based on information internal to the dataset, can be used.

The relevance of our work must be placed in the context of current discussions questioning the necessity for dropout imputation (section 1.3.2). On one hand, as scRNA-seq technologies improve, data sparsity is expected to be reduced. On the other hand, tools have been developed to overcome this excessive sparsity, meaning that dropout imputation is not always required for any type of downstream analysis [149]. In this context, we propose that our method is particularly relevant for studies where the expression of lowly expressed genes of especial interest to the researcher.

5.2.1 Limitations

Besides the discussed limitations of our global transcriptional regulatory model, we note that the ensemble approach included in our R package is limited in the number of methods it can compare. This is due to two factors: the high number of existing dropout imputation approaches, and the fact that, for performance comparison, the code of each tool must be integrated in our package. We have strived to select popular methods with distinct approaches to the imputation task, but the growing number of proposed methods complicates this task.

5.2.2 Future prospects

Our results show that imputation methods taking advantage of information contained in the dataset at hand can outperform our network-based method for well detected genes. This suggests that the global transcriptional regulatory network is not always successful at capturing dataset-specific relationships. Efforts to overcome this – for instance, by combining the gene-gene relationships observed in the dataset with the backbone of the global transcriptional network – could greatly improve the performance of our network-based method.

The ADImpute R package is well suited for the use of any type of user-provided network. By default, the package uses the global transcriptional regulatory network described in this dissertation, which is only valid for human. In the future, ADImpute's applicability could be extended by incorporating networks from other organisms.

5.3 Gene expression regulation changes during ageing

The results described in Chapter 4 demonstrate the advantage of using our model of gene-gene relationships to address the complex and multifactorial problem of human ageing. This is particularly highlighted by our findings of age-related changes in gene-gene relationships both within and between functional modules. Within-module changes, particularly towards an age-related decrease of gene-gene relationships, can be linked to reports of stoichiometry loss affecting protein complexes with age [150]. On the other hand, between-module changes can be associated to affected crosstalk between cellular functions. This is well exemplified in Figure 4.4.A, for a subnetwork of genes involved in mTOR signalling and cellular functions under its regulation.

5.3.1 Limitations

While recent literature focuses on the investigation of age-related 'transcriptional noise', corresponding to random fluctuations in gene expression levels across individual cells (section 1.4.2), this work is based on bulk RNA-seq, where such noise would be averaged out across the different cells.

Additionally, our global transcriptional regulatory model has been trained on a limited (although vast) set of transcriptomes. For this reason, we cannot assume that all possible gene-gene relationships have been captured. Thus, cases of gene-gene relationship loss cannot be directly interpreted as deregulation.

We further note that our study is based on bulk RNA-seq data, which does not allow for a dissection of which cell populations contribute to the observed changes in gene-gene relationships and the disentangling of the contribution of cell composition changes with age.

Finally, by focusing on the transcriptome exclusively, our study cannot distinguish between adaptive responses to ageing (for instance, in response to accumulated damage) and events contributing to further cellular and organismal deterioration with progressing age.

5.3.2 Future prospects

Potential integration of other information layers for causal information, discussed in section 5.1.2, would allow for more mechanistical insights into the alteration of gene-gene relationships. An open question that this would help address is whether the age-related changes in gene-gene relationships that we observe – in particular gene-gene relationship loss – stem from a loss of transcriptional regulation of specific genes or rather from a lack of ability of the cell to synchronize different cellular programmes.

Additionally, with the accumulation of scRNA-seq data across different tissues and age ranges, it will become possible to address the limitations posed by the use of bulk RNA-seq data and identify cell-type-specific changes in gene-gene relationships.

5.4 Concluding remarks

Differential expression analysis is commonly employed to quantify and identify differentially expressed genes across different biological conditions and diseases. Despite the usefulness of this approach, it ignores the complexity of gene-gene relationships. The work described in this dissertation shows that such gene-gene relationships can be inferred from example datasets and subsequently applied to a wide range of cell- and tissue types. This work has further exemplified the power of this approach by using the inferred gene-gene relationships to aid in the processing of scRNA-seq data and investigate the age-related breakdown of gene-gene relationships.

We expect this work to contribute to a better understanding of gene expression coordination in health and its breakdown in disease.

Bibliography

[1] O. Porrua and D. Libri, “Transcription termination and the control of the transcriptome: why, where and how to stop,” *Nature Reviews Molecular Cell Biology*, vol. 16, no. 3, pp. 190–202, Mar. 2015, doi: 10.1038/nrm3943.

[2] B. Alberts, A. Johnson, and J. Lewis, “From DNA to RNA,” in *Molecular Biology of the Cell*, 4th ed., New York, NY: Garland Science, 2002.

[3] S. A. Lambert *et al.*, “The Human Transcription Factors.,” *Cell*, vol. 172, no. 4, pp. 650–665, Feb. 2018, doi: 10.1016/j.cell.2018.01.029.

[4] J.-P. Kruse and W. Gu, “Modes of p53 regulation,” *Cell*, vol. 137, no. 4, pp. 609–622, 2009.

[5] E. Morgunova and J. Taipale, “Structural perspective of cooperative transcription factor binding,” *Current opinion in structural biology*, vol. 47, pp. 1–8, 2017.

[6] P. A. Bommarito and R. C. Fry, “Chapter 2-1 - The Role of DNA Methylation in Gene Regulation,” in *Toxicoepigenetics*, S. D. McCullough and D. C. Dolinoy, Eds. Academic Press, 2019, pp. 127–151. doi: 10.1016/B978-0-12-812433-8.00005-8.

[7] I. Kovesdi, R. Reichel, and J. R. Nevins, “Role of an adenovirus E2 promoter binding factor in E1A-mediated coordinate gene control.,” *Proceedings of the National Academy of Sciences*, vol. 84, no. 8, pp. 2180–2184, Apr. 1987, doi: 10.1073/pnas.84.8.2180.

[8] P. H. Tate and A. P. Bird, “Effects of DNA methylation on DNA-binding proteins and gene expression,” *Current Opinion in Genetics & Development*, vol. 3, no. 2, pp. 226–231, Jan. 1993, doi: 10.1016/0959-437X(93)90027-M.

