UNIVERSITÉ DU LUXEMBOURG

PhD- FSTM-2023-047
The Faculty of Science, Technology and Medicine

DISSERTATION

Defence held on 14/06/2021 in Esch-sur-Alzette

to obtain the degree of

DOCTEUR DE L'UNIVERSITÉ DU LUXEMBOURG

EN BIOLOGIE

by

Nikola Maria DE LANGE
Born on 10. June 1992 in Kleve, Germany

DEVELOPMENT OF THE BIOINFORMATICS PIPELINE DREMFLOW FOR THE IDENTIFICATION OF CELL-TYPE AND TIME POINT SPECIFIC TRANSCRIPTIONAL REGULATORS

Dissertation defence committee

Dr Roland Krause, dissertation supervisor
Université du Luxembourg

Dr Emma Schymanski, Chairwoman
Associate Professor, Université du Luxembourg

Dr Lasse Sinkkonen
Université du Luxembourg

Dr Ivan G. Costa
Professor, University Hospital RWTH Aachen

Dr Vera van Noort,
Professor, Universiteit Leiden

Development of the bioinformatics pipeline DREMflow for the identification of cell-type and time point specific transcriptional regulators

A dissertation by

Nikola Maria de Lange

submitted to the University of Luxembourg

in partial fulfillment of the requirements for the degree of

DOCTEUR DE L'UNIVERSITÉ DU LUXEMBOURG EN BIOLOGIE

Dissertation defense committee:

Prof. Dr. Emma Schymanski, Chair of the committee

Prof. Dr. Vera van Noort

Prof. Dr. Ivan G. Costa (CET Committee Member)

Dr. Lasse Sinkkonen (CET Committee Member)

Dr. Roland Krause, Dissertation Supervisor

Affidavit

I hereby declare that this thesis entitled “Development of the bioinformatics pipeline DREM-flow for the identification of cell-type and time point specific transcriptional regulators” has been written independently and without any other sources than cited.

Acknowledgments

I want to express my sincere gratitude to my PhD supervisor, Roland Krause, who guided me with a lot of patience and understanding throughout the years and my PI Reinhard Schneider for his valuable advices during our monthly meetings.

I would like to thank my committee, Prof Emma Schymanski, Prof Vera van Noort, Prof Ivan Costa and Dr Lasse Sinkkonen, for agreeing to read and evaluate my thesis. A special thanks to my CET members, Prof Ivan Costa and Dr Lasse Sinkkonen, who took time of their busy schedule to guide me throughout the years.

And I would like to thank my support system of family and friends, who always reminded me of who I am and what I am capable of. To my parents, who supported me already for my whole life. I would not be here without them. To my sister and best friend, Isabelle, who has always been there for me, no matter what. She pointed in the right direction when ever I felt lost. To my friends, Susana, Shaman and Axel, who where my bubble during the pandemic. And to my friend Michela, who made time for writing sessions with me when I needed it the most.

Last but not least, I would like to thank my grandpa. Although you are not around anymore, you inspired me to be here.

Abstract

A detailed understanding of the mechanism that drive cell differentiation of stem cells into a desired cell type provides opportunities to study diseases and disease progression in patient derived cells and enable the development of new therapy approaches. The main challenge in this directed differentiation is the identification of the essential transcriptional regulators involved that are specific to a cell type or lineage and the inference of the underlying gene regulatory network.

Transcription factor activity during cell differentiation can be measured through gene expression and chromatin accessibility, ideally jointly over time. Integrated time course regulatory analysis yields more detailed gene regulatory networks than expression data alone. Due to the large number of parameters and tools employed in such analysis, computational workflows help to manage the inherent complexity of such analyses.

This thesis describes Dynamics Regulatory Events Miner Snakemake workflow (DREMflow) which combines temporally-resolved RNA-seq and ATAC-seq data to identify cell type and time point specific gene regulatory networks. DREMflow builds on the Differentially Regulatory Events Miner (DREM), the workflow management system Snakemake and the package manager Mamba. It includes the processing starting from sequencing reads, quality control reports and parameters as well as additional downstream analyses for the inference of key transcription factors during differentiation.

DREMflow is applied to multiple data sets obtained during the differentiation of midbrain dopaminergic neurons as well as blood cells and compared to TimeReg, a pipeline with similar aims. The expansion to accommodate for single-cell data is explored.

Results from other studies were reproduced and extended, identifying additional key transcriptional regulators. LBX1 was found as key regulator in differentiation of midbrain dopaminergic neurons while exploring different settings of the pipeline. Members of the AP-1 family of transcription factors were identified in all blood cell differentiation data sets. The comparison to TimeReg resulted in DREMflow being more sensitive in the identification of known transcriptional regulators in macrophages. Computationally, DREMflow outperforms TimeReg as well.

DREMflow enables users to perform time-resolved multi-omics analysis reproducibly with minimal setup and configuration.

Table of contents

List of Figures

List of Tables

List of abbreviations

Abbreviation	
AP-1	activator protein-1
ASCL1	Achaete-Scute Complex-Like 1
ATAC-seq	Assay for transposase-accessible chromatin using sequencing
BFU-E	burst forming unit-erythroid
BRN2	Brain-2
bZIP	basic leuzin zipper factors
CEBP	CCAAT Enhancer Binding Protein
CEBPA	CCAAT Enhancer Binding Protein Alpha
CFU-E	colony-forming unit erythroid
ChIP-seq	Chromatin immunoprecipitation followed by sequencing
CLP	common lymphoid progenitors
CMP	common myeloid progenitors
CRS	cis-regulation score
DBD	DNA Binding Domain
DEG	differentially expressed genes
DNAse-seq	DNAse I coupled with deep sequencing
DREM	Dynamics Regulatory Events Miner

Abbreviation	
EBF1	Early B Cell Factor 1
EGR1/2/3/4	Early Growth Response Factor 1/2/3/4
EN1/2	Engrail 1/2
ESCs	embryonic stem cells
FACS	flow-cytometry activated cell sorting
FOXA2	Forkhead Box A2
FP	floor plate
FPKM	Fragments Per Kilobase of transcript per Million reads mapped
GATA1	GATA Binding Protein 1
GATA3	GATA Binding Protein 3
GFI1	Growth Factor Independent 1 Transcriptional Repressor
GMP	granylocyte-macrophage progenitors
GO	Gene Ontology
GTFs	general transcription factors
H3K27Ac	histone H3 lysine 27 acetylation
H3K4me3	histone H3 lysine 4 trimethylation
HI-C	High-throughput chromosome conformation capture
HMM	Hidden Markov models
HOX	Homeobox
HPC	High Performance Computing
HSCs	hematopoietic stem cells

Abbreviation	
HSPCs	hematopoietic stem and progenitor cells
iDREM	interactive Dynamics Regulatory Events Miner
IOHMM	Input-Output Hidden Markov models
iPSCs	induced pluripotent stem cells
IRF	interferon regulatory factors
KLF1	Kruppel-Like Factor 1
KLF4	Kruppel-Like Factor 4
LEF1	Lymphoid Enhancer Binding Factor 1
LMO2	LIM Domain Only 2
LMX1A/B	LIM homeobox Transcription Factor 1 Alpha/Beta
MAF	C-Maf Proto-Oncogene
mDAN	midbrain dopaminergic neurons
MEP	megakaryocyte-erythroid progenitors
mESC	mouse embryonic stem cells
MHB	midbrain-hindbrain boundary
MYC	MYC Proto-Oncogene, BHLH Transcription Factor
MYT1	Myelin Transcription Factor 1
NANOG	Nanog Homeobox
NEUROD1	Neuronal Differentiation 1
NR4A2	Orphan Nuclear Receptor Nurr1
NSCs	neural stem cells

Abbreviation	
OTX2	Orthodenticle Homeobox 2
PAX5	Paired Box 5
PCA	principle component analysis
PetaFLOP/s	Peta Floating Point Operations per Second
PIC	preinitiation complex
POU5F1	POU Class 5 Homeobox 1
PWM	position weight matrix
RA	retinoic acid
RPKM	Reads Per Kilobase of transcript per Million reads mapped
RUNX1	RUNX Family Transcription Factor 1
SHH	Sonic Hedgehog
smNPCs	small molecule neuronal stem cells
SOX2	SRY-Box Transcription Factor 2
SP1	Specificity Protein 1
SPI1	Spi-1 Proto-Oncogene
TAL1	T-Cell Acute Lymphocytic Leukemia Protein 1
TF	transcription factors
TFBS	transcription factor binding site
TG	target genes
TPM	Transcripts Per Million
TSS	transcription start site
ULHPC	University of Luxembourg High Performance Computing

1. Introduction

The ability to differentiate stem cells into a desired cell type gives rise to new therapies such as cell replacement and individualized drug development by using patient cells to study their specific condition *in vitro* (Figure 1.1). While many protocols for directed cell differentiation and cellular reprogramming are known, the safety of aforementioned therapies depends on the exact knowledge of the differentiation mechanisms driven by master regulating transcription factors and regulatory elements. Many important regulators have been identified but their specific roles and functions as well as a complete census are subject to research.

Studying cell differentiation is not only valuable for cellular reprogramming but also essential to understand the functions of cells and cell complexes. In addition, many diseases such as cancer or degenerative diseases can be traced back to abnormalities during cell differentiation. Understanding the process can increase insights into diseases.

1.1. Cell differentiation

Cell differentiation is the process of unspecialized stem cells committing to highly specialized cell types during embryonic development and also continuously inside the adult body (Sánchez Alvarado and Yamanaka 2014; Zakrzewski et al. 2019).

Stem cells are unspecialized cells, e.g. oocytes and can be classified into totipotent, pluripotent, multipotent, oligopotent and unipotent. All cells are derived from a fertilized

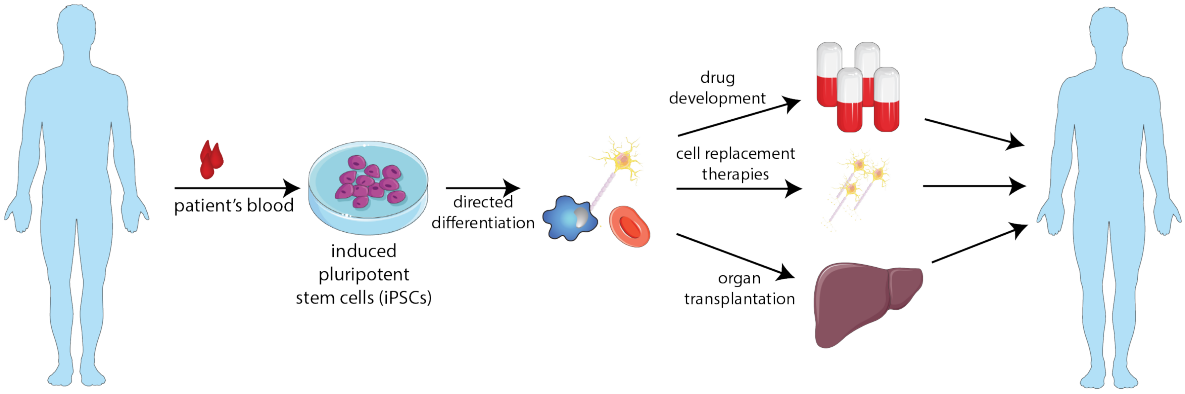


Figure 1.1.: Schematic overview of applications in regenerative medicine. Cells from patients can be regprogrammed to induced pluripotent stem cells and differentiated into a relevant cell type. These cells can be used for drug development and cell replacement therapy. The figure was generated using Servier Medical Art, provided by Servier, licensed under a Creative Commons Attribution 3.0 unported license and modified.

oocyte that through repeated cell division differentiates into highly specialized cells (Bloushtain-Qimron et al. 2009). A totipotent stem cell, e.g. a zygote, has the potential to differentiate into the three germ layers and the placenta. Pluripotent stem cells like embryonic stem cells (ESCs) are able to form cells of the three germ layers, ectoderm, mesoderm and endoderm, but not the placenta (Figure 1.2) (Zakrzewski et al. 2019; Mohr, De Pablo, and Palecek 2006).

The ectoderm germ layer gives rise to neurons, astrocytes, oligodendrocytes, keratinocytes and melanocytes initiated by key signaling pathways such as Wnt, SHH and retionic-acid (RA). It is the most distal layer and can be classified into ectoderm for skin and pigment cells and neuroectoderm for neural cells (Figure 1.2).

Mesoderm and endoderm lineages are relatively close with both relying on activin signaling. Mesoderm gives rise to cardiac, endothelial and hematopoietic stem cells, while endoderm develops into hepatocytes and □cells (Efthymiou et al. 2014).

Somatic or adult stem cells are multipotent. They are harbored in many mature tissues and are crucial for regeneration and homeostasis. Examples are hematopoietic stem cells (HSCs) as progenitors for blood cells and neural stem cells (NSCs) for neural cells

(Sánchez Alvarado and Yamanaka 2014; Zakrzewski et al. 2019; Efthymiou et al. 2014). Oligopotency describes the ability to differentiate into several cell types like myeloid stem cells. HSCs can still differentiate into all blood cells, while myeloid stem cells are comitted to white blood cells (Zakrzewski et al. 2019). Unipotent stem cells are committed to one cell type but retain the ability to repeatedly divide, making them valuable for cell therapy or regenerative medicine (Zakrzewski et al. 2019).

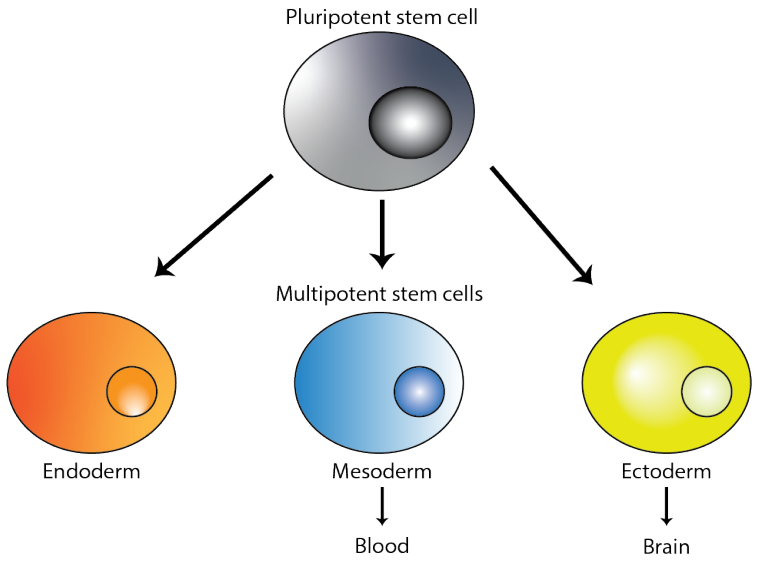


Figure 1.2.: Simplified overview of cell differentiation from pluripotent stem cells. Pluripotent cells differentiate into multipotent cells from endoderm, mesoderm and ectoderm. Multipotent stem cells can differentiate into highly specified cells, for example blood cells like macrophages and brain cells like neurons. The graphic was adapted from Zakrzewski et al. (2019).

With the discovery of cellular reprogramming it was possible to transform somatic cells into induced pluripotent stem cells (iPCS), given rise to another model to study cell differentiation (Takahashi and Yamanaka 2006; Sánchez Alvarado and Yamanaka 2014).

Cell differentiation progresses in several stages before resulting in highly specialized cells and is driven by changes in chromatin states and the resulting effects on gene expression (Bloushtain-Qimron et al. 2009; T. Chen and Dent 2014; Sánchez Alvarado and Yamanaka 2014). This process is not considered irreversible anymore (Takahashi and Yamanaka 2006; Vierbuchen et al. 2010). By asymmetric cell division, one daughter

cell maintains pluripotency while the other daughter cell becomes more specific (Morrison and Kimble 2006). Pluripotency in ESCs is accompanied by a generally open chromatin landscape (T. Chen and Dent 2014). To maintain the pluripotent state of a stem cell the transcription factors (TF) SRY-Box Transcription Factor 2 (SOX2), POU Class 5 Homeobox 1 (POU5F1), also known as OCT4 and Nanog Homeobox (NANOG), known as core pluripotency factors, are active (Takahashi and Yamanaka 2013). Transcription factors are proteins that bind to the DNA at a specific sequence and regulate gene expression. A detailed introduction of TF will follow in Section 1.2.1. Lineage-specific genes are repressed to preserve pluripotency in stem cells.

Cell differentiation occurs opposed to the maintenance of pluripotency by the regulation of cell type specific genes influenced by the micro-environment and triggered by signaling cascades (Efthymiou et al. 2014). To enable the transcription, the open chromatin state in stem cells is remodeled to a more compact and lineage-specific state. A combination of signaling pathways and transcription factor activity enables the expression of cell type specific genes while pluripotency genes are silenced (T. I. Lee and Young 2013).

Signaling molecules such as morphogens play a crucial role in the process of differentiation. Morphogens form a concentration gradient to facilitate the activation of signalling pathways. Examples for such molecules are Hedgehog (Hh) or Wingless (Wg)/Wnt proteins (Tabata 2001; Efthymiou et al. 2014).

The interplay of transcription factors and regulatory elements and signalung gradients to control gene regulation is in general a robust system that can react to perturbations by replacing a missing gene within a gene regulatory network by a gene covering the same function (MacNeil and Walhout 2011). However, it can be sensitive to mutations in regulatory sequences on the DNA. Such mutations can hinder transcription factor binding, prevent gene expression and alter cell differentiation leading to diseases such as cancer (T. I. Lee and Young 2013).

1.1.1. Cellular identity

Cellular phenotypes are defined by gene expression and regulating functions that are specific for a cell type or cell state (Basso et al. 2005). The chromatin state, gene expression, protein abundance and a cell specific response to stimuli are measure to determine the cell identity (Abdolhosseini et al. 2019; Mincarelli et al. 2018). The use of high-throughput technologies proved to be valuable to identify the cell state, when RNA-seq replaced microarrays (Zhao et al. 2014). As suggested by Samantha A. Morris, lineage and function of a cell can be considered as well for the concept of cell identity (Morris 2019). Understanding the cellular phenotype is necessary to distinguish between health and disease (Basso et al. 2005). Cell differentiation was thought to be irreversible for a long time and therefore the cell identity was defined as a static state. This changed with the discovery of cell conversion and reprogramming (Morrison and Kimble 2006; Takahashi and Yamanaka 2006).

1.1.2. Cellular reprogramming

Changing the cellular identity of a cell can be achieved through environmental factors as well as mechanical manipulation of the cell. In 1938, Spemann proposed the concept of nuclear transplantation (Reviewed in Sánchez Alvarado and Yamanaka (2014)), which proved to be successful in 1952, when blastula cell nuclei were transplanted into frog oocytes (Briggs and King 1952). The breakthrough came in 1981, when Martin Evans and Gail Martin generated infinitely proliferating mouse ESCs cells by culturing the inner cell mass of blastocysts (Evans and Kaufman 1981; Martin 1981).

The next milestone in the field was achieved in 2006, when Kazutoshi Takahashi and Shinya Yamanaka discovered that expression of only four transcription factors induces pluripotency in mouse fibroblasts. Named after their discoverer, the yamanaka factors were identified as OCT3/4 (now POU5F1), SOX2, Kruppel-Like Factor 4 (KLF4) and MYC

Proto-Oncogene, BHLH Transcription Factor (MYC or c-MYC) and by simultaneously introducing them to fibroblasts, they reprogrammed the cells into embryonic stem cell-like cells (Takahashi and Yamanaka 2006; Sánchez Alvarado and Yamanaka 2014). The resulting cells since then have been called induced pluripotent stem cells (iPSCs).

The discovery of the Yamanaka factors to reprogram fibroblasts into iPCS eventually led to the development of cell conversion, enabling the reprogramming from fibroblasts into neurons and other cell types (Vierbuchen et al. 2010).

Patient derived iPSCs give options to study rare diseases and in the long term provide the possibility for cell replacement therapies but also carry a risk of uncontrolled cell division if the differentiation protocol is not precise. While there are already many known protocols for the formation of progenitors to each germ layer, the use of transcription factors for the directed differentiation have an influence on the phenotype *in vivo* (Zakrzewski et al. 2019).

1.1.3. Differentiation into dopaminergic neurons

Cellular reprogramming and induced differentiation are widely used to obtain specific cell types to study the process of differentiation or diseases. Neurodegenerative diseases such as Parkinson's Disease (PD) were difficult to study for many years. PD is associated with the loss of dopaminergic neuron in the substantia nigra, a region in the midbrain. Obtaining cells from patients' brains to study the progression of the disease especially in early stages is not possible. Cells become available postmortem. A popular in vitro model to study neurodegenerative diseases was the neuroblastome cell line SH-SY5Y, that was discovered in 1978 (Biedler et al. 1978). While SH-SY5Y cells showed neuronlike features they were not suitable to study neuronal functions (Feles et al. 2022). After the discovery of iPSCs in 2006, a new possibility to generate and study neurons and eventually dopaminergic neurons emerged.

The long road to a protocol for dopaminergic neurons started in 2010 when the successful cellular reprogramming from mouse embryonic fibroblasts to functional neurons was achieved by a combinatorial expression of Brain-2 (BRN2), Achaete-Scute Complex-Like 1 (ASCL1) and Myelin Transcription Factor 1 (MYT1). They are referred to as BAM factors (Vierbuchen et al. 2010). For humans, the TF Neuronal Differentiation 1 (NEUROD1) was added (Pang et al. 2011). Already in 2011 lineage reprogramming from fibroblasts to midbrain dopaminergic neurons (mDAN) was succesful (Pfisterer et al. 2011) but the functional capacity did not compare to tissue cells.

In 2014 Grealish et al. discovered a protocol to match the functional capacity of fetal tissue with iPSC-derived mDAN (Grealish et al. 2014). One year later, several differentiation protocols for the differentiation of midbrain dopaminergic neurons (mDAN) were discussed by Arenas at al, but not all mechanisms underlying the differentiation were understood (E. Arenas, Denham, and Villaescusa 2015).

The formation of mDAN occurs during the development of the midbrain. Without going into detail, first the neural tube is formed, then it progressed to develop into spinal cord, hindbrain, midbrain and forebrain. The formation of the neural tube is governed by the floor plate (FP) which serves as organizing center and signaling center. The other signaling center required for mDAN differentiation is the midbrain-hindbrain boundary (MHB). Neurogenesis of mDAN depends on these signaling centers and expression of mDAN specific TFs (Ernest Arenas 2014; M. Wang et al. 2020). Sonic Hedgehog (SHH)-Forkhead Box A2 (FOXA2) signaling regulates neurogenesis through FOXA2 expression and promotes mDAN differentiation in the floor plate (M. Wang et al. 2020). The Wnt signaling pathway facilitates the expression of midbrain specific TFs such as Orthodenticle Homeobox 2 (OTX2), Engrail 1/2 (EN1/2), Orphan Nuclear Receptor Nurr1 (NR4A2) and LIM homeobox Transcription Factor 1 Alpha/Beta (LMX1A/B) (Ernest Arenas 2014; M. Wang et al. 2020). A simplified version of the gene regulatory network is shown in Figure 4.4.

The discovery of iPSCs was a milestone for the research in PD. iPSCs did not only provide a possibility to study the disease *in vitro* from patient derived fibroblasts but also an oppor-

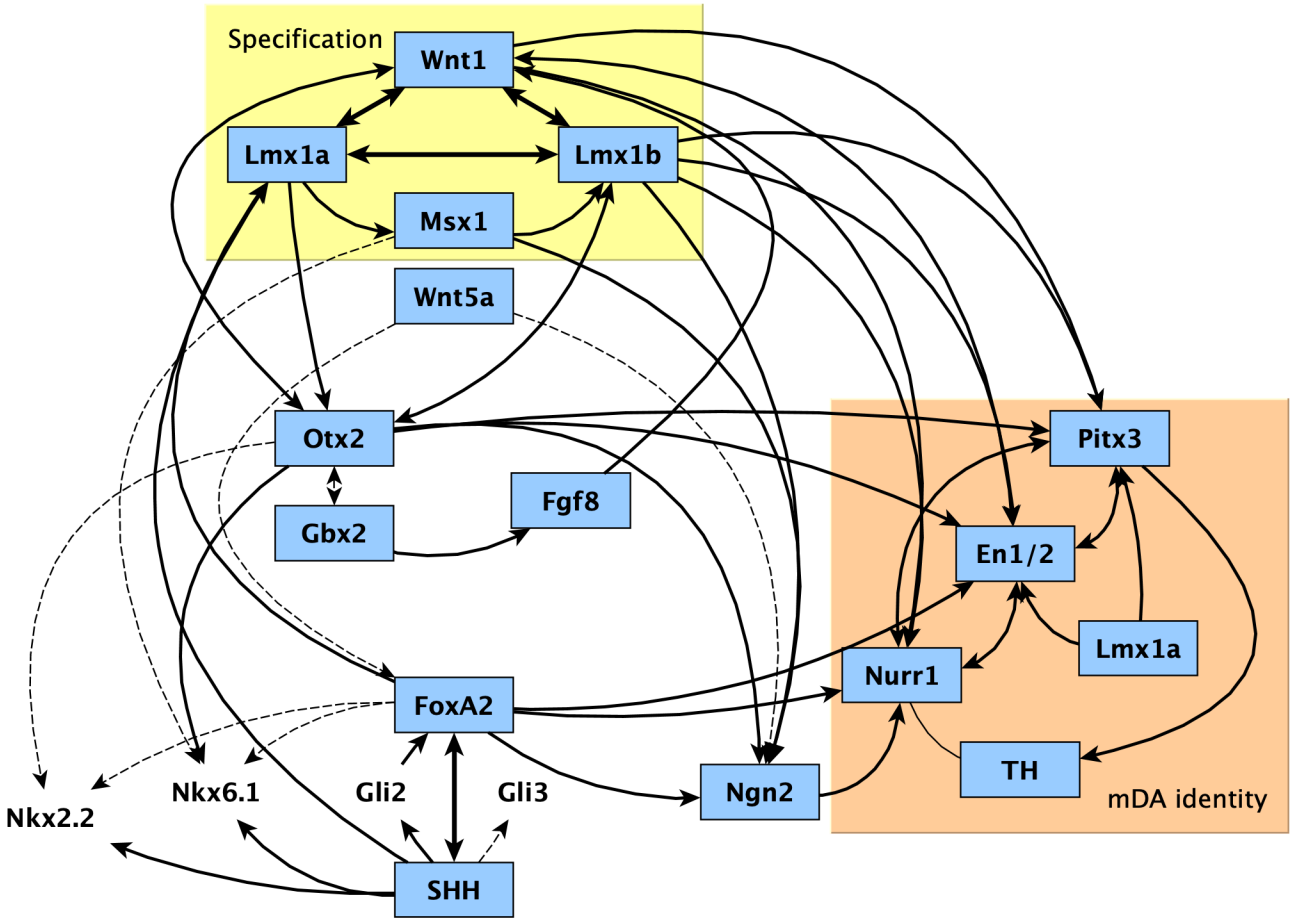


Figure 1.3.: Gene regulatory network in mDAN neurogenesis. Adapted from E. Arenas, Denham, and Villaescusa (2015) and M. Wang et al. (2020). Solid lines represent activation while dashed lines represent inhibition. Otx2, Gbx2 and Fgf8 are related to the midbrain hindbrain boundary. Lmx1a, Lmx1b, Msx1 and Wnt1 form the Wnt1-Lmx1a/b-Msx1 network during floorplate specification and regulate each other and Otx2. Nkx2.2 and Nkx6.1 are transcription factors for the basal midbrain and therefore need to be repressed by Msx1, Otx2 and FoxA2. Nurr1, En1 and Pitx3 are postmitotic marker for mDAN and regulate each other. FoxA2 builds the SHH-FoxA2 network either directly or through Gli2 and Gli3.

tunity to work eventually towards cell replacement therapies (Hiller et al. 2022). Dopamine progenitor cells derived from iPSCs are implanted already in 2020 in a therapeutic context (Schweitzer et al. 2020).

1.1.4. Differentiation of blood cells

With over 50 years of intensive studies, hematopoietic stem cells can be considered as best characterized stem cells (Zakrzewski et al. 2019). They reside in the bone marrow, which makes them more easily accessible than other precursor cells (Weiskopf et al. 2016).

Hematopoiesis (also hemopoiesis) describes the continues replenishing of the blood cells through differentiation of HSC into highly specialized cells. The HSC reside primarily in the bone marrow Pinho and Frenette (2019). A fine balance between self-renewal and differentiation ensures the maintenance of the HSC pool. The mechanisms and signals for this maintenance are referred to as HSC niche (Pinho and Frenette 2019).

During development hematopoiesis starts in the yolk sac blood islands and continues in the fetal liver and organs. However, tissue resident macrophages can develop directly from yolk sac progenitors(Weiskopf et al. 2016).

HSC commit to myeloid and lymphoid cell types. The myeloid lineage starts with the common myeloid progenitors (CMP) and the lymphoid lineage with the common lymphoid progenitors (CLP) (Weiskopf et al. 2016; Rosenbauer and Tenen 2007; Ramirez et al. 2017; Álvarez-Errico et al. 2015). Following the CMP, the granylocyte-macrophage progeniters (GMP) proceed to differentiate into monocytes, neutrophils, eosinophils, mast cells and basophils, while the megakaryocyte-erythroid progenitors (MEP) differentiate further into erythrocytes and platelets (Figure 1.4) (Álvarez-Errico et al. 2015; Ramirez et al. 2017; Ludwig et al. 2019; Lara-Astiaso et al. 2014; Dzierzak and Philipsen 2013). This is the deterministic model of hematopoiesis. An alternative model, called the stochastic model

was proposed by Novak and Stewart and suggests that depending on cell environmental conditions, any cell type can directly arise from HSCs. The model suggests that the differentiation is determined by the availability of differentiation factors Fisher (2002). An example are the tissue resident macrophages (Weiskopf et al. 2016). Notta et al. (2016) came to similar conclusions, proposing a two tier model that suggests a transition from multipotent HSC to unipotent myeloid, lymphocyte and erythrocyte cells (Notta et al. 2016).

In the early stages of hematopoiesis, multipotency related TFs in HSCs like RUNX Family Transcription Factor 1 (RUNX1) and GATA Binding Protein 3 (GATA3) should be inactive for differentiation, while CCAAT Enhancer Binding Protein Alpha (CEBPA), Early B Cell Factor 1 (EBF1) and Paired Box 5 (PAX5) are required to be active for myeloid cell differentiation (Álvarez-Errico et al. 2015; Ramirez et al. 2017). Transcription factors T-Cell Acute Lymphocytic Leukemia Protein 1 (TAL1) and LIM Domain Only 2 (LMO2) were identified as essential regulators for hemopoiesis (Dzierzak and Philipsen 2013). Spi-1 Proto-Oncogene (SPI1, also known as PU.1) serves as divider between CLP commitment and CMP commitment. Low expression levels are associated with the lymphoid lineage and high expression of SPI1 with the myeloid lineage (Dzierzak and Philipsen 2013; Jego et al. 2014). SPI1 is specifically an early TF in hematopoiesis (Nagamura-Inoue, Tamura, and Ozato 2001) and acts as pioneer factor, a type of early transcription factor that has the ability to bind to closed chromatin to enable other TFs to bind to the DNA (Atlasi and Stunnenberg 2017).

Macrophage differentiate often but not exclusively from monocytes (Novak and Stewart 1991; Fisher 2002). A long list of transcription factors involved in macrophage differentiation, proliferation and activation can be gathered across literature (more details in Section 3.4.2). SPI1 is mentioned most often and apart from the early role in myeloid differentiation it was identified as regulator during the transition from GMP to monocytes and macrophages. Early Growth Response Factor 1 (EGR1) was found as transcriptional regulator for the transition from monocytes to macrophages as well(Nagamura-Inoue, Tamura, and Ozato 2001). Already in 1993 FOS-related and JUN-related TFs

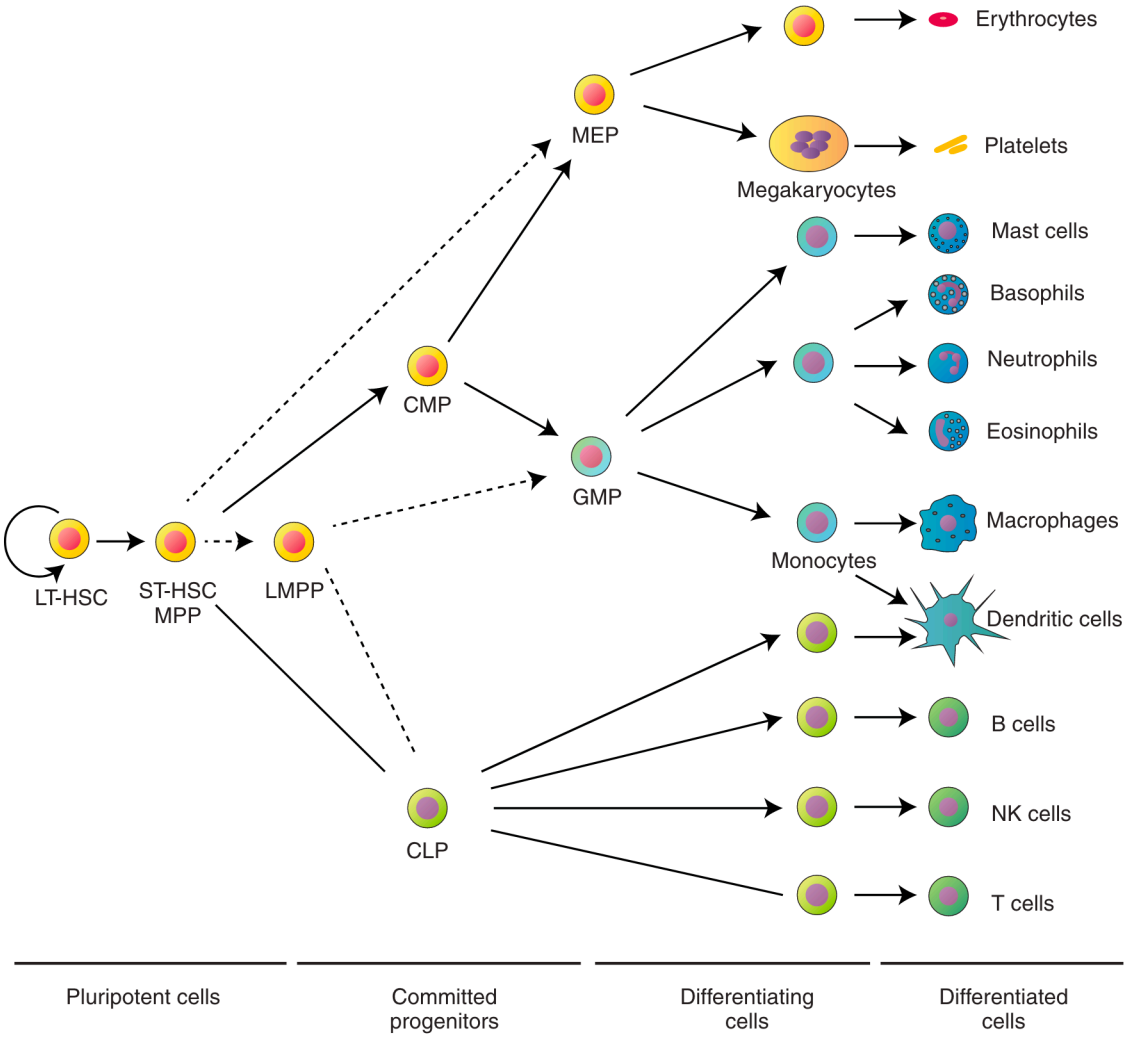


Figure 1.4.: Possible mechanisms of hematopoiesis as suggested by Dzierzak and Philipsen (2013). Long-term hemapoietic stem cells (LT-HSC) self renew, short-term hemapoietic stem cells differentiation progresses with common myeloid progenitors (CMP) or common lymphocyte progenitors (CLP). CMP further differentiate into granulocyte-macrophage progenitor (GMP) and megakaryocyte-erythroid progenitor (MEP). The dotted arrows represent an alternative model suggesting that MEPs can directly generate from ST-HPCs and lymphoid-primed multipotential progenitor (LMPP) can give rise to GMPs and CLPs.

were found to promote myeloid differentiation but the specific transitions were not specified (Lord et al. 1992). Hume and Himes summarized in their review that SPI1, TFs of the CCAAT Enhancer Binding Protein family (CEBP), C-Maf Proto-Oncogene (MAF), Specificity Protein 1 (SP1) and EGR1 are required for gene expression in mature macrophages (D. A. Hume and Himes 2003). The before mentioned TFs are best known in the context of macrophage differentiation, but many more are involved (David A. Hume, Summers, and Rehli 2016; Veremeyko et al. 2018; Pundhir et al. 2018; Nagamura-Inoue, Tamura, and Ozato 2001).

Monopoiesis, the differentiation into monocytes, is often explained in the context of macrophage differentiation, therefore many TFs such as SPI1 and CEBP TFs overlap. In addition, Hume and Himes found Homeobox (HOX) TFs in the context of monopoiesis. (D. A. Hume and Himes 2003). Specific for monocytes are interferon regulatory factors (IRF), especially IRF8 (Pundhir et al. 2018).

Although neutrophils are the most abundant leukocytes, the differentiation into neutrophils is not studied in the same extent as differentiation into monocytes and macrophages. Ai and Udalova wrote a comprehensive review and highlighted nine TFs during the differentiation from HSCs to circulating neutrophils. CEBPA and SPI1 are regulators during the whole process of differentiation, MYC was found to be an early regulator and AML1 an early and a late regulator. For the transition between later stage, from pre-neutrophils to immature, mature and circulating neutrophils, CEBPE, Growth Factor Independent 1 Transcriptional Repressor (GFI1), Lymphoid Enhancer Binding Factor 1 (LEF1), CEBPB and CEBPG were highlighted (Ai and Udalova 2020).

Transcriptional regulators in erythropiesis included early myeloid TFs such as SPI1 and CEBPA. Emerging from CMP the earliest erythroid committed progenitors are burst forming unit-erythroid (BFU-E), followed by colony-forming unit erythroid (CFU-E) in differentiation. They are more mature and more abundant than BFU-E (Dzierzak and Philipsen 2013). BFU-E differentiation further into proerythroblasts, basophilic erythroblasts, polychromic erythroblasts, orthochromatic erythroblasts until they reach maturation as reticu-

locytes and erythrocytes (Ludwig et al. 2019; Dzierzak and Philipsen 2013).

Krueppel-Like Factor 1 (KLF1) is the best known TF for terminal erythropoiesis, but GATA Binding Protein 1 (GATA1) was found to be involved in many regulatory processes and is often observed together with TAL1 (Cheng et al. 2009; Dzierzak and Philipsen 2013).

1.2. Gene regulation

Gene regulation is the essential mechanism for cell differentiation, homeostasis and proliferation (T. Chen and Dent 2014; Sánchez Alvarado and Yamanaka 2014). The expression of genes is a complex process that includes transcription initiation, elongation and termination as well as mRNA processing and nuclear export (Figure 1.5)(Buccitelli and Selbach 2020).

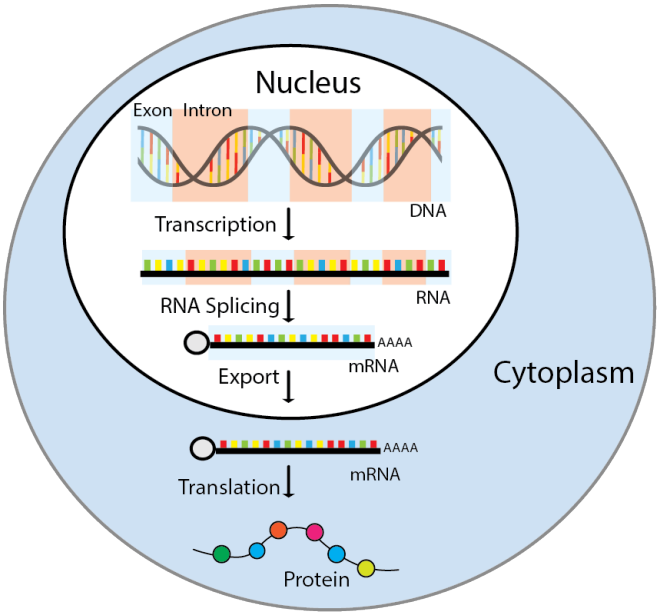


Figure 1.5.: Schematic overview of gene expression in the cell. DNA is transcriped into RNA. RNA undergoes splicing and post-transcriptional modifications resulting in mRNA inside the nucleus. After nuclear export, mRNA is translated into proteins in the cytoplasm.

A finely tuned network of cell-type specific transcription factors, signalling pathways and

post-translational modifications of proteins regulates the intercellular mechanisms necessary for gene expression (Schuettengruber et al. 2017; Buccitelli and Selbach 2020). To enable transcription factor binding, often changes in the chromatin state are required.

Chromatin is densely packed around nucleosomes consisting of histone proteins in the nucleus (Tsompana and Buck 2014). Nucleosome reorganization is needed to allow access of transcription factors to the DNA (Hager, McNally, and Misteli 2009). Pioneer factors, a type of transcription factor that has the ability to bind to closed chromatin, can increase the chromatin accessibility (Atlasi and Stunnenberg 2017).

Transcription of genes into protein-coding or non-coding RNA is facilitated by the binding of transcription factors to regulatory elements to recruit RNA polymerase II to the transcription start site (TSS) of target genes (Figure 1.6 A-C) (Maston, Evans, and Green 2006). Often not only one transcription factor is involved but several work jointly together (T. I. Lee and Young 2013; Hager, McNally, and Misteli 2009). It is a highly dynamic process and the binding of TFs to regulatory elements can last from a few seconds up to several minutes (Hager, McNally, and Misteli 2009; Coulon et al. 2013). Regulatory elements are regions in the genome that transcription factors interact with. Types of regulatory elements are insulators, silencers and enhancers, which are distal regulatory elements and promotors close to TSS as well as the core promoter at the TSS. The group of general transcription factors (GTFs) initiate transcription by assembling on the core promoter at the TSS building the preinitiation complex (PIC) to recruit RNA polymerase II (Figure 1.6 A) (Maston, Evans, and Green 2006; Hager, McNally, and Misteli 2009). The RNA polymerase II transcribes the DNA to RNA. The process is paused after transcription of 20-50bp and can transition to either elongation or termination of transcription (Figure 1.6 B-C) (T. I. Lee and Young 2013). The resulting RNA is subject to post-transcriptional modification such as splicing in the nucleus to generate mRNA followed by nuclear export and translation to proteins. Distal regulatory elements like enhancers regulate gene expression through interaction of promotor regions. These interactions are facilitated by DNA looping (Maston, Evans, and Green 2006). Regulatory elements can be identified through histone modifications.

In active, engaged enhancer regions nucleosomes with histone H3 lysine 27 acetylation (H3K27Ac) can be found while histone H3 lysine 4 trimethylation (H3K4me3) was linked to TSS (Figure 1.6 F)(Atlasi and Stunnenberg 2017; T. I. Lee and Young 2013).

Another mechanism involved in gene regulation is the interaction of microRNAs (miRNAs) by acting as post-transcriptional regulators of mRNA primarily through mRNA degradation or translational repression, so that mRNA is not further translated to proteins O'Brien et al. (2018). In some cases miRNAs can also regulate transcription or act as translational activators (O'Brien et al. 2018).

1.2.1. Classes of transcription factors

Transcription factors are proteins that regulate gene expression by binding to the DNA in a sequence-specific manner with DNA-binding domains (DBD). DBDs are structures in a folded protein that provide a surface to get in contact with DNA sequences (Harrison 1991). The specific sequence is called *motif* (reviewed in Lambert et al. (2018)). Classification of TFs is based on their specific DBD using the taxonomy introduced by Harrison (1991) as ground work (Lambert et al. 2018; Luscombe et al. 2000). Wingender et al. (2018) extended the classification based on DBD and proposed TFClass, a framework that takes the motif into account for subclassification, resulting in five levels. Level 1 categorizes TFs into superclasses that are defined based on the DBD topology, such as Zinc coordinating DNA-binding domains and helix-turn-helix domain (Wingender, Schoeps, and Dönitz 2013; Wingender et al. 2018). Level 2 represents classes that describe groups of TF with a similar structure due to similarities in the specific sequence. Examples are nuclear receptors with C4 zinc fingers for the superclass of Zinc coordinating DNA-binding domains or homeo domain factors for helix-turn-helix TFs (Wingender, Schoeps, and Dönitz 2013). Classes can be subcategorized into families and subfamilies, presenting level 3 and 4. The closer the motifs are between TFs, the closer they are grouped together. The name is often based on a well studied member of the family, for example Jun-related family for

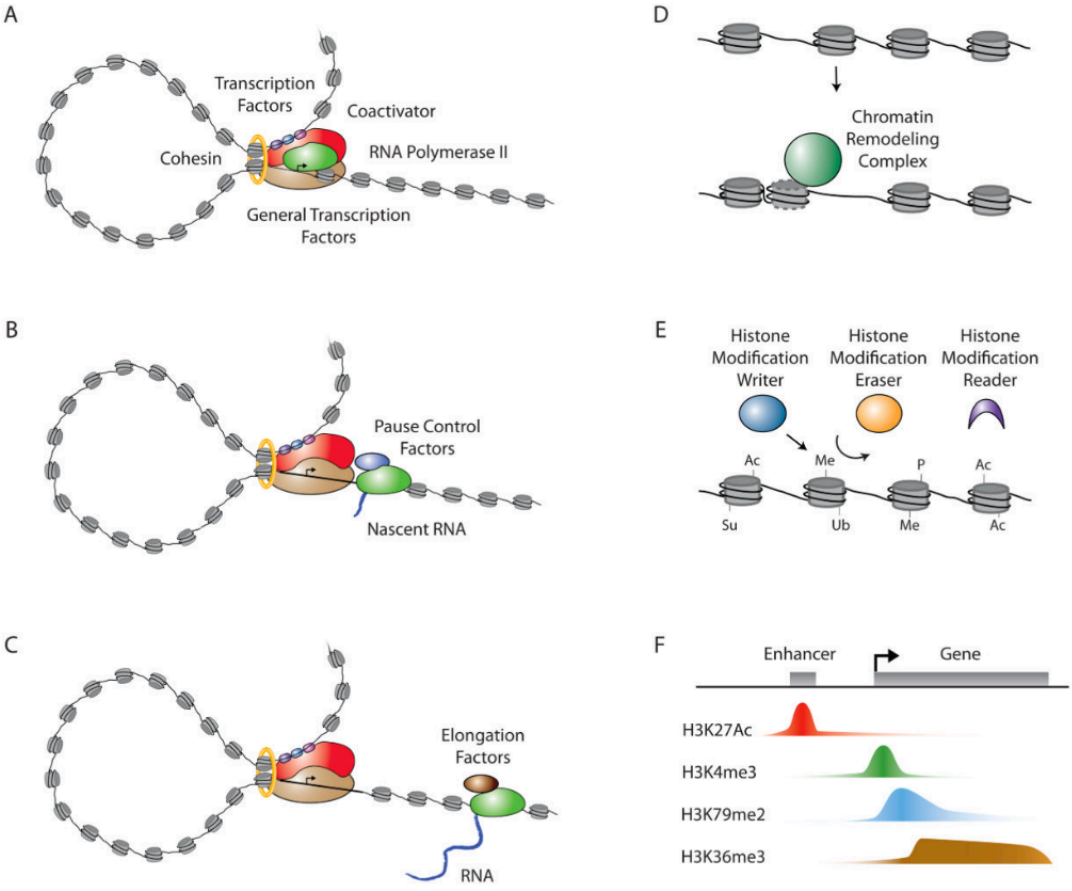


Figure 1.6.: Regulation of gene expression. (A) TF binding to regulatory elements to form the pre-initiation complex. DNA looping is stabilized by cohesin. (B) Transcription by RNA polymerase II is initiated and paused. (C) Elongation is enabled by elongation factors. (D) To allow transcription factors to bind to DNA sequences, chromatin remodeling by remodeling complexes is required. (E) Transcriptional regulation depends on histone modifications. Specific proteins can add, remove and bind thorugh these modifications. They are called writers, erasers and readers. (F) Specific histone modifications can be associated to transcribed genes. Histone H3 lysine 27 acetylation (H3K27Ac) at enhancers, histone H3 lysine 4 trimethylation (H3K4me3) at the transcription start site and histone H3 lysine 79 dimethylation (H3K79me2) and histone H3 lysine 36 trimethylation (H3K36me3) further downstream (T. I. Lee and Young 2013)

basic leuzin zipper factors (bZIP), a class of the basic domain superclass. Level 5 takes into account the gene product and is called genera (Wingender et al. 2018). Binding of helix-turn-helix and bZIP factors to DNA is illustrated in Figure 1.7 A. Only a limited sequence of basepairs is involved in the interaction between the DBD and DNA.

Not all commonly used family names follow the classification of Wingender et al. (2018). The activator protein-1 (AP-1) family of TFs represents bZIP class transcription factors including Jun, Fos, ATF and Maf families of TFs Wu et al. (2021).

A motif is most commonly represented as position weight matrix (PWM) assigning a weight to each base at the positions of a motif (Nishida, Frith, and Nakai 2009). Each element of a PWM is defined by the log likelihood ratio of a specific base being observed at the position of a motif. This ratio is estimated using data from known binding sites Nishida, Frith, and Nakai (2009). In early versions of JASPAR, the information was gathered through SELEX experiments and reviews while from 2009 on PWMs were generated as well from ChIP-seq experiments Portales-Casamar et al. (2010).

The graphical representation of a motif is the sequence logo that shows the probability of a base at a certain position by the height of the letter and is proportional to the binding energy Nishida, Frith, and Nakai (2009). Comparing motifs by sequencing logos can give an overview of the similarities between TFs from the same family (Figure 1.7 B). SPI1 and SPIB for example show a similar preference for the sequence GGAAGT at position 9-14 and 8-13, respectively.

1.2.2. Measuring gene expression levels

The identification of changes in gene regulation is an essential part of understanding mechanisms underlying cell differentiation. Changes in gene expression between conditions or time points are indicative for gene regulation.

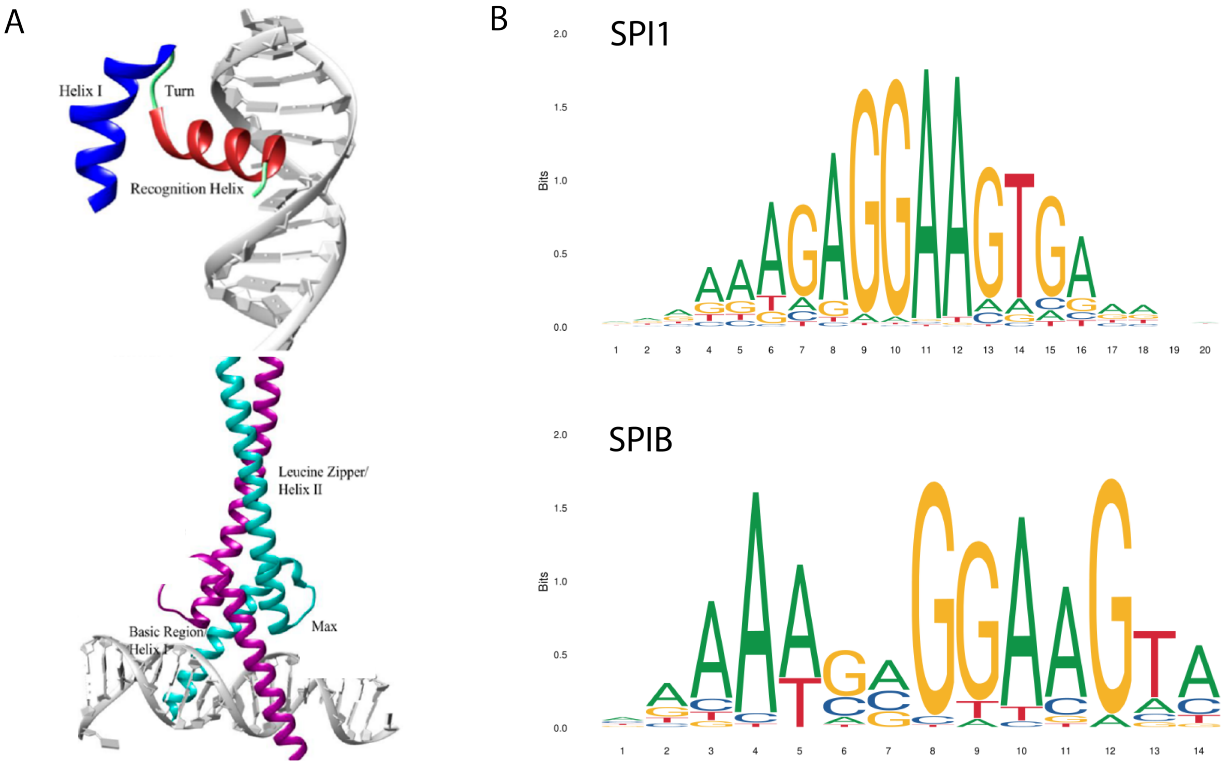


Figure 1.7.: Transcription factor DNA binding domains and sequence logos. (A) Examples of a helix-turn-helix DNA Binding Domains and a bZIP and the interaction with DNA (Yesudhas et al. 2017). (B) Sequence logos for SPI1 and SPIB taken from the JASPAR database. Both TFs share the same preference for the sequence GGAAGT.

In the past, DNA microarrays were commonly used to identify gene expression. According to the review by Zhao et al. (2014), it was still the preferred method in 2014 but was said to be replaced by RNA-seq soon (Zhao et al. 2014).

In 2008 Mortazavi et al. (2008) introduced RNA-seq, a high-throughput sequencing methods for the quantification of gene expression. In 2009 it was already described as more precise than microarrays (Fu et al. 2009). In RNA-seq, transcripts are converted to cDNA fragments. Sequencing adapters are attached to either one or both ends of a fragment. The fragments are sequenced to gain reads (Z. Wang, Gerstein, and Snyder 2009; Zhao et al. 2014; Weber 2015).

Due to the limitations in accuracy of expression measurements for low abundant transcripts in microarrays, the restriction to investigate only genes that have probes designed for and on the other side the high-throughput in RNA-seq, the later became invaluable for measuring gene expression levels and is now the method of choice (Zhao et al. 2014).

The processing of the obtained reads is done with bioinformatics tools that will be discussed in Section 1.3.1 for quantification of gene expression levels and comparison between different experimental conditions.

1.2.3. Identification of open chromatin regions and chromatin interaction

As described in Section 1.2, transcription factors regulate gene expression through interaction with regulatory elements. To understand the ongoing regulatory interactions and mechanisms in a cell, several high-throughput sequencing methods can be utilized.

Chromatin immunoprecipitation followed by sequencing (ChIP-seq) can detect interactions of transcriptional regulators on the DNA and profile histone modifications and nucleosomes (Johnson et al. 2007; Park 2009). In ChIP-seq, formaldehyde is used to cross-link transcription factors to DNA followed by sample fragmentation and immunoprecipitation

using an antibody specific to a protein of interest. Crosslinks are reversed afterwards and the DNA fragments are sequenced. In case the histone modifications and nucleosome positions are targeted for identification, fragmentation is done without cross linking (Park 2009).

DNAse I coupled with deep sequencing (DNAse-seq) and Assay for transposase-accessible chromatin using sequencing (ATAC-seq) aim to measure chromatin accessibility across the genome. The landscape of chromatin is highly dynamic and is an important indicator for regulatory events. DNAse-seq utilizes the enzyme DNAse I to cleave the DNA at open chromatin regions. Resulting fragments are sequenced. Regions with high read counts are detected as open chromatin (L. Song and Crawford 2010). During ATAC-seq, the prokaryotic transposase Tn5 integrates adapters into open chromatin regions through cleavage. As described before for DNAse-seq, the resulting fragments are sequenced and quantified in the same manner (Figure 1.8) (Buenrostro et al. 2013; Klemm, Shipony, and Greenleaf 2019). ATAC-seq requires fewer cells and the sample preparation is faster in comparison to DNAse-seq (Luo, Gribskov, and Wang 2022).

High-throughput chromosome conformation capture (HI-C) is used to identify the 3D structure of the genome generating genome-wide interaction maps (Belton et al. 2012; Luo, Gribskov, and Wang 2022). In Hi-C, after the DNA is cross-linked, it is digested into small fragments and after labeling the fragments are subsequently ligated together. Ligation junctions are marked through this method and after sequencing, conclusions about interactions between DNA fragments can be made. In comparison to other methods it can capture even long range interactions between genes and their regulatory elements (Belton et al. 2012).

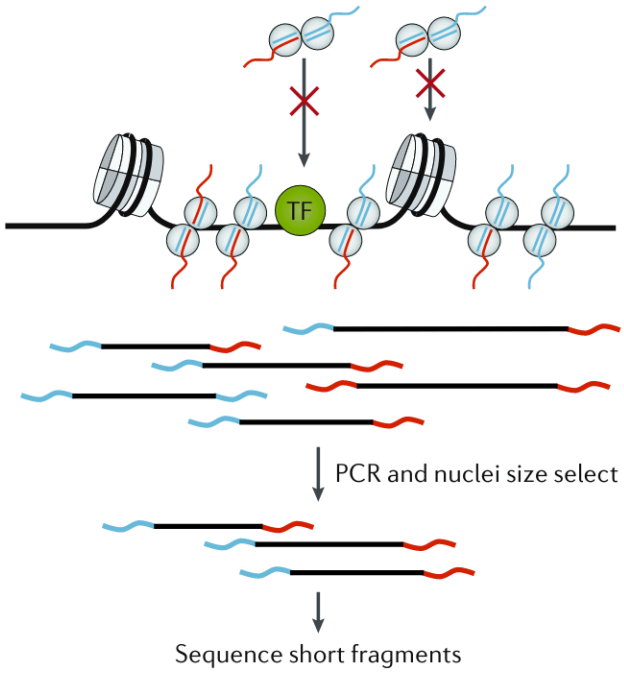


Figure 1.8.: Principle of transposase (Tn5) cleavage and adapter insertion during ATAC-seq. Tn5 is able cleave and simultaneously insert adapters in accessible regions. Fragments are selected for size. Short fragments are sequenced (Klemm, Shipony, and Greenleaf 2019).

1.3. Bioinformatics for cell differentiation and gene regulatory networks

Next generation sequencing (NGS) methods, such as RNA-seq and ATAC-seq (described in Section 1.2.2 and Section 1.2.3) are high-throughput methods and generate gigabytes of data that needs to be processed and analyzed. Many bioinformatics tools are developed for NGS and most analyses follow a similar workflow, starting with the alignment to a reference genome (Langmead and Salzberg 2012) after assessing the overall quality with tools like FastQC (Andrews 2010). Sequencing reads are aligned to a reference genome to identify their position in the genome with a quality score of the alignment. Depending on the sequencing data, different tools are needed. BWA and bowtie are common for the alignment of ChIP-seq and ATAC-seq, while STAR was specifically developed for RNA-seq (H. Li and Durbin 2010; Langmead and Salzberg 2012; Dobin et al. 2012). Quantification steps follow after the alignment.

1.3.1. Quantification of gene expression and chromatin accessibility

Quantification of gene expression combines the aligned reads and a reference annotation. By counting the number of reads for a specific genomic feature, such as a genes or exons, a count matrix is generated. HTSeq-count, Cufflinks and featureCounts are well known methods for quantification after alignment (Jin, Wan, and Liu 2017; Liao, Smyth, and Shi 2014; Liao, Smyth, and Shi 2019). Low quality reads can be excluded by the quality score from the alignment (Liao, Smyth, and Shi 2014). Gene or feature counts are not comparable between samples or even genes within the same sample and require normalization. Three normalization units are known: reads Per Kilobase of transcript per Million reads mapped (RPKM), Fragments Per Kilobase of transcript per Million fragments mapped (FPKM) and Transcripts Per Million (TPM) (Zhao, Ye, and Stanton 2020; Jin, Wan, and Liu 2017). Since the RPKM measure has a statistical bias, TPM is recommend as

the measure for transcript abundance with RNA-seq data (Wagner, Kin, and Lynch 2012) and is calculated as

$$TPM = 10^6 * \frac{reads\ mapped\ to\ transcript/transcript\ length}{Sum(reads\ mapped\ to\ transcript/transcript\ length)}$$

Considering the dynamics of gene regulation, time-course experiments are well suited to capture temporal changes. For comparison between different time points and samples differential expression analysis is applied to identified genes that have a statistically different expression between time points (Weber 2015). These genes are called differentially expressed genes (DEG) and the established method for identification is the use of *DESeq2*, an R package for differential expression analysis (Love, Huber, and Anders 2014). In addition, genes can be clustered into groups of co-expressed genes, classifying genes with the same expression into the same group for downstream analyses like gene ontology (GO) enrichment. In case of time-course experiments, the clustering should take the temporal dependencies into account. Distance based clustering methods like k-means clustering do not consider temporal dependencies, instead model-based clustering is recommended. Schliep et al. proposed Hidden Markov models (HMM) in 2005 for microarray data (Schliep et al. 2005). In 2007 Ernst et al. used Input-Output Hidden Markov models (IOHMM) to extend the use of HMMs for modeling gene expression data to combine it with static motif data (Ernst et al. 2007). This work built the first instance of the Dynamic Regulatory Events Miner (DREM).

Chromatin accessibility data can be quantified similar to gene expression data. Instead of genomic features like genes, peaks are used as area of interest. Peaks are regions in the genomes, where many reads are aligned, suggesting that the chromatin is accessible there. The most commonly used method for peak calling, the process of identifying the peak regions, is MACS2. It is not the best choice for peak calling on ATAC-seq data, since it was originally developed for ChIP-seq and extended to handle DNAse-seq data(Y.

Zhang et al. 2008). Genrich (https://github.com/jsh58/Genrich) offers a dedicated ATAC-seq mode and offers filtering according to mapping quality, exclusion of mitochondrial RNA and user defined genomic regions, for example blacklist regions, which are genomic regions that would provide biased or erroneous alignment (Amemiya, Kundaje, and Boyle 2019). To account for the Tn5 enzyme specific cleavage, Genrich creates intervals that are centered on the Tn5 cut sites, making it the preferred for ATAC-seq analysis. Peak counts and differentially expressed peaks can be computed in the same way as described above for RNA-seq. In addition to the open chromatin regions, transcription factors binding at regulatory elements can be inferred from ATAC-seq by identifying footprints. Footprints are drops in reads at a peak due to transcription factors binding during the cleavage (Z. Li et al. 2019). TFBS in footprints can be identified using known motifs of TFs. Several databases and motif collections exist. JASPAR focuses on having well curated PWMs, HOCOMOCO is species focused and HOMER is all-inclusive (Ambrosini et al. 2020). The comparison between motifs from different data bases by Ambrosini et al. (2020) highlights the similar motif problem of TF subfamilies. Members from the ETS-subfamily have indistinguishable DNA binding sequences assigned (Ambrosini et al. 2020).

1.3.2. Integration of chromatin accessibility and gene expression to identify gene regulatory networks

Transcriptional regulation and changes in chromatin accessibility to facilitate gene regulation are highly dynamic (Hager, McNally, and Misteli 2009). These dynamic processes are often visualized as Gene Regulatory Networks (GRN). Approaches to build GRNs comprise logical models, such as boolean networks, probabilistic boolean networks, inference of particular network properties and continues models like continuous linear models, models of TF activity, ordinary differential equations (ODEs) or regulated flux balance analysis (Karlebach and Shamir 2008).

ARACNe was an early and well known tool to infer GRNs and was based on mutual in-

formation theory. It was originally build for static microarray data and later adjusted and improved for RNA-seq (Margolin et al. 2006; Lachmann et al. 2016).

Static analysis of gene expression data can not capture the dynamic nature and would only provide a snapshot of a current state. While one snapshot would not suffice, by generating time series data, this series of snapshots could give a more detailed picture of cellular processes (Duren et al. 2020; Bar-Joseph, Gitter, and Simon 2012). Many methods rely on time series gene expression data for the inference of GRNs. *ppcor* and *LEAP* are correlation based methods to identify GRNs from gene expression data, *GENIE3* and *dynGENIE3* infer GRNs based in regression (Kim 2015; Specht and Li 2017; Huynh-Thu and Geurts 2018). Monocle3 takes single cell data and infers pseudotime through unsupervised learning (Cao et al. 2018). All of these methods focus only on gene expression.

Given the essential role of epigenetics and the chromatin state during gene regulation during differentiation, methods to identify gene regulatory networks should not only include gene expression but especially chromatin accessibility information (L. Liu et al. 2019; Luo, Gribskov, and Wang 2022).

The Integrated System for Motif Actitivity Response Analysis (ISMARA) provides an online tool that requires either gene expression (RNA-seq/ micro-array) or chromatin state (ChIP-seq) data to infer GRNs but not a combination of both (Balwierz et al. 2014). Jung and del Sol proposed Moni, a computational method that integrates epigenetic, transcriptomic and protein-protein interactions, but the method does not consider the time aspect of gene regulation (Jung and Del Sol 2020).

From the development of ATAC-seq in 2013 the interest in this method as well as combining ATAC-seq with RNA-seq increased strongly. A bibliometric review by Luo, Gribskov and Wang identified almost over 100 articles in PubMed using both ATAC-seq and RNA-seq in 2021 (Luo, Gribskov, and Wang 2022). DREM was one of the earliest methods to incorporated static TF-gene interactions ti. build regulatory networks, even before ATAC-seq was even developed (Ernst et al. 2007).

In 2018, Gerard et al introduced EPIC-DREM, a machine learning pipeline that combine epigenomics and transciptomics time-series data based on paired RNA-seq and ChIP-seq samples. Transcriptional regulators were identified through time-point specific TF-gene scores by associating them with time-point specific changes in gene expression (Gérard et al. 2018; Schmidt et al. 2019). There is no complete implementation of the previously mentioned EPIC-DREM pipeline, making it difficult for users to perform the same kind of analysis on their own time-series epigenomics and transcriptomics data. The concept of EPIC-DREM including preprocessing is visualized in Figure 1.9. Given paired time series epigenomics and transcriptomics data from a differentiation experiment, the data is processed according to commonly used bioinformatics workflows as described in Section 1.3.1. EPIC-DREM describes process of computing TF-gene links through TEPIC in footprinting data and calculating a threshold to decide which TF-gene links have enough statistical relevance to be included as input for DREM (Schmidt et al. 2019). The genes are clustered according to gene expression profiles using IOHMM, to identify bifurcation events. Taking the information from the TF-gene links, DREM explains these bifurcation events by assigning TFs to the newly identified clusters after the split. Each TF is assigned according to a Split Score, which is comparable to a p-value. Based on the TF-gene links splits can additionally be added or removed from a model (Ernst et al. 2007).

While the joined analysis of gene expression and chromatin accessibility data increases, many methods focus on static data (Berest et al. 2019; K. Zhang et al. 2019). Ludwig et al. introduced a pseudo time course by identifying intermediate stages during erythropoiesis by FACS sorting, but they lack a combined analysis step (Ludwig et al. 2019). Ramirez et al. provide a time-series of paired ATAC-seq and RNA-seq samples but the analysis is separated with a manual overlap of the results (Ramirez et al. 2017).

In 2020, Duren et al. published their method PECA2/TimeReg that integrates time series ATAC-seq and RNA-seq data including the inference of transcription factor modules and GO enrichment analysis of target genes. The main tool was implemented in Matlab while being called in shell scripts (Duren et al. 2017, 2020). Duren et al. infer static time-point

specific connections between TFs and TGs using a cis-regulation score (CRS) that is based on the chromatin accessibility data combined with the expression data within their tool called PECA2 (Duren et al. 2017). The time component is added with TimeReg by comparing scores of differentially upregulated genes between time points and select the TFs that also have a significant higher CRS as possible main regulators to connect the time-point specific gene regulatory networks (Duren et al. 2020).

With the increased application of single-cell, there is a high demand in methods combining single-cell ATAC-seq and single-cell RNA-seq but there are still limited options. *FigR* identifies activators and repressors, *sciCAN* address the heterogeneity from single-cell data with the integration through unsupervised learning and *scDART* focusing on integrating unmatched single-cell ATAC-seq and RNA-seq (Kartha et al. 2022; Z. Zhang, Yang, and Zhang 2022; Xu, Begoli, and McCord 2022). Despite the increased generation of single-cell data, bulk ATAC-seq and RNA-seq data is still generated (Q. Song et al. 2022; R. Li, Grimm, and Wade 2021; Perrin et al. 2021; Kiani et al. 2022).

Given the large amount of data that needs to be processed during multi-omics analysis, workflow management systems like Snakemake became invaluable (Koster and Rahmann 2012; Larsonneur et al. 2018). They enable the combination of necessary pre-processing tools for high-throughput sequencing data and allow users with limited computational skills to perform bioinformatics analysis. In combination with environment management systems like Mamba (https://mamba.readthedocs.io/en/latest/index.html) a reproducible and closed setup is guaranteed. Snakemake, Nextflow and BigDataScript provide an easy-to-develop and easy-to-use framework, although other pipeline frameworks have the advantage in performance (Leipzig 2016). Among comparable workflow management systems, Snakemake gave the best overall performance in a direct comparison (Larsonneur et al. 2018). An effective but time intensive selection of a workflow management systems was described by Jackson et al. and resulted in Nextflow as their choice with the emphasize on positive experience with Snakemake (Jackson, Kavoussanakis, and Wallace 2021). In addition, the choice of Snakemake over others was based on the experience at the Lux-

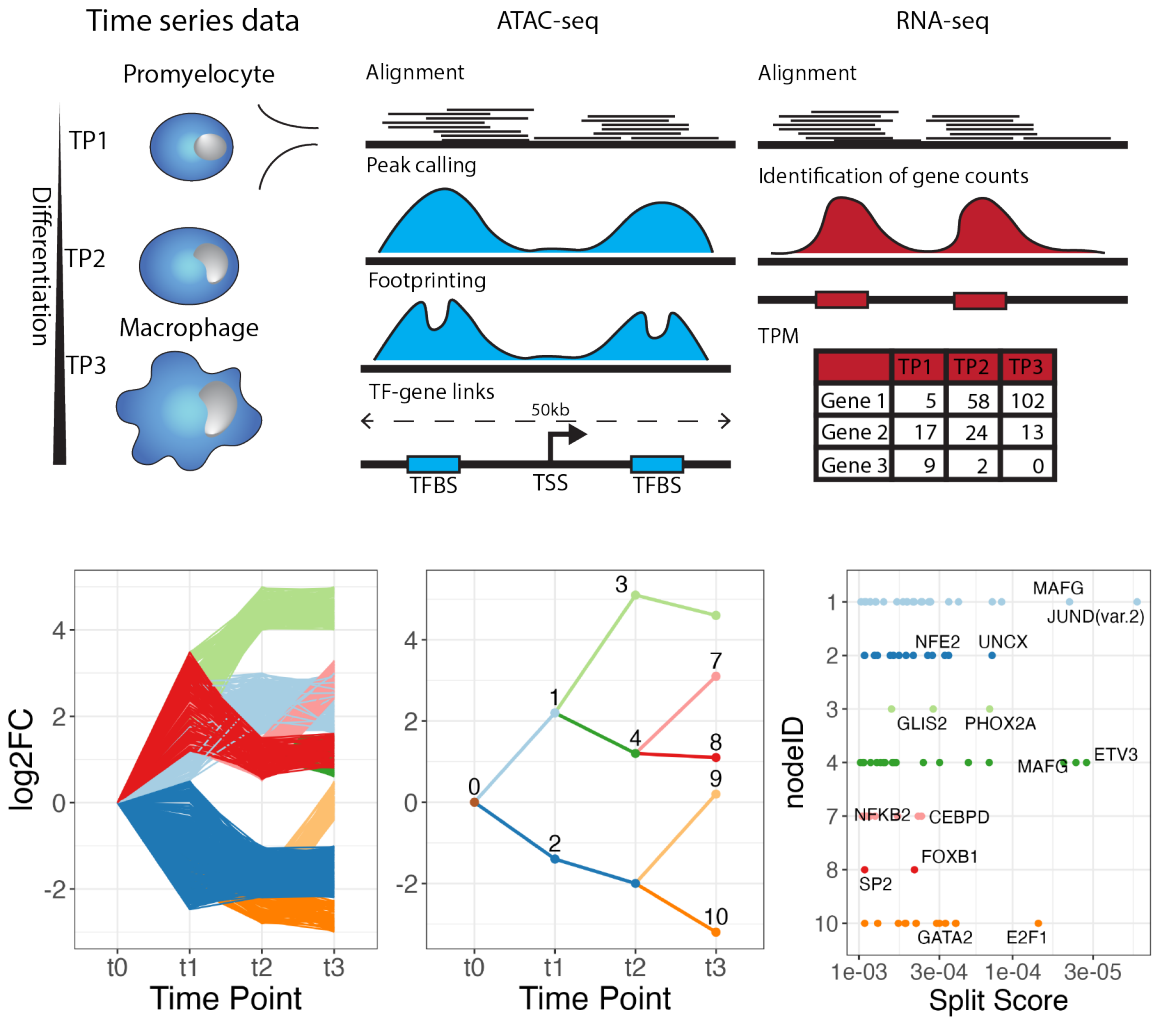


Figure 1.9.: Simplified example of EPIC-DREM workflow with preprocessing. EPIC-DREM requires paired time series transcriptomics and epigenetics data from differentiation experiments and infers transcription factor-gene links from the epigenetic data. Combined with the identified gene counts from transcriptomics data, a bifurcation network is built according to the expression profiles. Transcription factors are assigned as key regulators inferred from the TF-gene links data to explain the expression changes resulting in bifurcation events.

embourg Centre for Systems Biomedicine with Snakemake resulting in tools like *IMP* and *binny* (Narayanasamy et al. 2016; Hickl et al. 2021).

2. Scope and Aims of Thesis

The inference of gene regulatory networks can be vastly improved by the integration of chromatin occupancy with gene expression across several time points. Studies in the field measure gene expression and open chromatin and combine them manually, analyzing each data separately and searching for overlaps, which leads to cherry-picking of results on one hand and missing relevant signals on the other.

Amongst others, Gerard et al demonstrated the success with machine learning framework EPIC-DREM but a complete workflow was not their primary aim. TimeReg by Duren et al is the only pipeline that promises similar generality and inference but differs in many aspects, including choices of tools and databases in the pipeline, integration and the algorithm eventually used for inference of transcription factors.

The aims for this thesis are

- an implementation of EPIC-DREM using the workflow management system Snakemake and the environment management system mamba, including ATAC-seq data as well as H3K27ac ChIP-seq signals. The new DREMflow pipeline should extend EPIC-DREM with additional normalization, filtering and integration of omics data to optimize the performance and the results and provide an automated selection step of regulators in time series differentiation data. Comprehensive and automated visualization and downstream analysis are to be added to aid the user in a selection of candidates for validation in the lab. Current data management and reproducible research consideration are to be supported.

- The midbrain dopaminergic neuron data set by Ramos et al. in the context of the PARK-QC Doctoral Training Unit was the primary use case for this work. DREMflow is to be applied on this data set primarily.
- Application of DREMflow to several relevant data sets of cell differentiation to demonstrate its applicability in different biological contexts. The predictive ability is to be evaluated on suitable data sets from the literature, serving as gold standards.
- Expansion of the scope of the DREMflow to deal with non-human organisms.
- Comparison to the other computational method integrating time series chromatin accessibility data and gene expression data.
- Demonstration or discussion of the adaptability of DREMflow on the single nuclei ATAC-seq data and a pseudo time series data and other methods with individual cell identity resolution.

3. Materials and Methods

3.1. Application of EPIC-DREM in differentiation of dopaminergic neurons

EPIC-DREM as originally described by Gérard et al. (2018) was applied to midbrain dopaminergic neuron differentiation data from Ramos et al. (2023). The RNA sequences were preprocessed with the paleomix pipeline by Dr. Aurélien Ginolhac from the Bioinformatic Core of the Department of Life Science and Medicine, University of Luxembourg. This included trimming sequencing adapters with AdapterRemoval, filtering of ribosomal RNA with SortMeRNA and the alignment to a reference genome using the Spliced Transcripts Alignment to a Reference (STAR) (Dobin et al. 2012). The identification of gene counts was performed by Dr. Borja Gomez Ramos from the Luxembourg Center of Systems Biomedicine and the Department of Life Science and Medicine at the University of Luxembourg with featureCounts of the RSubreader package (Liao, Smyth, and Shi 2019). Genome version GRCh38 patch 12 was used as reference genome and Gencode human release 31 as gene annotation for the RNA-seq preprocessing.

The ATAC-seq data was preprocessed in the same way as RNA-seq with the paleomix pipeline by Dr. Aurélien Ginolhac. Afterwards I used Genrich to perform peak calling over replicates, excluding low quality reads (<Q30) and mRNA in the process. Footprints were identified using HINT-ATAC from the Regulatory Genomics Toolbox (RGT) Z. Li et al. (2023) and TF-gene links were inferred with TEPIC (Schmidt et al. 2019). EPIC-DREM

was originally applied to ChIP-seq data. When I first applied TEPIC to ATAC-seq, it resulted in an overestimation of TF-gene links, the p-value threshold had to be set to 0.01 instead of 0.05 as recommended by the developers and after trying different p-values. Due to the nature of TEPIC being only available for Python 2.7 at the time of the first analysis, this step was performed in a Snakemake rule instead of a script, since Python 2.7 was deprecated.

TF-gene links and gene counts converted to transcripts per kilobase million (TPM) were used to retrieve important regulators in human mDAN differentiation with the DREM (Schulz et al. 2012). DREM was run in batch mode. The resulting model together with the TF-gene links and gene counts were used in the DREM graphical user interface to retrieve the visual output of the calculated model. This step was later circumvented using the interactive Dynamics Regulatory Events Miner (iDREM) (Ding et al. 2018) to allow a continues command line execution.

Model visualizations and gene ontology (GO) enrichment of biological processes on target genes were performed in R with *ggplot* and the *clusterProfiler* package(Wu et al. 2021; Yu et al. 2012). GO enrichment was also performed on the significant TFs for each node. Details of the visualization can be found in the DREMflow source code.

After the implementation of DREMflow the analysis was repeated with the same settings as the original analysis.

3.2. Implementation of DREMflow

3.2.1. Setup and installations

DREMflow is implemented in Snakemake (Koster and Rahmann 2012) and utilizes mamba (https://mamba.readthedocs.io/en/latest/index.html) to install all tools in closed environments with their dependencies. In Snakemake, each analysis step is written as a *rule*.

The connection between rules is inferred via the defined input and output of each rule. It is possible to have rules without input or output, e.g. the first and last rule. Snakemake will execute all rules necessary to reach the final output. Defining the input and output with *wildcards* allows Snakemake to execute the same rule for several input files, for example for all time points. Snakemake infers file names to replace the wildcards with. In case of changed settings or input files, only affected rules will be re-executed. Settings are defined in a configuration file. A Snakemake pipeline does not necessarily have to be executed from the first rule. If input for intermediate rules is provided that is sufficient to acquire the final results, the pipeline can start from this intermediate rule instead.

The reference genome and annotation are installed via a Snakemake wrapper, a predefined rule from the Snakemake wrapper repository (https://snakemake-wrappers.readthedocs.io/en/stable/index.html), according to the specification of version and build by the user in the configuration settings. The annotation is limited to only protein coding genes. Indexes for mapping to the reference genome are generated by preparation rules. Tools that are not available in mamba, such as TEPIC and Regulatory Genomics Toolbox, are installed individually by setup rules. To comply with data management guidelines, symlinks to the raw data are set and serve as input to DREMflow. After installation of Snakemake and mamba, the user has to provide the time points, replicates, reads and the path to the raw reads for ATAC-seq and RNA-seq separately in a spreadsheet. The execution of the pipeline requires only a one line command.

3.2.2. ATAC-seq processing

Quality control on the reads is performed with FastQC (Andrews 2010). ATAC-seq sample processing starts with mapping to the reference genome with BWA (H. Li and Durbin 2010). Afterwards peak calling over replicates is performed with Genrich (https://github.com/jsh58/Genrich). By default the parameters are set to `-m 30 -j -a 500 -g 15 -l 15 -d 50`. The peaks are merged with bedtools (Quinlan and Hall 2010) over replicates to retrieve

counts for all samples in the open chromatin regions. Using DESeq2, Principle Component Analysis (PCA), a dimensionality reduction method, is performed and differential peaks are computed with these counts (Love, Huber, and Anders 2014). Footprints are identified with the Regulatory Genomics Toolbox using HINT-ATAC Z. Li et al. (2023). Since the peaks were called over replicates, the aligned files were merged with samtools to create the required input file (Danecek et al. 2021). The identified footprints serve as candidate regions for TEPIC to identify TF-gene links. In a 50kb window (`-w 50000`) around the transcription start site (TSS), TEPIC identifies motifs in the specified candidate region to infer if TFs are linked to the respective TSS (Schmidt et al. 2019). The chosen p-value cut-off was set to 0.01 after having overestimation with the default p-value cut-off. This can be done for either individual TFs, the *combined* set of known motifs, or a set of TF clusters according to the similarity in the motif. There are 96 TF clusters overall grouping motifs according to their TF family and similarity. Each cluster is represented by a combined motif that is similar to all motifs included in the cluster. If the combined set is used, acquired TF-gene links are filtered using the RNA-seq data. Only links of TFs that are expressed according to the expression data are kept as input for the Dynamics Regulatory Events Miner (DREM) (Schulz et al. 2012). To assess the activity of TFs, motif matching and differential footprinting is done with RGT. Default settings of all tools in DREMflow can be found in the appendix. Peaks are quantified using the featureCount function of the RSubreader package (Liao, Smyth, and Shi 2019) and differential peaks are identified with DESeq2 (Love, Huber, and Anders 2014).

3.2.3. RNA-seq processing

RNA-seq reads are mapped to the reference genome with STAR (Dobin et al. 2012) with the DREMflow default parameters

```
--outSAMunmapped Within
--outSAMtype "BAM SortedByCoordinate"
```

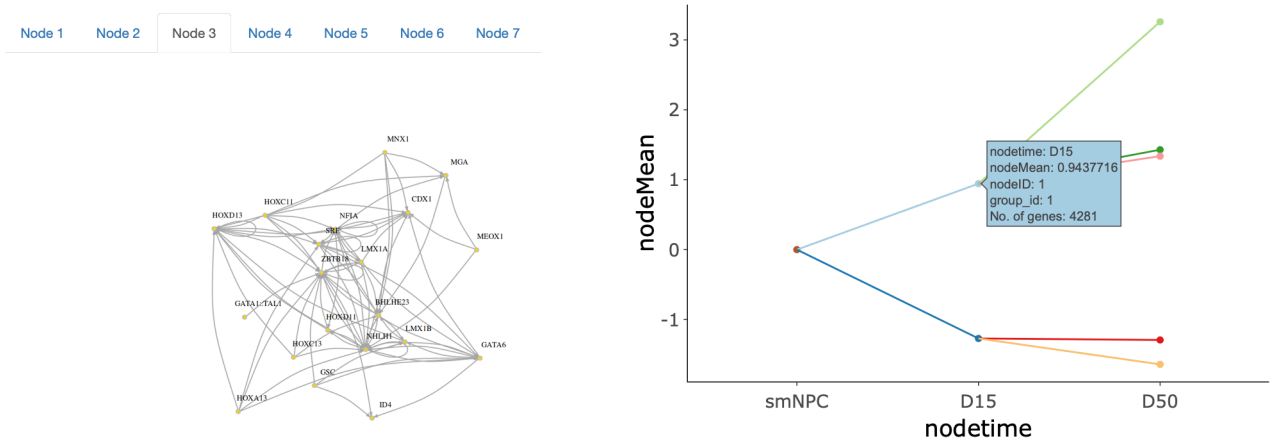
```
--twopassMode Basic
--limitOutSJcollapsed 1000000
--limitSjdbInsertNsj 1000000
--outFilterMultimapNmax 100
--outFilterMismatchNmax 33
--outFilterMismatchNoverLmax 0.3
--seedSearchStartLmax 14
--alignSJoverhangMin 15
--alignEndsType Local
--outFilterMatchNminOverLread 0
--outFilterScoreMinOverLread 0.3
--winAnchorMultimapNmax 50
--alignSJDBoverhangMin 3
--quantMode GeneCounts
--outSAMstrandField intronMotif
--outFilterType BySJout.
```

The resulting mapped reads are quantified as gene counts for protein coding genes with the featureCount function of the RSubreader package with a minimum mapping quality of 30 (`minMQS = 30`) from the GTF annotation file(Liao, Smyth, and Shi 2019). The PCA and sample-to-sample distance are computed on the normalized gene counts with DESeq2. After normalization TPM are calculated and filtered for expression (Love, Huber, and Anders 2014). By default genes are included if they have a TPM>1 in at least one sample. This threshold can be set by the user.