[9] M. C. Lorincz, D. R. Dickerson, M. Schmitt, and M. Groudine, “Intragenic DNA methylation alters chromatin structure and elongation efficiency in mammalian cells.,” *Nat Struct Mol Biol*, vol. 11, no. 11, pp. 1068–1075, Nov. 2004, doi: 10.1038/nsmb840.

[10] A. Hellman and A. Chess, “Gene body-specific methylation on the active X chromosome.,” *Science*, vol. 315, no. 5815, pp. 1141–1143, Feb. 2007, doi: 10.1126/science.1136352.

[11] M. P. Ball *et al.*, “Targeted and genome-scale strategies reveal gene-body methylation signatures in human cells.,” *Nat Biotechnol*, vol. 27, no. 4, pp. 361–368, Apr. 2009, doi: 10.1038/nbt.1533.

[12] R. Lister *et al.*, “Human DNA methylomes at base resolution show widespread epigenomic differences.,” *Nature*, vol. 462, no. 7271, pp. 315–322, Nov. 2009, doi: 10.1038/nature08514.

[13] T. A. Rauch, X. Wu, X. Zhong, A. D. Riggs, and G. P. Pfeifer, “A human B cell methylome at 100−base pair resolution,” *Proceedings of the National Academy of Sciences*, vol. 106, no. 3, pp. 671–678, Jan. 2009, doi: 10.1073/pnas.0812399106.

[14] M. Shogren-Knaak, H. Ishii, J.-M. Sun, M. J. Pazin, J. R. Davie, and C. L. Peterson, “Histone H4-K16 acetylation controls chromatin structure and protein interactions.,” *Science*, vol. 311, no. 5762, pp. 844–847, Feb. 2006, doi: 10.1126/science.1124000.

[15] E. R. Gibney and C. M. Nolan, “Epigenetics and gene expression.,” *Heredity (Edinb)*, vol. 105, no. 1, pp. 4–13, Jul. 2010, doi: 10.1038/hdy.2010.54.

[16] M. Lawrence, S. Daujat, and R. Schneider, “Lateral Thinking: How Histone Modifications Regulate Gene Expression.,” *Trends Genet*, vol. 32, no. 1, pp. 42–56, Jan. 2016, doi: 10.1016/j.tig.2015.10.007.

[17] K. D. Robertson, “DNA methylation and chromatin–unraveling the tangled web,” *Oncogene*, vol. 21, no. 35, pp. 5361–5379, 2002.

[18] Y. Ge and B. T. Porse, “The functional consequences of intron retention: alternative splicing coupled to NMD as a regulator of gene expression,” *Bioessays*, vol. 36, no. 3, pp. 236–243, 2014.

[19] J. O’Brien, H. Hayder, Y. Zayed, and C. Peng, “Overview of microRNA biogenesis, mechanisms of actions, and circulation,” *Frontiers in endocrinology*, vol. 9, p. 402, 2018.

[20] P. Langfelder and S. Horvath, “WGCNA: an R package for weighted correlation network analysis,” *BMC Bioinformatics*, vol. 9, no. 1, p. 559, Dec. 2008, doi: 10.1186/1471-2105-9-559.

[21] D. Degli Esposti *et al.*, “Co-expression network analysis identifies gonad- and embryo-associated protein modules in the sentinel species Gammarus fossarum,” *Scientific Reports*, vol. 9, no. 1, p. 7862, May 2019, doi: 10.1038/s41598-019-44203-5.

[22] A. A. Margolin *et al.*, “ARACNE: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context,” 2006, vol. 7, no. 1, pp. 1–15.

[23] R. Steuer, J. Kurths, C. O. Daub, J. Weise, and J. Selbig, “The mutual information: Detecting and evaluating dependencies between variables,” *Bioinformatics*, vol. 18, no. suppl_2, pp. S231–S240, Oct. 2002, doi: 10.1093/bioinformatics/18.suppl_2.S231.

[24] L. Song, P. Langfelder, and S. Horvath, “Comparison of co-expression measures: mutual information, correlation, and model based indices,” *BMC Bioinformatics*, vol. 13, no. 1, p. 328, Dec. 2012, doi: 10.1186/1471-2105-13-328.

[25] N. Meinshausen and B. Yu, “Lasso-type recovery of sparse representations for high-dimensional data,” 2006. doi: 10.21236/ada472998.

[26] R. Bonneau *et al.*, “The Inferelator: an algorithm for learning parsimonious regulatory networks from systems-biology data sets de novo,” *Genome biology*, vol. 7, no. 5, pp. 1–16, 2006.

[27] M. Seifert, B. Friedrich, and A. Beyer, “Importance of rare gene copy number alterations for personalized tumor characterization and survival analysis,” *Genome Biology*, vol. 17, no. 1, p. 204, Oct. 2016, doi: 10.1186/s13059-016-1058-1.

[28] N. Meinshausen and P. Bühlmann, “Stability selection,” *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 72, no. 4, pp. 417–473, 2010.

[29] A.-C. Haury, F. Mordelet, P. Vera-Licona, and J.-P. Vert, “TIGRESS: Trustful Inference of Gene REgulation using Stability Selection,” *BMC Systems Biology*, vol. 6, no. 1, p. 145, Nov. 2012, doi: 10.1186/1752-0509-6-145.

[30] G. Robertson *et al.*, “Genome-wide profiles of STAT1 DNA association using chromatin immunoprecipitation and massively parallel sequencing,” *Nature methods*, vol. 4, no. 8, pp. 651–657, 2007.

[31] Z. Bar-Joseph *et al.*, “Computational discovery of gene modules and regulatory networks,” *Nature Biotechnology*, vol. 21, no. 11, pp. 1337–1342, Nov. 2003, doi: 10.1038/nbt890.

[32] A. M. Ackermann, Z. Wang, J. Schug, A. Naji, and K. H. Kaestner, “Integration of ATAC-seq and RNA-seq identifies human alpha cell and beta cell signature genes,” *Molecular metabolism*, vol. 5, no. 3, pp. 233–244, 2016.

[33] J. Qin, Y. Hu, F. Xu, H. K. Yalamanchili, and J. Wang, “Inferring gene regulatory networks by integrating ChIP-seq/chip and transcriptome data via LASSO-type regularization methods,” *Methods*, vol. 67, no. 3, pp. 294–303, 2014.

[34] T. Håndstad, M. Rye, R. Močnik, F. Drabløs, and P. Sætrom, “Cell-type specificity of ChIP-predicted transcription factor binding sites,” *BMC Genomics*, vol. 13, no. 1, p. 372, Aug. 2012, doi: 10.1186/1471-2164-13-372.