3.2.4. Derivation of TF-TF networks

The highest 20 ranked TF are selected to build the TF-TF networks for each split node based on the TF-gene links identified with TEPIC. The TF-TF network nodes are TFs that

regulate any other high ranked TF on the same split node according to the affinity data. The TF-TF network is based on the ranking and the inferred connections from TEPIC, thus not significant TFs according to the Split Score are included. The networks are visualized with the *igraph* package and individual networks can be inspected by browsing through the tabs for each node Figure 3.1.



(a) TF-TF network can be inspected by browsing through tabs.

(b) The interactive elements in the split node network give information about the mean expression in the node and number of target genes going through the node.

Figure 3.1.: Example for TF-TF networks and the use of tabs to browse through the data and an interactive model overview in the HTML results.

3.2.5. Model description, target gene clusters and transcription factors

The split node network is computed by the interactive Dynamics Regulatory Events Miner (iDREM)(Ding et al. 2018). As a side product to the regular DREM output, iDREM stores the model as JSON. This allows the circumvention of the graphical user interface and a fluent computational processing of the network model data ensuring reproducibility. The downstream analysis includes visualization, selection of top regulators in the network and gene ontology enrichment. The overall output from DREMflow shows the results from PCA and the difference between samples for ATAC-seq and RNA-seq data. A table lists the general statistics such as peak counts, differentially expressed peaks and DEG. The

model overview is split in three parts. The first part shows the average expression profile of all target genes going through a node shown at each time point. The later the time points, the more nodes exist given the splits occurring, which create subsets of the gene sets assigned to the previous nodes. The second part shows the overall split node network. The initial node and nodes after a split are labeled. The nodes after a split have TFs assigned according to a Split Score. Those TFs are assigned as regulators to explain the expression change, causing the split in one direction or the other. The third part shows the significant TF according to their Split Score for each node that has TFs assigned. Not all TFs are assigned as significant and not all nodes after a split need to have significant TF. This depends on the data, especially the time point specific TF-gene links. To provide more details, an interactive version of the split node network and the significant TFs are provided for the user (Figure 3.1).

3.2.5.1. Selection of top regulators

The term *top regulators* refers to TFs that are assumed to play an important role according to the computed network. They are selected according to their presence in the list of significant TF at each node after a split. A TF is considered as top regulators if it appears in more than one node among the first n significant TFs. The default of n is 10. Data sets with few time points might require an adjustment. The cutoff for significance is a Split Score lower than 0.001.

A heatmap providing the number of occurrence at each time point as well as the total number of occurrences in the model is computed for the top regulators. In addition, the expression change and the estimated target genes are visualized for the top regulators over all time points.

A table with the number of occurrences and the best ranking at each time point gives an overview at which time the top regulators were inferred as highly ranked in the system.

In addition, the two TF with the lowest Split Score are added in the list of selected TF for evaluation of activity differences between time points.

3.2.5.2. Target gene clusters and Gene Ontology enrichment

To better understand the trajectories of the nodes as shown in the model overview, the expression profile of all genes in a node are plotted node wise, which also provides a visual estimation of the number of target genes assigned to a node. These clusters are filtered for DEG and GO enrichment of biological processes is performed on the remaining gene set at each node with the clusterProfiler package (Wu et al. 2021; Yu et al. 2012).

After the selection of top regulators, their specific target genes are investigated as well with GO enrichment.

The first 20 terms if available for each cluster and top regulator are visualized. Tabs for individual nodes and TF in the HTML report allow an organized inspection of all results (as shown for TF-TF networks Figure 3.1).

3.2.5.3. Additional information

After the GO enrichment and the identification of TF-TF networks, the results provide an overview of the activity of TFBS for selected top regulators. The combined openness score and protection score for TFBS of top regulator are combined, normalized and visualized for each time point, resulting in a heatmap that shows time point specific activity.

For each major rule in the pipeline benchmarking files are written by Snakemake containing information about runtime and memory load. The cumulative runtimes are plotted in sequential order for the main rules together with the required maximum memory providing the user and overview of computational requirements and resources used.

In addition, intermediate results are saved in an serialized R data object (RDS), containing the DREM model results, list of selected top regulators, target gene list, quality control

measures such as number of differentially expressed peaks and genes, number of footprints and the benchmark information.

This information can be built on in additional analyses and is a major improvement as no additional efforts for navigating complex objects or files is required.

3.2.6. Computational requirements and comparison

Required tools and dependencies are installed via mamba. The environments created through mamba at the first execution of the pipeline require a total disk usage of 70GB and create approximately 350000 files.

The user needs to install Snakemake, conda and mamba to run DREMflow. The recommendation is the installation of miniconda and the creation of a Snakemake specific environment just containing Snakemake, conda and mamba.

The CPU and memory requirements depend on the processing and analysis sets. Generally, given the large amount of data, an execution on a high performance cluster is necessary. The setup of DREMflow is adjusted for a slurm workload manager, a commonly used queuing system on high performance computing clusters (Yoo, Jette, and Grondona 2003).

For the execution on a cluster, Snakemake requires a cluster profile. This is set up together with the rules and the environments for each rule. The cluster profile contains the configuration to submit jobs to the cluster via a slurm workload manager and the individual settings for each rule. The default is set to 1:00 hour run time with 1 core. For rules requiring more resources, this default values are replaced with specific values. The memory depends on the number of cores for each rule.

The University of Luxembourg High Performance Computing (ULHPC) provides two Supercomputers, Iris and Aion. Iris consists of overall 196 compute nodes, providing 5824 compute cores with total of 53 TB RAM. The peak performance is at approximately 1.072

Peta Floating Point Operations per Second (PetaFLOP/s). In addition, Iris features big memory nodes that were only needed during the implementation phase of EPIC-DREM (Section 3.1). The regular compute nodes have 28 cores and 128GB of RAM. DREMflow was developed and tested on regular Iris compute nodes.

Aion is the latest Supercomputer of the ULHPC with 318 compute nodes, providing 40704 compute cores with a total of 81TB RAM. The peak performance was estimated at 1.7 PetaFLOP/s. In contrast to Iris nodes' that have 4GB RAM per core, Aion nodes have only 2GB RAM per core. Since all data sets were re-analyzed on Aion, the cluster specific settings had to be adjust accounting for this difference in RAM per core.

The maximum memory requirement to execute the pipeline is 75 GB of RAM to calculate the TF-gene links. The maximum number of cores required in the default setting for parallel execution is 12. Considering the memory requirements, up to 28 cores are needed for TEPIC.

The data for the comparison of computational resources is taken from the saved benchmark information provided as output by DREMflow.

3.3. Application of DREMflow to differentiation data

DREMflow requires paired ATAC-seq and RNA-seq data with at least two time points as input data. There are several publicly available data sets meeting the requirement (Table 3.1).

Data set	Time points	Cell type	Notes

Table 3.1.: Publicly available data sets meeting the requirements for DREMflow.

Data set	Time points	Cell type	Notes
Ramos et al. (2023)	4	Human mDAN	Non-dopaminergic neuron cell mix available
Ramirez et al. (2017)	8	Human macrophages, neutrophils and monocytes	Monocyte-derived macrophages with five time points available
Duren et al. (2020)	5	Mix of mouse neuronal cells	No targeted differentiation into one specific cell type
Ludwig et al. (2019)	8	Erythrocytes and progenitors	Pseudo time points through FACS sorting
Hor et al. (2019)	5	Mouse cortex cell mix	Mice were sacrificed after periods of sleep deprivation. No replicates
Xiang et al. (2017)	3	Human cortical organoid and medial ganglionic eminence organoid	Unclear replicates with two versions for replicate 2 taken seven days later than replicate 1
van der Raadt et al. (2019)	3	Human fibroblasts, iPSCs and iNeurons	Overexpression of ONECUT TFs and Bclxl. Two replicates

Results of DREMflow applied to the first four data sets are shown in the following chapters. Ramirez et al. provided three cell lines and eight time points each of well studied cell types of human hemopoiesis, making it a valuable data set for building and evaluating the pipeline (Ramirez et al. 2017). EPIC-DREM was first applied to Ramos et al. two cell lines (see Section 3.1), therefore DREMflow was tested on the same data set to demonstrate reproducibility and robustness of the pipeline (Ramos et al. 2023). The data set of Duren et al. was selected to compare DREMflow to TimeReg, a competing method developed by Duren et al. and to add a non-human data set to analyze (Duren et al. 2017, 2020). The erythropoiesis data set was chosen to demonstrate how an extension of DREMflow to single cell data could work using a pseudotime series (Ludwig et al. 2019). The data from Xiang et al. and van der Raadt et al. were used in the early developmental phase of DREMflow to build the pipeline (Xiang et al. 2017; van der Raadt et al. 2019). These preliminary results are not shown. All data sets are publicly available.

3.3.1. Differentiation into human midbrain dopaminergic neurons

Human induced pluripotent stem cells (hiPSCs) were differentiated towards mDANs. Through flow-cytometry activated cell sorting (FACS) mDANs were purified from the heterogeneous neuronal cell mix. The mDAN time course consists of four time points, namely small molecule neuronal stem cells (smNPCs), day 15, 30 and 50. The time course for the heterogeneous neuronal cell mix derived from the hiPSCs, from here on referred to as non-mDANs, only consisted of three time points, smNPC, day 15 and 50. All samples were provided in triplicates for RNA-seq and ATAC-seq (Ramos et al. 2023). The data is available at https://egaarchive.org/, under the accession number EGAD00001009288

The data was analyzed multiple times in DREMflow with different versions of annotations, tools and settings in addition to the analysis with scripts as described in Section 3.1. Initially, DREMflow settings were strongly adjusted to the EPIC-DREM analysis settings (Sec-

tion 3.1). Annotation Gencode version 29 and PWMs from 2018 (version 2.1, JASPAR) were used for TEPIC. Second, the default DREMflow settings were used with Ensemble annotation release 108 and the newest combined set of PWMs (version 3.0) resulting in 708 TFs. A third time the analysis was rerun with the clustered PWMs from 2022 (version 3.0). A fourth time DREMflow was run with default settings and the 2018 PWMs (version 2.1, JASPAR), but the newest Ensembl annotation release 108 was limited to contain only protein coding genes.

3.3.2. Retinoic acid-induced mouse embryonic stem cell differentiation

Duren et al. (2017) induced differentiation of mouse ESCs with retinoic acid (RA) treatment. Samples were taken at days 0, 2, 4, 10 and 20 (d0. d2, d4, d10, d20). Day 0 is referred to as mESC. The analysis was performed on triplicates at each time point , although the number of available replicates for day 0 was seven, since DREMflow requires the same number of replicates at each time point. The number of splits allowed was set to four. The log2FC was set to 2 instead of 1 to ensure comparability to the results by Duren et al (Duren et al. 2020). The data was downloaded from the NCBI Gene Expression Omnibus (GEO; https://www.ncbi.nlm.nih.gov/geo/) under the identifier GSE136312.

3.3.3. Myeloid cell differentiation

Promyelocytes from the HL-60 cell line were differentiated into macrophages, neutrophils, monocytes and monocyte derived macrophages. Samples were taken in triplicates at 3h, 6h, 12h, 24h, 48h, 96h and 120h. Including the HL-60 promyelocyte as starting point (referred to as HL60), the data provided eight time points for the analysis. DREMflow was applied to the macrophage, neutrophil and monocyte data sets (Ramirez et al. 2017). The number of splits allowed at each node was kept at the default setting of two. PWMs from 2018 (version 2.1, JASPAR) were used. The data was taken from GEO: GSE79046.

Table 3.2.: Time point corresponding populations in erythropoiesis

Time point	Abbreviation	Description
P1	MyP	myeloid progenitor cells
P2	CFU-E	colony-forming unit erythroid cells
P3	ProE1	proerythroblasts
P4	ProE2	proerythroblasts
P5	BasoE	basophilic erythroblasts
P6	PolyE	polychromatic erythroblasts
P7	OrthoE	orthochromatic erythroblasts
P8	Orth/Ret	orthochromatic erythroblasts enriched with reticulocytes

3.3.4. Time course of human adult erythropoiesis

CD34+ hematopoietic stem and progenitor cells (HSPCs) from three healthy donors were differentiated to obtain enucleated reticulocytes (Ludwig et al. 2019). Eight populations were identified through FACS by known erythroid markers and considered as pseudotime points for the analysis in the order P1-P8 (Table 3.2). The most mature cell type identified were orthochromatic erythroblasts enriched with reticulocytes. Only two splits were allowed out of each node. The samples were available in triplicates. The data is available at GEO: GSE115684.

3.4. Comparison to PECA2/TimeReg

The comparison to PECA2/TimeReg was done on the macrophage data set from Ramirez et al. to ensure the use of neutral data (Ramirez et al. 2017) and the differentiation time series from RA-induced mESCs from Duren et al (Section 3.3.2) that provide a list of modules specified with GO ontology enrichment and a list of driver TFs to compare to top regulators from DREMflow (Duren et al. 2020).

3.4.1. Application of PECA2/TimeReg

The macrophage ATAC-seq data was aligned to the Gencode Genome sequence, primary assembly (GRCh38), since PECA2 could not be executed with Ensembl style chromosome names. The TPM computed with DREMflow were used as input for PECA2 to calculate the cis-regulation scores (CRS), which are comparable to TF-gene links and that are needed for TimeReg. Only protein coding genes with a log2FC>1 were included. PECA2 was executed for each time point separately in scripts on Iris to generate the input for TimeReg. The standard configuration of TimeReg was used and the number of modules was set to one at the first time point, two at the second time point and three at the remaining time points.

3.4.2. List of transcription factors for macrophage differentiation and function

For the comparison of DREMflow and TimeReg, a list of TFs known in literature for macrophage differentiation was obtained by manually sighting publications from pubmed to create a gold standard. Search terms were “macrophage differentiation” and “transcription factors”. The search was limited to reviews from 2020 onwards. Using these criteria, only three relevant publications were found S. Chen et al. (2020). An extension to all type of articles resulted in Table 3.3.

This list consists of following TFs.

Table 3.3.: Transcription factors known in the context of macrophage differentiation

TF	TF family/class	Key references
SPI1, SPIB, ERG, ELF2, ELF1, ELF4, ETS2, FLI1, ETV3, ETV5	ETS family	D. A. Hume and Himes (2003), Dekkers et al. (2019), S. Chen et al. (2020) David A. Hume, Summers, and Rehli (2016)
EGR1, EGR2, EGR3, EGR4, SP1, GATA6, MZF1	EGR family Zinc finger transcription factors	Dekkers et al. (2019), Jego et al. (2014)
CEBPA, CEBPE, CEBPG	CCAAT/ Enhancer-Binding Protein	D. A. Hume and Himes (2003), David A. Hume, Summers, and Rehli (2016)
JUN, JUND, JUNB, FOS, FOSB, FOSL1, FOSL2	AP-1	Dekkers et al. (2019), Lord et al. (1992)
HOXA10, HOXA9, HOXB3, HOXB8, HOXB7	Homeobox	D. A. Hume and Himes (2003)
IRF5, IRF4, IRF8	Interferon regulatory factors	Pundhir et al. (2018)

TF	TF family/class	Key references
RUNX1 (AML1), RUNX3	RUNX family	Nagamura-Inoue, Tamura, and Ozato (2001), Ai and Udalova (2020)
ERK1, EKR2	MAP kinase family	D. A. Hume and Himes (2003)
TFEB, TFEC, TFEC, MITF, BHLH40	Basic Helix-Loop-Helix	D. A. Hume and Himes (2003)
MYB, MYC	Remaining proto-oncogenes	Nagamura-Inoue, Tamura, and Ozato (2001)
FOXO1, FOXO3, FOXP3	Forkhead family	Santoni de Sio et al. (2017)
STAT1, STAT5, STAT6	Signal Transducer And Activator Of Transcription	Nagamura-Inoue, Tamura, and Ozato (2001)
MAFB, SPIC	Miscellaneous	Jego et al. (2014), Hamada et al. (2020)

The findings are also supported by highly cited reviews in the last decade Bencheikh et al. (2019).

TFs that were mentioned in several publications and supported by recent reviews were included in this list. The list was considered complete when no new TFs were mentioned in several publications. Since this is a manual selection there is no guarantee for biological completeness.

A search at https://amigo.geneontology.org/ provided a list of proteins related to *macrophage differentiation* and *regulation of macrophage differentiation*. An overlap with the TFs included in the analysis resulted in CEBPA, CEBPE, GATA3, SPI1, NR3C1, GATA2, HSF1 and NKX2-3 being added to or confirmed on the list from Table 3.3 [Ashburner et al. (2000); Carbon et al. (2009)]. Search results from https://mogrify.net/ for the transition from monocyte immature derived dendritic cell to macrophages added SLC11A1, GCLC, MAFB, ETS2, MITF DBP, SPI1 and MYC to the list or confirmed the ones that were already included due to the literature search (The FANTOM Consortium et al. 2016).

3.5. Code availability

DREMflow is currently available under https://gitlab.lcsb.uni.lu/drem/snakemake-epic-drem.

This thesis was completely written in Quarto, which is an open source software for scientific publishing build on pandoc and it is licensed under the GNU GPL v2. Figures in the results chapters and most figures in the appendix are directly generated from code to ensure reproducibility throughout the thesis.

4. Multi-omics analysis of mDAN differentiation as primary use case for DREMflow

The midbrain dopaminergic neuron data set from Ramos et al. (2023)(Section B.2) was the first application of EPIC-DREM on ATAC-seq data. In the context of the PARK-QC doctoral training unit I provided Borja Gomez Ramos with a list of possible transcriptional regulations of differentiation into dopaminergic neurons identified with the EPIC-DREM framework. This first analysis was the motivation to implement DREMflow, as it showed that a bioinformatics pipeline to automate this kind of analysis was missing. The protocol applied by Ramos et al. for human midbrain dopaminergic neuron differentiation resulted in paired time course epigenomic and transcriptomic samples for mDAN, purified by FACS sorting and a heterogeneous neuronal cell mix not limited to one specific cell type (non-mDAN). Although the main interest was the identification of mDAN specific transcriptional regulators, the analysis of non-mDAN was performed to compare identified top regulators from the mDAN data set to those found for the heterogeneous neuronal cell mix.

DREMflow combines information from time course chromatin accessibility data and paired samples of gene expression data to build time point-specific GRN. The main steps in the pipeline are the inference of TF-gene links based on the open chromatin regions in the samples via TEPIC (Schmidt et al. 2017, 2019) and the integration of these predictions with the expression changes over time via iDREM (Ding et al. 2018). According to the ob-

served expression changes, iDREM identifies split events to cluster co-expressed genes together and assigns TFs regulating the expression changes based on the predictions from TEPIC (Schulz et al. 2012; Ding et al. 2018; Schmidt et al. 2019).

4.1. Optimization of DREMflow on human midbrain dopaminergic neuron differentiation data

As the development of DREMflow was motivated by the mDAN time course data, it was used extensively in its testing, implementation and optimization. The mDAN time course consists of four time points, with samples taken from smNPC, at day 15, day 30 and day 50 and was analyzed multiple times with different settings often leading to new options for the pipeline execution. While most of the intermediate steps are not shown in the scope of this thesis, four main results are highlighted here.

1. The fully implemented DREMflow pipeline was adjusted to use exactly the same settings as were used in the scripts to reproduce the initial results. This adjustment was required since the normalization was not included initially and the filtering of TF-gene links and TG used different parameters. In addition, the reference genome version (GRCh38, patch 1) and the annotation version 29 from Gencode (Ensembl release 94) were from 2018, since the data was analyzed first then and led to compatibility problems with the default DREMflow settings that were meant for Ensembl annotations.
2. The data was analyzed with the default DREMflow settings including the Ensembl reference genome and annotation release 108 and the newest set of PWMs available for TEPIC (version 3.0) to test if the same results as achieved under 1. could be confirmed with the newest annotation version and an extended PWM set comprising 708 TF motifs overall.

3. After seeing diverging results between the first two analyses, the option of using TF clusters instead of individual TFs as input for TEPIC to identify cluster-gene links was explored as possibility to account for similar motifs shared between TFs of the same family.
4. The option of using only protein coding genes from the annotation was implemented and applied since non protein coding genes could be accounted for changes between annotation version.

These four result sets will be presented in the following and illustrate the difficulties of implementing a pipeline in the face of updating annotations.

4.1.1. DREMflow with early settings identifies known mDAN differentiation TF as top regulators

After DREMflow was fully implemented as a Snakemake pipeline, the adjusted version of DREMflow as described in Section 3.1 and above was applied to the mDAN differentiation data and identified LMX1A, LMX1B, EN1 and NR4A2 as drivers of expression change in human mDAN differentiation (Figure A.4). The regulatory network consists of 26 nodes with 327 TFs predicted as significant in at least one node of the network. The 20 nodes after a split contain information about the significant regulators and co-expressed target genes. Four splits were allowed out of each split node since the default setting of two splits could not capture the data well given that the time series only contained four time points (Figure A.1). Approximately 21000 genes were included in the analysis with the default cut-off of log2FC>1 (see Section 3.1). Highlighted are highly ranked TFs known for mDAN differentiation and TFs that were identified as regulators by Borja Gomez Ramos in the scope of his thesis and his publication. Details about the selection and further validation experiments are described in Ramos et al. (2023) . The TFs were ranked highly in the split node network even in case of a high Split Score (>=0.001)(Table A.1), which would be considered not significant. This suggests that the Split Score cut-off can be adjusted

if necessary to include TFs based on their ranking as well, since the Split Score strongly depends on the identified number of TF-gene links and the number of co-expressed target genes in a node.

GO enrichment of target gene clusters for nodes after a split showed a separation between upregulated and downregulated genes. Upregulated genes were enriched for neuron differentiation related terms such as *axon development* and *synapse organization* while downregulated nodes were enriched for *ncRNA processing* and *nuclear division*.

GO enrichment for significant TFs at a node did not result in specific terms and included *myeloid cell differentiation* and *regulation of neurogenesis*.

In the scope of his thesis and the publication, Dr. Borja Gomez Ramos selected candidates for the transcriptional regulation of differentiation into dopaminergic neurons by prioritizing TFs that have a high number of target genes and are under high regulatory load. The role of LBX1 and NHLH1 as transcriptional regulators during mDAN differentiation was confirmed through knockdown and overexpression experiments (Ramos et al. 2023). As described, the candidates were selected manually instead of using the automatic selection from DREMflow that considers recurring significant TFs the ranked 15 or higher among all nodes. Through the automatic selection only EN1 appears in the top regulators with two occurrences (Figure A.2). Also identified are ASCL1, a known reprogramming factor for the generation of dopaminergic neurons, and MSX1 that is directly regulated by LMX1A and LMX1B in the midbrain specific gene regulatory network (E. Arenas, Denham, and Villaescusa 2015). These results demonstrate that an automatic selection by DREMflow can identify some but not all important regulators in a system.

4.1.1.1. Regulators of interest

Eleven TFs were further profiled as regulators of interest in the network. LMX1A, LMX1B, SOX4, NR4A2 and EN1 are TFs known for mDAN neuron differentiation (E. Arenas, Denham, and Villaescusa 2015; Veenvliet et al. 2013; Hermanson 2003; Sherf et al. 2015)

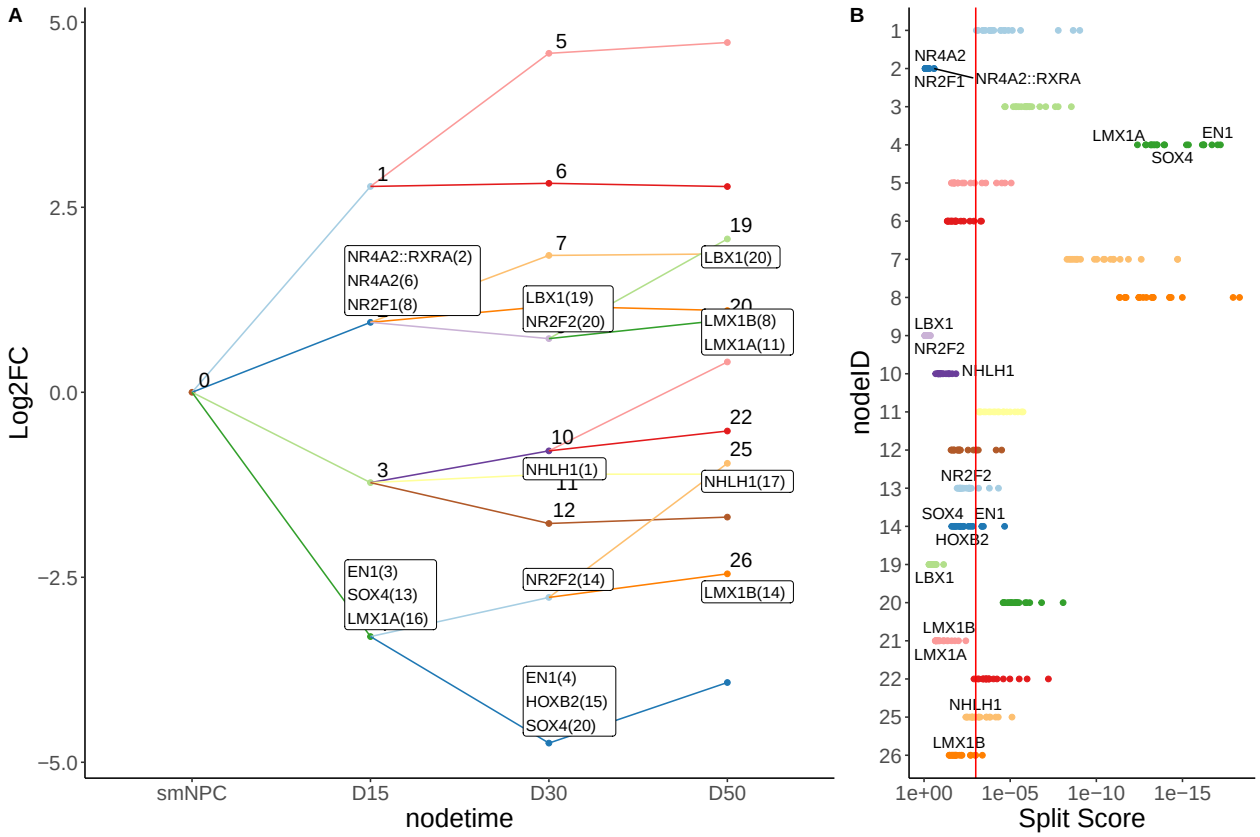


Figure 4.1.: Initial DREM model overview of mDA differentiation. (A) DREMflow computed split node network of co-expressed genes. Labeled nodes are the nodes after split nodes. The text boxes display known and newly identified TFs for mDAN differentiation with the rankings at that node. (B) TFs assigned to split nodes according to a split score. The red vertical line indicates the recommended Split Score cut-off of 0.001. LBX1 and LMX1A/B are highlighted as manual selected regulators despite not being considered as significant according to their Split Score.

while the remaining ones were included as candidates of interest. For EN1 a strong upregulation is observed over all time points with a log2FC>8 on D50. The observed expression change does not only support the important role of EN1 for the differentiation but also the maintenance of mDAN (Figure A.2 A)(E. Arenas, Denham, and Villaescusa 2015). LBX1 was clustered close to SOX4 and LMX1B, which could suggest a similar role in the network. The lowest number of estimated target genes was observed for NR2F1 and NR2F2 with less than 4500 at three out of four time points, while the number of TF-gene links for NHLH1, NHLH2, LMX1A, LMX1B, LBX1, HOXB2 and EN1 were high in comparison to the other regulators of interest with 9000 and more, implying an important role in the network (Figure A.2 B).

4.1.2. Different results with DREMflow using GRCh38 release 108 and updated position weight matrices

After successfully reproducing the results acquired by scripts with the adjusted DREMflow version, the data set was analyzed with the default DREMflow settings. The pipeline was executed on the mDAN differentiation data with Ensembl GRCh38 release 108 and showed diverging results in comparison to the adjusted settings (Figure A.4). The selected candidate TFs as well as known mDAN neuron differentiation TFs were not ranked highly anymore. The only exception was EN1 that was identified again among the recurring top regulators (Figure A.4 B). Also, DLX1 and PBX1 were selected as top regulators. DLX1 has an important role in neurogenesis while PBX1 is known specifically for dopaminergic neuron differentiation (Ghanem et al. 2012; Villaescusa et al. 2016). The early high activity observed for PBX1 correlates with its activating and repressing function during embryogenesis (Figure A.4 C), while the late high activity observed for EN1 confirms the late role in mDAN differentiation and maintenance (Villaescusa et al. 2016; E. Arenas, Denham, and Villaescusa 2015). The top regulators identified in the network using the newest annotation version and the extended set of TF motifs still includes known TFs in the context of mDAN differentiation, however the results from the earliest DREMflow run could not be reproduced.

The results imply a lack of robustness towards changing annotation versions, but not only the annotation release changed but also the position weight matrices (PWMs) for TEPIC. With the 3.0 addition from April 26th, 2022, TEPIC only provides a combined input of JASPAR, Hocomoco and Kellis motifs resulting in 708 TFs in total compared to the 432 previously used TFs that include only TFs from JASPAR. The difference in annotation resulted in an overall increase in target genes (Figure 4.3 A), leading to a different split node network as before, while the change of PWMs identified approximately 50% more TF-gene links.

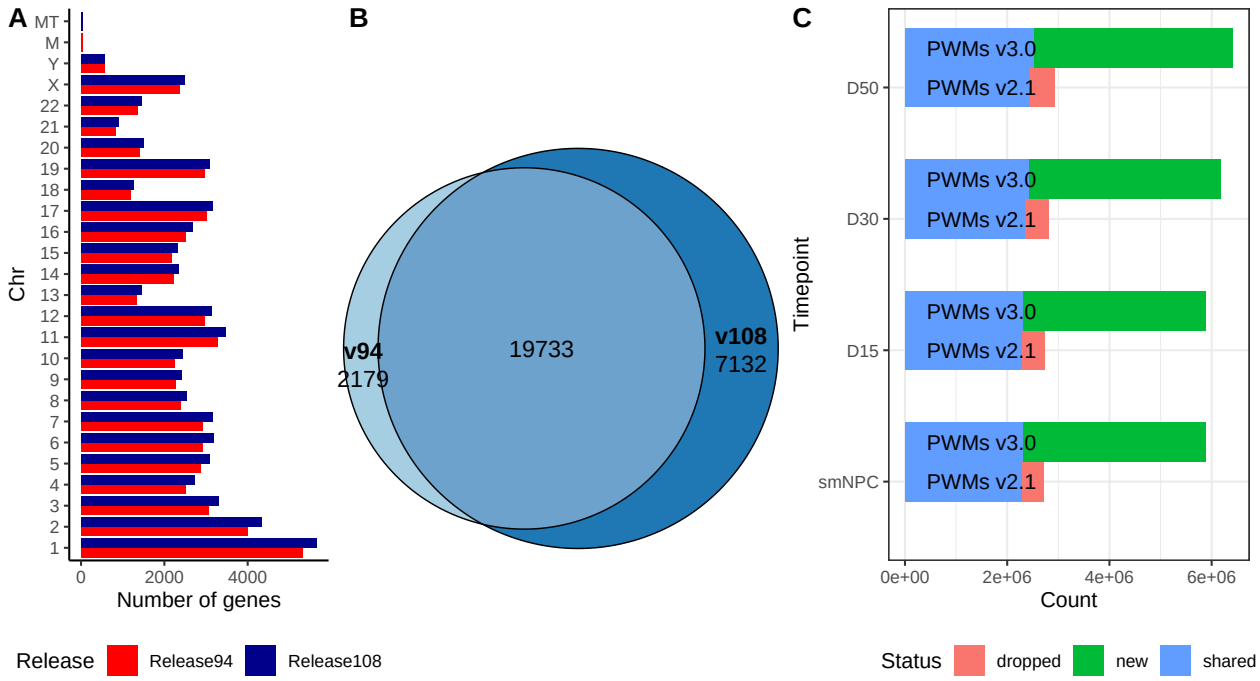


Figure 4.3.: Differences between annotation release 94 and 108 and PMWs version 2.1 and 3.0. (A) Number of genes at each chromosome for the respective annotation release. Release 108 includes more genes. (B) Intersection of target genes included with release 94 (v94) and release 108 (v108). They share 68% of target genes. (C) Comparison between PWMs version 2.1 and 3.0 in terms of identified number of TF-gene links for the mDAN data set. Red are new links, blue links are shared and green links were only found with version 2.1 (previous).

4.1.2.1. Differences between initial settings and most recent annotation and motifs

The differences in annotation and motifs was further investigated to identify the origin of the lack of robustness. A comparison between the resulting target genes and TF-gene links using the annotation release 94 or 108 was done. While Ensembl annotation release 108 introduced more genes at each chromosome than release 94 (Figure 4.3 A), the number of target genes for the calculation increased with approximately 7000 more genes as input for DREM (Figure 4.3 B). The change from only using JASPAR PWMs for the calculation of TF-gene links to combined input from several databases (JASPAR, Hocomoco and Kellis) resulted in twice the amount of links associated within the footprints. This effect was driven by the new TFs, that made up more than half of the identified TF-gene links (Figure 4.3 C). An attempt to decrease the number by reducing the peaks with more stringent peak calling parameters was not successful (results not shown).
Another approach to reduce the number of TF-gene links would be filtering according to the number of transcription factor binding sites. A time point specific removal of 10% TFs with least TFBS would result in a reduction of TF-gene links of approximately 25% overall (Figure A.3) but might also prevent the discovery of strong regulators with less TFBS and strongly depends on the chromatin accessibility data. The high number of TF-gene links and the difference of over 7000 genes led to the exploration of other approaches. The first one was the usage of TF clusters instead of individual TFs while the second approach was the exclusion on non-protein coding genes in the analysis.

4.1.3. Using transcription factor clusters confirms TFs of the same family as LBX1 and LMX1A among top regulators in dopaminergic neurons

Instead of using individual TFs as input for TEPIC, the option to use TF clusters that combine TFs from the same TF family with similar binding motifs, was implemented and

tested on the mDAN data set. With 96 TF clusters overall this reduced the input of TF-gene links significantly.

The cluster containing LMX1A, LMX1B, EN1 and LBX1 was assigned to upregulated and downregulated nodes. From the previous results it can be inferred that EN1 would be assigned to the downregulated nodes while LBX1 would be assigned to upregulated nodes (Figure A.5). The cluster containing SOX4 was found at downregulated nodes, which correlates with the previous results. Although the cluster with LMX1A, LMX1B, EN1 and LBX1 is the largest cluster with 79 TFs, the number of estimated target genes was only at approximately 10000. This numbers suggest, that the motif representing all TFs in the cluster is not suitable to identify all TFBS in the data.

The results from the clustering approach proved to be problematic since it reduced the previously detailed information to only show a tendency by identifying certain TF families as main regulator families in the split node network. It can only be inferred that EN1 is actually on the downregulated nodes and LBX1 on the upregulated nodes, since no TF is individually assigned. The filtering of TFs according to expression is also removed. The option of using clusters is implemented in the pipeline. It should only be used in addition to the combined TFs approach.

4.1.4. LBX1, EN1, LMX1A and LMX1B are highly ranked regulators for protein coding genes

Since the difference in number of target genes was large with only 68% shared (Figure 4.3 B), the option of reducing the annotation to only include protein coding genes was explored. The results with a combined set of 708 TFs was comparable to the adjusted DREMflow run described above.

Attempting to reproduce the initially observed results, the previous set of PWMs including only motifs from JASPAR was chosen as input. Although the ranking was not identical, LMX1A, LMX1B, EN1 and LBX1 were identified as highly ranked regulators in the split

node network again. The heterodimer NR4A2::RXRA was ranked 2nd at node 4 (in the earliest network labeled node 2). EN1 was assigned to the most downregulated node on D15. Interestingly, LBX1 now was assigned to node 5 and NHLH1 and LMX1B were found highly ranked on the upregulated node 6. The identified network consists of 31 nodes, 22 of those after a split, and includes almost 8000 target genes. This is a reduction of more than 50% compared to the initial run. Given the size of the data set, the first fifteen TFs were considered for the selection of top regulators as described in Section 3.2.5.1. ELK3, ETV1 and more members of the ETS family of transcription factors were include as top regulators. LBX1 was identified as top regulator although the number of estimated target genes is low with less than 4500 in comparison to members of the ETS family having more than 6000. NHLH1 was not identified as a top regulator through the automated selection.

Despite the reduction of target genes similar results as before were observed for the GO enrichment on node specific target genes. Upregulated nodes were enriched for *axogenesis regulation* and *synapse organization* while downregulated nodes were enriched for *ncRNA processing* and *chromosome segregation*. Given the unspecific results observed before, GO enrichment on significant TF at each node was not performed again. Instead, GO enrichment was performed on the target genes that are predicted to be regulated only by selected top regulators. This narrowed the GO enrichment to TF specific target genes instead of all genes assigned to a node to see the specificity of the top regulators in the systems with regards to the target genes they are predicted to regulate. The target genes of ETS family members were enriched for *ncRNA processing* and *ribosome biogenesis*, mostly including downregulated genes. The target genes of LBX1 had *cell-cell signaling by wnt* and *Wnt signaling pathway* as enriched terms. These findings support the selection of LBX1 playing an important role as a regulator for dopaminergic neuron differentiation, since the Wnt signaling pathway is essential in the development of mDAN E. Arenas, Denham, and Villaescusa (2015). Neuron differentiation related terms were found for target genes of REST, INSM1, MTF1, TGIF1 and TGIF2.

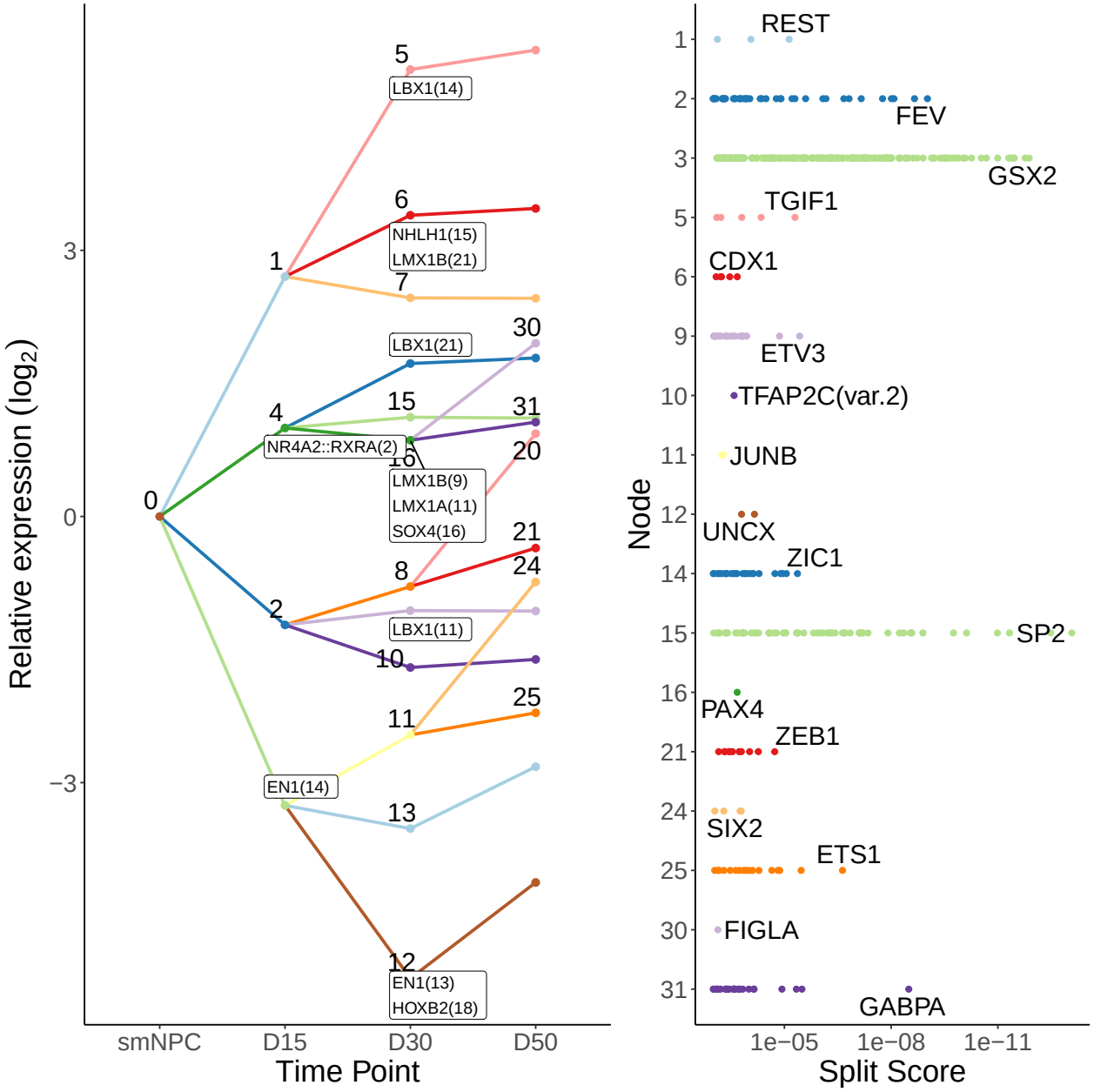


Figure 4.4.: Split node network for mDAN neurons for protein coding genes. (A) DREMflow computed split node network of co-expressed genes. Highlighted are known TFs and selected candidates for further validation. (B) Significant TFs identified to regulate a bifurcation event. TFs are assigned to the nodes after a split. Labeled are the highest ranked significant TFs assigned to a node.

4.2. Application to non-dopaminergic neurons

As a comparison to mDAN, the non-dopaminergic neuron data set was analyzed as well. With only three time points overall, the non-mDAN resulted in the smallest split node network. Four paths were allowed out of each node to adjust for the small number of samples but only two paths were identified at the first split node and three out of node 1. With this the network contains seven nodes, all of them after a split but only five having TFs assigned as significant regulators (Figure 4.5). Overall, 6000 target genes are included. (Figure A.6) Surprisingly, LMX1A and LMX1B were highly connected in the TF-TF network for node 3, despite both of them being a mDAN specific transcriptional regulator (Figure A.7).

The results from GO enrichment are similar to the ones from mDAN. Upregulated nodes contain neuron differentiation related terms while downregulated nodes are enriched for chromosome segregation and DNA replication. From the identified top regulators, only target genes of NRF1 and FOXP3, both the only regulators at node 4 and 7 respectively, are enriched for neuron related terms.

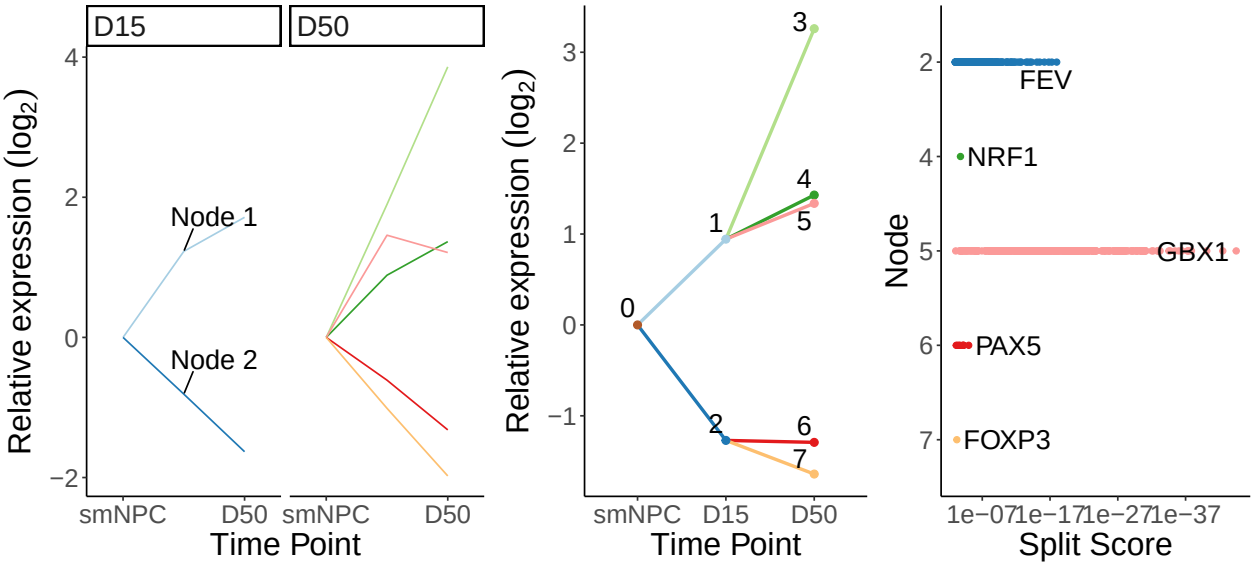


Figure 4.5.: Split node network for non mDA neurons. (A) Average expression profiles for the computed target gene clusters at each time point. (B) DREMflow computed split node network of co-expressed genes. (C) Significant TF assigned to nodes as regulators for the expression change.

4.2.1. Comparison to dopaminergic neurons

Despite an overlap of 66% of target genes (Figure 4.6 A), the non-mDAN model without a day 30 sample resulted in a strikingly small network with only seven nodes compared to the 31 nodes identified from the mDAN data. This difference can not be explained by the number of TF-gene links per time point, since the non-mDAN have more TF-gene links if D30 is exclude (Figure 4.6 B) but might be connected to the TF-links identified in the data. Approximately 50% are shared for day 15 and day 30.

Apart from TFs of the ETS family no overlap of regulators was observed between the networks.

The difference between mDAN and non-mDAN demonstrates the impact one time point more can have on the interpretability of the results and shows the impact of the TF-gene link data.

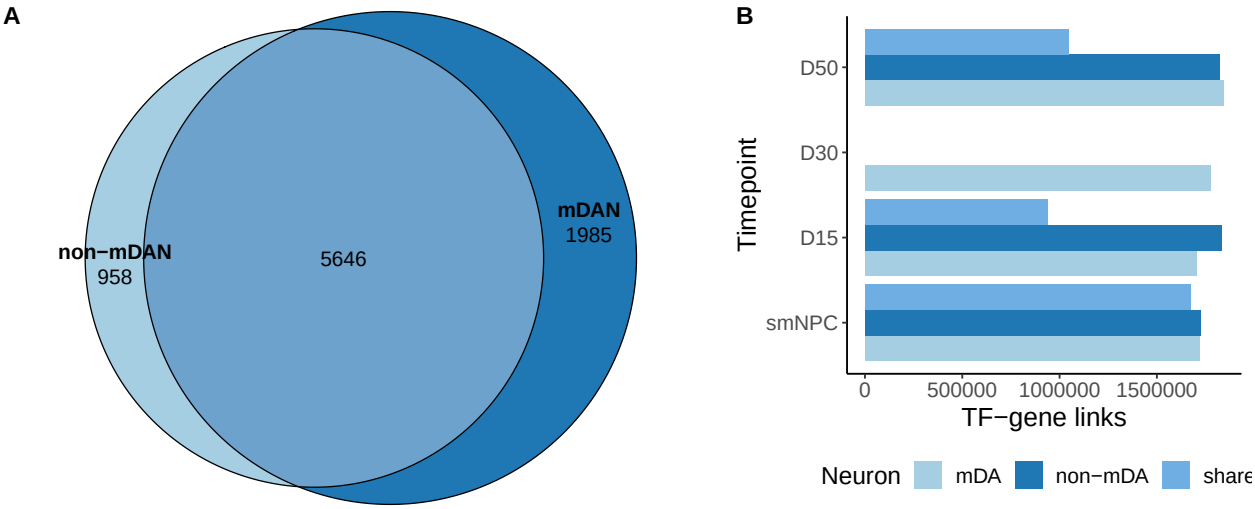


Figure 4.6.: Intersection of target genes and TF-gene links non-mDAN and mDAN. (A) Number of shared and individual target genes. (B) Number of shared and specific TF-gene links for mDAN and non-mDAN at each time point.

4.3. Discussion

The implementation of DREMflow emerged from an adaption of the EPIC-DREM framework (Gérard et al. 2018) to analyze the midbrain dopaminergic neuron data and support the selection of possible transcriptional regulators in mDAN differentiation in the context of the PARK-QC Doctoral Training Unit. The development of this multi-omics data integration was strongly based on the mDAN data set in the early process of implementation and it is by far the data set that was analyzed with all developmental versions of DREMflow most often. The adjustment from H3K27ac ChIP-seq signals to ATAC-seq provided some challenges, e.g. an overestimation of TF-gene links in TEPIC and opportunities with the inclusion of time point specific binding activity inferred from the ATAC-seq signal.

Early results identifying known transcriptional regulators for mDAN differentiation like EN1, LMX1A and LMX1B suggest a good capture of the data by the early DREMflow version. With the implementation in Snakemake and the updated annotation and PWMs, the increased number of TF-gene links brought up the question of robustness of DREMflow to future versions of the annotation. With the annotation of new genes, the observed differences in the results are not themselves surprising. Given a change of 25% of the target genes with the newest annotation and over 50% of TF-gene links with added PWMs, the difference observed in the model were expected but changed especially the ranking of known TFs. The filtering of TF-gene links was discarded, since HINT-ATAC, that was used for footprinting, captures TFBS well.

The use of TF clusters instead of individual TFs confirmed the assumption that the TFs with a similar motif influence assignment of TF-gene links and the statistical analysis in DREM, favoring TFs with the same motif. The analysis based on TF clusters did not provide a lot of detail and a single cluster contained most of the known regulators and selected candidate TFs, suggesting that the use of TF clusters only provides a broad overview but looses many details and is not suitable for the identification of top regulators in a system. In addition, it lacks the comprehensive output that was implemented for individual TFs

based on the selection of top regulators such as expression profiles and activity scores.

The reduction of annotation to contain only protein coding genes provided similar results as the first analysis. In addition, the set of PWMs was reduced to only include PWMs from JASPAR, since JASPAR is a common standard. Unfortunately the newest PWMs sets in TEPIC only offer combined sets of HOCOMOCO, JASPAR and Kellis PWMs, so an older set from 2018 was used. Inclusion of only protein coding genes is likely to increase the robustness of changes in annotation versions, since a rapid increase of protein coding genes is unlikely. However, future annotation updates will still change results most likely. For reproducibility, it is recommended that the annotation version should be frozen. In the case of the mDAN data set, I updated the version because the annotation from the earliest run was four years older than the most recent annotation and major changes could be observed. This was not only due to the version change but also due to the annotation database. The reference data is taken from Ensembl database by the wrapper (as explained in Section 3.2.1) , which provides the most comprehensive set of genes. In the earliest version, a Gencode annotation was used (see Section 3.1). Since the quantification of gene counts can differ up to 50% between different annotation databases (Zhao and Zhang 2015), this could have had an influence as well.

Considering the calculation of the Split Score based on the hypergeometric distribution, not only the reduced number of included TFs but also the reduced number of included target genes, when the data is limited to only protein coding genes, improves the Split Scores and therefore the prediction.

The results from DREMflow reflect the biology well with the identification of known TFs and LBX1 was confirmed as candidate even considering newest annotation input data. However, NHLH1 would not have been detected via the selection of highly ranked recurring top regulators in the systems, demonstrating that not all details are captured while focusing on strict values such as a Split Score cut-off. Only the literature research and the accompanying ChIP-seq experiments to identify super enhancers led to the identificatio of NHLH1 as candidate (Ramos et al. 2023). While the results for the mDAN differentia-

tions are promising, it is not the optimal time series data set to study differentiation events in general. The time points were chosen according to the maturity of mDAN neurons (Ramos et al. 2023). Additional time points and shorter time spans between the data points could improve the predictions, considering the rapid changes in chromatin accessibility and short binding period of TFs (Hager, McNally, and Misteli 2009). Wnt signaling among the enriched terms for LBX1 target genes further supports an important role of this TF in dopaminergic neuron differentiation, since Wnt signaling activates a cascade of regulatory mechanisms essential for mDAN neurogenesis (Ernest Arenas 2014; M. Wang et al. 2020).

The comparison to non-mDAN shows the the impact of one time point on the data. While almost 6000 target genes are shared and the number of TF-gene links is similar across the time points, the split node network identified for non-mDAN includes only seven nodes. Especially the time span of 35 days between day 15 and day 50 appears large and raises the question of biological interpretability.

5. Application of DREMflow to myeloid differentiation

Among the suitable data sets meeting the requirements described in Section 3.3, the myeloid differentiation lineages were selected to validate if DREMflow identifies known regulators and possible candidates for further validation.

Ramirez et al. (2017) used the HL-60 promyelocyte cell line to study differentiation into macrophages, monocytes, neutrophils and monocyte-derived macrophages and identify TFs involved. Samples were taking over a 5-day time span on eight time points. LPS stimulated samples were not included in the analysis with DREMflow, since LPS stimulus changes the epigenetic landscape as the results from Ramirez have shown (Ramirez et al. 2017). Ramirez et al. combined time points in the ATAC-seq data into early, intermediate and late stages and used FIMO to identify a set of TF motifs from JASPAR to consider for the analysis. For the construction of gene regulatory networks, only TFs that were differentially expressed were included, limiting the set of transcriptional regulators to 232 overall. The transcription factors were linked to the nearest gene TSS based on the data derived from footprints for a distance of +/- 15kb around the TSS. With this approach Ramirez et al. identified SPI1, EGR, GFI1 and CEPBA among the top regulators (Ramirez et al. 2017).

5.1. Known myeloid transcriptional regulators highlighted in macrophage differentiation

Macrophages are well studied and many transcriptional regulators are identified, making it a model differentiation line to test DREMflow. Often differentiated from monocytes, macrophages can also emerge directly from myeloid progenitor cells without going through the stage of monocytes. Ramirez et al. provided both, promyelocyte and monocyte-derived macrophages. The results achieved with DREMflow for the promyelocyte-derived macrophage differentiation are described in this section. In comparison to the monocyte-derived macrophages eight time points instead of five were available without LPS stimulus (Ramirez et al. 2017).

Ramirez found SPI1 and EGR family TFs, MAFB and JUND most relevant specifically in macrophage differentiation (Ramirez et al. 2017).

Out of the three myeloid cell differentiation data sets, the macrophage data was the largest with approximately 10000 target genes and 6.6 million predicted TF-gene links included in the model. 90000 peaks were identified across all samples.

5.1.1. EGR1 identified as prominent regulator in macrophage differentiation

The split node network includes 89 nodes in total with 48 nodes after a split containing the unique information about target genes in the node and significant TFs assigned. Among the 39 recurring top regulators, EGR1 and EGR3 as well as SPI1 were found (Figure 5.1). The number of differentially expressed genes and peaks increases over time. The number of peaks and footprints are similar over the later time points with early differentiation time points showing only half the number compared to the control and last time point (Table A.2). With the consideration of highest two ranked TF as top regulators in addition to

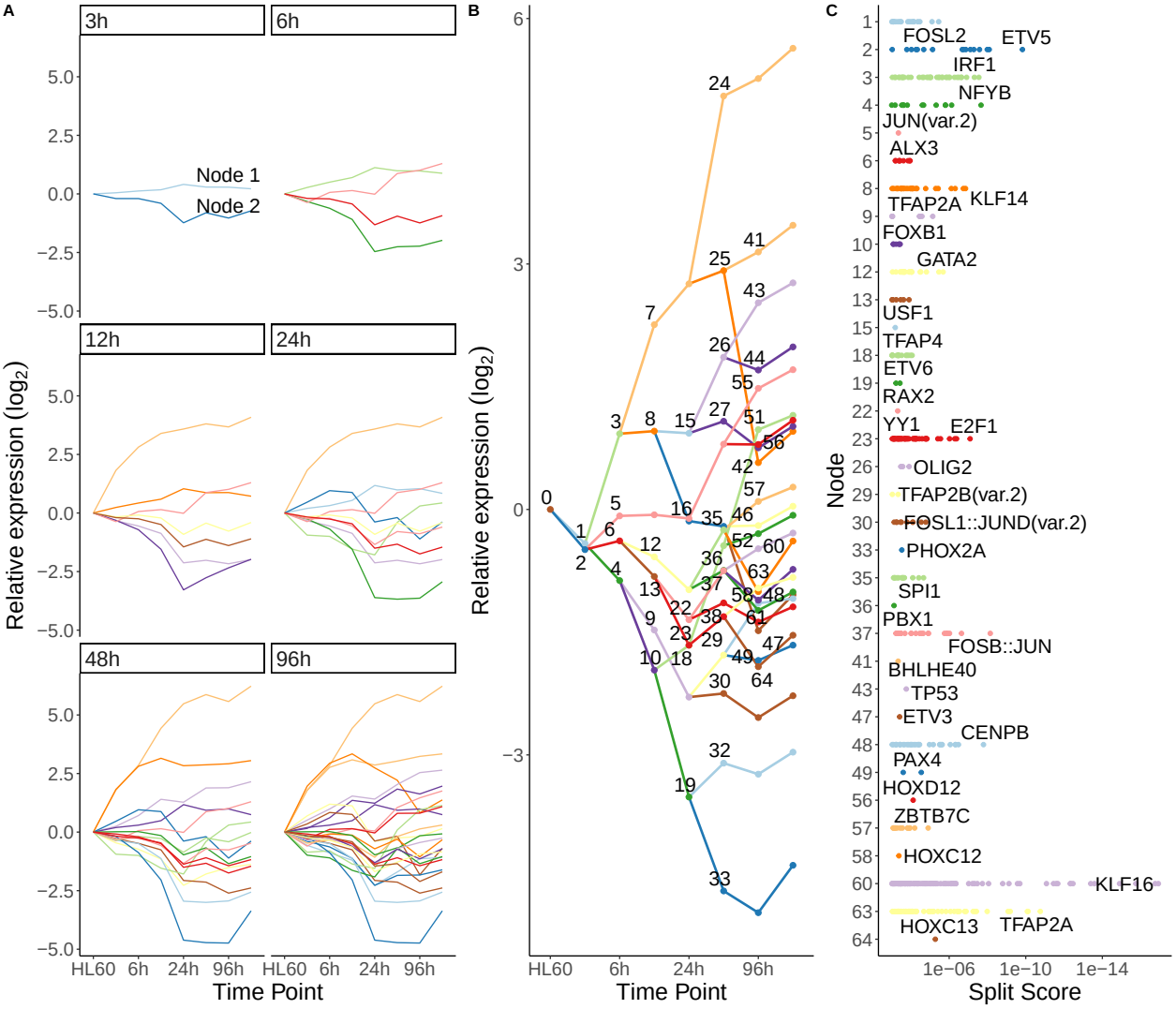


Figure 5.1.: Overview of macrophage split node networks and associated top regulators. (A) Average expression profiles of target gene clusters at each time point. (B) DREMflow computed split node network of co-expressed genes. Labeled nodes are the nodes after a bifurcation event. (C) Significant TFs regulating split events. Each dot represents a TF assigned to a split. The highest ranked TF is labeled.

the recurring TFs, the macrophage data provides a list of 79 top regulators over all. SP8 appeared four times in the top 10 significant TFs, making it the most prominent TF in the network. While no specific role in the context of myeloid differentiation is known, it has a similar motif to the EGR transcription factors that are well characterized in myeloid differentiation (Ramirez et al. 2017; Veremeyko et al. 2018). SPI1 was selected as well as top regulator. Surprisingly, it is assigned to the network at 6h as well as again at 48h. While the early role of SPI1 in hemopoiesis is known (Dzierzak and Philipsen 2013), the assignment to later time points is supported by literature as well as regulator in the transition from GMP to monocytes and macrophages (Figure 5.2) (Nagamura-Inoue, Tamura, and Ozato 2001). These results suggest that DREMflow has a good resolution with quality chromatin accessibility data. EGR TFs were upregulated early while the number of estimated target genes was low (<3000) during the early time points from 3h to 12h.

5.1.2. Upregulated target genes are enriched for hemopoiesis related terms

GO enrichment for target gene clusters reveals terms on upregulated nodes strongly related to macrophage and myeloid cells that includes nodes 1, 3, 15, 24, 25 and 41. The downregulated gene clusters are enriched for general terms such as *DNA packaging* and *rRNA processing*. GO enrichment for target genes of top regulators resulted in macrophage related terms for the heterodimers BATF::JUN and FOSL2::JUNB and the TFs JUN, TFAP4 and IRF1. Target genes of EGR TFs and SP8 were enriched for *ncRNA processing* and *chromosome segregation*. The regulating role of JUN and FOS related TFs was already described in 1991 (Lord et al. 1992).

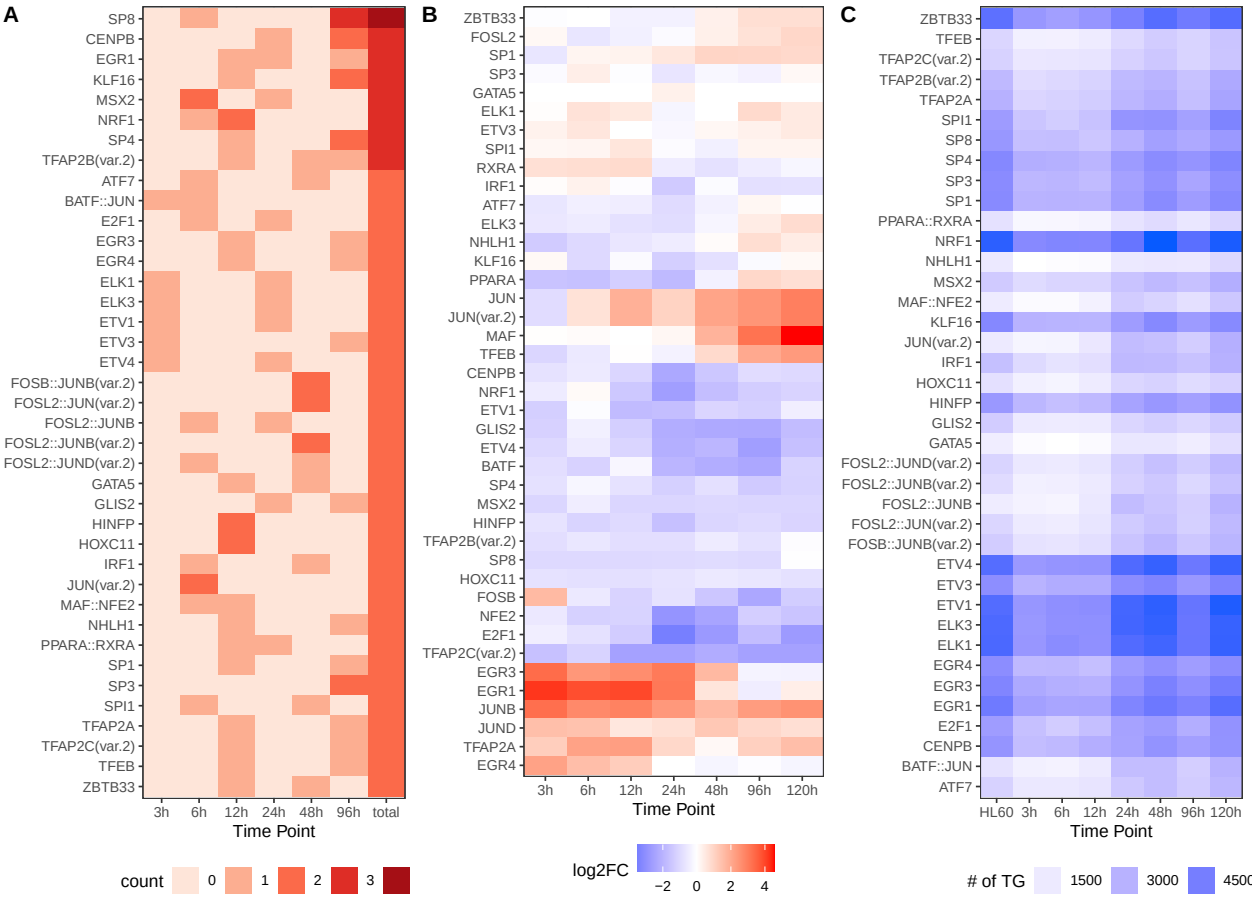


Figure 5.2.: Top recurring TFs identified for macrophage differentiation. (A) Recurring TFs selected as top regulators. The heatmap shows number of appearances per time point and int total in the top 10 ranked significant TFs over all nodes. (B) Relative expression of recurring top regulators as log2FC. Heterodimers are separated into individual TFs (C) Estimated number of target genes for each of the top regulators.

5.2. Application to identify key regulators in neutrophil differentiation

For neutrophils, Ramirez identified RUNX1, SPI1, CEBPA, ERG1, ERG2, GFI1, STAT6 and MAFB as transcriptional regulators(Ramirez et al. 2017).

The neutrophil data contained approximately 7000 target genes and 4.3 million TF-gene links, making it the second in terms of target genes but last comparing the TF-gene links. 70000 peaks were identified across all samples.

5.2.1. JUN and EGR2 but not SPI1 highlighted in neutrophils

The split node network identified for neutrophils shows a decrease in expression change and regulation after 24h (Figure 5.3 A). Overall it consists of 66 nodes, only 30 nodes after a split with no splits occurring at the 48h time points (Figure 5.3 B). Among the highly ranked regulators CLOCK, NHLH1 and TFDP1 were identified (Figure 5.3 C). The short list of selected top regulators included FOSL2 and JUN and EGR2, with ELK1, ELK3 and EGR2 being strongly up regulated at the 120h. The identification of FOSL2 and JUN can be accounted for by their regulatory function in myeloid differentiation in general (Lord et al. 1992).

Despite the low number of nodes and top regulators, GO enrichment identified node 7, node 13, node 14, node 25 and node 26 as neutrophil specific nodes. All nodes are upregulated apart from node 13. Non of the target gene sets for selected top regulators was enriched for terms related to neutrophils, myeloids or immune function.

Although the network for neutrophils is sparse, EGR2 was recognized as top regulator by DREMflow identifying only one regulator highlighted by Ramirez et al (Ramirez et al. 2017). Given the similar expression profile to EGR2, ELK1 and ELK3 would be interesting candidates for further investigation. Interestingly, the role of ELK1 as an essential TF for

the commitment to neutrophils was recently implied by Dong et al (Dong et al. 2022) suggesting to further look into ELK3 as well. With FOXA1 being upregulated at 3h on a central position in the TF-TF network at node 5, it could also be considered as candidate to look further into in terms of neutrophil differentiation.

5.3. ETS family TFs found to be transcriptional regulators in monocytes

Monopoiesis describes the differentiation of myeloid cells into monocytes. Monocytes can serve as macrophage precursors and therefore share a transcriptional and regulational landscape with them. Ramirez at al highlighted SPI1, STAT6, MAFB and JUND as transcriptional regulators in monocytes (Ramirez et al. 2017).

With 6800 target genes and 61000 peaks over all samples the monocyte data set can be considered the smallest, although TEPIC identified 1 million more TF-gene links than for neutrophils, resulting in 5.3 million links overall.

5.3.1. EGR2 and SPIB are early regulators in monocyte differentiation

In the split node network comprising 79 nodes, HEY1, KLF14 and MAX were the most prominent TFs in the top 10 ranked significant TFs (Figure 5.4 A and B). EGR TFs and SP1 and SPIB are among the top regulators with at least two assignments to a node (Figure 5.4 B). Interestingly most splits occur before time point 24h, were the least TF-gene links were identified from the chromatin accessibility data suggesting that most regulation happens early and fast in monocyte differentiation (Figure 5.4 C). Among the identified top regulators are those described by Ramirez et al. such as the EGR TFs (Ramirez et al. 2017) but also others known in the context of monocyte differentiation such as SPIB, which has a similar motif to SPI1 (Jego et al. 2014). SPI1 is ranked second on node 8.

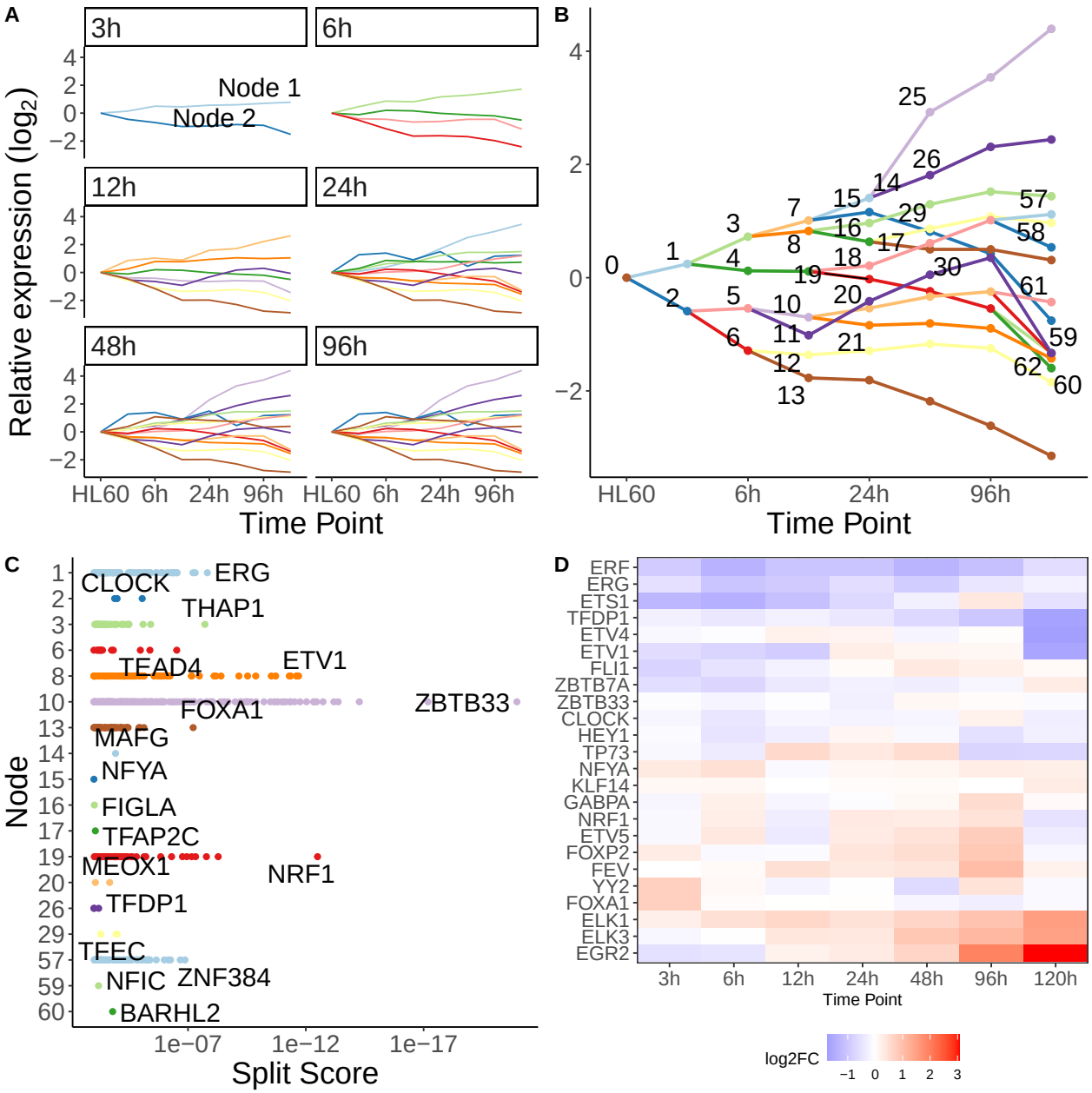


Figure 5.3.: Application of DREMflow to neutrophil differentiation. (A) Average expression profile of target genes assigned to each node displayed at each time point for neutrophil differentiation. (B) DREMflow computed split node network of co-expressed genes. (C) Significant transcriptional regulators responsible for bifurcation events displayed according to their split score at the assigned node. Labeled are the highest ranked TFs. (D) Expression of recurring top regulators as log2FC.

Terms such as *leukocyte migration* and *regulation of inflammatory response* were enriched in node 1, node 3 and node 45. Node 46 is enriched for *myeloid cell differentiation* containing EGR2 not as regulator but as target gene. Since the GO enrichment is limited to DEGs in a cluster, many clusters are quite sparse, especially for downregulated nodes.

Target genes of IRF1, CEBPA and SPI1 are enriched for *leukocyte migration* and *myeloid cell differentiation*. Previous studies have shown the importance of those TFs in the context of myeloid cell differentiation, suggesting that DREMflow captures the data well and identifies important regulators in the network (Pundhir et al. 2018).

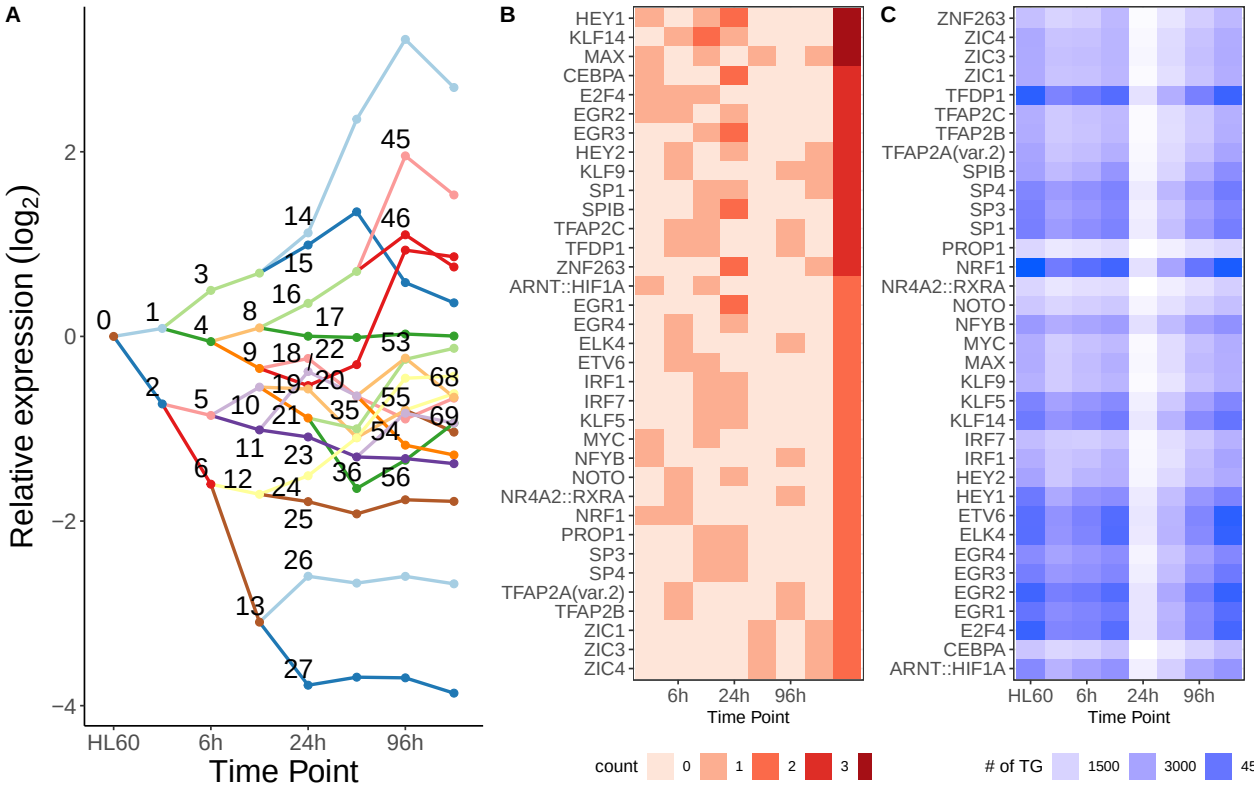


Figure 5.4.: Monocyte results overview. (A) Split node network identified for monocyte differentiation. (B) Number of appearances of selected top regulators in the top 10 highest ranked TFs at each time point and overall. (C) Number of estimated target genes for selected top regulators in the monocyte data set.

5.4. Comparison between transcriptional regulators in macrophages, neutrophils and monocytes

The three analyzed myeloid cell differentiation data sets all share the same origin in promyelocytes providing a common ground for comparison.

5.4.1. Few shared transcriptional regulators between specific myeloid cell commitment

Out of the top regulators for each cell commitment, only three were shared between all of them despite the same origin. The three TFs are TFAP2C, KLF14 and NRF1. DREMflow identified the most TFs for the macrophage data set with 79, 55 of them specifically found in macrophages. 69 top regulators were found in monocytes, 32 of those only in monocytes, and from the 40 top regulators in the neutrophil differentiation line, 22 were specific for neutrophils compared to the other myeloid cell lines. Monocytes and macrophages shared 11, neutrophils and macrophages shared 10 and monocyte and neutrophils shared five top regulators (Figure 5.5 A).

EGR1, EGR3 and EGR4 were exclusively identified as top regulators in macrophages. FLI1, EGR2 were found in neutrophils and SPIB in monocytes.

The normalized time point-specific binding activity score for the five shared top regulators differed between the three cell lines. The activity score represents the strength of a TF binding in comparison to the other time points. In macrophages, the TFs are highly active at HL60 and have a low activity until 48h with the exception being NRF1 (Figure 5.5 B). The activity shows the same results as seen for TF-gene links in monocytes at 24h, suggesting closed chromatin after one day, with an increase of activity after four days (Figure 5.5 C).

All three cell lines shared 4000 target genes and 1.5 million TF-gene links with NRF1 having the most TF-gene links over all time points. Although EGR2 was only found among

the top regulators in neutrophils, the three cell lines share approximately 12000 TF-gene links from EGR2.

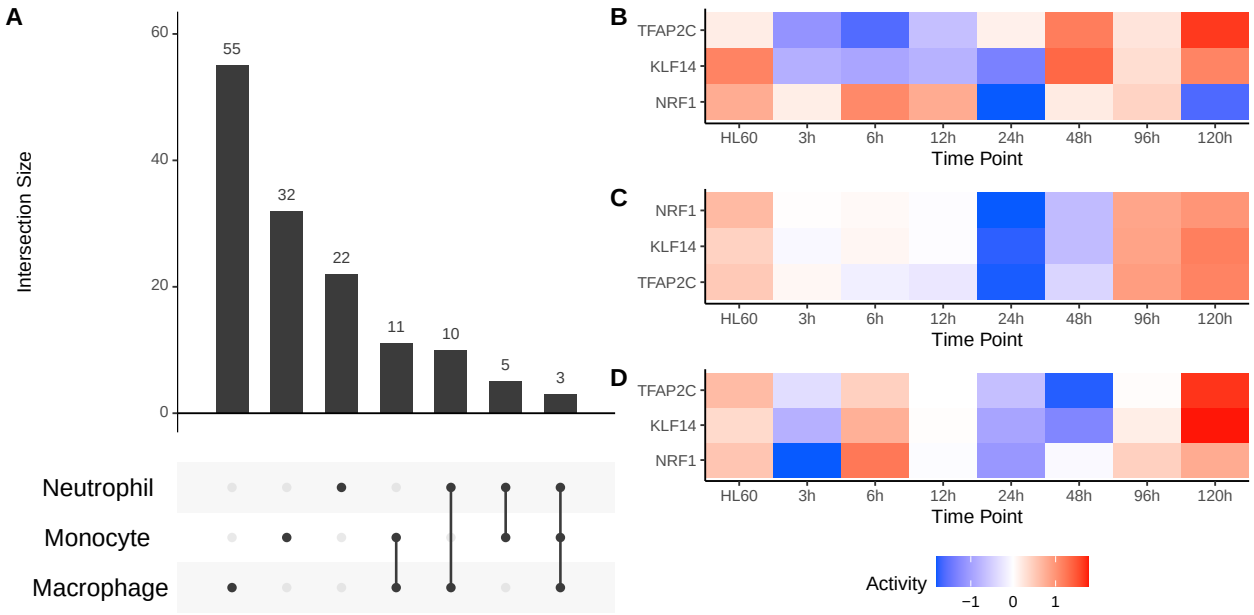


Figure 5.5.: Shared and unique TFs for the three myeloid differentiation cell lines. (A) Intersection of top regulators between the different myeloid cell lines. There are three top regulators shared between all of them. (B) Cell specific activity for the three shared TF in macrophage differentiation (B), monocyte differentiation (C) and neutrophil differentiation (D).

5.5. Discussion

The analysis of the data sets provided by Ramirez et al. of myeloid cell differentiation from promyelocytes into three distinct cell types demonstrated the biological relevance of the results achieved by DREMflow and identified known transcriptional regulators in myeloid cell differentiation. The data sets were well suited for testing the pipeline considering the following criteria.

Myeloid cell differentiation is well studied with hematopoietic stem cells considered as the best characterized stem cells (Zakrzewski et al. 2019). Literature research provided a list of 60 TFs known in the context of macrophage differentiation, activation and proliferation alone. Although DREMflow did not detect all of the TFs from the literature list, no publication or review mentioned all of them, suggesting that it depends on the method of identification of relevant regulators.

All cell types were differentiated from the same origin, enabling a comparison of how well DREMflow captures details in specific cell lines or just identifies general transcription factors in myeloid cells. Eight time points with a difference of 3h up to 24h between the time points allowed an investigation of myeloid cell differentiation at a high resolution.

The largest data set from macrophages provided most splits and transcriptional regulators identified by DREMflow. Among the TFs selected for recurrence were EGR1 and SPI1, TFs that described in detail by Ramirez et al. and identified as the main regulators for differentiation into macrophages (Ramirez et al. 2017). Other top regulators were heterodimers of the AP-1 like TFs, including FOSL2::JUND that are associated with differentiation and activation of macrophages (Dekkers et al. 2019).

The neutrophil data resulted in the smallest split node network including only 67 nodes and 40 identified top regulators when the recurring TFs and highest two ranked TFs are combined. There was an overlap of one TF with the results from Ramirez et al. identifying EGR2 as top regulator(Ramirez et al. 2017). Given the identification of other regulators known from literature such as ELK1 (Dong et al. 2022) and AP-1 family members FOSL2 and JUN (D. A. Hume and Himes 2003), DREMflow captures the transcriptional landscape well.

Interestingly Ramirez et al. mentioned STAT6 as transcriptional regulator (Ramirez et al. 2017), although JASPAR only contains a STAT6 motif for mice (https://jaspar.genereg.net search for STAT6).

For monocytes, DREMflow identified overall 69 TFs as top regulators in the network includ-

ing an overlap of 16 TFs with macrophage top regulators. Since macrophages typically arise from monocytes, although in this data set they were directly derived from promyelocyte, more similarities to macrophages than to neutrophils are expected. The 24h time point of the monocyte data set can be considered the turning point in differentiation with major changes in chromatin. While Ramirez et al. describe the first 24h of differentiation as uneventful, DREMflow suggests that most regulation is occurring in this time span with little change in chromatin accessibility needed.