[35] F. Tang *et al.*, “mRNA-Seq whole-transcriptome analysis of a single cell,” *Nature methods*, vol. 6, no. 5, pp. 377–382, 2009.

[36] S. Islam *et al.*, “Quantitative single-cell RNA-seq with unique molecular identifiers,” *Nature Methods*, vol. 11, no. 2, pp. 163–166, Feb. 2014, doi: 10.1038/nmeth.2772.

[37] T. Hashimshony, F. Wagner, N. Sher, and I. Yanai, “CEL-Seq: single-cell RNA-Seq by multiplexed linear amplification,” *Cell reports*, vol. 2, no. 3, pp. 666–673, 2012.

[38] S. Picelli, Å. K. Björklund, O. R. Faridani, S. Sagasser, G. Winberg, and R. Sandberg, “Smart-seq2 for sensitive full-length transcriptome profiling in single cells,” *Nature methods*, vol. 10, no. 11, pp. 1096–1098, 2013.

[39] C. Ziegenhain *et al.*, "Comparative Analysis of Single-Cell RNA Sequencing Methods," *Molecular Cell*, vol. 65, no. 4, pp. 631-643.e4, Feb. 2017, doi: 10.1016/j.molcel.2017.01.023.

[40] A. Haque, J. Engel, S. A. Teichmann, and T. Lönnberg, "A practical guide to single-cell RNA-sequencing for biomedical research and clinical applications," *Genome Medicine*, vol. 9, no. 1, p. 75, Aug. 2017, doi: 10.1186/s13073-017-0467-4.

[41] B. Hwang, J. H. Lee, and D. Bang, "Single-cell RNA sequencing technologies and bioinformaticspipelines," *Experimental & Molecular Medicine*, vol. 50, no. 8, pp. 1–14, Aug. 2018, doi: 10.1038/s12276-018-0071-8.

[42] D. O'Neil, H. Glowatz, and M. Schlumpberger, "Ribosomal RNA Depletion for Efficient Use of RNA-Seq Capacity," *Current Protocols in Molecular Biology*, vol. 103, no. 1, p. 4.19.1-4.19.8, Jul. 2013, doi: 10.1002/0471142727.mb0419s103.

[43] E. Z. Macosko *et al.*, "Highly Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets," *Cell*, vol. 161, no. 5, pp. 1202–1214, May 2015, doi: 10.1016/j.cell.2015.05.002.

[44] A. M. Klein *et al.*, "Droplet Barcoding for Single-Cell Transcriptomics Applied to Embryonic Stem Cells," *Cell*, vol. 161, no. 5, pp. 1187–1201, May 2015, doi: 10.1016/j.cell.2015.04.044.

[45] T. Kivioja *et al.*, "Counting absolute numbers of molecules using unique molecular identifiers," *Nature Methods*, vol. 9, no. 1, pp. 72–74, Jan. 2012, doi: 10.1038/nmeth.1778.

[46] M. Hagemann-Jensen *et al.*, "Single-cell RNA counting at allele and isoform resolution using Smart-seq3," *Nature Biotechnology*, vol. 38, no. 6, pp. 708–714, Jun. 2020, doi: 10.1038/s41587-020-0497-0.

[47] O. Stegle, S. A. Teichmann, and J. C. Marioni, "Computational and analytical challenges in single-cell transcriptomics," *Nature Reviews Genetics*, vol. 16, no. 3, pp. 133–145, Mar. 2015, doi: 10.1038/nrg3833.

[48] V. Svensson, "Droplet scRNA-seq is not zero-inflated," *Nature Biotechnology*, vol. 38, no. 2, pp. 147–150, 2020.

[49] Y. Cao, S. Kitanovski, R. Küppers, and D. Hoffmann, "UMI or not UMI, that is the question for scRNA-seq zero-inflation," *Nature Biotechnology*, vol. 39, no. 2, pp. 158–159, 2021.

[50] V. Svensson, "Reply to: UMI or not UMI, that is the question for scRNA-seq zero-inflation," *Nature Biotechnology*, vol. 39, no. 2, pp. 160–160, 2021.

[51] D. Risso, F. Perraudeau, S. Gribkova, S. Dudoit, and J.-P. Vert, "ZINB-WaVE: A general and flexible method for signal extraction from single-cell RNA-seq data," *BioRxiv*, vol. 2017, p. 125112, 2017.

[52] E. Pierson and C. Yau, "ZIFA: Dimensionality reduction for zero-inflated single-cell gene expression analysis," *Genome biology*, vol. 16, no. 1, pp. 1–10, 2015.

[53] S. Harper, "Economic and social implications of aging societies," *Science*, vol. 346, no. 6209, pp. 587–591, 2014.

[54] S. Frenk and J. Houseley, "Gene expression hallmarks of cellular ageing," *Biogerontology*, vol. 19, no. 6, pp. 547–566, 2018.

[55] J. P. de Magalhães, J. Curado, and G. M. Church, "Meta-analysis of age-related gene expression profiles identifies common signatures of aging.," *Bioinformatics*, vol. 25, no. 7, pp. 875–881, Apr. 2009, doi: 10.1093/bioinformatics/btp073.

[56] M. J. Peters *et al.*, "The transcriptional landscape of age in human peripheral blood.," *Nat Commun*, vol. 6, p. 8570, Oct. 2015, doi: 10.1038/ncomms9570.

[57] A. Kumar *et al.*, "Age-associated changes in gene expression in human brain and isolated neurons.," *Neurobiol Aging*, vol. 34, no. 4, pp. 1199–1209, Apr. 2013, doi: 10.1016/j.neurobiolaging.2012.10.021.

[58] L. W. Harries *et al.*, "Advancing age is associated with gene expression changes resembling mTOR inhibition: evidence from two human populations.," *Mech Ageing Dev*, vol. 133, no. 8, pp. 556–562, Aug. 2012, doi: 10.1016/j.mad.2012.07.003.

[59] R. Bahar *et al.*, “Increased cell-to-cell variation in gene expression in ageing mouse heart.,” *Nature*, vol. 441, no. 7096, pp. 1011–1014, Jun. 2006, doi: 10.1038/nature04844.