Considering the same origin in HL60 promyelocyte cells similarities between the three cell lines are expected, but DREMflow still captures individual details such as ELK1 as top regulator for neutrophils, EGR1 and EGR3 in macrophages and the early transcriptional regulation in monocytes. Overall, DREMflow identified regulators that were highlighted in the analysis by Ramirez et al, but also selected other TFs whose role in myeloid differentiation was supported by literature.

6. Comparison of DREMflow to TimeReg

Duren et al. developed two methods that together infer time point specific transcriptional regulators from paired chromatin accessibility and gene expression data. PECA2 identifies a trans regulation score (TRS) from the ATAC-seq signal (Duren et al. 2017). TF-TG networks based on the TRS connecting TF with TG are divided into sub-networks, referred to as *modules*, by non-negative matrix factorization. These modules are connected over the different time points via driver TFs by TimeReg (Duren et al. 2020). Via PECA2, known and novel motifs are identified with Homer and scored according to the signal from the chromatin accessibility data. However, the list of included TFs to build the GRNs is based on a pre-selected list of TFs that is overlapped with the expression data to include only expressed TFs. An explanation about the pre-selected list is missing.

Duren et al. provide mouse neuronal differentiation that was induced by retinoic acid treatment (Duren et al. 2020). For the comparison to PECA2 and TimeReg this set was also analyzed with DREMflow.

In contrast to the previously investigated data sets with the exception of non-dopaminergic neurons that were targeting differentiation into a specific cell type, this data set provides a neuronal cell mix and does not present the expected data type. The performance of DREMflow on a resulting cell mix instead of a targeted cell type is evaluated.

The comparison between DREMflow and TimeReg was done on two data sets. First the

data set from Duren et al. was analyzed with DREMflow and the identified TFs were compared directly to the publication (Duren et al. 2020) as well as literature. Second a comparison was done on data that had no prior connections to either DREMflow or TimeReg. The macrophage data set from Ramirez et al. was chosen, because macrophage differentiation is well studied and a list of 60 TFs known in the context to macrophage differentiation and function was identified from literature (Ramirez et al. 2017; Dekkers et al. 2019; D. A. Hume and Himes 2003; David A. Hume, Summers, and Rehli 2016; Jego et al. 2014; H. Li et al. 2018; Nagamura-Inoue, Tamura, and Ozato 2001; Pundhir et al. 2018).

6.1. Identification of distinct clusters and driver regulators in neuronal cell mix

Duren et al generated and analyzed a time series of five time points from differentiation of mESC induced by RA treatment. Samples were taken at day 0 (mESC), and after two, four, 10 and 20 days (d2, d4, d10, d20). Genes with at least a two-fold expression change and a maximum expression level greater than 10 were included in their analysis. Identified were distinct modules enriched for response to retinoic acid at d2. The modules of d20 enriched neurons, mesodorm and endoderm, neural stem cell and glia. Driver TFs were only identified for the transition from mESCs to day 2 and day 4 to day 10 samples and are listed in Table B.2.

Table 6.1.: Drivers identified by TimeReg in the RE-induced mESCs differentiation

Transition	Driver TFs
mESC to d2 module 1	HOX, PBX, POU3F2/3, ASCL1, NR2F1, RARB
mESC to d2 module 2	RXRA, GATA4/6, SOX17, FOXA2
mESC to d2 module3	PAX6, HOXA1, DBX1, IRX3, PBX2
d4 module 3 to d10 module 4	OLIG1, OLIG2, SOX10, SOX8

6.1.1. DREMflow highlights E2F transcription factors in neurogenesis

The same time series data of the mESC derived neuronal cells was analyzed by DREMflow to investigate if the same main regulators can be identified by both methods. In comparison to the previously analyzed data sets, the combined PWMs from 2022 were used since TEPIC only identified TF-gene links for 21 TFs using the version 2.1 PWMs from JASPAR (results not shown).

The analysis resulted in a 45 node comprising split node network that included approximately 5500 target genes. Since Duren et al. set a threshold of log2FC >2, the same was applied in DREMflow. 23 nodes after a split were identified, resulting in the same number of target gene clusters in the network with those of later time points being subsets of the clusters in earlier time points (Figure 6.1). Overall, 45 TFs were identified as recurring top regulators in among the highest ranked 15 significant TFs. There was no overlap with the main regulators from Duren et al. REST and NANOG were ranked first on node 2 and node 3. E2F members were the highest ranked regulators on node 14, node 16 and node 35. E2F4 was assigned to five nodes. Interestingly, Retinoic Acid Receptor Alpha (RARA) was not identified as top regulator despite the differentiation being induced by retinoic acid. It was ranked 12th on node 6. The number of identified TF-gene links for RARA was one of the lowest with 8540. In comparison, E2F2 and E2F4 had approximately 40000.

Despite the identification of different top regulators as Duren et al, GO enrichment identified similar functional gene clusters. Node 1 and node 4 were enriched for *synapse organization*, node 2 for *axonogenesis* and node 3 for *nuclear division* and *chromosome integration*. Node 7, one of the sub-clusters from node 1, was enriched for *negative regulation of the immune system* and node 18 resulted in *gliogenesis* and *glial cell differentiation*, which is comparable to module 4 from Duren et al. (Duren et al. 2020).

GO enrichment of target genes of identified top regulators showed that most TG linked to E2F transcription factors were enriched for *synapse organization*, suggesting an important role of these TF in neurogenesis. Targets of REST, SP2, TCFL5 and ZBTB33 were

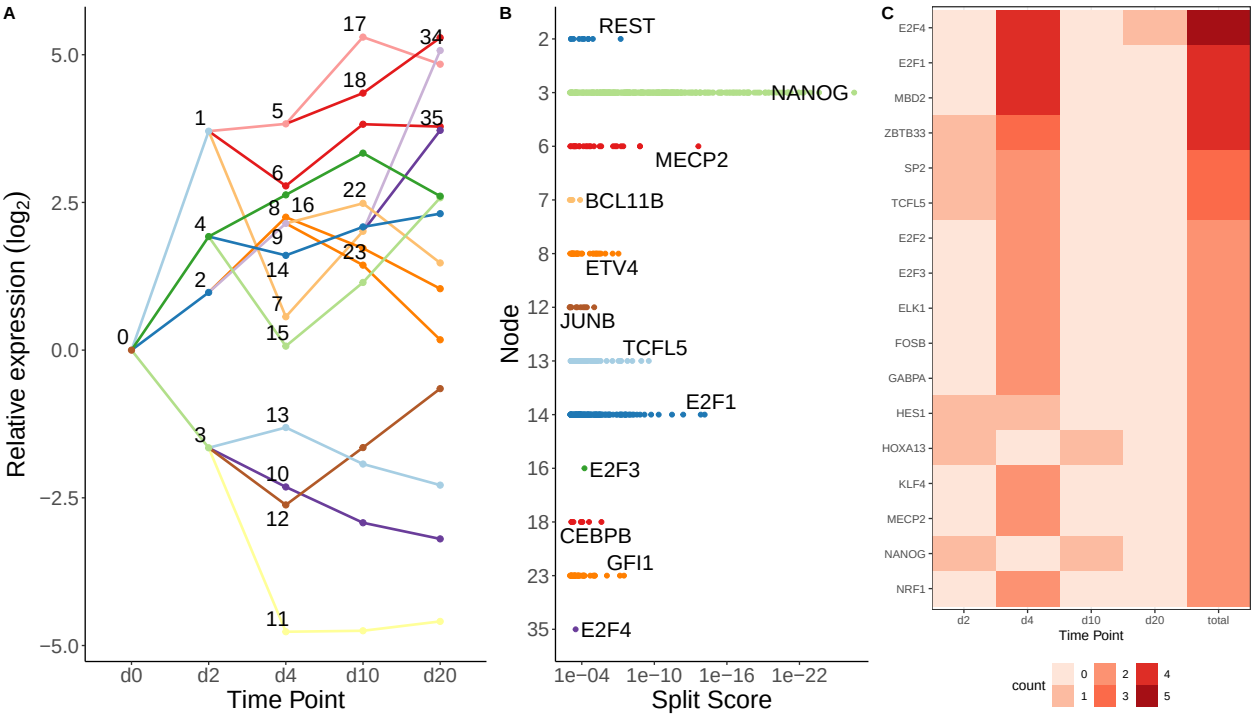


Figure 6.1.: Mouse ESC neuronal fate differentiation. (A) Split node network identified by DREMflow for RA-induced mESC differentiated into a neuronal fate. (B) Identified significant regulators responsible for bifurcation events for each node displayed according to their split score. (C) Recurring top regulators and their number of appearances at each time point and overall in the network.

enriched for *axonogenesis*.

The important role of E2F TFs in neurogenesis especially early during differentiation is supported by several studies (D. X. Liu and Greene 2001; Ghanem et al. 2012; Fong et al. 2022; Julian et al. 2016). NANOG is a known pluripotency factor (Takahashi and Yamanaka 2006). Identifying NANOG early on shows the high share of pluripotent cells that undergo self-renewal instead of differentiation.

6.2. TimeReg identifies JUN as driver TF for the macrophage data set

After comparing the results on the data set from Duren et al, the macrophage data was selected to serve as neutral data for comparing the two pipelines. For this, PECA2 and TimeReg were used to analyze the data from Ramirez et al (Ramirez et al. 2017). Since the preprocessing is not included in PECA2, the aligned ATAC-seq files and the identified TPM were taken from the intermediate output from DREMflow. After previously adjusting the default DREM settings to match the filtering described by Duren et al. to only include genes with a log2FC>2, in this comparison the settings for PECA2 and TimeReg were matched to the default DREM setting, meaning the cutoff was at log2FC>1.

Application of TimeReg to the macrophage data resulted in a sparse TF-TG heatmap (Figure A.10). Out of approximately 3 million possible connections between the 335 TFs and 900 TG, only 97000 connections were found with a TRS score > 0 (Figure A.10). Since the number of connections did not compare to the results shown in Duren et al. (Duren et al. 2020), TimeReg was applied to all myeloid data sets which resulted in equally sparse heatmaps at each time point (data not shown). To ensure the correct execution of PECA2 and TimeReg, the mDAN data set from Ramos et al. and the RA-induced mESC data from Duren et al. were analyzed as well with PECA2 and TimeReg (Duren et al. 2020; Ramos et al. 2023). The resulting TF-TG heatmap for day 15 of the mDAN data set

displayed approximately 3.5 million connections with a TRS > 0 out of 6.7 million possible connections.

The execution of TimeReg required the user to specify the number of modules to be identified at each time point. Based on Duren et al and the use of different numbers of modules, the final setting was two modules at 3h and three modules at any other time point. The module network identified by TimeReg is shown in Figure 6.2 A. Since DREM explains major expression changes with bifurcation events, the number of clusters identified at later time points can increase exponentially, resulting in more clusters than TimeReg. In macrophage differentiation all splits resulted in 24 clusters at 120h.

TimeReg infers the developmental trajectory between the modules and assigns driver TFs as important regulators that connect the modules. Drivers are inferred only considering upregulated gene and the TF-TG scores for upregulated genes. For the macrophage data set, driver TFs were identified for only three modules. Connecting 3h-2 and 6h-1, six driver TFs were assigned, from 12h-1 to 24h-1, TOX2 was identified and from 24h-2 to 48h-2, 13 driver TFs were assigned, including JUN, ATF3, FOXO1, SOX4 and BHLH41.

JUN functionally cooperates with SP1 and NFKB to activate the promotor of the macrophage inflammatory protein-2 (MIP-2) (K.-W. Lee et al. 2005), ATF3 is an important TF in the response to interferon in macrophages (Labzin et al. 2015), FOXO1 is required to maintain the functionality (Yan et al. 2020), SOX4 overexpression in myeloid differentiation can lead to oncogenic activity (H. Zhang et al. 2013) and BHLH41 was connected to the self-renewal of alveolar macrophages (Rauschmeier et al. 2019).

TimeReg identified known regulators in but not only specific to myeloid differentiation.

6.2.1. Intersection with identified regulators from DREMflow

Of the 437 TFs in the DREMflow setup and the 335 TFs in TimeReg, 173 TFs are shared by both. Among those shared are JUN, HOXA2 and ZBTB7C were identified as top regulators

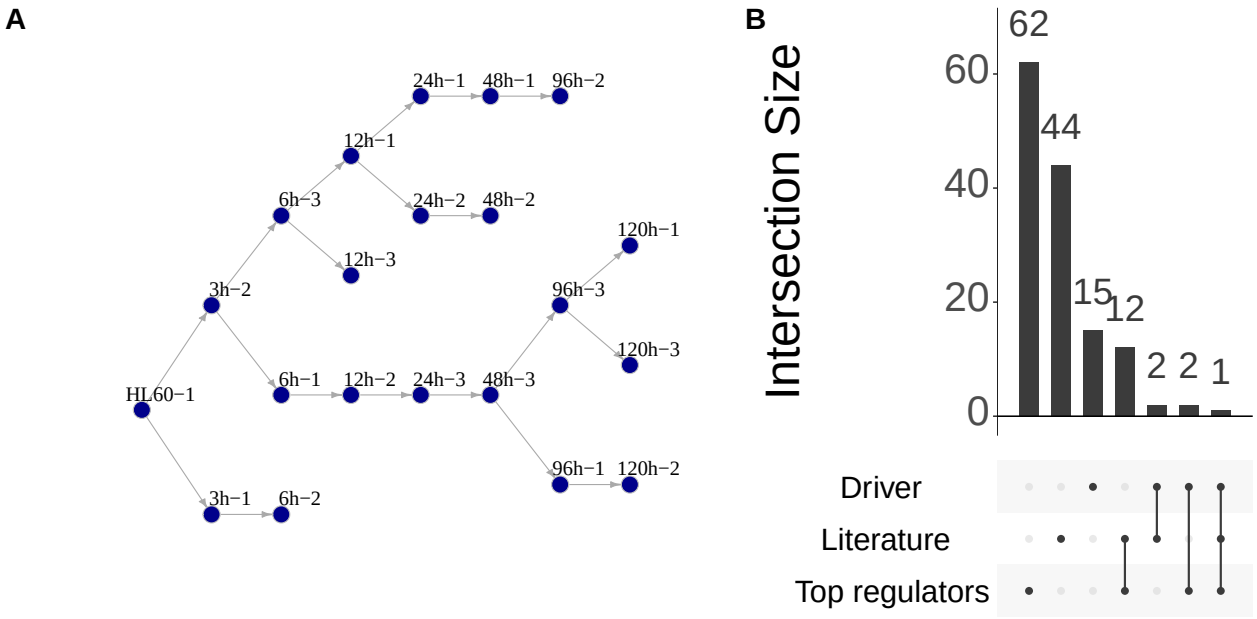


Figure 6.2.: Comparison of DREMflow and TimeReg. (A) Ancestor-descendant mapping from TimeReg on the macrophage data. The number after the time point represents the identified module. (B) Intersection TFs between TimeReg and DREMflow and TFs from the literature list. Heterodimers are excluded in this graphic. Drivers are associated with TimeReg while Top Regulators are associated with DREMflow.

by DREMflow and driver TFs by TimeReg. Since DREMflow includes heterodimers as well, that include JUN, the overlap could be extended to five TFs.

Instead of focusing solely on the overlap between the two methods, a list of 60 TFs related to macrophage differentiation, proliferation and activation was identified from literature (Section 3.4.2) and the list of identified top regulators as well as the list of identified driver TFs was compared to the literature derived set of TFs. If TFs from literature also occur as heterodimers, 73 TFs from DREMflow overlap with the list. From these 73 TFs, 22 were identified as top regulators, resulting in a sensitivity of 0.3 for DREMflow. The overlap from TimeReg TFs with the literature list is 53, with an overlap of three driver TFs, resulting in a sensitivity of 0.075. The difference in proportion of true positives between DREMflow and TimeReg was significant (Fisher's exact test, p-value = 0.0081).

Despite both methods identifying top regulators in myeloid differentiation the overlap is small with five TFs overall. One reason is the identification of motifs to create a TF set to include in the analysis. TimeReg uses Homer, that takes motifs from Homer, Kellis and JASPAR to create a set of TFs for each data set. DREMflow takes the motifs from JASPAR only. Considering the inclusion of several databases, the lower number of TFs included in TimeReg compared to the set of TFs in DREMflow is surprising.

Since different names are used for the same TFs across databases and TFs from the same TF family can have similar motifs, it is possible that the overlap between the overall number of TFs included might be greater than identified in Figure 6.2 B.

6.2.2. DREMflow offers flexibility in comparison to TimeReg

Comparing both tools in terms of execution and usability, DREMflow offers flexibility and detailed visual output to increase the interpretability of the results, while TimeReg provides a very clear short list of driver regulators. Looking at descriptive features, DREMflow not only provides preprocessing and downstream analysis but also avoids compatibility problems with diverging reference genome and gene annotation versions (Table 6.2).

DREMflow is programmed exclusively in open source programming languages, TimeReg is mostly written in Matlab and therefore requires the purchase of a license. The installation for DREMflow, especially in the context of a HPC cluster, is kept to a minimum and can be performed without administrator support or special access permissions. TimeReg requires the installation of Homer, Matlab, Bedtools and Samtools (Duren et al. 2020). Although often bioinformatic tools are provided on HPC clusters as modules that can be loaded by the individual user, Homer had to be installed manually on the ULHPC system. Duren et al. added a module load statement in a later version to the PECA2 script to account for the tool loading. This statement proved to be difficult, since the module loading differs between systems and this unspecific statement would cause errors during execution. Given the length of the script and the execution time of PECA2 with up to 5h for a sample, an error at the end of the script was time costly. A rerun was only possible from the beginning, unless the user manually disassembles the script into different parts for each step performed by PECA2.

Table 6.2.: Comparison between DREMflow and PECA2/TimeReg based on different features.

Features	DREMflow	PECA2/TimeReg
Preprocessing	FastQC, Alignment DESeq2, featureCounts	Not included
Reference genome and gene annotation	User specified version	Specific version only
Module and node assignment	All TG and some TFs by significance	All TFs and all TG Few driver TFs

Features	DREMflow	PECA2/TimeReg
Output	DREM model for GUI Results HTML Downstream analysis Top regulators GO enrichment TF-TF networks Visualization	Modules TRS heatmaps Driver regulators Ancestor-descendant relationship
Programming languages	Shell, Python, R	Shell, Matlab
Installation requirements	Snakemake, conda and mamba	Homer, Matlab, Bedtools, Samtools
Rerun	From any point of the pipeline	From the beginning
Flexibility	Addition of rules Replacement of rules	No flexibility

Application of PECA2 and TimeReg to different data sets revealed that the setup of Homer that is necessary for the analysis requires the deletion of intermediate files after an analysis created by Homer and a repeated run of the preparation step for each data set.

In terms of computational performance both methods are comparable. The preprocessing includes the most time consuming steps such as reference genome indexing and alignment, which are therefore the same in both approaches. The TF-TG score calculation by PECA2 was running for up to 5h including the peak calling and footprinting steps. For the analyzed data sets, DREMflow performed faster on the execution of those steps including TEPIC, that is the equivalent to PECA2 TF-TG score identification. The run times of DREMflow are discussed in Chapter 8 in detail. TimeReg outperforms iDREM in terms of speed with only a few minutes until the GRN is computed compared to 20 min for the macrophage data set. Taken all steps together there is not noticeable difference in the

overall execution time. The advantage that DREMflow offers is the specific cluster setting for each rule and the flexibility in terms of rerunning the pipeline or changing parameters and simplified exchange of rules by more experienced users.

6.3. Discussion

DREMflow and TimeReg were compared with regards to biological relevance on two data sets and computationally based on the features and computational requirements. The first data set was the data set TimeReg was built on was on a Mouse embryonic stem cell differentiation that resulted in a neuronal cell mix. While TimeReg achieved an identification of four distinct modules at the latest time point with clear separation in function and cell types, DREMflow could not reproduce the same results. The clustering of co-expressed genes and the subsequent GO enrichment identified distinct clusters for synapse organization, axogenesis but also immune related clusters. These results are comparable to the TimeReg identified modules, although Duren et al. were able to identify the immune related cluster more precisely as glial cells.

Despite identifying neurogenesis related TFs in the model, DREMflow did not include RARA among top recurring regulators, most likely to the low number of TF-gene links identified from the ATAC-seq data. Since RARA was the only retinoic acid related TFs included in the PWMs, this posed another limit to observe retinoic acid related TFs among the top regulators.

The identified top regulators had an overlap of one TF. The TF-gene link provided only 21 TFs overall for the computation of the split node network. This was clearly a problem related to TEPIC. The threshold calculation for EPIC-DREM was designed for ChIP-seq data (Gérard et al. 2018). The application to ATAC-seq resulted in an overestimation of TF-gene links that was fixed with the adjustment of the p-value, leaving it up to the user to decide on the stringency of the this cutoff. This creates a seemingly arbitrary selection of TF-gene links, if the p-value for two different data types are not comparable.

The use of PWMs version 2.1 from 2018 might create another problem. Motifs are constantly updated, if new data is available. Only identifying interactions for 21 TFs suggests that the mouse PWMs from 2018 are outdated.

In addition, the RA-induced mESC data was the only mouse data set and it was a neuronal cell mix, suggesting that DREMflow either does not capture mouse data well or has difficulties identifying top regulators in a cell mix.

The macrophage data chosen as neutral data set for comparison provided favorable results for DREMflow. PECA2/TimeReg were not able to identify many high TRS, which are required for the division into modules and the assignment of driver TFs. Interestingly the number of TFs included in TimeReg was lower although the identification of motifs was based on several data bases including JASPAR, Homer and Kellis instead of only JASPAR like DREMflow. This difference can be explained by the chosen peak calling method. DREMflow included long peaks, while the peak calling with MACS2 in PECA2 is restricted to small area peaks (Duren et al. 2017, 2020).

Despite the sparse TRS heatmap TimeReg did well in identifying 20 TFs from the data as drivers although the overlap with macrophage related TFs from literature was low.

The approach of driver identification has some flaws, since it is only focusing on upregulated genes. Driver TFs are transcriptional regulators that have a significant higher TRS for upregulated genes than on non-upregulated genes and driver TFs have to be upregulated at least 1.5 fold between the time points (Duren et al. 2020). Investigating transcriptional regulation only on upregulated genes is restrictive. This restriction is reflected in the sensitivity of 0.04 of TimeReg identifying only three known TFs taken the literature derived list as gold standard.

DREMflow performed significantly better identifying 23 TFs which resulted in a sensitivity of 0.27. While the results from a biological perspective capture the macrophage differentiation better than TimeReg, DREMflow has major advantages computationally as well.

Preprocessing and postprocessing of the data are included, providing quality control measures and interpretable results. PECA2 includes peak calling but does not provide information about the number of peaks identified in the data or the number of footprints or motifs. TimeReg provides more output files but lacks the downstream analysis that was presented by Duren et al. such as GO enrichment for the most specific genes for each cluster. Unfortunately the TF-TG heatmap is only provided as PNG and not as table with specific scores.

In comparison, DREMflow provides a comprehensive output HTML with quality control measures, downstream analysis and a short list of transcriptional regulators including GO enrichment and visualization of the GRN as directed graphs. Considering the minimal installation requirements and the flexibility, DREMflow clearly outperforms PECA2/TimeReg.

7. Adjustment of DREMflow to other experimental setups

As described in the section Section 3.3, the input criteria for DREMflow are strict, requiring paired chromatin accessibility data for at least three time points. The implementation of DREMflow in Snakemake allows for adjustment to other input types. In this chapter an example of adjustment was shown.

DREMflow is applied to analyze the erythropoiesis data from Ludwig et al. that provides a pseudo time series acquired by FACS sorting instead of actual time points as input (Ludwig et al. 2019). While this data did not require computational adjustments of the pipeline, it is representative for a possible integration of single cell data using a pseudo time series identified from individual cell types.

7.1. Using erythrocyte differentiation stages as pseudo time points

Instead of creating a time series through sampling at several time points, Ludwig et al. performed FACS sorting to identify cell populations and built a pseudo time series according to the cell types found during the stages of erythropoiesis. The identified cell types are described in Table 3.2 in Section 3.3.4. Ludwig et al. identified co-expressed gene with k-means clustering, identifying seven cluster in total, and performed GO enrichment on

the gene clusters resulting in a *DNA processing*-related cluster that includes early upregulated genes and *heme biosynthetic* processes for genes that were upregulated in later stages. They identified GATA1, TAL1 and KLF1 as main regulators of erythropoiesis but also emphasized the high accessibility found for SPI1, CEBPA, RUNX1, FOSL1, NFE2 and FOXD4.

While this data set does not fulfill the requirements for DREMflow per se, by assuming that the erythropoiesis stages and known erythrocyte progenitor cell types can be used to build a time series, the data set can be used as input for DREMflow without any need for adjustment or extension of rules. This presents an interesting opportunity to test DREMflows ability to infer transcriptional regulators with cell-type specific TF-gene links on a large data set.

7.1.1. AP-1 family heterodimers identified as top regulators in erythropoiesis

The time course data of human erythropoiesis with eight differentiation stages resulted in a split node network of 105 nodes with 10843 included target genes. Overall the model contains 59 nodes after a split (Figure 7.1) with 71 TF defined as top regulators for recurrence and is the largest model identified with DREMflow.

FOXP3 together with the heterodimers FOSL2::JUNB, FOSL2::JUND and FOSB::JUNP appeared five times over the whole network among the ten highest ranked significant regulators (Figure A.11 A). While AP-1 family members are known regulators for myeloid differentiation, FOXP3 is specifically found in T cells and not erythrocytes (Lord et al. 1992; Santoni de Sio et al. 2017; B-H Yang et al. 2016). Among TFs appearing four times more heterodimers that include TFs of the AP-1 family were found; FOS::JUN, FOSL1::JUND and JUN::JUND. FOX related TFs and GATA related TFs appeared three times among the ten highest ranked significant regulators.

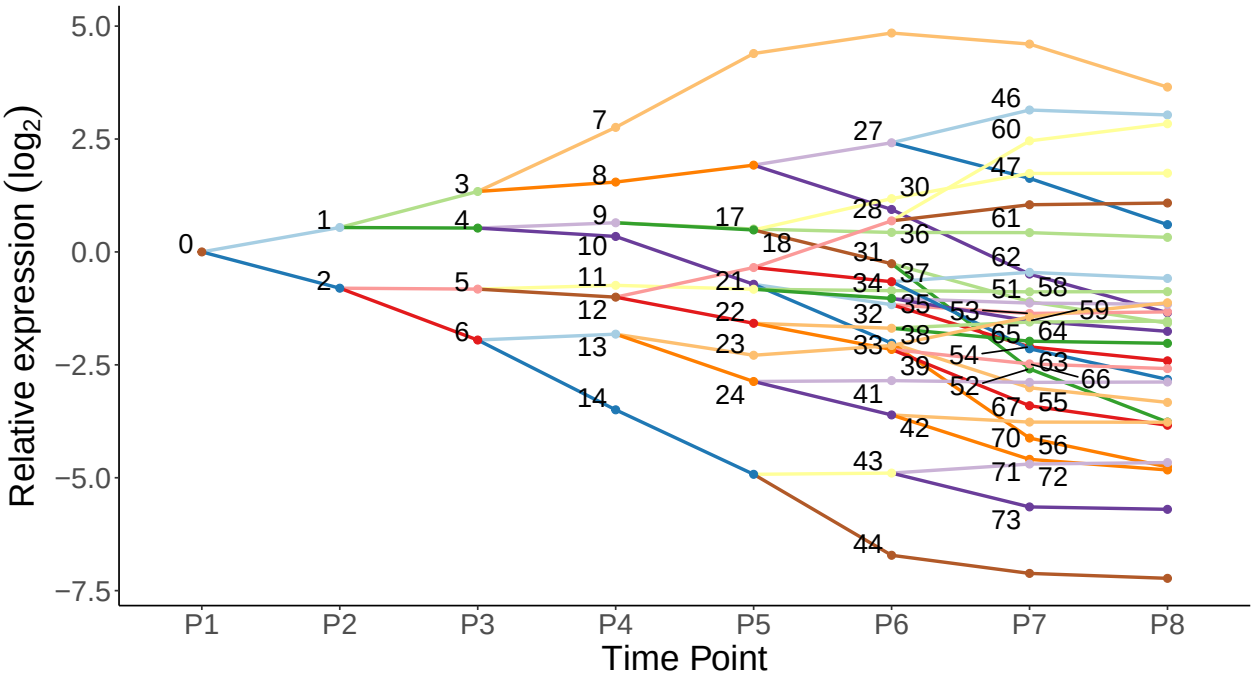


Figure 7.1.: Erythropoiesis split node network comprising 114 nodes for the erythrocyte pseudo time series differentiation data. The time points are cell populations identified through FACS sorting representing differentiation stages.

Given the high amount of recurring TFs, 71 overall, the selected transcriptional regulators can be separated into three groups according to their expression profiles; Upregulated TFs, downregulated TFs and TFs that show stable expression, so no noticeable expression change over all stages. The last group included 17 TFs that showed a stable expression over differentiation stages with almost no upregulation or downregulation visible. Among those were GATA6, GATA3, EGR4 and FOXO6. TPM values show that they are barely expressed at all with TPM < 0.5. FOXO3 and FOXO4 are clearly upregulated at later stages (log2FC >2) while GATA2 and BATF are downregulated (log2FC < -6). GATA1 and TAL1, that are assigned as heterodimer GATA1::TAL1 twice among the ten highest ranked significant regulators, were grouped with the upregulated genes (Figure A.11 B).

The number of target genes The highest number of target genes was identified for EGR2, ZBTB33, TFDP1 and NRF1 and ETS family members with over 6500 at P1. Overall a decrease of number of predicted target genes was observed over the progression of differentiation stages (Figure A.11 C). Interestingly, the number of TF-gene links for heterodimers

that include AP-1 TFs is low over all with only 4500 at P3 and 3000 or less at other stages. This data suggests that these heterodimers are ranked high because they regulate a high number of relevant target genes.

GATA1::TAL1 was identified as top regulator by DREMflow, identifying on of the factors highlighted by Ludwig et al. (2019). KLF1 was not found to be a regulator, because the KLF1 motif was not included in the analysis.

7.1.2. Target genes for AP-1 TFs were enriched for myeloid cell differentiation

Gene ontology enrichment on the differentially expressed genes of each node resulted in general myeloid related terms in downregulated nodes (Figure 7.2). Node 2, node 43 and node 73 are enriched for *lymphocyte*, *leukocyte* and *myeloid cell differentiation*, while node 3, node 7, node 8 and node 24 show *heme synthesis* specific terms (Figure A.12). The only significant TFs assigned to node 7 and node 8 are LEF1 and heat shock factor (HSF2). Both are not recurring but ranked first. Node 7, an upregulated node, is enriched for *erythrocyte differentiation* and *homeostasis*. Genes assigned to node 13 are enriched for *T-cell differentiation,* which could explain FOXP3, a T-cell specific TF, being ranked highly in the network.

GO enrichment on specific target gene clusters of identified top regulators showed that *myeloid differentiation* is enriched in all FOS/JUN heterodimers. GATA1::TAL1 target genes are enriched for *myeloid cell differentiation* as well and *regulation of hemopoiesis* (Figure 7.2). Target genes of the ETS family members such as ELK1, ELK3, ETV1, ETV4, ETV5, ELF4 were enriched for ncRNA processing and DNA replication. GO enrichment of LEF1 target genes resulted the terms *hemostasis* and *coagulation* while for HSF2 *lymphocyte differentiation* and *B cell activation* were found.

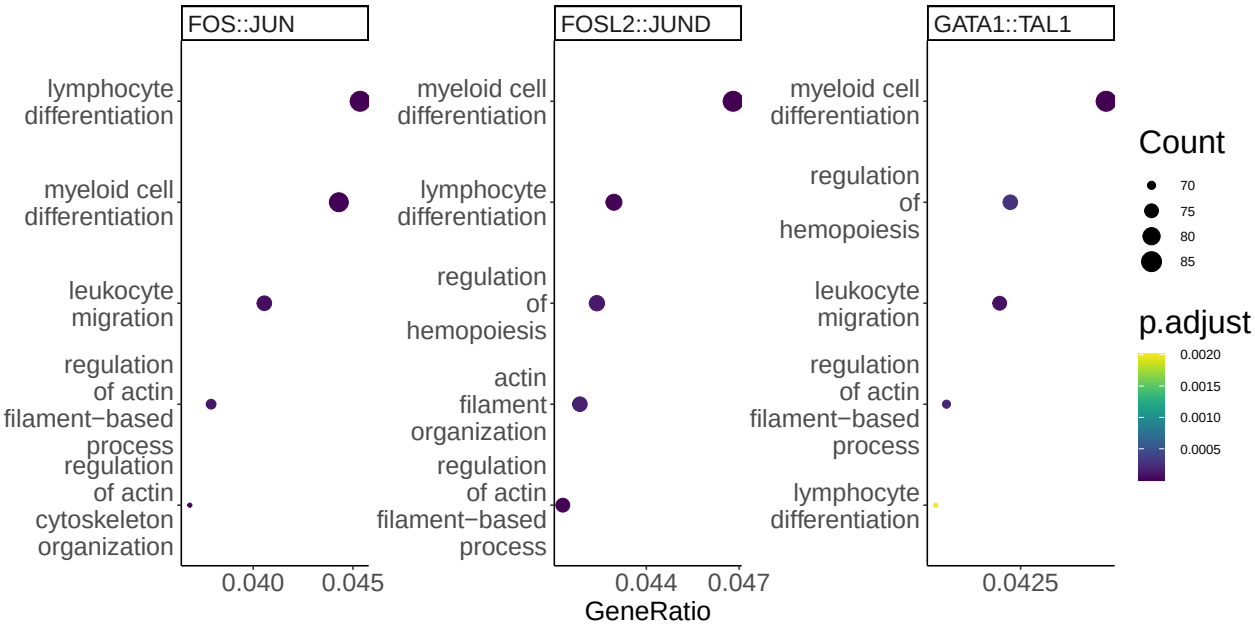


Figure 7.2.: GO enrichment on target genes heterodimers FOS::JUN, FOSL1::JUND and GATA::TAL1.

7.2. Discussion

An alternative experimental setup is paired single-cell data from one time point only and the generation of a time course like data set through inference of a pseudo time series from the identified cell types. The FACS sorted erythropoiesis data from Ludwig at al. demonstrated that DREMflow can identify important regulators from individual cell types instead of a time series. The FACS sorting approach has an advantage over single-cell data, since single-cell ATAC-seq often results in sparse count matrices (H. Chen et al. 2019). The FACS sorted pseudo time series provided high quality heterogeneous data, which resulted in a complex split node network with 71 TFs assigned. The identification of AP-1 family TFs as key regulators is supported by their well known role as transcriptional regulators in the context of myeloid differentiation (Lord et al. 1992). Given the different approaches, Ludwig et al. identifying cluster with k-means clustering and DREMflow with IOHMM, the number of clusters for GO enrichment was different. Although there was an overlap with the results from Ludwig et al, e.g. the identification of nodes being enriched for heme synthesis, overall the GO enrichment for all nodes showed repetitive results with

terms like *myeloid differentiation* appearing often, especially in downregulated nodes. The enrichment for *heme synthesis* and *erythrocyte differentiation* in upregulated nodes suggests that the key regulators specifically for erythropoiesis can be found assigned to those nodes. While HSF2 was found to have an either activating or repressing effect, depending on the ratio of its spliced isoforms (Leppä et al. 1997), LEF1 was found in the context of FOXP3 positive T regulatory cells (Bi-Huei Yang et al. 2019). A possible explanation for the prominence of T cell related factors in the network could be the ability of human red blood cells to regulate neighboring cells like T cells and modulate their growth (Antunes et al. 2011; Arosa, Pereira, and Fonseca 2004). LEF1 could still be an interesting candidate in the context of erythropoiesis.

Overall the results captured erythropoiesis well, with GATA1,TAL1 and AP-1 family TFs among the key regulators and suggesting to investigate LEF1 in with regards to erythrocyte differentiation.

Extending this experimental setup in the context of single-cell data, it was shown that it is possible to infer the gene regulatory network from pseudo time series data. Already preprocessed single-cell data with an inferred pseudotime series can be analyzed in the same manner as the FACS sorted erythropoiesis data set. While DREMflow would benefit greatly from an extension to infer pseudo time series from single-cell experiments on its own, the analysis demonstrated that it is not only limited to bulk sequencing data.

8. Computational performance of DREMflow on differentiation data sets

The value of a computational pipeline for high-throughput sequencing data does not only depend on the biological relevance of the results but also the computational performance handling several 100 GB of data, intermediate data and resources needed for each step. This data load requires high performance computing and a framework that enables to efficiently use of the resources available. DREMflow is implemented to run on HPC clusters with the slurm workload manager (Yoo, Jette, and Grondona 2003). In case another system is used, Snakemake provides a selection of profiles to submit pipelines with dependencies, including a generic profile that works on almost any HPC cluster (https://github.com/Snakemake-Profiles).

The application to several differentiation data sets revealed strengths as well as limitations of DREMflow that can be taken as reference to guide further development of the pipeline and future perspectives.

8.1. Runtime and memory

High performance computing clusters use workload management systems that schedule jobs according to resources and priority. If no resources are available, the jobs are assigned to a queue. Without queuing DREMflow can analyze the myeloid data sets with

eight time points in triplicates on an HPC cluster in under 10 hours (Figure B.8). The cumulative runtimes consider the maximum runtime from all samples for each step without queuing time and the required memory represents the maximum RAM from all samples for each step. For visualization purposes, some steps that are run in parallel are shown sequentially while others are omitted in the graphic. The differential footprinting across all samples that identifies time point specific activity at TFBS is executed in parallel to all steps following the regular footprinting. Since it is a time consuming step and required for the results, it was added in the graphical overview and the cumulative runtime. The alignment of RNA-seq was omitted. In all analyzed data sets the RNA-seq alignment with STAR that runs in parallel to the ATAC-seq alignment with BWA was faster with the longest runtime being 30 min.

The most time consuming step was the ATAC-seq mapping while the indexing of the reference genome and TEPIC depending on the data set required the most memory.

During the analysis of the monocyte data, the 24h sample required more memory than the other samples with 96GB over four cores. Using eight cores which is the default setting in DREMflow, TEPIC was running until the set time limit for the rule was exceeded but did not provide results or a memory related error message. Running the sample with four cores did only provide results in 50% of the cases, so the setting was reduced to two cores only. With two cores results were provided in ten out of ten test runs. The lower number of cores for parallelization explains the difference in runtime for the monocyte data for TEPIC with 30 min instead of 10 min observed in the other myeloid differentiation data sets. Given the total length of the analysis, this difference is neglectable. Surprisingly the inference of TF-gene links with TEPIC for the RA-induced mESC data showed the opposite trend. The longest runtime was 10h although the number of cores was increased to 12 without causing an out-of-memory error. This long runtime was only after the PWMs was changed to the combined set from 2022. Before, the runtime was approximately 20 minutes for TEPIC.

Overall, the analysis of the RA-induced mESC data set was the longest with close to 30

hours. The data presents an outlier in comparison to the other analyzed data sets. The ATAC-seq alignment took 11h and TEPIC 10h. Since this is a mouse data set, it uses a different PWMs input, which could explain the runtime, although the number of included PWMs is lower for mouse than human.

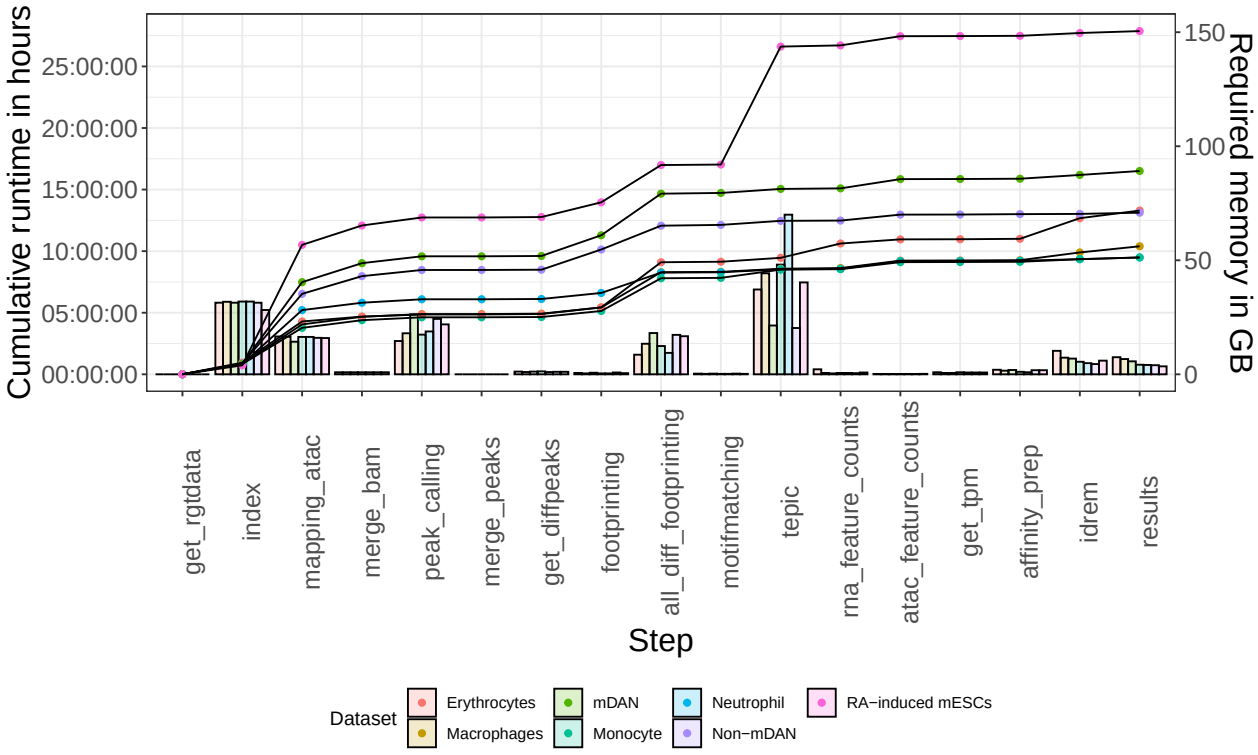


Figure 8.1.: Cumulative runtimes for each rule and maximum memory required. Main rules were considered for the cumulative runtime. This is an approximation, since some rules run in parallel or overlap. The maximum runtime and memory requirements are displayed for the analyses described in the previous chapters.