[60] L. A. Warren, D. J. Rossi, G. R. Schiebinger, I. L. Weissman, S. K. Kim, and S. R. Quake, “Transcriptional instability is not a universal attribute of aging,” *Aging Cell*, vol. 6, no. 6, pp. 775–782, Dec. 2007, doi: 10.1111/j.1474-9726.2007.00337.x.

[61] M. Enge *et al.*, “Single-Cell Analysis of Human Pancreas Reveals Transcriptional Signatures of Aging and Somatic Mutation Patterns,” *Cell*, vol. 171, no. 2, pp. 321-330.e14, Oct. 2017, doi: 10.1016/j.cell.2017.09.004.

[62] I. Angelidis *et al.*, “An atlas of the aging lung mapped by single cell transcriptomics and deep tissue proteomics,” *Nature Communications*, vol. 10, no. 1, p. 963, Feb. 2019, doi: 10.1038/s41467-019-08831-9.

[63] G. E. W. Marti, S. Chu, and S. R. Quake, “Aging causes changes in transcriptional noise across a diverse set of cell types,” *bioRxiv*, p. 2022.06.23.497402, Jan. 2022, doi: 10.1101/2022.06.23.497402.

[64] L. K. Southworth, A. B. Owen, and S. K. Kim, “Aging Mice Show a Decreasing Correlation of Gene Expression within Genetic Modules,” *PLOS Genetics*, vol. 5, no. 12, p. e1000776, Dec. 2009, doi: 10.1371/journal.pgen.1000776.

[65] O. Levy *et al.*, “Age-related loss of gene-to-gene transcriptional coordination among single cells.,” *Nat Metab*, vol. 2, no. 11, pp. 1305–1315, Nov. 2020, doi: 10.1038/s42255-020-00304-4.

[66] M. Zampieri, F. Ciccarone, R. Calabrese, C. Franceschi, A. Bürkle, and P. Caiafa, “Reconfiguration of DNA methylation in aging,” *Mechanisms of Ageing and Development*, vol. 151, pp. 60–70, Nov. 2015, doi: 10.1016/j.mad.2015.02.002.

[67] L. N. Booth and A. Brunet, “The Aging Epigenome,” *Molecular Cell*, vol. 62, no. 5, pp. 728–744, Jun. 2016, doi: 10.1016/j.molcel.2016.05.013.

[68] C. Nikopoulou, S. Parekh, and P. Tessarz, “Ageing and sources of transcriptional heterogeneity,” vol. 400, no. 7, pp. 867–878, 2019, doi: 10.1515/hsz-2018-0449.

[69] P. Cheung *et al.*, “Single-Cell Chromatin Modification Profiling Reveals Increased Epigenetic Variations with Aging,” *Cell*, vol. 173, no. 6, pp. 1385-1397.e14, May 2018, doi: 10.1016/j.cell.2018.03.079.

[70] C. P. Martinez-Jimenez *et al.*, “Aging increases cell-to-cell transcriptional variability upon immune stimulation,” *Science*, vol. 355, no. 6332, pp. 1433–1436, Mar. 2017, doi: 10.1126/science.aah4115.

[71] J. D. Silverman, K. Roche, S. Mukherjee, and L. A. David, “Naught all zeros in sequence count data are the same,” *Comput Struct Biotechnol J*, vol. 18, pp. 2789–2798, Sep. 2020, doi: 10.1016/j.csbj.2020.09.014.

[72] Z. Zhang, F. Cui, C. Wang, L. Zhao, and Q. Zou, “Goals and approaches for each processing step for single-cell RNA sequencing data.,” *Brief Bioinform*, vol. 22, no. 4, Jul. 2021, doi: 10.1093/bib/bbaa314.

[73] T. Andrews and M. Hemberg, “False signals induced by single-cell imputation [version 2; peer review: 4 approved],” *F1000Research*, vol. 7, no. 1740, 2019, doi: 10.12688/f1000research.16613.2.

[74] D. van Dijk *et al.*, “Recovering Gene Interactions from Single-Cell Data Using Data Diffusion,” *Cell*, vol. 174, no. 3, pp. 716-729.e27, Jul. 2018, doi: 10.1016/j.cell.2018.05.061.

[75] W. V. Li and J. J. Li, “An accurate and robust imputation method scImpute for single-cell RNA-seq data,” *Nat. Commun.*, vol. 9, no. 1, p. 997, Mar. 2018, doi: 10.1038/s41467-018-03405-7.

[76] W. Gong, I.-Y. Kwak, P. Pota, N. Koyano-Nakagawa, and D. J. Garry, “DrImpute: imputing dropout events in single cell RNA sequencing data,” *BMC Bioinformatics*, vol. 19, no. 1, p. 220, Jun. 2018, doi: 10.1186/s12859-018-2226-y.

[77] M. Moussa and I. I. Măndoiu, “Locality Sensitive Imputation for Single Cell RNA-Seq Data,” *Journal of Computational Biology*, vol. 26, no. 8, pp. 822–835, Feb. 2019, doi: 10.1089/cmb.2018.0236.

[78] F. Wagner, Y. Yan, and I. Yanai, “K-nearest neighbor smoothing for high-throughput single-cell RNA-Seq data,” *bioRxiv*, p. 217737, Jan. 2018, doi: 10.1101/217737.

[79] M. Huang *et al.*, “SAVER: gene expression recovery for single-cell RNA sequencing,” *Nat. Methods*, vol. 15, no. 7, pp. 539–542, Jul. 2018, doi: 10.1038/s41592-018-0033-z.

[80] W. Wu, Y. Liu, Q. Dai, X. Yan, and Z. Wang, “G2S3: A gene graph-based imputation method for single-cell RNA sequencing data,” *PLOS Computational Biology*, vol. 17, no. 5, May 2021, doi: 10.1371/journal.pcbi.1009029.

[81] T. Peng, Q. Zhu, P. Yin, and K. Tan, “SCRABBLE: single-cell RNA-seq imputation constrained by bulk RNA-seq data,” *Genome Biology*, vol. 20, no. 1, p. 88, May 2019, doi: 10.1186/s13059-019-1681-8.

[82] N. Meinshausen and P. Bühlmann, “Stability selection,” *J. R. Stat. Soc. Series B Stat. Methodol.*, vol. 72, no. 4, pp. 417–473, 2010, doi: 10.1111/j.1467-9868.2010.00740.x.