8.2. Strengths and limitations of DREMflow

While the previous chapters have demonstrated the biological relevance of the results acquired with DREMflow and competitiveness with other methods, it has some major limitations. Before I point out the limitations, first I want to highlight the strength.

DREMflow is a pipeline, not only a tool, and includes every step of a bioinformatics analysis

to achieve the gene regulatory networks identified by iDREM. In addition to the assigned transcriptional regulators, it DREMflow provides a list of recurring highly ranked TFs, their properties such as expression profiles, TF-gene links and activity that support the selection of candidate TFs in the context of the investigated differentiation. Being implemented in Snakemake with mamba environments, it is flexible and provides environments that can be exported and reused for any other data set. The environments are defined without strict version and could potentially result in different dependencies when versions are updated for mamba, but experience during development has shown that this will less likely resulting in dependency issues following those updates. The specific version of each tool installed in an environment is saved in a file generated by Snakemake during the first execution of the pipeline. As in other workflow management systems like Galaxy, rules can be removed or added by adjusting the input and output files, giving an opportunity to include additional steps like extra filtering. The installation of DREMflow is simple by just cloning or downloading the git repository, making it locally usable. That means the user can run DREMflow on any Ubuntu system without the need to upload their data anywhere. This and the generation of symlinks to the raw data files are strengths also in the context of data management and data protection rules.

As mentioned in the beginning of this chapter, efficient execution of a pipeline depends on the handling of resources that are available. DREMflow allows individual settings for each rule that can be optimized by the user for their system.

The results in Section 7.1 demonstrated how pseudo time series data can be used instead of an actual time course. While this presents an option on how to handle single cell data, so far the complete pipeline is only implemented for bulk RNA-seq and bulk ATAC-seq data .

One of the problematic steps is the inference of TF-gene links with TEPIC. The threshold calculation of a cutoff for the scores of TF-gene links was originally done for ChIP-seq data. The adjustment of the p-value for ATAC-seq data is a quick fix to prevent overestimation of TF-gene links, but the adjustment is left to the user and could seem arbitrary.

Despite DREMflow being flexible in terms of changing rules and parameters the number of replicates needs to be the same for all samples to not result in an error in the pipeline. The peak calling step currently depends on this, since the peaks are called over replicates. This can be problematic if one replicate needs to be discarded for quality reasons, resulting in the same replicate being discarded for all time points and reducing the overall number of replicates.

Results depend on the choice of the PWMs set. The PWMs input is restricted due to the later sets only providing combined or clustered PWMs but not a set for a single database. With the decision to only use JASPAR PWMs, as a result, the PWMs were from 2018.

The model from non-mDANs has shown the impact if one time point less and a time span of 35 days between time points. The results from DREMflow, as from any time series data, strongly depend on the chosen time points and distance between time points. More time points are better but costly and many time points in some cases can result in a dense split node network that is difficult to interpret as the data from erythropoiesis has shown (Section 7.1).

8.3. Discussion

The results from the runtime comparison demonstrate that DREMflow is capable of analyzing a data set in under 10 hours in an optimal case and up to 30 hours in for the RA-induced mESC data set. The analysis can take longer if jobs are put to wait by the workload management system, for example if not enough resources are available. This is a reason why memory and CPU requirements should be set for each step of the pipeline and not for the entirety of a bioinformatics workflow. Waiting time can be reduced as well, if only required resources are requested. More cores and memory often come with a longer queuing time due to availability. These cluster setting should be carefully adjusted to the needs like the number of cores provided for TEPIC for different data sets. Requesting more cores means blocking more resources, but if it reduces the time of the computation

significantly, the benefits outweigh the cost. This advantage of individual adjustment is not a property of DREMflow alone but the benefit of implementing pipelines with workflow management systems like Snakemake in general (Koster and Rahmann 2012).

The longest step could still be optimized by exchanging BWA-MEM with BWA-MEM2, an improved version of the alignment tool. Depending on the data the alignment would be accelerated up to 3.1 times (Vasimuddin et al. 2019).

One limitation of DREMflow is that the preprocessing is only implemented for bulk RNA-seq and ATAC-seq, while there are methods already developed to integrate scRNA-seq and scATAC-seq (Kartha et al. 2022; Xu, Begoli, and McCord 2022). However, the implementation in the Snakemake framework enables the use of pseudotime series from preprocessed single-cell data by supplying intermediate input files at the peak calling and gene expression quantification step. In Chapter 7 a full extension to single-cell data was already discussed. The consideration of the experimental setup, either building networks for each cell type or using pseudo time series, play a crucial role in the extension of the pipeline. While the first idea would provide cell type-specific networks over time that could be compared between the individual cell types, a pseudo time series would remove the need for actual time series data and provide and option to analyze already available data sets for paired scATAC-seq and scRNA-seq (Cao et al. 2018; Ranzoni et al. 2021). Implementing both options would provide a pipeline that can handle several different experimental setups.

TEPIC originally being developed for ChIP-seq data created issues with the p-value cutoff during early runs of the pipeline and resulted in the overestimation of TF-gene links. The developers of TEPIC explained this overestimation with the comparison to the background signal and suggested the quick fix with the adjustment of the p-value tp 0.01 instead of 0.05. The lower p-value corrected for the overestimation. An extension of TEPIC to include an ATAC-seq mode would be a future solution. Since the inference of TF-gene links is the crucial step in the pipeline, another consideration would be the replacement with Hi-C data to take into account even long range interactions (Belton et al. 2012).

The choice of PWMs to use for TEPIC is given to the user. The options are limited, since the third and newest PWMs only includes combined or clustered sets. Each database has advantages and limitations and while a combination of all might be the logical choice, many TFs of the same family share similar motifs. Including TFs with many similar motifs can push top regulators down in the ranking. An option for a user supplied PWMs set could accommodate for all preferences with regards to databases and included TF motifs.

9. Conclusion and perspectives

Inference of gene regulatory networks through integration of chromatin accessibility and gene expression high throughput data is often done in ad-hoc fashion and not in defined algorithmic way that can be automated and compared across parameter ranges or different data sets. Gerard et al. presented the machine learning framework EPIC-DREM that provides a statistical integration through DREM, but only included the integration. DREMflow presents a full implementation of EPIC-DREM with Snakemake and mamba. Expanding the analysis to ATAC-seq provided a broader field of use and additional intermediate steps aim to optimize the pipeline. The automated selection provides a list of top regulators in the data set. The results from all analyzed data sets have demonstrated that the selection works well, although it is not able to identify all regulators previously discovered. The motifs might not be included, as seen for KLF1 in erythropoiesis, or other TFs from the same family with similar motifs might have a higher expression, leading to a better ranking.

The comprehensive visualization supported by downstream analysis such as GO enrichment and the generation of TF-TF networks based on ranking can aid the manual inspection of the results. The visualization was chosen in a way to guide the user to identify key regulators in the system a specific nodes and time points. While the automated selection of top regulators provides the user with the most prominent TFs in the network, a manual inspection of the results could possible identify unique regulators in a system.

The comparison to the only other computational method integrating time series bulk ATAC-seq and bulk RNA-seq demonstrated that DREMflow is competitive in terms of biological relevance and appliance. Although the demand in methods for single-cell data is rising,

there are still recent studies combining bulk high throughput sequencing measures, suggesting that DREMflow brings value to the field and the use of a pseudotime series identified from FACS sorting demonstrated that DREMflow can handle different experimental setups, with a possibility to apply the pipeline to pseudotime series identified from single-cell data sets.

In conclusion, DREMflow captures transcriptional regulators in the biological context well and outperforms TimeReg in terms of identification of transcriptional regulators, user friendliness, versatility and access to intermediate results.

For the future, further optimization of the pipeline is planned. One minor improvement will be the inference of TF-gene links for both, combined and clustered PWMs in parallel, resulting in split node networks with individual transcriptional regulators that can be directly compared to networks for TF families. This can help to identify specific TFs from a transcription factor family that might be overlooked when TFs with a similar motif are favored.

Another perspective would be the embedding of interactive graphs in the HTML output with the R package *shiny.* This would enable users to change parameters like the number of highly ranked significant regulators for recurring TFs, the number of appearances to be included as top regulator or the number of TFs in the TF-TF networks.

A major development for DREMflow will be the extension to include preprocessing of single-cell data to compete with the already existing and upcoming methods in the field. This does not require a remake of the whole pipeline since many elements and the environments of DREMflow can be reused. An additional adjustment to accommodate for the sparsity of count matrices and unmatched data will be needed to improve predictions. The inference of pseudotime from single-cell data is a common approach, therefore the extension will likely be in that direction, building up on the results from the application to FACS sorted identified pseudotime series.

References

Abdolhosseini, Farzad, Behrooz Azarkhalili, Abbas Maazallahi, Aryan Kamal, Seyed Abolfazl Motahari, Ali Sharifi-Zarchi, and Hamidreza Chitsaz. 2019. “Cell Identity Codes: Understanding Cell Identity from Gene Expression Profiles Using Deep Neural Networks.” *Scientific Reports* 9 (1): 2342. https://doi.org/10.1038/s41598-019-38798-y.

Ai, Zhichao, and Irina A. Udalova. 2020. “Transcriptional Regulation of Neutrophil Differentiation and Function During Inflammation.” *Journal of Leukocyte Biology* 107 (3): 419–30. https://doi.org/10.1002/JLB.1RU1219-504RR.

Álvarez-Errico, Damiana, Roser Vento-Tormo, Michael Sieweke, and Esteban Ballestar. 2015. “Epigenetic Control of Myeloid Cell Differentiation, Identity and Function.” *Nature Reviews Immunology* 15 (1): 7–17. https://doi.org/10.1038/nri3777.

Ambrosini, Giovanna, Ilya Vorontsov, Dmitry Penzar, Romain Groux, Oriol Fornes, Daria D. Nikolaeva, Benoit Ballester, et al. 2020. “Insights Gained from a Comprehensive All-Against-All Transcription Factor Binding Motif Benchmarking Study.” *Genome Biology* 21 (1): 114. https://doi.org/10.1186/s13059-020-01996-3.

Amemiya, Haley M., Anshul Kundaje, and Alan P. Boyle. 2019. “The ENCODE Blacklist: Identification of Problematic Regions of the Genome.” *Scientific Reports* 9 (1): 9354. https://doi.org/10.1038/s41598-019-45839-z.

Andrews, S. 2010. *FastQC: A Quality Control Tool for High Throughput Sequence Data [Online].* http://www.bioinformatics.babraham.ac.uk/projects/fastqc/.

Antunes, Ricardo F, Cláudia Brandão, Margarida Maia, and Fernando A Arosa. 2011. “Red Blood Cells Release Factors with Growth and Survival Bioactivities for Normal

and Leukemic T Cells." *Immunology & Cell Biology* 89 (1): 111–21. https://doi.org/10.1038/icb.2010.60.

Arenas, E., M. Denham, and J. C. Villaescusa. 2015. "How to Make a Midbrain Dopaminergic Neuron." *Development* 142 (11): 1918–36. https://doi.org/10.1242/dev.097394.

Arenas, Ernest. 2014. "Wnt Signaling in Midbrain Dopaminergic Neuron Development and Regenerative Medicine for Parkinson's Disease." *Journal of Molecular Cell Biology* 6 (1): 42–53. https://doi.org/10.1093/jmcb/mju001.

Arosa, Fernando, Carlos Pereira, and Ana Fonseca. 2004. "Red Blood Cells as Modulators of T Cell Growth and Survival." *Current Pharmaceutical Design* 10 (2): 191–201. https://doi.org/10.2174/1381612043453432.

Ashburner, Michael, Catherine A. Ball, Judith A. Blake, David Botstein, Heather Butler, J. Michael Cherry, Allan P. Davis, et al. 2000. "Gene Ontology: Tool for the Unification of Biology." *Nature Genetics* 25 (1): 25–29. https://doi.org/10.1038/75556.

Atlasi, Yaser, and Hendrik G. Stunnenberg. 2017. "The Interplay of Epigenetic Marks During Stem Cell Differentiation and Development." *Nature Reviews Genetics* 18 (11): 643–58. https://doi.org/10.1038/nrg.2017.57.

Balwierz, Piotr J., Mikhail Pachkov, Phil Arnold, Andreas J. Gruber, Mihaela Zavolan, and Erik Van Nimwegen. 2014. "ISMARA: Automated Modeling of Genomic Signals as a Democracy of Regulatory Motifs." *Genome Research* 24 (5): 869–84. https://doi.org/10.1101/gr.169508.113.

Bar-Joseph, Ziv, Anthony Gitter, and Itamar Simon. 2012. "Studying and Modelling Dynamic Biological Processes Using Time-Series Gene Expression Data." *Nature Reviews Genetics* 13 (8): 552–64. https://doi.org/10.1038/nrg3244.

Basso, Katia, Adam A Margolin, Gustavo Stolovitzky, Ulf Klein, Riccardo Dalla-Favera, and Andrea Califano. 2005. "Reverse Engineering of Regulatory Networks in Human B Cells." *Nature Genetics* 37 (4): 382–90. https://doi.org/10.1038/ng1532.

Behmoaras, Jacques, Gurjeet Bhangal, Jennifer Smith, Kylie McDonald, Brenda Mutch, Ping Chin Lai, Jan Domin, et al. 2008. "Jund is a determinant of macrophage activation and is associated with glomerulonephritis susceptibility." *Nature Genetics* 40 (5): 553–

59. https://doi.org/10.1038/ng.137.

Belton, Jon-Matthew, Rachel Patton McCord, Johan Harmen Gibcus, Natalia Naumova, Ye Zhan, and Job Dekker. 2012. "Hi–C: A Comprehensive Technique to Capture the Conformation of Genomes." *Methods* 58 (3): 268–76. https://doi.org/10.1016/j.ymeth.2012.05.001.

Bencheikh, Laura, M'Boyba Khadija Diop, Julie Rivière, Aygun Imanci, Gerard Pierron, Sylvie Souquere, Audrey Naimo, et al. 2019. "Dynamic Gene Regulation by Nuclear Colony-Stimulating Factor 1 Receptor in Human Monocytes and Macrophages." *Nature Communications* 10 (1): 1935. https://doi.org/10.1038/s41467-019-09970-9.

Berest, Ivan, Christian Arnold, Armando Reyes-Palomares, Giovanni Palla, Kasper Dindler Rasmussen, Holly Giles, Peter-Martin Bruch, et al. 2019. "Quantification of Differential Transcription Factor Activity and Multiomics-Based Classification into Activators and Repressors: diffTF." *Cell Reports* 29 (10): 3147–3159.e12. https://doi.org/10.1016/j.celrep.2019.10.106.

Biedler, J. L., S. Roffler-Tarlov, M. Schachner, and L. S. Freedman. 1978. "Multiple neurotransmitter synthesis by human neuroblastoma cell lines and clones." *Cancer Research* 38 (11 Pt 1): 3751–57.

Bloushtain-Qimron, Noga, Jun Yao, Michail Shipitsin, Reo Maruyama, and Kornelia Polyak. 2009. "Epigenetic Patterns of Embryonic and Adult Stem Cells." *Cell Cycle* 8 (6): 809–17. https://doi.org/10.4161/cc.8.6.7938.

Bohmann, Dirk, Timothy J. Bos, Arie Admon, Tetsuji Nishimura, Peter K. Vogt, and Robert Tjian. 1987. "Human Proto-Oncogene c- *Jun* Encodes a DNA Binding Protein with Structural and Functional Properties Pf Transcription Factor AP-1." *Science* 238 (4832): 1386–92. https://doi.org/10.1126/science.2825349.

Briggs, Robert, and Thomas J. King. 1952. "Transplantation of Living Nuclei from Blastula Cells into Enucleated Frogs' Eggs." *Proceedings of the National Academy of Sciences* 38 (5): 455–63. https://doi.org/10.1073/pnas.38.5.455.

Buccitelli, Christopher, and Matthias Selbach. 2020. "mRNAs, Proteins and the Emerging Principles of Gene Expression Control." *Nature Reviews Genetics* 21 (10): 630–44.

https://doi.org/10.1038/s41576-020-0258-4.

Buenrostro, Jason D, Paul G Giresi, Lisa C Zaba, Howard Y Chang, and William J Greenleaf. 2013. "Transposition of Native Chromatin for Fast and Sensitive Epigenomic Profiling of Open Chromatin, DNA-Binding Proteins and Nucleosome Position." *Nature Methods* 10 (12): 1213–18. https://doi.org/10.1038/nmeth.2688.

Cao, Junyue, Darren A. Cusanovich, Vijay Ramani, Delasa Aghamirzaie, Hannah A. Pliner, Andrew J. Hill, Riza M. Daza, et al. 2018. "Joint Profiling of Chromatin Accessibility and Gene Expression in Thousands of Single Cells." *Science* 361 (6409): 1380–85. https://doi.org/10.1126/science.aau0730.

Carbon, Seth, Amelia Ireland, Christopher J. Mungall, ShengQiang Shu, Brad Marshall, Suzanna Lewis, the AmiGO Hub, and the Web Presence Working Group. 2009. "AmiGO: Online Access to Ontology and Annotation Data." *Bioinformatics* 25 (2): 288–89. https://doi.org/10.1093/bioinformatics/btn615.

Catalanotto, Caterina, Carlo Cogoni, and Giuseppe Zardo. 2016. "MicroRNA in Control of Gene Expression: An Overview of Nuclear Functions." *International Journal of Molecular Sciences* 17 (10): 1712. https://doi.org/10.3390/ijms17101712.

Chen, Huidong, Caleb Lareau, Tommaso Andreani, Michael E. Vinyard, Sara P. Garcia, Kendell Clement, Miguel A. Andrade-Navarro, Jason D. Buenrostro, and Luca Pinello. 2019. "Assessment of Computational Methods for the Analysis of Single-Cell ATAC-Seq Data." *Genome Biology* 20 (1): 241. https://doi.org/10.1186/s13059-019-1854-5.

Chen, Siyuan, Jing Yang, Yuquan Wei, and Xiawei Wei. 2020. "Epigenetic regulation of macrophages: from homeostasis maintenance to host defense." *Cellular & Molecular Immunology* 17 (1): 36–49. https://doi.org/10.1038/s41423-019-0315-0.

Chen, Taiping, and Sharon Y. R. Dent. 2014. "Chromatin Modifiers and Remodellers: Regulators of Cellular Differentiation." *Nature Reviews Genetics* 15 (2): 93–106. https://doi.org/10.1038/nrg3607.

Cheng, Yong, Weisheng Wu, Swathi Ashok Kumar, Duonan Yu, Wulan Deng, Tamara Tripic, David C. King, et al. 2009. "Erythroid GATA1 Function Revealed by Genome-

Wide Analysis of Transcription Factor Occupancy, Histone Modifications, and mRNA Expression." *Genome Research* 19 (12): 2172–84. https://doi.org/10.1101/gr.098921.109.

Coulon, Antoine, Carson C. Chow, Robert H. Singer, and Daniel R. Larson. 2013. "Eukaryotic Transcriptional Dynamics: From Single Molecules to Cell Populations." *Nature Reviews Genetics* 14 (8): 572–84. https://doi.org/10.1038/nrg3484.

Danecek, Petr, James K Bonfield, Jennifer Liddle, John Marshall, Valeriu Ohan, Martin O Pollard, Andrew Whitwham, et al. 2021. "Twelve Years of SAMtools and BCFtools." *GigaScience* 10 (2): giab008. https://doi.org/10.1093/gigascience/giab008.

Dekkers, Koen F., Annette E. Neele, J. Wouter Jukema, Bastiaan T. Heijmans, and Menno P. J. de Winther. 2019. "Human Monocyte-to-Macrophage Differentiation Involves Highly Localized Gain and Loss of DNA Methylation at Transcription Factor Binding Sites." *Epigenetics & Chromatin* 12 (1): 34. https://doi.org/10.1186/s13072-019-0279-4.

Ding, Jun, James S. Hagood, Namasivayam Ambalavanan, Naftali Kaminski, and Ziv Bar-Joseph. 2018. "iDREM: Interactive visualization of dynamic regulatory networks." *PLoS computational biology* 14 (3): e1006019. https://doi.org/10.1371/journal.pcbi.1006019.

Dobin, Alexander, Carrie A. Davis, Felix Schlesinger, Jorg Drenkow, Chris Zaleski, Sonali Jha, Philippe Batut, Mark Chaisson, and Thomas R. Gingeras. 2012. "STAR: Ultrafast Universal RNA-Seq Aligner." *Bioinformatics* 29 (1): 15–21. https://doi.org/10.1093/bioinformatics/bts635.

Dong, Yong, Yimeng Zhang, Yongping Zhang, Xu Pan, Ju Bai, Yijin Chen, Ya Zhou, et al. 2022. "Dissecting the Process of Human Neutrophil Lineage Determination by Using Alpha-Lipoic Acid Inducing Neutrophil Deficiency Model." *Redox Biology* 54 (August): 102392. https://doi.org/10.1016/j.redox.2022.102392.

Duren, Zhana, Xi Chen, Rui Jiang, Yong Wang, and Wing Hung Wong. 2017. "Modeling Gene Regulation from Paired Expression and Chromatin Accessibility Data." *Proceedings of the National Academy of Sciences* 114 (25): E4914–23. https://doi.org/10.107

3/pnas.1704553114.

Duren, Zhana, Xi Chen, Jingxue Xin, Yong Wang, and Wing Wong. 2020. “Time Course Regulatory Analysis Based on Paired Expression and Chromatin Accessibility Data.” *Genome Research*, March, gr.257063.119. https://doi.org/10.1101/gr.257063.119.

Dzierzak, E., and S. Philipsen. 2013. “Erythropoiesis: Development and Differentiation.” *Cold Spring Harbor Perspectives in Medicine* 3 (4): a011601–1. https://doi.org/10.1101/cshperspect.a011601.

Efthymiou, Anastasia G, Guibin Chen, Mahendra Rao, Guokai Chen, and Manfred Boehm. 2014. “Self-Renewal and Cell Lineage Differentiation Strategies in Human Embryonic Stem Cells and Induced Pluripotent Stem Cells.” *Expert Opinion on Biological Therapy* 14 (9): 1333–44. https://doi.org/10.1517/14712598.2014.922533.

Ernst, Jason, Oded Vainas, Christopher T Harbison, Itamar Simon, and Ziv Bar-Joseph. 2007. “Reconstructing Dynamic Regulatory Maps.” *Molecular Systems Biology* 3 (1): 74. https://doi.org/10.1038/msb4100115.

Evans, M. J., and M. H. Kaufman. 1981. “Establishment in Culture of Pluripotential Cells from Mouse Embryos.” *Nature* 292 (5819): 154–56. https://doi.org/10.1038/292154a0.

Ewels, Philip, Måns Magnusson, Sverker Lundin, and Max Käller. 2016. “MultiQC: Summarize Analysis Results for Multiple Tools and Samples in a Single Report.” *Bioinformatics* 32 (19): 3047–48. https://doi.org/10.1093/bioinformatics/btw354.

Feles, Sebastian, Christian Overath, Sina Reichardt, Sebastian Diegeler, Claudia Schmitz, Jessica Kronenberg, Christa Baumstark-Khan, Ruth Hemmersbach, Christine E. Hellweg, and Christian Liemersdorf. 2022. “Streamlining Culture Conditions for the Neuroblastoma Cell Line SH-SY5Y: A Prerequisite for Functional Studies.” *Methods and Protocols* 5 (4): 58. https://doi.org/10.3390/mps5040058.

Fisher, Amanda G. 2002. “Cellular Identity and Lineage Choice.” *Nature Reviews Immunology* 2 (12): 977–82. https://doi.org/10.1038/nri958.

Fong, Bensun C., Imane Chakroun, Mohamed Ariff Iqbal, Smitha Paul, Joseph Bastasic, Daniel O’Neil, Edward Yakubovich, et al. 2022. “The Rb/E2F Axis Is a Key Regulator of

the Molecular Signatures Instructing the Quiescent and Activated Adult Neural Stem Cell State." *Cell Reports* 41 (5): 111578. https://doi.org/10.1016/j.celrep.2022.111578.

Fu, Xing, Ning Fu, Song Guo, Zheng Yan, Ying Xu, Hao Hu, Corinna Menzel, et al. 2009. "Estimating Accuracy of RNA-Seq and Microarrays with Proteomics." *BMC Genomics* 10 (1): 161. https://doi.org/10.1186/1471-2164-10-161.

Gans, Ian, Ellen I. Hartig, Shusen Zhu, Andrea R. Tilden, Lucie N. Hutchins, Nathaniel J. Maki, Joel H. Graber, and James A. Coffman. 2020. "Klf9 Is a Key Feedforward Regulator of the Transcriptomic Response to Glucocorticoid Receptor Activity." *Scientific Reports* 10 (1): 11415. https://doi.org/10.1038/s41598-020-68040-z.

Gérard, Deborah, Florian Schmidt, Aurélien Ginolhac, Martine Schmitz, Rashi Halder, Peter Ebert, Marcel H. Schulz, Thomas Sauter, and Lasse Sinkkonen. 2018. "Temporal enhancer profiling of parallel lineages identifies AHR and GLIS1 as regulators of mesenchymal multipotency." *Nucleic Acids Research*, December. https://doi.org/10.1093/nar/gky1240.

Ghanem, N., M. G. Andrusiak, D. Svoboda, S. M. Al Lafi, L. M. Julian, K. A. McClellan, Y. De Repentigny, et al. 2012. "The Rb/E2F Pathway Modulates Neurogenesis Through Direct Regulation of the Dlx1/Dlx2 Bigene Cluster." *Journal of Neuroscience* 32 (24): 8219–30. https://doi.org/10.1523/JNEUROSCI.1344-12.2012.

Grealish, Shane, Elsa Diguet, Agnete Kirkeby, Bengt Mattsson, Andreas Heuer, Yann Bramoulle, Nadja Van Camp, et al. 2014. "Human ESC-Derived Dopamine Neurons Show Similar Preclinical Efficacy and Potency to Fetal Neurons When Grafted in a Rat Model of Parkinson's Disease." *Cell Stem Cell* 15 (5): 653–65. https://doi.org/10.1016/j.stem.2014.09.017.

Grüning, Björn, Ryan Dale, Andreas Sjödin, Brad A. Chapman, Jillian Rowe, Christopher H. Tomkins-Tinch, Renan Valieris, and Johannes Köster. 2018. "Bioconda: Sustainable and Comprehensive Software Distribution for the Life Sciences." *Nature Methods* 15 (7): 475–76. https://doi.org/10.1038/s41592-018-0046-7.

Gusmao, Eduardo G., Christoph Dieterich, Martin Zenke, and Ivan G. Costa. 2014. "De-

tection of Active Transcription Factor Binding Sites with the Combination of DNase Hypersensitivity and Histone Modifications." *Bioinformatics* 30 (22): 3143–51. https://doi.org/10.1093/bioinformatics/btu519.

Haas, Simon, Andreas Trumpp, and Michael D. Milsom. 2018. "Causes and Consequences of Hematopoietic Stem Cell Heterogeneity." *Cell Stem Cell* 22 (5): 627–38. https://doi.org/10.1016/j.stem.2018.04.003.

Hager, Gordon L., James G. McNally, and Tom Misteli. 2009. "Transcription Dynamics." *Molecular Cell* 35 (6): 741–53. https://doi.org/10.1016/j.molcel.2009.09.005.

Hamada, Michito, Yuki Tsunakawa, Hyojung Jeon, Manoj Kumar Yadav, and Satoru Takahashi. 2020. "Role of MafB in macrophages." *Experimental Animals* 69 (1): 1–10. https://doi.org/10.1538/expanim.19-0076.

Harrison, Stephen C. 1991. "A Structural Taxonomy of DNA-Binding Domains." *Nature* 353 (6346): 715–19. https://doi.org/10.1038/353715a0.

Hermanson, E. 2003. "Nurr1 Regulates Dopamine Synthesis and Storage in MN9D Dopamine Cells." *Experimental Cell Research* 288 (2): 324–34. https://doi.org/10.1016/S0014-4827(03)00216-7.

Hickl, Oskar, Pedro Queirós, Paul Wilmes, Patrick May, and Anna Heintz-Buschart. 2021. "Binny: An Automated Binning Algorithm to Recover High-Quality Genomes from Complex Metagenomic Datasets." https://doi.org/10.1101/2021.12.22.473795.

Hiller, Benjamin M., David J. Marmion, Cayla A. Thompson, Nathaniel A. Elliott, Howard Federoff, Patrik Brundin, Virginia B. Mattis, Christopher W. McMahon, and Jeffrey H. Kordower. 2022. "Optimizing Maturity and Dose of iPSC-Derived Dopamine Progenitor Cell Therapy for Parkinson's Disease." *Npj Regenerative Medicine* 7 (1): 24. https://doi.org/10.1038/s41536-022-00221-y.

Hor, Charlotte N., Jake Yeung, Maxime Jan, Yann Emmenegger, Jeffrey Hubbard, Ioannis Xenarios, Felix Naef, and Paul Franken. 2019. "Sleep–wake-Driven and Circadian Contributions to Daily Rhythms in Gene Expression and Chromatin Accessibility in the Murine Cortex." *Proceedings of the National Academy of Sciences* 116 (51): 25773–83. https://doi.org/10.1073/pnas.1910590116.

Hume, D. A., and S. R. Himes. 2003. “Transcription Factors That Regulate Macrophage Development and Function.” In, edited by Siamon Gordon, 158:11–40. Berlin, Heidelberg: Springer Berlin Heidelberg. http://link.springer.com/10.1007/978-3-642-55742-2_2.

Hume, David A., Kim M. Summers, and Michael Rehli. 2016. “Transcriptional Regulation and Macrophage Differentiation.” Edited by Siamon Gordon. *Microbiology Spectrum* 4 (3). https://doi.org/10.1128/microbiolspec.MCHD-0024-2015.

Huynh-Thu, Vân Anh, and Pierre Geurts. 2018. “dynGENIE3: Dynamical GENIE3 for the Inference of Gene Networks from Time Series Expression Data.” *Scientific Reports* 8 (1): 3384. https://doi.org/10.1038/s41598-018-21715-0.

Jackson, Michael, Kostas Kavoussanakis, and Edward W. J. Wallace. 2021. “Using Prototyping to Choose a Bioinformatics Workflow Management System.” Edited by Francis Ouellette. *PLOS Computational Biology* 17 (2): e1008622. https://doi.org/10.1371/journal.pcbi.1008622.

Jego, G, D Lanneau, A De Thonel, K Berthenet, A Hazoumé, N Droin, A Hamman, et al. 2014. “Dual Regulation of SPI1/PU.1 Transcription Factor by Heat Shock Factor 1 (HSF1) During Macrophage Differentiation of Monocytes.” *Leukemia* 28 (8): 1676–86. https://doi.org/10.1038/leu.2014.63.

Jin, Haijing, Ying-Wooi Wan, and Zhandong Liu. 2017. “Comprehensive Evaluation of RNA-Seq Quantification Methods for Linearity.” *BMC Bioinformatics* 18 (S4): 117. https://doi.org/10.1186/s12859-017-1526-y.

Johnson, David S., Ali Mortazavi, Richard M. Myers, and Barbara Wold. 2007. “Genome-Wide Mapping of in Vivo Protein-DNA Interactions.” *Science* 316 (5830): 1497–1502. https://doi.org/10.1126/science.1141319.

Julian, L M, Y Liu, C A Pakenham, D Dugal-Tessier, V Ruzhynsky, S Bae, S-Y Tsai, G Leone, R S Slack, and A Blais. 2016. “Tissue-Specific Targeting of Cell Fate Regulatory Genes by E2f Factors.” *Cell Death & Differentiation* 23 (4): 565–75. https://doi.org/10.1038/cdd.2015.36.

Jung, Sascha, and Antonio Del Sol. 2020. “Multiomics Data Integration Unveils Core Tran-

scriptional Regulatory Networks Governing Cell-Type Identity." *Npj Systems Biology and Applications* 6 (1): 26. https://doi.org/10.1038/s41540-020-00148-4.

Karlebach, Guy, and Ron Shamir. 2008. "Modelling and Analysis of Gene Regulatory Networks." *Nature Reviews Molecular Cell Biology* 9 (10): 770–80. https://doi.org/10.1038/nrm2503.

Kartha, Vinay K., Fabiana M. Duarte, Yan Hu, Sai Ma, Jennifer G. Chew, Caleb A. Lareau, Andrew Earl, et al. 2022. "Functional Inference of Gene Regulation Using Single-Cell Multi-Omics." *Cell Genomics* 2 (9): 100166. https://doi.org/10.1016/j.xgen.2022.100166.

Kiani, Karun, Eric M Sanford, Yogesh Goyal, and Arjun Raj. 2022. "Changes in Chromatin Accessibility Are Not Concordant with Transcriptional Changes for Single-Factor Perturbations." *Molecular Systems Biology* 18 (9). https://doi.org/10.15252/msb.202210979.

Kim, Seongho. 2015. "Ppcor: An R Package for a Fast Calculation to Semi-Partial Correlation Coefficients." *Communications for Statistical Applications and Methods* 22 (6): 665–74. https://doi.org/10.5351/CSAM.2015.22.6.665.

Klemm, Sandy L., Zohar Shipony, and William J. Greenleaf. 2019. "Chromatin Accessibility and the Regulatory Epigenome." *Nature Reviews Genetics* 20 (4): 207–20. https://doi.org/10.1038/s41576-018-0089-8.

Knudsen, Kasper Jermiin, Matilda Rehn, Marie Sigurd Hasemann, Nicolas Rapin, Frederik Otzen Bagger, Ewa Ohlsson, Anton Willer, et al. 2015. "ERG Promotes the Maintenance of Hematopoietic Stem Cells by Restricting Their Differentiation." *Genes & Development* 29 (18): 1915–29. https://doi.org/10.1101/gad.268409.115.

Koster, J., and S. Rahmann. 2012. "Snakemake–a Scalable Bioinformatics Workflow Engine." *Bioinformatics* 28 (19): 2520–22. https://doi.org/10.1093/bioinformatics/bts480.

Labzin, Larisa I., Susanne V. Schmidt, Seth L. Masters, Marc Beyer, Wolfgang Krebs, Kathrin Klee, Rainer Stahl, et al. 2015. "ATF3 Is a Key Regulator of Macrophage IFN Responses." *Journal of Immunology (Baltimore, Md.: 1950)* 195 (9): 4446–55.

https://doi.org/10.4049/jimmunol.1500204.

Lachmann, Alexander, Federico M. Giorgi, Gonzalo Lopez, and Andrea Califano. 2016. "ARACNe-AP: Gene Network Reverse Engineering Through Adaptive Partitioning Inference of Mutual Information." *Bioinformatics* 32 (14): 2233–35. https://doi.org/10.1093/bioinformatics/btw216.

Lambert, Samuel A., Arttu Jolma, Laura F. Campitelli, Pratyush K. Das, Yimeng Yin, Mihai Albu, Xiaoting Chen, Jussi Taipale, Timothy R. Hughes, and Matthew T. Weirauch. 2018. "The Human Transcription Factors." *Cell* 172 (4): 650–65. https://doi.org/10.1016/j.cell.2018.01.029.

Langmead, Ben, and Steven L Salzberg. 2012. "Fast Gapped-Read Alignment with Bowtie 2." *Nature Methods* 9 (4): 357–59. https://doi.org/10.1038/nmeth.1923.

Lara-Astiaso, D., A. Weiner, E. Lorenzo-Vivas, I. Zaretsky, D. A. Jaitin, E. David, H. Keren-Shaul, et al. 2014. "Chromatin State Dynamics During Blood Formation." *Science* 345 (6199): 943–49. https://doi.org/10.1126/science.1256271.

Larsonneur, Elise, Jonathan Mercier, Nicolas Wiart, Edith Le Floch, Olivier Delhomme, and Vincent Meyer. 2018. "2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)." In, 2773–75. Madrid, Spain: IEEE. https://doi.org/10.1109/BIBM.2018.8621141.

Lawrence, Toby, and Gioacchino Natoli. 2011. "Transcriptional Regulation of Macrophage Polarization: Enabling Diversity with Identity." *Nature Reviews Immunology* 11 (11): 750–61. https://doi.org/10.1038/nri3088.

Lee, K.-W., Y. Lee, H.-J. Kwon, and D.-S. Kim. 2005. "Sp1-associated activation of macrophage inflammatory protein-2 promoter by CpG-oligodeoxynucleotide and lipopolysaccharide." *Cellular and molecular life sciences: CMLS* 62 (2): 188–98. https://doi.org/10.1007/s00018-004-4399-y.

Lee, Tong Ihn, and Richard A. Young. 2013. "Transcriptional Regulation and Its Misregulation in Disease." *Cell* 152 (6): 1237–51. https://doi.org/10.1016/j.cell.2013.02.014.

Leipzig, Jeremy. 2016. "A Review of Bioinformatic Pipeline Frameworks." *Briefings in Bioinformatics*, March, bbw020. https://doi.org/10.1093/bib/bbw020.

Leppä, Sirpa, Lila Pirkkala, Helena Saarento, Kevin D. Sarge, and Lea Sistonen. 1997. "Overexpression of HSF2-β Inhibits Hemin-Induced Heat Shock Gene Expression and Erythroid Differentiation in K562 Cells." *Journal of Biological Chemistry* 272 (24): 15293–98. https://doi.org/10.1074/jbc.272.24.15293.

Li, Guanglan, Wenke Hao, and Wenxue Hu. 2020. "Transcription Factor PU.1 and Immune Cell Differentiation (Review)." *International Journal of Molecular Medicine* 46 (6): 1943–50. https://doi.org/10.3892/ijmm.2020.4763.

Li, Heng, and Richard Durbin. 2010. "Fast and Accurate Long-Read Alignment with Burrows–Wheeler Transform." *Bioinformatics* 26 (5): 589–95. https://doi.org/10.1093/bioinformatics/btp698.

Li, Heng, Ting Jiang, Meng-Qi Li, Xi-Long Zheng, and Guo-Jun Zhao. 2018. "Transcriptional Regulation of Macrophages Polarization by MicroRNAs." *Frontiers in Immunology* 9 (May): 1175. https://doi.org/10.3389/fimmu.2018.01175.

Li, Ruifang, Sara A. Grimm, and Paul A. Wade. 2021. "A Simple and Robust Method for Simultaneous Dual-Omics Profiling with Limited Numbers of Cells." *Cell Reports Methods* 1 (3): 100041. https://doi.org/10.1016/j.crmeth.2021.100041.

Li, Zhijian, Chao-Chung Kuo, Fabio Ticconi, Mina Shaigan, Julia Gehrmann, Eduardo Gade Gusmao, Manuel Allhoff, Martin Manolov, Martin Zenke, and Ivan G. Costa. 2023. "RGT: A Toolbox for the Integrative Analysis of High Throughput Regulatory Genomics Data." *BMC Bioinformatics* 24 (1): 79. https://doi.org/10.1186/s12859-023-05184-5.

Li, Zhijian, Marcel H. Schulz, Thomas Look, Matthias Begemann, Martin Zenke, and Ivan G. Costa. 2019. "Identification of Transcription Factor Binding Sites Using ATAC-Seq." *Genome Biology* 20 (1). https://doi.org/10.1186/s13059-019-1642-2.

Liao, Yang, Gordon K. Smyth, and Wei Shi. 2014. "featureCounts: An Efficient General Purpose Program for Assigning Sequence Reads to Genomic Features." *Bioinformatics* 30 (7): 923–30. https://doi.org/10.1093/bioinformatics/btt656.

Liao, Yang, Gordon K Smyth, and Wei Shi. 2019. "The r Package Rsubread Is Easier, Faster, Cheaper and Better for Alignment and Quantification of RNA Sequencing

Reads." *Nucleic Acids Research* 47 (8): e47–47. https://doi.org/10.1093/nar/gkz114.

Liu, David X, and Lloyd A Greene. 2001. "Regulation of Neuronal Survival and Death by E2F-Dependent Gene Repression and Derepression." *Neuron* 32 (3): 425–38. https://doi.org/10.1016/S0896-6273(01)00495-0.

Liu, Longqi, Lizhi Leng, Chuanyu Liu, Changfu Lu, Yue Yuan, Liang Wu, Fei Gong, et al. 2019. "An Integrated Chromatin Accessibility and Transcriptome Landscape of Human Pre-Implantation Embryos." *Nature Communications* 10 (1): 364. https://doi.org/10.1038/s41467-018-08244-0.

Lord, Kenneth A, Abbas Abdollahi, Barbara Hoffman-Liebermann, and Dan A Liebermann. 1992. "Proto-Oncogenes of the Fos/Jun Family of Transcription Factors Are Positive Regulators of Myeloid Differentiation." *Molecular and Cellular Biology*.

Love, Michael I, Wolfgang Huber, and Simon Anders. 2014. "Moderated Estimation of Fold Change and Dispersion for RNA-Seq Data with DESeq2." *Genome Biology* 15 (12): 550. https://doi.org/10.1186/s13059-014-0550-8.

Ludwig, Leif S., Caleb A. Lareau, Erik L. Bao, Satish K. Nandakumar, Christoph Muus, Jacob C. Ulirsch, Kaitavjeet Chowdhary, et al. 2019. "Transcriptional States and Chromatin Accessibility Underlying Human Erythropoiesis." *Cell Reports* 27 (11): 3228–3240.e7. https://doi.org/10.1016/j.celrep.2019.05.046.

Luo, Liheng, Michael Gribskov, and Sufang Wang. 2022. "Bibliometric Review of ATAC-Seq and Its Application in Gene Expression." *Briefings in Bioinformatics* 23 (3): bbac061. https://doi.org/10.1093/bib/bbac061.

Luscombe, Nicholas M., Susan E. Austin, Helen M. Berman, and Janet M. Thornton. 2000. "An Overview of the Structures of Protein-DNA Complexes." *Genome Biology* 1 (1): reviews001.1. https://doi.org/10.1186/gb-2000-1-1-reviews001.

MacNeil, Lesley T., and Albertha J. M. Walhout. 2011. "Gene Regulatory Networks and the Role of Robustness and Stochasticity in the Control of Gene Expression." *Genome Research* 21 (5): 645–57. https://doi.org/10.1101/gr.097378.109.

Margolin, Adam A, Ilya Nemenman, Katia Basso, Chris Wiggins, Gustavo Stolovitzky, Ric-

cardo Dalla Favera, and Andrea Califano. 2006. “ARACNE: An Algorithm for the Reconstruction of Gene Regulatory Networks in a Mammalian Cellular Context.” *BMC Bioinformatics* 7 (S1): S7. https://doi.org/10.1186/1471-2105-7-S1-S7.

Martin, G R. 1981. “Isolation of a Pluripotent Cell Line from Early Mouse Embryos Cultured in Medium Conditioned by Teratocarcinoma Stem Cells.” *Proceedings of the National Academy of Sciences* 78 (12): 7634–38. https://doi.org/10.1073/pnas.78.12.7634.

Maston, Glenn A., Sara K. Evans, and Michael R. Green. 2006. “Transcriptional Regulatory Elements in the Human Genome.” *Annual Review of Genomics and Human Genetics* 7 (1): 29–59. https://doi.org/10.1146/annurev.genom.7.080505.115623.

Mincarelli, Laura, Ashleigh Lister, James Lipscombe, and Iain C. Macaulay. 2018. “Defining Cell Identity with Single-Cell Omics.” *PROTEOMICS* 18 (18): 1700312. https://doi.org/10.1002/pmic.201700312.

Mohr, Jeffrey C., Juan J. De Pablo, and Sean P. Palecek. 2006. “3-D Microwell Culture of Human Embryonic Stem Cells.” *Biomaterials* 27 (36): 6032–42. https://doi.org/10.1016/j.biomaterials.2006.07.012.

Morris, Samantha A. 2019. “The Evolving Concept of Cell Identity in the Single Cell Era.” Edited by Allon Klein and Barbara Treutlein. *Development* 146 (12): dev169748. https://doi.org/10.1242/dev.169748.

Morrison, Sean J., and Judith Kimble. 2006. “Asymmetric and Symmetric Stem-Cell Divisions in Development and Cancer.” *Nature* 441 (7097): 1068–74. https://doi.org/10.1038/nature04956.

Mortazavi, Ali, Brian A Williams, Kenneth McCue, Lorian Schaeffer, and Barbara Wold. 2008. “Mapping and Quantifying Mammalian Transcriptomes by RNA-Seq.” *Nature Methods* 5 (7): 621–28. https://doi.org/10.1038/nmeth.1226.

Nagamura-Inoue, Tokiko, Tomohiko Tamura, and Keiko Ozato. 2001. “Transcription Factors That Regulate Growth and Differentiation of Myeloid Cells.” *International Reviews of Immunology* 20 (1): 83–105. https://doi.org/10.3109/08830180109056724.

Narayanasamy, Shaman, Yohan Jarosz, Emilie E. L. Muller, Anna Heintz-Buschart, Malte Herold, Anne Kaysen, Cédric C. Laczny, Nicolás Pinel, Patrick May, and Paul Wilmes.

2016. “IMP: A Pipeline for Reproducible Reference-Independent Integrated Metagenomic and Metatranscriptomic Analyses.” *Genome Biology* 17 (1): 260. https://doi.org/10.1186/s13059-016-1116-8.

Nishida, Keishin, Martin C. Frith, and Kenta Nakai. 2009. “Pseudocounts for Transcription Factor Binding Sites.” *Nucleic Acids Research* 37 (3): 939–44. https://doi.org/10.1093/nar/gkn1019.

Notta, Faiyaz, Sasan Zandi, Naoya Takayama, Stephanie Dobson, Olga I. Gan, Gavin Wilson, Kerstin B. Kaufmann, et al. 2016. “Distinct routes of lineage development reshape the human blood hierarchy across ontogeny.” *Science (New York, N.Y.)* 351 (6269): aab2116. https://doi.org/10.1126/science.aab2116.

Novak, J. P., and C. C. Stewart. 1991. “Stochastic Versus Deterministic in Haemopoiesis: What Is What?” *British Journal of Haematology* 78 (2): 149–54. https://doi.org/10.1111/j.1365-2141.1991.tb04409.x.

O’Brien, Jacob, Heyam Hayder, Yara Zayed, and Chun Peng. 2018. “Overview of MicroRNA Biogenesis, Mechanisms of Actions, and Circulation.” *Frontiers in Endocrinology* 9 (August): 402. https://doi.org/10.3389/fendo.2018.00402.

Orecchioni, Marco, Yanal Ghosheh, Akula Bala Pramod, and Klaus Ley. 2019. “Macrophage Polarization: Different Gene Signatures in M1(LPS+) Vs. Classically and M2(LPS–) Vs. Alternatively Activated Macrophages.” *Frontiers in Immunology* 10 (May): 1084. https://doi.org/10.3389/fimmu.2019.01084.

Pang, Zhiping P., Nan Yang, Thomas Vierbuchen, Austin Ostermeier, Daniel R. Fuentes, Troy Q. Yang, Ami Citri, et al. 2011. “Induction of Human Neuronal Cells by Defined Transcription Factors.” *Nature* 476 (7359): 220–23. https://doi.org/10.1038/nature10202.

Pape, Utz J., Sven Rahmann, and Martin Vingron. 2008. “Natural Similarity Measures Between Position Frequency Matrices with an Application to Clustering.” *Bioinformatics* 24 (3): 350–57. https://doi.org/10.1093/bioinformatics/btm610.

Park, Peter J. 2009. “ChIP–seq: Advantages and Challenges of a Maturing Technology.” *Nature Reviews Genetics* 10 (10): 669–80. https://doi.org/10.1038/nrg2641.

Perrin, Hannah J., Kevin W. Currin, Swarooparani Vadlamudi, Gautam K. Pandey, Kenneth K. Ng, Martin Wabitsch, Markku Laakso, Michael I. Love, and Karen L. Mohlke. 2021. "Chromatin Accessibility and Gene Expression During Adipocyte Differentiation Identify Context-Dependent Effects at Cardiometabolic GWAS Loci." Edited by Chris Cotsapas. *PLOS Genetics* 17 (10): e1009865. https://doi.org/10.1371/journal.pgen.1009865.

Pfisterer, Ulrich, Agnete Kirkeby, Olof Torper, James Wood, Jenny Nelander, Audrey Dufour, Anders Björklund, Olle Lindvall, Johan Jakobsson, and Malin Parmar. 2011. "Direct Conversion of Human Fibroblasts to Dopaminergic Neurons." *Proceedings of the National Academy of Sciences* 108 (25): 10343–48. https://doi.org/10.1073/pnas.1105135108.

Pinho, Sandra, and Paul S. Frenette. 2019. "Haematopoietic Stem Cell Activity and Interactions with the Niche." *Nature Reviews Molecular Cell Biology* 20 (5): 303–20. https://doi.org/10.1038/s41580-019-0103-9.

Portales-Casamar, Elodie, Supat Thongjuea, Andrew T. Kwon, David Arenillas, Xiaobei Zhao, Eivind Valen, Dimas Yusuf, Boris Lenhard, Wyeth W. Wasserman, and Albin Sandelin. 2010. "JASPAR 2010: The Greatly Expanded Open-Access Database of Transcription Factor Binding Profiles." *Nucleic Acids Research* 38 (suppl_1): D105–10. https://doi.org/10.1093/nar/gkp950.

Pundhir, Sachin, Felicia Kathrine Bratt Lauridsen, Mikkel Bruhn Schuster, Janus Schou Jakobsen, Ying Ge, Erwin Marten Schoof, Nicolas Rapin, Johannes Waage, Marie Sigurd Hasemann, and Bo Torben Porse. 2018. "Enhancer and Transcription Factor Dynamics During Myeloid Differentiation Reveal an Early Differentiation Block in Cebpa Null Progenitors." *Cell Reports* 23 (9): 2744–57. https://doi.org/10.1016/j.celrep.2018.05.012.

Quinlan, Aaron R., and Ira M. Hall. 2010. "BEDTools: A Flexible Suite of Utilities for Comparing Genomic Features." *Bioinformatics* 26 (6): 841–42. https://doi.org/10.1093/bioinformatics/btq033.

Ramirez, Ricardo N., Nicole C. El-Ali, Mikayla Anne Mager, Dana Wyman, Ana Conesa,

and Ali Mortazavi. 2017. “Dynamic Gene Regulatory Networks of Human Myeloid Differentiation.” *Cell Systems* 4 (4): 416–429.e3. https://doi.org/10.1016/j.cels.2017.03.005.

Ramos, Borja Gomez, Jochen Ohnmacht, Nikola de Lange, Aurélien Ginolhac, Elena Valceschini, Aleksandar Rakovic, Rashi Halder, et al. 2023. “Multi-Omics Analysis Identifies LBX1 and NHLH1 as Central Regulators of Human Midbrain Dopaminergic Neuron Differentiation.” https://doi.org/10.1101/2023.01.27.525898.

Ranzoni, Anna Maria, Andrea Tangherloni, Ivan Berest, Simone Giovanni Riva, Brynelle Myers, Paulina M. Strzelecka, Jiarui Xu, et al. 2021. “Integrative Single-Cell RNA-Seq and ATAC-Seq Analysis of Human Developmental Hematopoiesis.” *Cell Stem Cell* 28 (3): 472–487.e7. https://doi.org/10.1016/j.stem.2020.11.015.

Rauschmeier, René, Charlotte Gustafsson, Annika Reinhardt, Noelia A-Gonzalez, Luigi Tortola, Dilay Cansever, Sethuraman Subramanian, et al. 2019. “Bhlhe40 and Bhlhe41 Transcription Factors Regulate Alveolar Macrophage Self-Renewal and Identity.” *The EMBO Journal* 38 (19). https://doi.org/10.15252/embj.2018101233.

Richardson, Edward T., Supriya Shukla, Nancy Nagy, W. Henry Boom, Rose C. Beck, Lan Zhou, Gary E. Landreth, and Clifford V. Harding. 2015. “ERK Signaling Is Essential for Macrophage Development.” Edited by Kevin D Bunting. *PLOS ONE* 10 (10): e0140064. https://doi.org/10.1371/journal.pone.0140064.

Rosenbauer, Frank, and Daniel G. Tenen. 2007. “Transcription Factors in Myeloid Development: Balancing Differentiation with Transformation.” *Nature Reviews Immunology* 7 (2): 105–17. https://doi.org/10.1038/nri2024.

Sánchez Alvarado, Alejandro, and Shinya Yamanaka. 2014. “Rethinking Differentiation: Stem Cells, Regeneration, and Plasticity.” *Cell* 157 (1): 110–19. https://doi.org/10.1016/j.cell.2014.02.041.

Sandelin, A. 2004. “JASPAR: An Open-Access Database for Eukaryotic Transcription Factor Binding Profiles.” *Nucleic Acids Research* 32 (90001): 91D–94. https://doi.org/10.1093/nar/gkh012.

Santoni de Sio, F. R., L. Passerini, M. M. Valente, F. Russo, L. Naldini, M. G. Roncarolo,

and R. Bacchetta. 2017. “Ectopic FOXP3 Expression Preserves Primitive Features Of Human Hematopoietic Stem Cells While Impairing Functional T Cell Differentiation.” *Scientific Reports* 7 (1): 15820. https://doi.org/10.1038/s41598-017-15689-8.

Schliep, A., I. G. Costa, C. Steinhoff, and A. Schonhuth. 2005. “Analyzing Gene Expression Time-Courses.” *IEEE/ACM Transactions on Computational Biology and Bioinformatics* 2 (3): 179–93. https://doi.org/10.1109/TCBB.2005.31.

Schmidt, Florian, Nina Gasparoni, Gilles Gasparoni, Kathrin Gianmoena, Cristina Cadenas, Julia K. Polansky, Peter Ebert, et al. 2017. “Combining Transcription Factor Binding Affinities with Open-Chromatin Data for Accurate Gene Expression Prediction.” *Nucleic Acids Research* 45 (1): 54–66. https://doi.org/10.1093/nar/gkw1061.

Schmidt, Florian, Fabian Kern, Peter Ebert, Nina Baumgarten, and Marcel H Schulz. 2019. “TEPIC 2—an Extended Framework for Transcription Factor Binding Prediction and Integrative Epigenomic Analysis.” Edited by Bonnie Berger. *Bioinformatics* 35 (9): 1608–9. https://doi.org/10.1093/bioinformatics/bty856.

Schuettengruber, Bernd, Henri-Marc Bourbon, Luciano Di Croce, and Giacomo Cavalli. 2017. “Genome Regulation by Polycomb and Trithorax: 70 Years and Counting.” *Cell* 171 (1): 34–57. https://doi.org/10.1016/j.cell.2017.08.002.

Schulz, Marcel H, William E Devanny, Anthony Gitter, Shan Zhong, Jason Ernst, and Ziv Bar-Joseph. 2012. “DREM 2.0: Improved Reconstruction of Dynamic Regulatory Networks from Time-Series Expression Data.” *BMC Systems Biology* 6 (1): 104. https://doi.org/10.1186/1752-0509-6-104.

Schweitzer, Jeffrey S., Bin Song, Todd M. Herrington, Tae-Yoon Park, Nayeon Lee, Sanghyeok Ko, Jeha Jeon, et al. 2020. “Personalized iPSC-Derived Dopamine Progenitor Cells for Parkinson’s Disease.” *New England Journal of Medicine* 382 (20): 1926–32. https://doi.org/10.1056/NEJMoa1915872.

Sherf, Orna, Limor Nashelsky Zolotov, Keren Liser, Hadas Tilleman, Vukasin M. Jovanovic, Ksenija Zega, Marin M. Jukic, and Claude Brodski. 2015. “Otx2 Requires Lmx1b to Control the Development of Mesodiencephalic Dopaminergic Neurons.” Edited by Renping Zhou. *PLOS ONE* 10 (10): e0139697. https://doi.org/10.1371/journal.pone

.0139697.

Silvennoinen, Katri, Nikola de Lange, Sara Zagaglia, Simona Balestrini, Ganna Androsova, Merel Wassenaar, Pauls Auce, et al. 2019. “Comparative Effectiveness of Antiepileptic Drugs in Juvenile Myoclonic Epilepsy.” *Epilepsia Open* 4 (3): 420–30. https://doi.org/10.1002/epi4.12349.

Song, Lingyun, and Gregory E. Crawford. 2010. “DNase-Seq: A High-Resolution Technique for Mapping Active Gene Regulatory Elements Across the Genome from Mammalian Cells.” *Cold Spring Harbor Protocols* 2010 (2): pdb.prot5384. https://doi.org/10.1101/pdb.prot5384.

Song, Qiao, Yuli Hou, Yiyin Zhang, Jing Liu, Yaqi Wang, Jingxuan Fu, Chi Zhang, et al. 2022. “Integrated Multi-Omics Approach Revealed Cellular Senescence Landscape.” *Nucleic Acids Research* 50 (19): 10947–63. https://doi.org/10.1093/nar/gkac885.

Specht, Alicia T, and Jun Li. 2017. “LEAP: Constructing Gene Co-Expression Networks for Single-Cell RNA-Sequencing Data Using Pseudotime Ordering.” Edited by Ziv Bar-Joseph. *Bioinformatics* 33 (5): 764–66. https://doi.org/10.1093/bioinformatics/btw729.

Spitz, François, and Eileen E. M. Furlong. 2012. “Transcription Factors: From Enhancer Binding to Developmental Control.” *Nature Reviews Genetics* 13 (9): 613–26. https://doi.org/10.1038/nrg3207.

T’Jonck, Wouter, Martin Guilliams, and Johnny Bonnardel. 2018. “Niche Signals and Transcription Factors Involved in Tissue-Resident Macrophage Development.” *Cellular Immunology* 330 (August): 43–53. https://doi.org/10.1016/j.cellimm.2018.02.005.

Tabata, Tetsuya. 2001. “Genetics of Morphogen Gradients.” *Nature Reviews Genetics* 2 (8): 620–30. https://doi.org/10.1038/35084577.

Takahashi, Kazutoshi, and Shinya Yamanaka. 2006. “Induction of Pluripotent Stem Cells from Mouse Embryonic and Adult Fibroblast Cultures by Defined Factors.” *Cell* 126 (4): 663–76. https://doi.org/10.1016/j.cell.2006.07.024.

———. 2013. “Induced Pluripotent Stem Cells in Medicine and Biology.” *Development* 140 (12): 2457–61. https://doi.org/10.1242/dev.092551.

The FANTOM Consortium, Owen J L Rackham, Jaber Firas, Hai Fang, Matt E Oates, Melissa L Holmes, Anja S Knaupp, et al. 2016. “A Predictive Computational Framework for Direct Reprogramming Between Human Cell Types.” *Nature Genetics* 48 (3): 331–35. https://doi.org/10.1038/ng.3487.

Tsompana, Maria, and Michael J Buck. 2014. “Chromatin Accessibility: A Window into the Genome.” *Epigenetics & Chromatin* 7 (1): 33. https://doi.org/10.1186/1756-8935-7-33.

van der Raadt, Jori, Sebastianus H C van Gestel, Nael Nadif Kasri, and Cornelis A Albers. 2019. “ONECUT Transcription Factors Induce Neuronal Characteristics and Remodel Chromatin Accessibility.” *Nucleic Acids Research* 47 (11): 5587–5602. https://doi.org/10.1093/nar/gkz273.

Vasimuddin, Md., Sanchit Misra, Heng Li, and Srinivas Aluru. 2019. “2019 IEEE International Parallel and Distributed Processing Symposium (IPDPS).” In, 314–24. Rio de Janeiro, Brazil: IEEE. https://doi.org/10.1109/IPDPS.2019.00041.

Veenvliet, Jesse V., Maria T. M. Alves dos Santos, Willemieke M. Kouwenhoven, Lars von Oerthel, Jamie L. Lim, Annemarie J. A. van der Linden, Marian J. A. Groot Koerkamp, Frank C. P. Holstege, and Marten P. Smidt. 2013. “Specification of Dopaminergic Subsets Involves Interplay of En1 and Pitx3.” *Development* 140 (16): 3373–84. https://doi.org/10.1242/dev.094565.

Veremeyko, Tatyana, Amanda W. Y. Yung, Daniel C. Anthony, Tatyana Strekalova, and Eugene D. Ponomarev. 2018. “Early Growth Response Gene-2 Is Essential for M1 and M2 Macrophage Activation and Plasticity by Modulation of the Transcription Factor CEBPβ.” *Frontiers in Immunology* 9 (November): 2515. https://doi.org/10.3389/fimmu.2018.02515.

Vierbuchen, Thomas, Austin Ostermeier, Zhiping P. Pang, Yuko Kokubu, Thomas C. Südhof, and Marius Wernig. 2010. “Direct Conversion of Fibroblasts to Functional Neurons by Defined Factors.” *Nature* 463 (7284): 1035–41. https://doi.org/10.1038/nature08797.

Villaescusa, J Carlos, Bingsi Li, Enrique M Toledo, Pia Rivetti di Val Cervo, Shanzheng

Yang, Simon RW Stott, Karol Kaiser, et al. 2016. “A PBX1 Transcriptional Network Controls Dopaminergic Neuron Development and Is Impaired in Parkinson’s Disease.” *The EMBO Journal* 35 (18): 1963–78. https://doi.org/10.15252/embj.201593725.

Wagner, Günter P., Koryu Kin, and Vincent J. Lynch. 2012. “Measurement of mRNA Abundance Using RNA-Seq Data: RPKM Measure Is Inconsistent Among Samples.” *Theory in Biosciences* 131 (4): 281–85. https://doi.org/10.1007/s12064-012-0162-3.

Wang, Mengmeng, King-Hwa Ling, Jun Tan, and Cheng-Biao Lu. 2020. “Development and Differentiation of Midbrain Dopaminergic Neuron: From Bench to Bedside.” *Cells* 9 (6): 1489. https://doi.org/10.3390/cells9061489.

Wang, Zhong, Mark Gerstein, and Michael Snyder. 2009. “RNA-Seq: A Revolutionary Tool for Transcriptomics.” *Nature Reviews Genetics* 10 (1): 57–63. https://doi.org/10.1038/nrg2484.

Weber, Andreas P. M. 2015. “Discovering New Biology Through RNA-Seq.” *Plant Physiology*, September, pp.01081.2015. https://doi.org/10.1104/pp.15.01081.

Weiskopf, Kipp, Peter J. Schnorr, Wendy W. Pang, Mark P. Chao, Akanksha Chhabra, Jun Seita, Mingye Feng, and Irving L. Weissman. 2016. “Myeloid Cell Origins, Differentiation, and Clinical Implications.” *Microbiology Spectrum* 4 (5). https://doi.org/10.1128/microbiolspec.MCHD-0031-2016.

Wingender, Edgar, Torsten Schoeps, and Jürgen Dönitz. 2013. “TFClass: An Expandable Hierarchical Classification of Human Transcription Factors.” *Nucleic Acids Research* 41 (D1): D165–70. https://doi.org/10.1093/nar/gks1123.

Wingender, Edgar, Torsten Schoeps, Martin Haubrock, Mathias Krull, and Jürgen Dönitz. 2018. “TFClass: Expanding the Classification of Human Transcription Factors to Their Mammalian Orthologs.” *Nucleic Acids Research* 46 (D1): D343–47. https://doi.org/10.1093/nar/gkx987.

Wu, Tianzhi, Erqiang Hu, Shuangbin Xu, Meijun Chen, Pingfan Guo, Zehan Dai, Tingze Feng, et al. 2021. “clusterProfiler 4.0: A Universal Enrichment Tool for Interpreting Omics Data.” *The Innovation* 2 (3): 100141. https://doi.org/10.1016/j.xinn.2021.100141.

Xiang, Yangfei, Yoshiaki Tanaka, Benjamin Patterson, Young-Jin Kang, Gubbi Govindaiah, Naomi Roselaar, Bilal Cakir, et al. 2017. "Fusion of Regionally Specified hPSC-Derived Organoids Models Human Brain Development and Interneuron Migration." *Cell Stem Cell* 21 (3): 383–398.e7. https://doi.org/10.1016/j.stem.2017.07.007.

Xu, Yang, Edmon Begoli, and Rachel Patton McCord. 2022. "sciCAN: Single-Cell Chromatin Accessibility and Gene Expression Data Integration via Cycle-Consistent Adversarial Network." *Npj Systems Biology and Applications* 8 (1): 33. https://doi.org/10.1038/s41540-022-00245-6.

Yan, Kai, Tian-Tian Da, Zhen-Hua Bian, Yi He, Meng-Chu Liu, Qing-Zhi Liu, Jie Long, et al. 2020. "Multi-Omics Analysis Identifies FoxO1 as a Regulator of Macrophage Function Through Metabolic Reprogramming." *Cell Death & Disease* 11 (9): 800. https://doi.org/10.1038/s41419-020-02982-0.

Yang, B-H, S Hagemann, P Mamareli, U Lauer, U Hoffmann, M Beckstette, L Föhse, et al. 2016. "Foxp3+ T Cells Expressing RORγt Represent a Stable Regulatory T-Cell Effector Lineage with Enhanced Suppressive Capacity During Intestinal Inflammation." *Mucosal Immunology* 9 (2): 444–57. https://doi.org/10.1038/mi.2015.74.

Yang, Bi-Huei, Ke Wang, Shuo Wan, Yan Liang, Xiaomei Yuan, Yi Dong, Sunglim Cho, et al. 2019. "TCF1 and LEF1 Control Treg Competitive Survival and Tfr Development to Prevent Autoimmune Diseases." *Cell Reports* 27 (12): 3629–3645.e6. https://doi.org/10.1016/j.celrep.2019.05.061.

Yesudhas, Dhanusha, Maria Batool, Muhammad Anwar, Suresh Panneerselvam, and Sangdun Choi. 2017. "Proteins Recognizing DNA: Structural Uniqueness and Versatility of DNA-Binding Domains in Stem Cell Transcription Factors." *Genes* 8 (8): 192. https://doi.org/10.3390/genes8080192.

Yoo, Andy B., Morris A. Jette, and Mark Grondona. 2003. "SLURM: Simple Linux Utility for Resource Management." In, edited by Dror Feitelson, Larry Rudolph, and Uwe Schwiegelshohn, 2862:44–60. Berlin, Heidelberg: Springer Berlin Heidelberg. http://link.springer.com/10.1007/10968987_3.

Yu, Guangchuang, Li-Gen Wang, Yanyan Han, and Qing-Yu He. 2012. "clusterProfiler:

An R Package for Comparing Biological Themes Among Gene Clusters." *OMICS: A Journal of Integrative Biology* 16 (5): 284–87. https://doi.org/10.1089/omi.2011.0118.

Zakrzewski, Wojciech, Maciej Dobrzyński, Maria Szymonowicz, and Zbigniew Rybak. 2019. "Stem Cells: Past, Present, and Future." *Stem Cell Research & Therapy* 10 (1): 68. https://doi.org/10.1186/s13287-019-1165-5.

Zhang, Hong, Meritxell Alberich-Jorda, Giovanni Amabile, Henry Yang, Philipp B. Staber, Annalisa Di Ruscio, Robert S. Welner, et al. 2013. "Sox4 Is a Key Oncogenic Target in C/EBPα Mutant Acute Myeloid Leukemia." *Cancer Cell* 24 (5): 575–88. https://doi.org/10.1016/j.ccr.2013.09.018.

Zhang, Kai, Mengchi Wang, Ying Zhao, and Wei Wang. 2019. "Taiji: System-Level Identification of Key Transcription Factors Reveals Transcriptional Waves in Mouse Embryonic Development." *Science Advances* 5 (3): eaav3262. https://doi.org/10.1126/sciadv.aav3262.

Zhang, Yong, Tao Liu, Clifford A Meyer, Jérôme Eeckhoute, David S Johnson, Bradley E Bernstein, Chad Nusbaum, et al. 2008. "Model-Based Analysis of ChIP-Seq (MACS)." *Genome Biology* 9 (9): R137. https://doi.org/10.1186/gb-2008-9-9-r137.

Zhang, Ziqi, Chengkai Yang, and Xiuwei Zhang. 2022. "scDART: Integrating Unmatched scRNA-Seq and scATAC-Seq Data and Learning Cross-Modality Relationship Simultaneously." *Genome Biology* 23 (1): 139. https://doi.org/10.1186/s13059-022-02706-x.

Zhao, Shanrong, Wai-Ping Fung-Leung, Anton Bittner, Karen Ngo, and Xuejun Liu. 2014. "Comparison of RNA-Seq and Microarray in Transcriptome Profiling of Activated T Cells." Edited by Shu-Dong Zhang. *PLoS ONE* 9 (1): e78644. https://doi.org/10.1371/journal.pone.0078644.

Zhao, Shanrong, Zhan Ye, and Robert Stanton. 2020. "Misuse of RPKM or TPM normalization when comparing across samples and sequencing protocols." *RNA (New York, N.Y.)* 26 (8): 903–9. https://doi.org/10.1261/rna.074922.120.

Zhao, Shanrong, and Baohong Zhang. 2015. "A Comprehensive Evaluation of Ensembl,

RefSeq, and UCSC Annotations in the Context of RNA-Seq Read Mapping and Gene Quantification." *BMC Genomics* 16 (1): 97. https://doi.org/10.1186/s12864-015-1308-8.

A. Supplementary Material

A.1. Figures and tables

Table A.1.: Ranking without consideration of split score and significance of main regulators and candidate TFs at each time point. NHLH1 is ranked 1st on D30, LBX1 ranks among top 20 at late time points.

TF	D15	D30	D50
LBX1	62	19	20
NHLH1	34	1	17
LMX1A	16	38	11
LMX1B	22	22	8
EN1	3	4	1
NR4A2::RXRA	2	67	51
NR4A2	6	34	42
NR2F1	8	26	84
NR2F2	28	14	47
HOXB2	56	15	36
SOX4	13	20	111

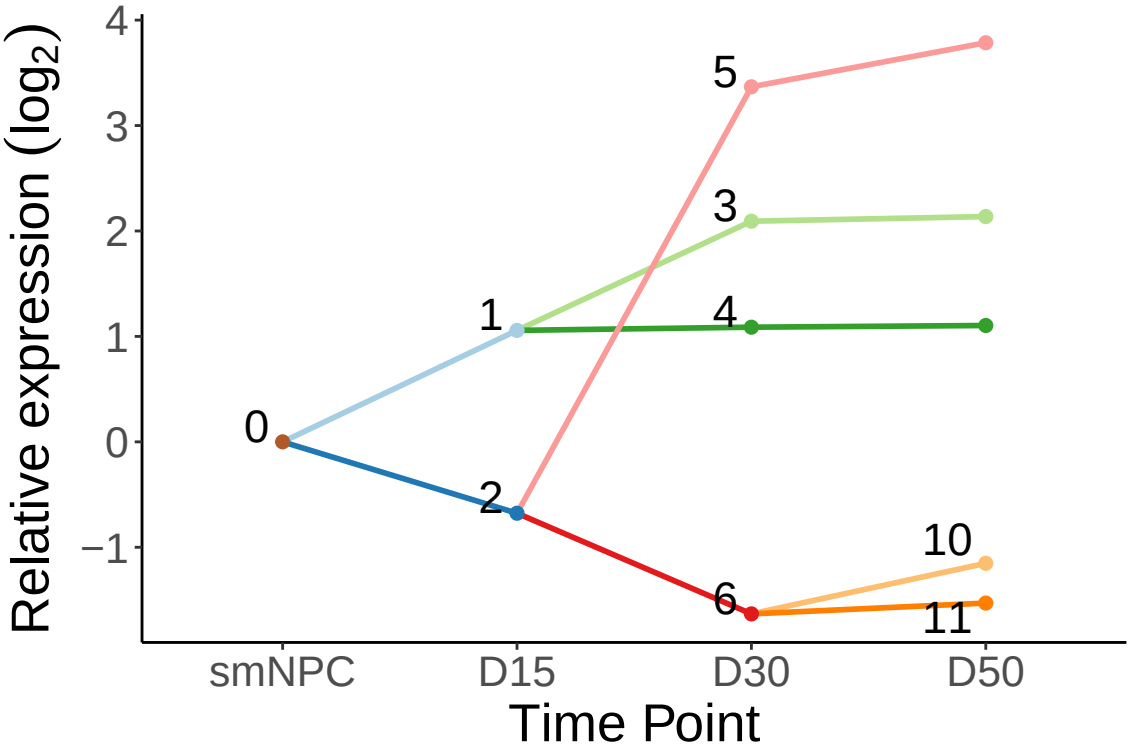


Figure A.1.: Split node network of human mDAN if only two paths are allowed at each split node showing the relative expression at each time point. With only four clusters at D30 and five clusters at D50 the data was not captured well.

Table A.2.: Chromatin accessibility overview for macrophages

Time point	# peaks	# footprints	Diff. peaks	DEG
HL60	52505	266489		
3h	31327	153285	1367	1745
6h	25420	134883	3316	2809
12h	34420	163247	5543	4392
24h	51060	233856	22173	5364
48h	42648	226356	25237	5667
96h	40262	202307	26267	6188
120h	66567	282204	35081	6006

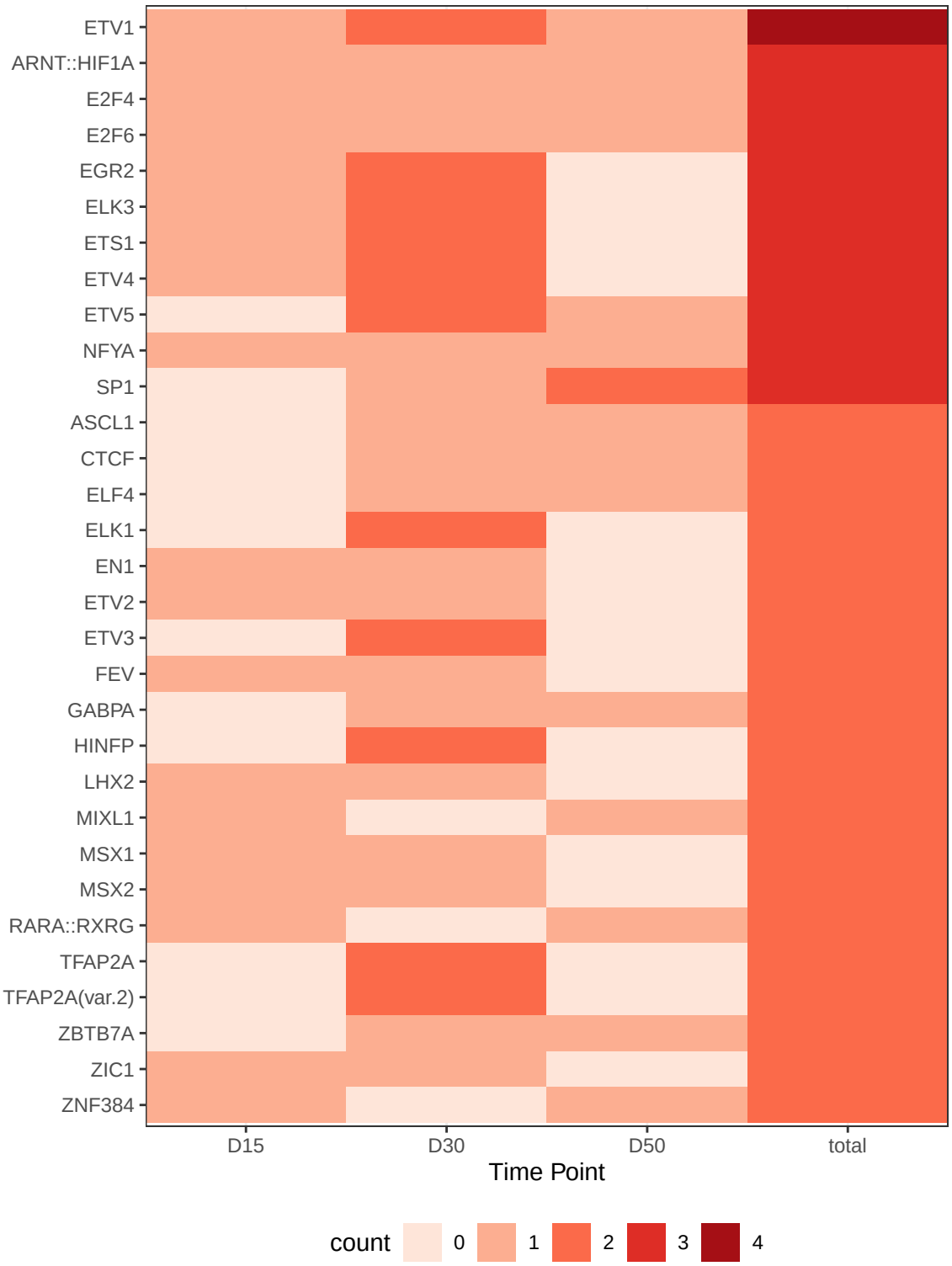


Figure A.2.: Recurring TFs identified with earliest DREMflow settings for mDAN neurons. Number of appearances of TFs among the top 15 significant TFs assigned to the nodes. EN1 and MSX1 were selected by DREMflow.

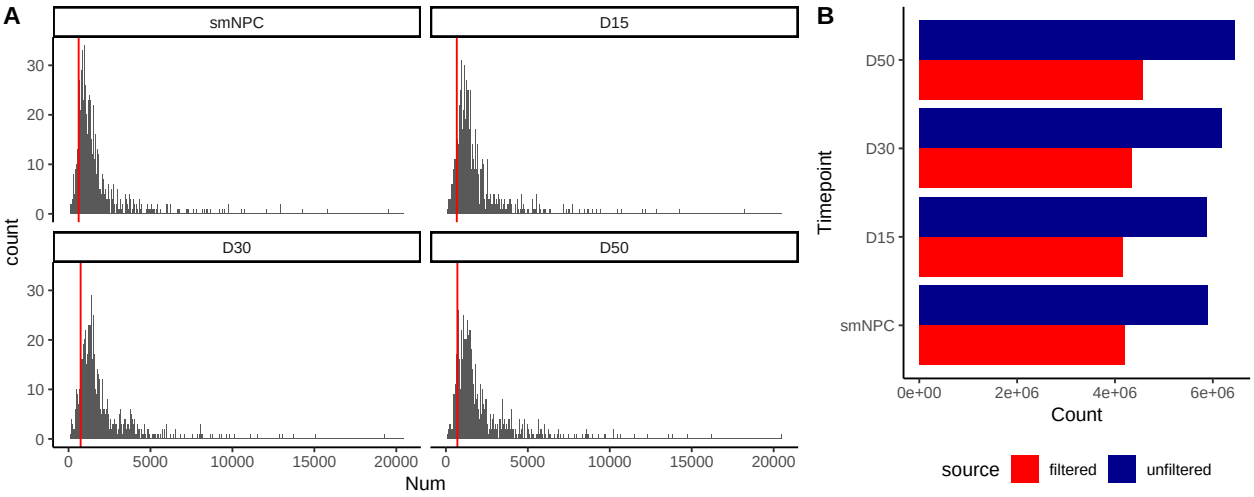


Figure A.3.: (A) Distribution of the number of TFBS identified for each TF by time point. The red vertical line marks the time wise cut-off of 10%. (B) Numbers of TF-gene links for filtered TFs according to their identified number of TFBS in the data and unfiltered data.

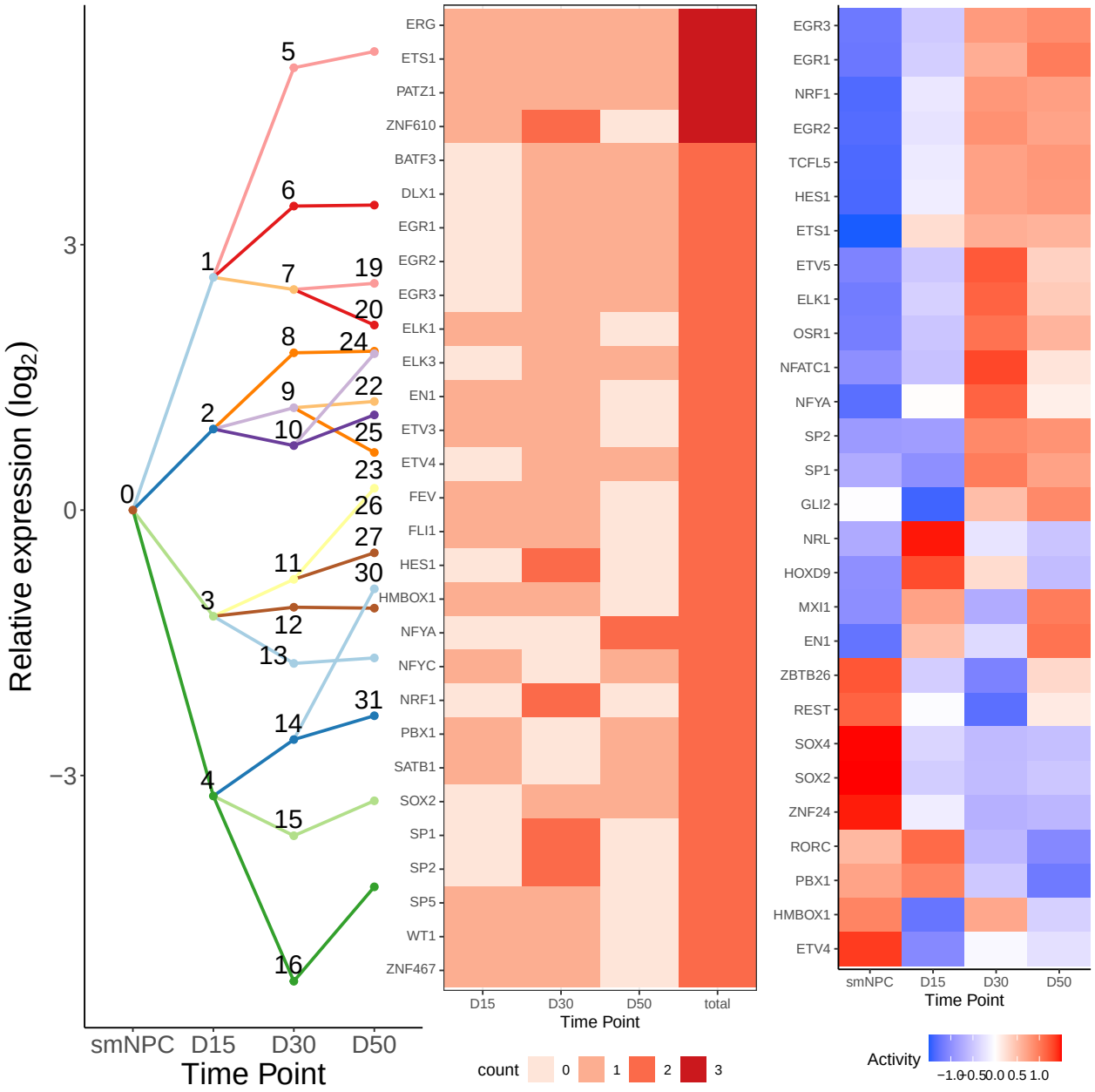


Figure A.4.: DREM model overview of mDA differentiation. (A) DREMflow computed split node network of co-expressed genes. Labelled nodes are the nodes after split nodes.(B) Top regulators with more than one occurence in the top 20 of all nodes. The color shows that number of occurences at each time point and overall. (C) TF activity for top regulators with at least 1000 TF binding sites.