[83] M. Hagemann-Jensen *et al.*, “Single-cell RNA counting at allele and isoform resolution using Smart-seq3,” *Nature Biotechnology*, vol. 38, no. 6, pp. 708–714, Jun. 2020, doi: 10.1038/s41587-020-0497-0.

[84] L.-F. Chu *et al.*, “Single-cell RNA-seq reveals novel regulators of human embryonic stem cell differentiation to definitive endoderm,” *Genome Biol.*, vol. 17, no. 1, p. 173, Aug. 2016, doi: 10.1186/s13059-016-1033-x.

[85] I. Tirosh *et al.*, “Single-cell RNA-seq supports a developmental hierarchy in human oligodendroglioma,” *Nature*, vol. 539, no. 7628, pp. 309–313, Nov. 2016, doi: 10.1038/nature20123.

[86] K. J. Travaglini *et al.*, “A molecular cell atlas of the human lung from single-cell RNA sequencing,” *Nature*, vol. 587, no. 7835, pp. 619–625, Nov. 2020, doi: 10.1038/s41586-020-2922-4.

[87] K. Bi *et al.*, “Tumor and immune reprogramming during immunotherapy in advanced renal cell carcinoma,” *Cancer Cell*, vol. 39, no. 5, pp. 649-661.e5, May 2021, doi: 10.1016/j.ccell.2021.02.015.

[88] L. Zhang and S. Zhang, “Comparison of computational methods for imputing single-cell RNA-sequencing data,” 2017, doi: 10.1101/241190.

[89] W. Hou, Z. Ji, H. Ji, and S. C. Hicks, “A systematic evaluation of single-cell RNA-sequencing imputation methods,” *Genome Biology*, vol. 21, no. 1, p. 218, Aug. 2020, doi: 10.1186/s13059-020-02132-x.

[90] C. Trapnell *et al.*, “The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells,” *Nat Biotechnol*, vol. 32, no. 4, pp. 381–386, Apr. 2014, doi: 10.1038/nbt.2859.

[91] K. Street *et al.*, “Slingshot: cell lineage and pseudotime inference for single-cell transcriptomics,” *BMC Genomics*, vol. 19, no. 1, p. 477, Jun. 2018, doi: 10.1186/s12864-018-4772-0.

[92] L. McInnes, J. Healy, N. Saul, and L. Grossberger, “UMAP: Uniform Manifold Approximation and Projection,” *Journal of Open Source Software*, vol. 3, p. 861, Sep. 2018, doi: 10.21105/joss.00861.

[93] A. Butler, P. Hoffman, P. Smibert, E. Papalexi, and R. Satija, “Integrating single-cell transcriptomic data across different conditions, technologies, and species,” *Nat. Biotechnol.*, vol. 36, no. 5, pp. 411–420, Jun. 2018, doi: 10.1038/nbt.4096.

[94] J. M. Vaquerizas, S. K. Kummerfeld, S. A. Teichmann, and N. M. Luscombe, “A census of human transcription factors: function, expression and evolution,” *Nat. Rev. Genet.*, vol. 10, no. 4, pp. 252–263, Apr. 2009, doi: 10.1038/nrg2538.

[95] A. Tugores *et al.*, “The epithelium-specific ETS protein EHF/ESE-3 is a context-dependent transcriptional repressor downstream of MAPK signaling cascades,” *J. Biol. Chem.*, vol. 276, no. 23, pp. 20397–20406, Jun. 2001, doi: 10.1074/jbc.M010930200.

[96] C. A. R. Boyd, “Review: Epithelial aspects of human placental trophoblast,” *Placenta*, vol. 34 Suppl, pp. S24-6, Mar. 2013, doi: 10.1016/j.placenta.2012.11.013.

[97] Y. Tomaru *et al.*, “A transient disruption of fibroblastic transcriptional regulatory network facilitates trans-differentiation,” *Nucleic Acids Research*, vol. 42, no. 14, pp. 8905–8913, Jul. 2014, doi: 10.1093/nar/gku567.

[98] D. Risso, F. Perraudeau, S. Gribkova, S. Dudoit, and J.-P. Vert, “A general and flexible method for signal extraction from single-cell RNA-seq data,” *Nature Communications*, vol. 9, no. 1, p. 284, Jan. 2018, doi: 10.1038/s41467-017-02554-5.

[99] J. Barretina *et al.*, “The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity,” *Nature*, vol. 483, no. 7391, pp. 603–607, Mar. 2012, doi: 10.1038/nature11003.

[100] C. Klijn *et al.*, “A comprehensive transcriptional portrait of human cancer cell lines,” *Nature Biotechnology*, vol. 33, no. 3, pp. 306–312, Mar. 2015, doi: 10.1038/nbt.3080.

[101] J. T. Leek and J. D. Storey, “Capturing Heterogeneity in Gene Expression Studies by Surrogate Variable Analysis,” *PLOS Genetics*, vol. 3, no. 9, p. e161, Sep. 2007, doi: 10.1371/journal.pgen.0030161.

[102] D. Talwar, A. Mongia, D. Sengupta, and A. Majumdar, “AutoImpute: Autoencoder based imputation of single-cell RNA-seq data,” *Scientific Reports*, vol. 8, no. 1, p. 16329, Nov. 2018, doi: 10.1038/s41598-018-34688-x.

[103] G. Finak *et al.*, “MAST: a flexible statistical framework for assessing transcriptional changes and characterizing heterogeneity in single-cell RNA sequencing data,” *Genome Biol*, vol. 16, pp. 278–278, Dec. 2015, doi: 10.1186/s13059-015-0844-5.

[104] A. Alexa, J. Rahnenführer, and T. Lengauer, “Improved scoring of functional groups from gene expression data by decorrelating GO graph structure,” *Bioinformatics*, vol. 22, no. 13, pp. 1600–1607, Apr. 2006, doi: 10.1093/bioinformatics/btl140.

[105] R. Elyanow, B. Dumitrascu, B. E. Engelhardt, and B. J. Raphael, “netNMF-sc: leveraging gene-gene interactions for imputation and dimensionality reduction in single-cell expression analysis,” *Genome Res*, vol. 30, no. 2, pp. 195–204, Feb. 2020, doi: 10.1101/gr.251603.119.

[106] C. López-Otín, M. A. Blasco, L. Partridge, M. Serrano, and G. Kroemer, “The Hallmarks of Aging,” *Cell*, vol. 153, no. 6, pp. 1194–1217, Jun. 2013, doi: 10.1016/j.cell.2013.05.039.

[107] J. Zhang and S. Zhang, “Modular Organization of Gene Regulatory Networks,” in *Encyclopedia of Systems Biology*, W. Dubitzky, O. Wolkenhauer, K.-H. Cho, and H. Yokota, Eds. New York, NY: Springer New York, 2013, pp. 1437–1441. doi: 10.1007/978-1-4419-9863-7_473.

[108] S. Aibar *et al.*, “SCENIC: single-cell regulatory network inference and clustering.,” *Nat Methods*, vol. 14, no. 11, pp. 1083–1086, Nov. 2017, doi: 10.1038/nmeth.4463.

[109] L. J. Carithers *et al.*, “A Novel Approach to High-Quality Postmortem Tissue Procurement: The GTEx Project,” *Biopreservation and Biobanking*, vol. 13, no. 5, pp. 311–319, Oct. 2015, doi: 10.1089/bio.2015.0032.

[110] M. Seifert, B. Friedrich, and A. Beyer, “Importance of rare gene copy number alterations for personalized tumor characterization and survival analysis,” *Genome Biology*, vol. 17, no. 1, p. 204, Oct. 2016, doi: 10.1186/s13059-016-1058-1.

[111] A. C. Leote, X. Wu, and A. Beyer, “Regulatory network-based imputation of dropouts in single-cell RNA sequencing data,” *PLOS Computational Biology*, vol. 18, no. 2, p. e1009849, Feb. 2022, doi: 10.1371/journal.pcbi.1009849.

[112] S. Köhler, S. Bauer, D. Horn, and P. N. Robinson, “Walking the Interactome for Prioritization of Candidate Disease Genes,” *The American Journal of Human Genetics*, vol. 82, no. 4, pp. 949–958, Apr. 2008, doi: 10.1016/j.ajhg.2008.02.013.

[113]A. Subramanian *et al.*, “Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles.,” *Proc Natl Acad Sci U S A*, vol. 102, no. 43, pp. 15545–15550, Oct. 2005, doi: 10.1073/pnas.0506580102.

[114]R. Kalluri and R. A. Weinberg, “The basics of epithelial-mesenchymal transition.,” *J Clin Invest*, vol. 119, no. 6, pp. 1420–1428, Jun. 2009, doi: 10.1172/JCI39104.

[115]M. Somel, P. Khaitovich, S. Bahn, S. Pääbo, and M. Lachmann, “Gene expression becomes heterogeneous with age,” *Current Biology*, vol. 16, no. 10, pp. R359–R360, May 2006, doi: 10.1016/j.cub.2006.04.024.

[116]V. R. Kedlian, H. M. Donertas, and J. M. Thornton, “The widespread increase in inter-individual variability of gene expression in the human brain with age,” *Aging (Albany NY)*, vol. 11, no. 8, pp. 2253–2280, /01/01, doi: 10.18632/aging.101912.

[117]T. Weichhart, “mTOR as Regulator of Lifespan, Aging, and Cellular Senescence: A Mini-Review,” *Gerontology*, vol. 64, no. 2, pp. 127–134, 2018, doi: 10.1159/000484629.

[118]S. A. Fernandes and C. Demetriades, “The Multifaceted Role of Nutrient Sensing and mTORC1 Signaling in Physiology and Aging,” *Frontiers in Aging*, vol. 2, 2021, [Online]. Available: https://www.frontiersin.org/articles/10.3389/fragi.2021.707372

[119]Y. Sancak, L. Bar-Peled, R. Zoncu, A. L. Markhard, S. Nada, and D. M. Sabatini, “Ragulator-Rag complex targets mTORC1 to the lysosomal surface and is necessary for its activation by amino acids.,” *Cell*, vol. 141, no. 2, pp. 290–303, Apr. 2010, doi: 10.1016/j.cell.2010.02.024.

[120]I. Ben-Sahra and B. D. Manning, “mTORC1 signaling and the metabolic control of cell growth,” *Current Opinion in Cell Biology*, vol. 45, pp. 72–82, Apr. 2017, doi: 10.1016/j.ceb.2017.02.012.

[121]D. J. Cipriano *et al.*, “Structure and regulation of the vacuolar ATPases.,” *Biochim Biophys Acta*, vol. 1777, no. 7–8, pp. 599–604, Aug. 2008, doi: 10.1016/j.bbabio.2008.03.013.

[122]R. Zoncu, L. Bar-Peled, A. Efeyan, S. Wang, Y. Sancak, and D. M. Sabatini, “mTORC1 senses lysosomal amino acids through an inside-out mechanism that requires the vacuolar H(+)-ATPase.,” *Science*, vol. 334, no. 6056, pp. 678–683, Nov. 2011, doi: 10.1126/science.1207056.

[123]M. Morita *et al.*, “mTORC1 Controls Mitochondrial Activity and Biogenesis through 4E-BP-Dependent Translational Regulation,” *Cell Metabolism*, vol. 18, no. 5, pp. 698–711, Nov. 2013, doi: 10.1016/j.cmet.2013.10.001.

[124]J. S. J. Anderson and R. Parker, “The 3′ to 5′ degradation of yeast mRNAs is a general mechanism for mRNA turnover that requires the SKI2 DEVH box protein and 3′ to 5′ exonucleases of the exosome complex,” *The EMBO Journal*, vol. 17, no. 5, pp. 1497–1506, Mar. 1998, doi: 10.1093/emboj/17.5.1497.

[125]K. Inoki, Y. Li, T. Zhu, J. Wu, and K.-L. Guan, “TSC2 is phosphorylated and inhibited by Akt and suppresses mTOR signalling,” *Nature Cell Biology*, vol. 4, no. 9, pp. 648–657, Sep. 2002, doi: 10.1038/ncb839.

[126]X. Gao *et al.*, “Tsc tumour suppressor proteins antagonize amino-acid–TOR signalling,” *Nature Cell Biology*, vol. 4, no. 9, pp. 699–704, Sep. 2002, doi: 10.1038/ncb847.

[127]A. R. Tee, D. C. Fingar, B. D. Manning, D. J. Kwiatkowski, L. C. Cantley, and J. Blenis, “Tuberous sclerosis complex-1 and -2 gene products function together to inhibit mammalian target of rapamycin (mTOR)-mediated downstream signaling,” *Proceedings of the National Academy of Sciences*, vol. 99, no. 21, pp. 13571–13576, Oct. 2002, doi: 10.1073/pnas.202476899.

[128]A. Blom, H. Pertoft, and E. Fries, “Inter-alpha-inhibitor is required for the formation of the hyaluronan-containing coat on fibroblasts and mesothelial cells.,” *J Biol Chem*, vol. 270, no. 17, pp. 9698–9701, Apr. 1995, doi: 10.1074/jbc.270.17.9698.

[129] L. D. McCullough *et al.*, “Exogenous inter-α inhibitor proteins prevent cell death and improve ischemic stroke outcomes in mice.,” *J Clin Invest*, vol. 131, no. 17, Sep. 2021, doi: 10.1172/JCI144898.

[130] K. Hager, J. Setzer, T. Vogl, J. Voit, and D. Platt, “Blood coagulation factors in the elderly,” *Archives of Gerontology and Geriatrics*, vol. 9, no. 3, pp. 277–282, Nov. 1989, doi: 10.1016/0167-4943(89)90047-2.

[131] J. Barretina *et al.*, “The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity.,” *Nature*, vol. 483, no. 7391, pp. 603–607, Mar. 2012, doi: 10.1038/nature11003.

[132] C. Klijn *et al.*, “A comprehensive transcriptional portrait of human cancer cell lines.,” *Nat Biotechnol*, vol. 33, no. 3, pp. 306–312, Mar. 2015, doi: 10.1038/nbt.3080.

[133] K. Charmpi, M. Chokkalingam, R. Johnen, and A. Beyer, “Optimizing network propagation for multi-omics data integration,” *PLOS Computational Biology*, vol. 17, no. 11, p. e1009161, Nov. 2021, doi: 10.1371/journal.pcbi.1009161.

[134] A. Liberzon, C. Birger, H. Thorvaldsdóttir, M. Ghandi, J. P. Mesirov, and P. Tamayo, “The Molecular Signatures Database (MSigDB) hallmark gene set collection.,” *Cell Syst*, vol. 1, no. 6, pp. 417–425, Dec. 2015, doi: 10.1016/j.cels.2015.12.004.

[135] M. E. Ritchie *et al.*, “limma powers differential expression analyses for RNA-sequencing and microarray studies,” *Nucleic Acids Research*, vol. 43, no. 7, pp. e47–e47, Apr. 2015, doi: 10.1093/nar/gkv007.

[136] C. C. Preston *et al.*, “Aging-induced alterations in gene transcripts and functional activity of mitochondrial oxidative phosphorylation complexes in the heart.,” *Mech Ageing Dev*, vol. 129, no. 6, pp. 304–312, Jun. 2008, doi: 10.1016/j.mad.2008.02.010.

[137] K. Boengler, M. Kosiol, M. Mayr, R. Schulz, and S. Rohrbach, “Mitochondria and ageing: role in heart, skeletal muscle and adipose tissue.,” *J Cachexia Sarcopenia Muscle*, vol. 8, no. 3, pp. 349–369, Jun. 2017, doi: 10.1002/jcsm.12178.

[138] H. Jiang, W.-J. Zhu, J. Li, Q.-J. Chen, W.-B. Liang, and Y.-Q. Gu, “Quantitative histological analysis and ultrastructure of the aging human testis,” *International Urology and Nephrology*, vol. 46, no. 5, pp. 879–885, May 2014, doi: 10.1007/s11255-013-0610-0.

[139] D. E. Harrison *et al.*, “Rapamycin fed late in life extends lifespan in genetically heterogeneous mice,” *Nature*, vol. 460, no. 7253, pp. 392–395, Jul. 2009, doi: 10.1038/nature08221.

[140] I. Bjedov *et al.*, “Mechanisms of life span extension by rapamycin in the fruit fly Drosophila melanogaster.,” *Cell Metab*, vol. 11, no. 1, pp. 35–46, Jan. 2010, doi: 10.1016/j.cmet.2009.11.010.

[141] E. L. Baar, K. A. Carbajal, I. M. Ong, and D. W. Lamming, “Sex- and tissue-specific changes in mTOR signaling with age in C57BL/6J mice.,” *Aging Cell*, vol. 15, no. 1, pp. 155–166, Feb. 2016, doi: 10.1111/acel.12425.

[142] B. D. Harris, M. Crow, S. Fischer, and J. Gillis, “Single-cell co-expression analysis reveals that transcriptional modules are shared across cell types in the brain,” *Cell Systems*, vol. 12, no. 7, pp. 748-756.e3, Jul. 2021, doi: 10.1016/j.cels.2021.04.010.

[143] M. Yang, B. R. Harrison, and D. E. Promislow, “In search of a Drosophila core cellular network with single-cell transcriptome data,” *G3 Genes| Genomes| Genetics*, 2022.

[144] E. Ravasz, A. L. Somera, D. A. Mongru, Z. N. Oltvai, and A.-L. Barabási, “Hierarchical Organization of Modularity in Metabolic Networks,” *Science*, vol. 297, no. 5586, pp. 1551–1555, Aug. 2002, doi: 10.1126/science.1073374.

[145] A. F. Møller and K. N. Natarajan, “Predicting gene regulatory networks from cell atlases,” *Life Sci. Alliance*, vol. 3, no. 11, p. e202000658, Nov. 2020, doi: 10.26508/lsa.202000658.

[146] A. Dixit *et al.*, “Perturb-Seq: dissecting molecular circuits with scalable single-cell RNA profiling of pooled genetic screens,” *cell*, vol. 167, no. 7, pp. 1853–1866, 2016.

[147] K. Glass, C. Huttenhower, J. Quackenbush, and G.-C. Yuan, “Passing messages between biological networks to refine predicted interactions,” *PloS one*, vol. 8, no. 5, p. e64832, 2013.

[148] A. R. Sonawane *et al.*, “Understanding tissue-specific gene regulation,” *Cell reports*, vol. 21, no. 4, pp. 1077–1088, 2017.

[149] D. Lähnemann *et al.*, “Eleven grand challenges in single-cell data science,” *Genome Biology*, vol. 21, no. 1, p. 31, Feb. 2020, doi: 10.1186/s13059-020-1926-6.

[150] E. Kelmer Sacramento *et al.*, “Reduced proteasome activity in the aging brain results in ribosome stoichiometry loss and aggregation.,” *Mol Syst Biol*, vol. 16, no. 6, p. e9596, Jun. 2020, doi: 10.15252/msb.20209596.

Ana Carolina Leote

Computational Biology of Ageing

Luxemburger Straße, 124-136, 2809,
50939 Cologne, Germany
☎ +491742060861
✉ ana.carolina.leote@uni-koeln.de

0000-0003-0879-328X
anacarolinaleote
ACarolinaLeote

Research Experience

Cologne Excellence Cluster on Cellular Stress Responses in Aging Associated Diseases (CECAD), Cologne, Germany

PhD candidate, Cellular Networks and Systems Biology Group (Oct 2017 - present)

"Transcriptional regulation invariance and breakdown with age and disease"

ADImpute Bioconductor package for single-cell RNA sequencing dropout imputation

Leote, A.C., Wu, X., Beyer, A. (2022) Regulatory network-based imputation of dropouts in single-cell RNA sequencing data. PLOS Computational Biology 18(2): e1009849.

scRNA-seq Data Analysis | Regularized Regression | Transcriptomic Network Inference | Network Topology Analysis | Random Forest | Stability Selection | Bioconductor Package Development (R)

Instituto de Medicina Molecular (iMM), Lisbon, Portugal

Researcher, Disease Transcriptomics Lab (Nov 2016 - Oct 2017)

Master student, Disease Transcriptomics Lab (Feb 2016 - Oct 2016)

"Genome-wide profiling of gene expression and alternative splicing in ageing across human tissues"

Bulk RNA-seq Data Analysis | Differential Expression Analysis | Alternative Splicing Analysis | Linear Regression | Principal Component Analysis | Cross-Validation (R)

Instituto de Tecnologia Química e Biológica (ITQB), Oeiras, Portugal

Intern, Protein Modeling Lab (Jul 2014 - Sep 2014)

"Investigation of the secondary structure of the small peptide Crotalphine"

Molecular Dynamics (Linux shell scripting)

University Education

Instituto Superior Técnico, Lisbon, Portugal

Integrated Master Degree in Biological Engineering (2011 - 2016)

Graded 15/20 for BSc. and 18/20 for MSc.
Part of Merit Board in 2014 - 2016
Course Representative in 2013 - 2016

Scientific Publications & Conferences

Leote, A.C., Wu, X., Beyer, A. (2022) Regulatory network-based imputation of dropouts in single-cell RNA sequencing data. PLOS Computational Biology 18(2): e1009849.

Debès, C., **Leote, A.C.**, Beyer, A., "Computational approaches for the systematic analysis of aging-associated molecular alterations" (2019) Drug Discovery Today: Disease Models 27:51-59.

Georgilis, A., Klotz, S., Hanley, C.J., Herranz, N., Weirich, B., Morancho, B., **Leote, A.C.**, (...), Gil J. "PTBP1-Mediated Alternative Splicing Regulates the Inflammatory Secretome and the Pro-tumorigenic Effects of Senescent Cells" (2018) Cancer Cell 34(1):85–102.e9.

Gallego-Paez, L.M., Bordone, M.C., **Leote, A.C.**, Saraiva-Agostinho N., Ascensão-Ferreira M., Barbosa-Morais N.L. "Alternative splicing: the pledge, the turn, and the prestige : The key role of alternative splicing in human biological systems." (2017) Human Genetics 136(9):1015-1042.

22nd EMBL PhD Symposium: The Roaring 20s: A New Decade for Life Sciences (Nov 2020) - Online poster presentation

9th CGA Graduate Symposium, Cologne, Germany (Jun 2021) – Online oral presentation

Big Data in Health PhD Summer School, London, UK (Jul 2019) – Oral presentation

ISMB/ECCB 2019, Basel, Switzerland (Jul 2019) – Poster presentation

Arbeitsgemeinschaft für Gen-Diagnostik annual meeting, Potsdam, Germany (Oct 2018) - Poster presentation

Healthy Ageing: From Molecules to Organisms, Hinxton, UK (Jan 2018)

Training & Awards

Big Data in Health PhD Summer School, London, UK (Jul 2019)

3rd poster prize winner, Arbeitsgemeinschaft für Gen-Diagnostik annual meeting, Potsdam, Germany (Oct 2018)

Cologne Graduate School for Ageing Research (CGA) (2017 - 2020)
Selected for a structured PhD programme (9 slots out of over 900 candidates)

2015 Summer School on Computational Biology, Coimbra, Portugal (Sep 2015)

Research-related Responsibilities

Bachelor and Master student co-supervisor

RNA-seq Co-Expression Analysis Across Tissues and Ageing (Feb 2021 - Oct 2021)
Genome-Wide Association Study of Biological vs Chronological Aging (Nov 2019 - Nov 2020)
Inference of age-associated copy number alterations from single-cell RNA-sequencing data (Nov 2018 - Sep 2019)
Inference of copy number variations in aging cells from scRNA-seq data (Sep 2018 - May 2019)
Method development for Copy Number Alteration detection from single cell RNA-sequencing data (Jan - Jun 2018)

Practical class teacher (University of Cologne)

Computational Biology, winter semester 2021/22
Bioinformatics I, winter semesters 2018/19, 2019/20 and 2020/21
Advanced Bioinformatics, summer semester 2018

Cologne Graduate School for Ageing Research Student Representative (2017-2019)

Skills

Programming Languages: R - daily usage | Python, C, MATLAB and F – basic programming skills | Unix shell

Machine Learning Methods: Linear Regression | Regularized Regression | Random Forest | Cross-Validation | Stability Selection

Data Analysis: Single-cell RNA-seq | Bulk RNA-seq | Principal Component Analysis | Transcriptomic Network Inference | Network Topology Analysis

Development: Bioconductor Package Development | Git | GitHub

Languages: Portuguese (Native) | English (Fluent) | German (A2) | French (Basic)

References

Prof. Dr. Andreas Beyer
Professor for Systems Biology, University of Cologne
http://cellnet.cecad.uni-koeln.de/
andreas.beyer@uni-koeln.de
+49 221-478 84429

Dr. Peter Tessarz
Max Planck Research Group Leader
CECAD Principal Investigator
www.tessarz-lab.com
ptessarz@age.mpg.de
+49 221 37970-680