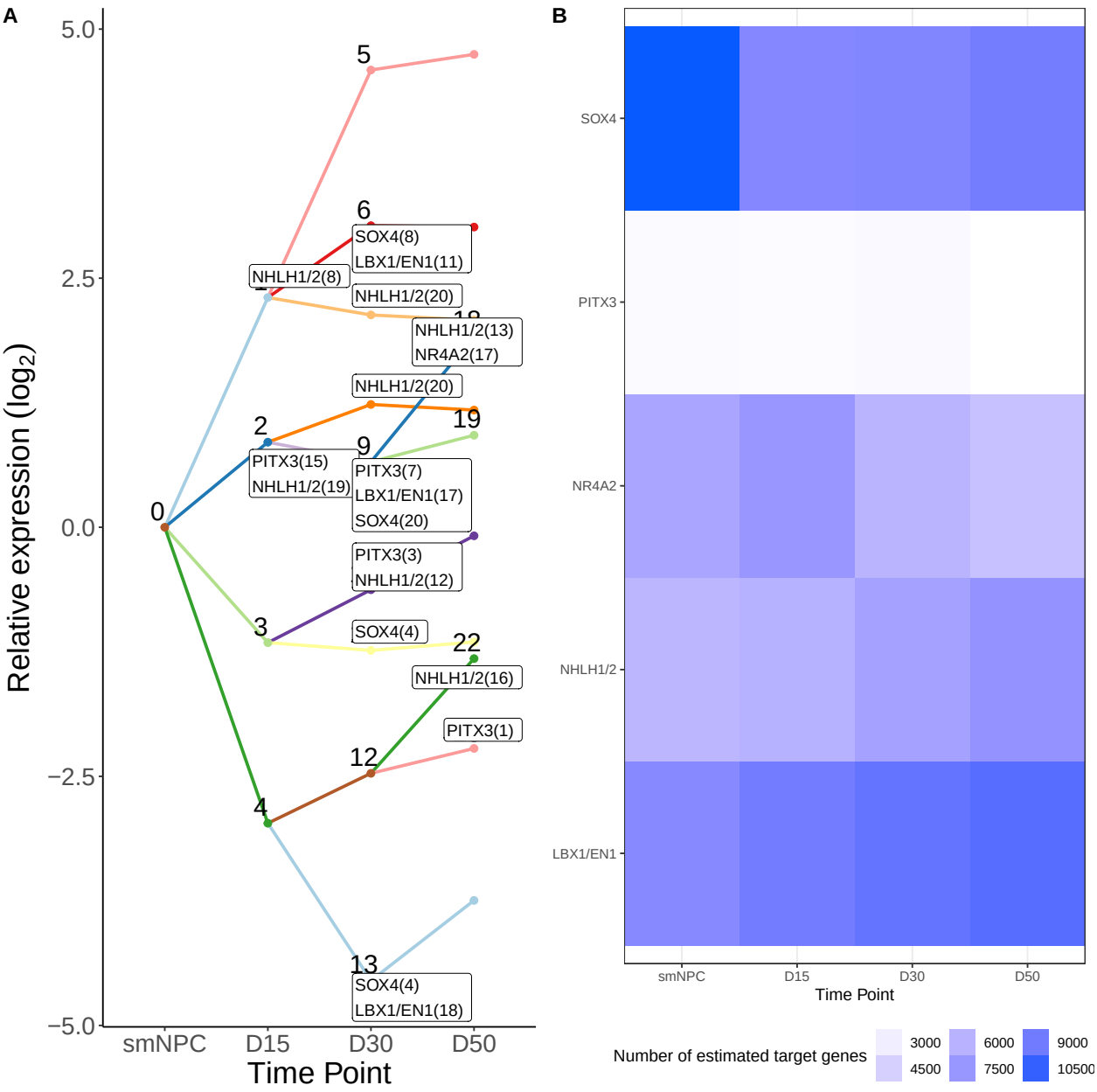


Figure A.5.: DREM model overview of mDA differentiation using TF clusters. (A) DREM-flow computed split node network of co-expressed genes using TF clusters. Each cluster is represented by specific TFs from this cluster. (B) Number of TF-gene links for each cluster.

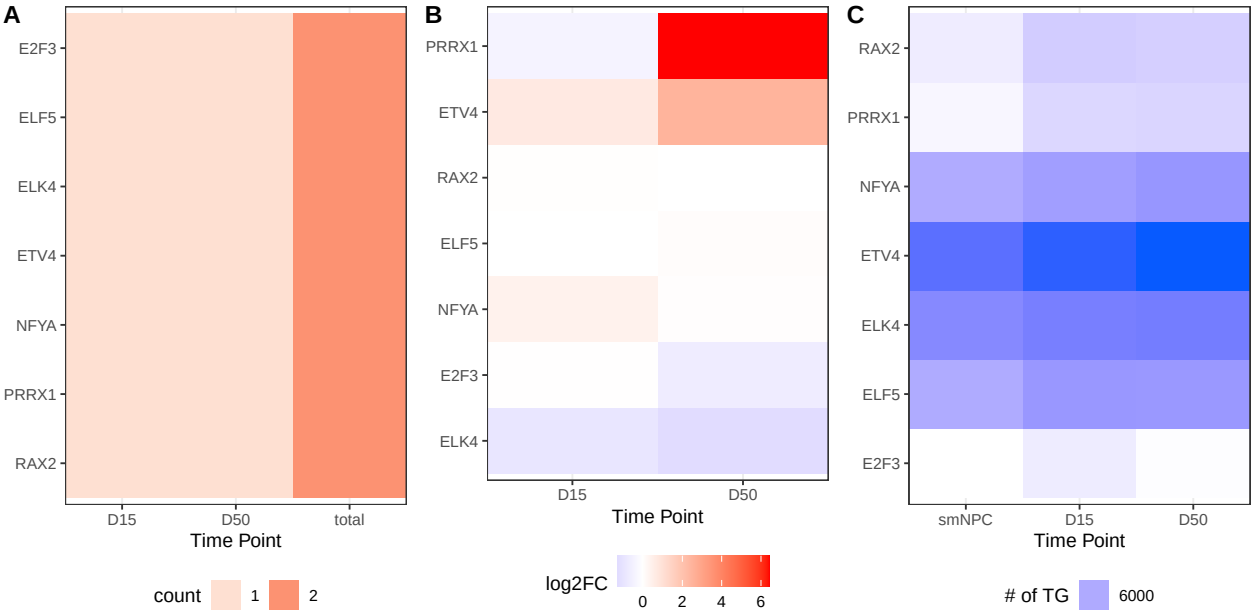


Figure A.6.: Recurring TF for the non-dopaminergic neuron mix. Only seven TFs were selected. (A) Number of appearances. None of the TFs appeared more than once at a timepoint. (B) Lof2FC of expression for the selected TFs. (C) Number of predicted target genes or the selected TFs.

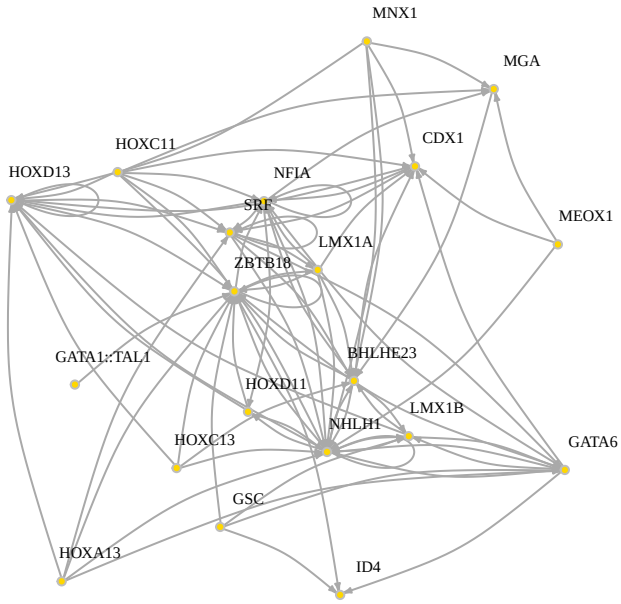


Figure A.7.: TF-TF network for the non-dopaminergic neurons. The network displays the connections identified for node 3 showing LMX1A and LMX1B in the network.

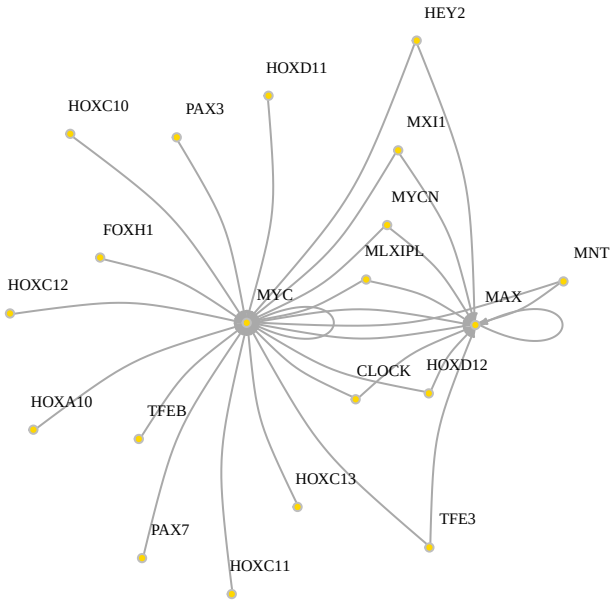


Figure A.8.: TF-TF network for neutrophil differentiation. The network displays the connections identified for TFs at node 5 showing the central position of MYC

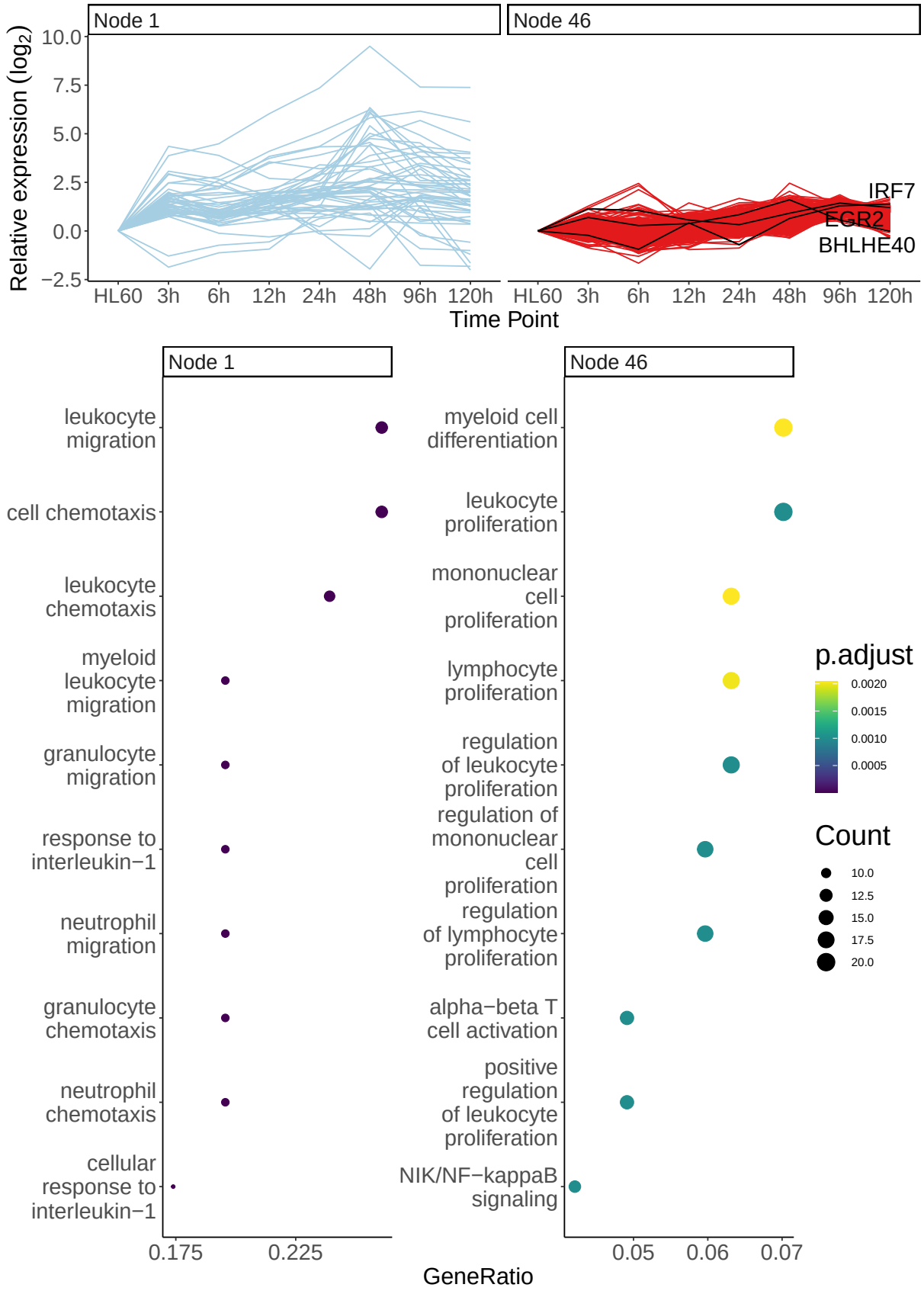


Figure A.9.: GO enrichment on selected nodes for monocytes. (A) Expression profiles of target genes of selected nodes. (B) Terms enriched on node 2, node 26 and node 47.

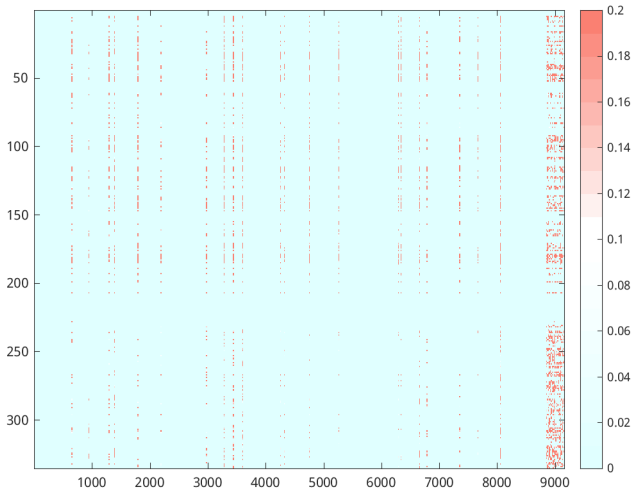


Figure A.10.: TF-TG heatmap from TimeReg at 3h. Only apprixmately 97000 connections out of 3 million theoretically possible connections were identified, resulting in a sparse matrix.

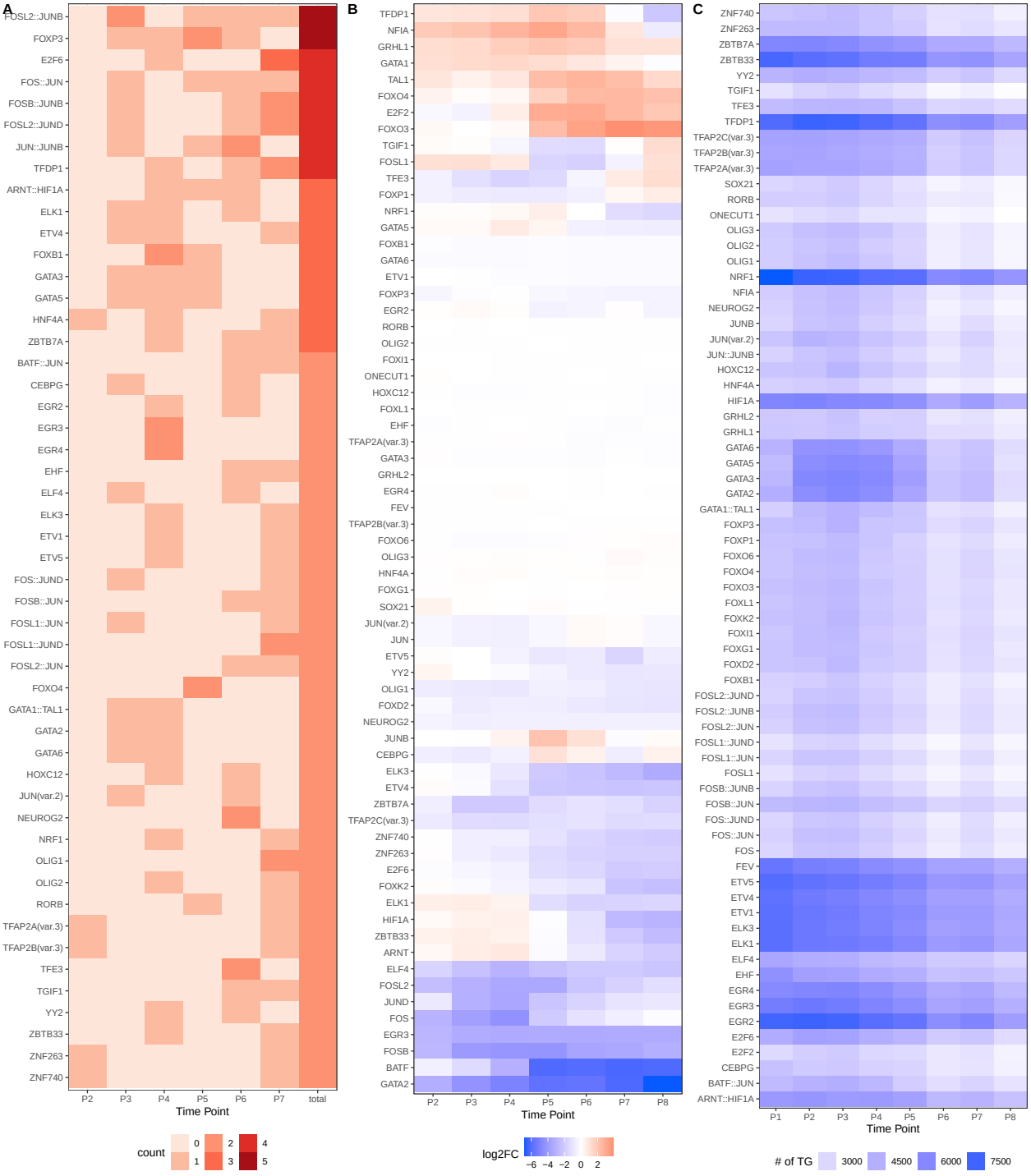


Figure A.11.: Overview of top regulators for erythropoiesis. (A) Number of appearances of top regulators for the 8 differentiation stages of erythrocytes. (B) Expression profile of top regulators. (C) Number of predicted target genes for each top regulator at the different stages.

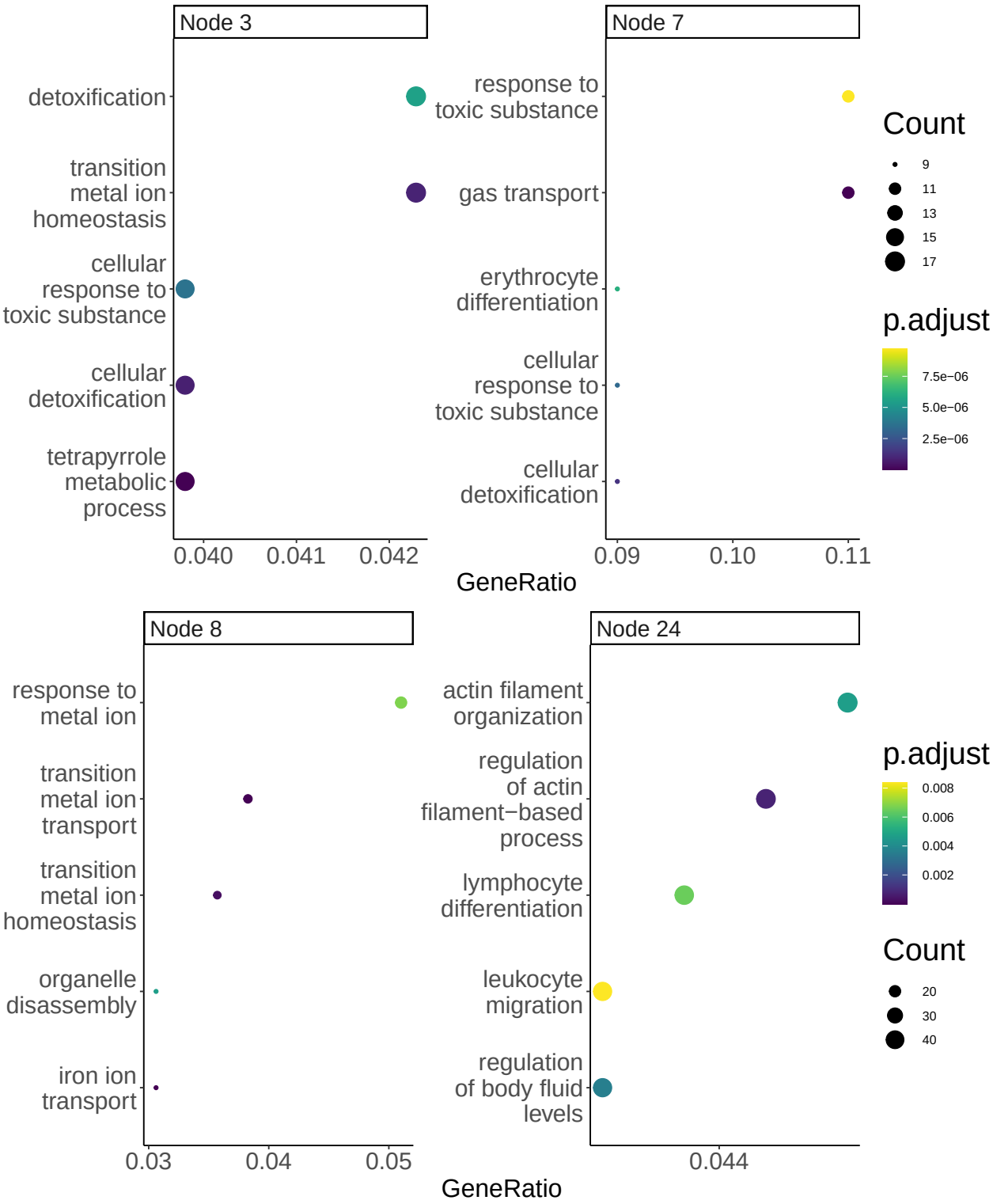


Figure A.12.: GO enrichment on selected nodes for erythrocytes. Terms enriched for node 3, node 7, node 44 and node 72 are shown.

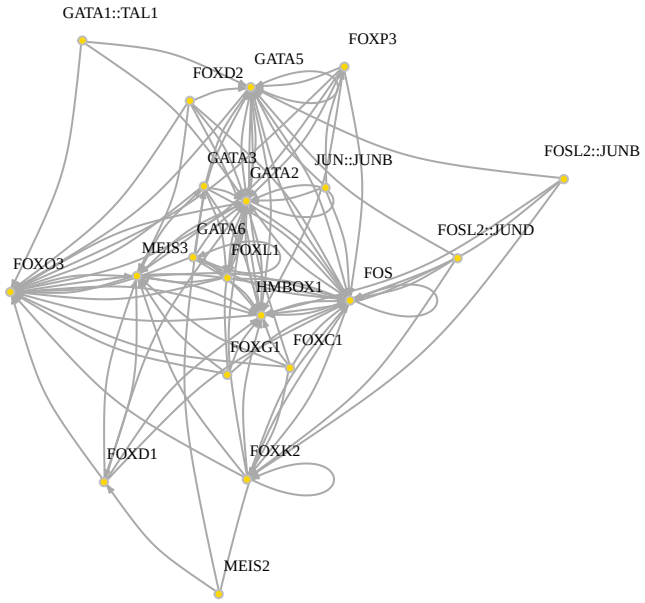


Figure A.13.: TF-TF network for erythrocytes for node 3 of the erythrocyte differentiation data set. GATA2 is highly connected.

B. Supplementary Material - Manuscripts

B.1. Manuscript - DREMflow

Manuscript to be submitted in Epigenetics & Chromatin as Methodology paper with the title **Identification of cell-type and time point specific transcriptional regulators with DREMflow** and focused on the results for the macrophage differentiation and the comparison to PECA2/TimeReg (Ramirez et al. 2017; Duren et al. 2017, 2020). This manuscript is in preparation.

Abstract

Background: Transcription factor activity during cell differentiation can be measured through gene expression and chromatin accessibility, ideally jointly over time. Integrated time course regulatory analysis yields more detailed gene regulatory networks than expression data alone. Due to the large number of parameters and tools employed in such analysis, computational workflows help to manage the inherent complexity of such analyses.

Results: Here, we describe Dynamics Regulatory Events Miner Snakemake workflow (DREMflow) which combines temporally-resolved RNA-seq and ATAC-seq data to identify

cell type and time point specific gene regulatory networks. DREMflow builds on the Dynamics Regulatory Events Miner (DREM), the workflow management system Snakemake and the package manager Mamba. It includes the processing starting from sequencing reads, quality control reports and parameters as well as additional downstream analyses for the inference of key transcription factors during differentiation. DREMflow is compared to results from studies that did not use a workflow engine as well as TransReg, a pipeline with similar aims.

Conclusions: DREMflow enables users to perform time-resolved multi-omics analysis reproducibly with minimal setup and configuration. The results compare favorably to individual analyses as well as other workflows.

Keywords: Chromatin accessibility, expression, ATAC-seq, RNA-seq, transcriptional regulation, transcription factors

Background

Cell differentiation is a prerequisite for animal life. Despite the enormous complexity of cellular organisation, we have identified key processes in transcriptional regulation as driving cellular identity. We have catalogued transcription factors and discovered the DNA motifs they bind to (Spitz and Furlong 2012).

We have learned to influence the differentiation by manipulation of transcription factors and to even revert cell identity. The prospects for human health that arise from this knowledge is substantial and ranges from treatment of rare diseases to rejuvenating therapies.

The knowledge we have acquired has raised many new questions and inspired scientists to develop technologies to study the role of gene expression in cell differentiation. Chromatin density limits the access of TFs to binding motifs and is key determinatent in epigentic gene regulation (Klemm, Shipony, and Greenleaf 2019).

The development of the Assay for transposase-accessible chromatin using sequencing (ATAC-seq) by Buenrostro et al. (2013) allows for an efficient measurement of measuring of a TF binding site can actually reached by the protein components.

The need to study the RNA-Seq and ATAC-Seq together in the context of Transcription Factor Binding Sites (TFBS) arises and it has been shown to increase our understanding of cell differentiation(L. Liu et al. 2019).

Such a *time course regulatory analysis* yields gene regulatory networks (GRN) (Duren et al. 2020). As an integrative analysis for both epigenomics and transcriptomics data Gérard et al. (2018) introduced EPIC-DREM, a machine learning pipeline that combines time-series gene expression data with time-point specific TF-gene scores to identify transcriptional regulators that can be linked to time specific changes in gene expression. EPIC-DREM uses TEPIC, a tool for transcription factor binding prediction, to compute TF-gene scores from epigenomics data such as ATAC-seq and ChIP-seq (Schmidt et al. 2019). These scores serve as input for the Dynamics Regulatory Events Miner (DREM) to explain the change in gene expression profiles by assigning TFs as top regulators which are most likely facilitating the observed changes. Gérard et al. (2018) provided detailed data and code for their epigenetic analysis of differentiation into adipocytes and osteoblasts but not a implementation of the EPIC-DREM pipeline that would allow users to perform these analyses on their own time-series epigenomics and transcriptomics data, to study effect of parameter choices or to take advantage of recent developments in package and workflow managment.

Comparable methods are few since many experimental designs only pair static gene expression and chromatin accessibility data (Berest et al. 2019; K. Zhang et al. 2019) or lack a combined analysis step (Ludwig et al. 2019). Taiji integrates gene expression and chromatin accessibility data by using publicly available histone modification data to infer chromatin interaction. The results are used to construct a network and the expression data is used to weigh the edges. Application of the PageRank test identifies key transcription regulators (K. Zhang et al. 2019). Taiji is capable of integrating Hi-C data but lacks

the time series aspect. Ludwig et al. analyze paired chromatin accessibility and gene expression time series data in parallel instead of integrating both methods like EPIC-DREM (Ludwig et al. 2019). Duren et al. presents TimeReg, which includes an integration of time series ATAC-seq and RNA-seq data including the inference of transcription factor modules and GO enrichment analysis of target genes. TimeReg differs conceptually from the EPIC-DREM pipeline (CHECK see below?). The time course regulatory analysis is performed primarily with Matlab(Duren et al. 2020). In addition to the complication arising from propitiatory software, it requires manual intervention for the installation as it does not employ a package management system.

Major types of cell differentiation have been studied with paired RNA-Seq and ATAC-Seq data. The development of promyelocytes into macrophages, neutrophils, monocytes and monocyte derived macrophages was studied over eight time points by Ramirez et al. (2017). Other cell types and setups have been studied and provide their data, e.g. (Hor et al. 2019; Xiang et al. 2017; van der Raadt et al. 2019; Ludwig et al. 2019; Ramos et al. 2023).

Bioinformatic analyses often require a variety of heterogeneous software in various stage of maturity to process the high-dimensional and complex omics data. Given the large volume of data produced, high performance computing resources are often necessary to conduct such analyses. A complete analysis workflow requires several bioinformatics tools that differ in execution environments as well as CPU and memory requirements. This frequently leads to individual scripts for each step in a pipeline and manual intervention for data management, which is laborious and error-prone. The simple alternative of concatenating all steps into a single script demand the maximal computing resources over the complete run-time and additional overhead for checkpointing which steps in a pipeline have already been produced.

Workflow management systems such as Snakemake have become invaluable for resource-efficient computational analyses (Koster and Rahmann 2012). They capture each step and allow to submit each step with the correct environment to a computing

node adequately tailored to the task.

Another problem for complex pipelines is the definition of specific version of the bioinformatics program in each step. Particularly popular, competitive tools are frequently enhanced, which might require the use of older versions in some projects to deliver computational reproducibility as well as operating system compatibility and managing library dependencies.

Conda is package manager that helps to manage project-focused, local installation of bioinformatics tools through a programmatic interface(Grüning et al. 2018). Snakemake is relying on *mamba* (https://mamba.readthedocs.io/en/latest/index.html) which for practical purposes can be considered an enhanced Conda version and defines executing environments for each step that can be installed by the user with a single command and – if configured correctly – without any additional intervention.

Here, we present Dynamics Regulatory Events Miner Snakemake workflow (DREMflow), a pipeline that combines the processing and the integration of transcriptomic and epigenomic data for the identification of transcriptional regulators in cell differentiation processes as described in Gerard et al. (Gérard et al. 2018). We focus on ATAC-seq data but retain the option to use ChIP-seq data as in the original pipeline. The implementation of DREMflow with Snakemake allows for a fluent execution of the code without manipulation of intermediate data and enables experienced users to customize it by changing or adding rules and tools. The analysis with our pipeline does not require manual inference at the file level and focuses user interactions on choosing parameters for the data set of interest. It is suitable for users with limited desire to implement all steps from scratch.

Results

Differentiation of human promyelocytes

Applied to the macrophage data from Ramirez et al. DREMflow identified SPI1 (Pu.1) as well as EGR1/2/3 among the top transcriptional regulators for the differentiation of human promyelocytes into macrophages, with each of the TFs having a rank of 10 or higher during at least two time points.

Quality Control

TODO: change qualitative statements to numbers The PCA on the RNA-seq counts showed separation of the time points, but it also placed replicate 2 distant from the other replicates. The same could not be observed in the ATAC-seq data apart from the last time point (Suppl. Fig.). The number of ATAC-seq reads at replicate 2 of 120h are substantially lower than for the other replicates, which can explain the outlier in the PCA (Suppl.Figure). Over all time points 12 genes were included with a TPM>1. At 120h the highest number of DEG compared to HL60 was identified (Table B.1).

Identification of footprints and peaks

The number of peaks between all time points varied with an average 45k. The average number of footprints identified is 200k, which is more as twice than the number reported by Ramirez, possibly due to updated databases and different footprinting tools. Similar to the DEG, the number of differential peaks increased over time (Table B.1).

Identification of top regulators

Applied DREMflow to the myeloid differentiation data from Ramirez et al. (Ramirez et al. 2017), DREMflow computed a bifurcate split node network comprising 90 nodes, 49 of those containing unique information over eight different time points (Figure B.1). The model includes 9820 target genes expressed at a log2FC greater than 1. Every TF was assigned a split score, comparable to a p-value at each split point. TFs with a split score < 0.001 are considered as significant resulting in a list of 281 different regulators over all nodes. The number of significant TF for each split point varied greatly (Figure B.1).

Alternative: We selected the ten highest ranked TFs at each node to identify recurring TFs but included the top two regulators irrespective of recurrence to be included as important regulators inferred in the model.

Recurring TF made up 39 TF while the list of highest two ranked TF had 40 excluding those that are considered as recurring. We focused on the following three characteristics of the selected recurring regulators; number of appearances overall and time point wise, the expression profile and the number of predicted target genes at each time point (Figure B.2).

Gene ontology enrichment of DEG in target gene clusters

For each split node a gene ontology (GO) enrichment of biological processes was performed on the target gene sets. Myeloid related terms such as “regulation of hemopoiesis” and “myeloid cell differentiation” were associated with upregulated nodes on earlier time points building a clear path from Node 1 to Node 43 (Figure B.3). Node 7, node 24 and node 25 are enriched for “ERK1 and ERK2 cascade (Figure B.6). The ERK signaling cascade is essential for macrophage differentiation (Richardson et al. 2015). General terms appear on downregulated nodes (Figure B.7). DREMflow is able to identify co-expressed gene clusters for distinct biological processes.

TimeReg

Duren et al. introduced their statistical approach to paired expression and chromatin accessibility (PECA) in 2017 on static data (Duren et al. 2017). Building on this, they presented TimeReg (time course regulatory analysis) in 2020 for the analysis of time series data (Duren et al. 2020). A major output of PECA2 is the identification of regulation strength of transcription factors to target genes, which is used to identifies core regulatory modules via non-negative matrix factorization at each time point. The time component is added via driver TF that primarily regulate expression change between the time points focusing solely on changes in regulation strength of transcription factors to up-regulated target genes. For the inference of GRN PECA2/Timereg excludes transcription factors as target genes (Duren et al. 2020). Both tools, PECA2 and TimeReg are available on Github and both are implemented in MatLab combined with a shell launch script. The pre-processing such as alignment and identification of gene counts is not included. During the execution we noticed that an Ensembl-style reference genome (without “chr” prefix) was not supported. For the comparison of the results we focused on the driver TF identified by TimeReg, since those add the time component to the GRN. Overall 20 driver TF were identified by TimeReg (Table B.2) with two of those overlapping the top regulators from the DREMflow analysis (Figure B.5). We noticed a lack of comparability between the results due to different methods used for identifying transcription factor binding motifs. Out of the 432 TF included in DREMflow and 349 TF from TimeReg only 188 overlapped overall with no heterodimers included in TimeReg. Different names being used for the same transcription factor still pose a challenge in the field of gene regulation. Instead of focusing on the biological results, we compared the features and user handling between the two tools (Table B.3).

Application to pseudotime series data

Execution time and computational requirements

RNA-seq and ATAC-seq date are processed in parallel until they are combined in the filtering of TF-gene affinities and the computation of the time integrative gene regulation mode. Overall ATAC-seq takes longer than RNA-seq with the alignment of ATAC-seq files is the most time consuming step followed by the BWA index building, which is only required once, and footprinting of each sample (Figure B.8).
The conda/mamba environments require approximately 80GB of storage for all environments.
Overall the DREMflow analyzed the macrophage data set in under 11h.

Discussion

The combination of epigenomics and transcriptomics time-series data in the context of differentiation and cell-state transition enables us to infer gene regulatory processes. The identification of transcriptional regulators is not only relevant to improve in-vitro disease models but will be crucial as well in the context of personalized medicine. The machine learning pipeline EPIC-DREM that was introduced by Gerard et al. provides a framework for this kind of multi-omics analysis (Gérard et al. 2018). The disadvantage of many tools in bioinformatics and in this context EPIC-DREM as well is the requirement of advanced computational skills to execute them or a lack of reproducibility in case of graphical user interfaces (GUI). Online tools might not be GDPR compliant. Heterogenous sequencing data requires different processing tools which leads to many scripts for one analysis. Analysis of high throughput data often requires computing clusters that complicate the installation and setup due to missing admin rights for the general user.
With DREMflow we provide a fully executable pipeline that addresses the above

mentioned issues. Implemented in the framework of Snakemake with conda/mamba environments DREMflow takes care of the installation of all necessary tools, intermediate steps and visualization of results.

Bulk-seq vs. single-cell

We applied DREMflow to the macrophage differentiation data set by Ramirez et al. to demonstrate the application and output (Ramirez et al. 2017). Macrophages are well studied in terms off differentiation and polarization (Richardson et al. 2015; D. A. Hume and Himes 2003; Orecchioni et al. 2019). DREMflow identified SPI1 and EGR2 among the most striking regulators in terms of appearance, ranking and expression profile. TF of AP-1 and ETS families were highlighted as well as significant regulators with high rankings. This concurs with current knowledge of transcriptional regulation in macrophages (Gans et al. 2020; D. A. Hume and Himes 2003; David A. Hume, Summers, and Rehli 2016; Jego et al. 2014; Labzin et al. 2015; Behmoaras et al. 2008). DREMflow defined co-expressed gene sets being enrichment for different biological processes with late strongly upregulated nodes being enriched for ERK1-ERK2 cascade, which is essential for macrophage differentiation (Richardson et al. 2015). Further investigation of TF in the ERK related nodes a could lead to the identification of additional regulators.

We compared DREMflow to TimeReg(Duren et al. 2020), the only method to our knowledge that performs a similar time series integration as DREMflow. While both methods aim to achieve the same results with the same kind of data, TimeReg adds the time series component at the end of building timepoint-wise GRN, while DREMflow includes this component by assigning the TF according to the changes in expression. The comparability proved to be difficult due to different names for transcription factors coming from different methods to identify transcription factor binding sites.

The 50kb window around transcription start sites used in TEPIC seems arbitrary and might be the bottle neck of the pipeline. For this reason we included differential footprinting in

the pipeline to enable the user to observe the TF activity and changes in activity between the different timepoints. In the future TEPIC will be replaced by Hi-C data integration.

Conclusion

- DREMflow identifies top regulators
- Objectives and conclusions have to match

our objective: providing a full reproducible and “easily” executable pipeline

DREMflow enables users to perform complex multi-omics analysis reproducibly with minimal setup and configuration.

single cell application

Methods

DREMflow

DREMflow utilizes the workflow management system Snakemake to combine the analysis steps into one executable and reproducible pipeline (Koster and Rahmann 2012). Snakemake uses mamba which minimize the installation requirements as the necessary tools are installed in separate conda environments(https://mamba.readthedocs.io/en/latest/index.html). In the Snakemake framework each step is defined as a *rule.* Based on the input and output files of all rules, Snakemake is able to connect and execute them in the correct order while taking into account the necessary dependencies such as annotation files or the reference genome. DREMflow takes FASTQ files of time series transcriptomic and epigenomic data as default input. It is possible to use preprocessed data such as narrowPeak files, BAM files or gene counts and start the pipeline at later stages. Compatibility issues might arise if the genome version used for the alignment is incompatible with the

annotation version used for the quantification of the expression data. Running DREMflow from FASTQ inputs delivers the most consistent results and is recommended.

B.1.0.0.1. * Set up of tools and reference data

Exact definition of the version of a tool required for each rule provides installation routines with minimal overhead for users of a pipeline.

Since not all tools are available via mamba, some installations is done via set up rules included in the pipeline. TEPIC(Schmidt et al. 2019) and iDREM(Ding et al. 2018) are thus installed directly to the project directory. Data required for the Regulatory Genomics Toolbox (Gusmao et al. 2014) is installed as well through a dedicated rule.

The reference genome and annotation files are acquired by Snakemake wrapper functions. This ensures a compatibility between preprocessed data throughout the whole analysis. The user provides species, build and release version. The 2bit version and chromosome sizes are derived from the reference genome via UCSC tools.

B.1.0.0.2. * Quality Control

Since two sets of different omics data are processed simultaneously, the workflow is split in two parts that are combined in the preparation of TF-gene affinities and the execution of DREM. The workflow starts with a basic quality control (QC) of the raw FASTQ files with FASTQC(Andrews 2010). DESeq2(Love, Huber, and Anders 2014) is used to compute PCA and sample-to-sample distance for RNA-seq and ATAC-seq data. The resulting figures are displayed in the final result document together with counts of genes, peaks, footprints and differentially expressed genes. The results from FASTQC, the mapping steps and the feature counts are summarized in a MultiQC report(Ewels et al. 2016).

B.1.0.0.3. * Gene expression

FASTQ files are aligned via STAR(Dobin et al. 2012). The resulting BAM files are further processed with the featureCounts function of the Rsubreader package to generate gene counts(Liao, Smyth, and Shi 2019). Gene counts are normalized with DESeq2 before the TPM (Transcripts Per Kilobase Million) are calculated(Love, Huber, and Anders 2014). Low abundant genes below a TPM threshold of 1 in all samples are excluded from further analysis.

Chromatin accessibility

The pipeline utilizes BWA-MEM for the alignment of ATAC-seq data(H. Li and Durbin 2010). To identify open chromatin regions peak calling is performed on the aligned ATAC-seq/ChIP-seq files. The peak caller used is Genrich (available at https://github.com/jsh58/Genrich), as it can perform peak calling over replicates. The BAM files are merged over replicates in parallel, as both peaks and alignment files are needed for the footprinting step. The merged BAM as well as the individual BAM files can be optionally converted to bigwig files for visualization. Then, footprinting is performed by the Regulatory Genomics Toolbox (RGT). With the footprints as regions of interest, TF-gene affinities are inferred by TEPIC (Schmidt et al. 2019). The resulting TF-gene affinities are filtered for expression, removing all TF-gene connections of TF that are not expressed at all in the data. In addition, differential footprinting is performed to determine the change in TF activity for different time points.

Time resolved integrative network

The resulting TF-gene affinities as well as the filtered expression data serves as input for DREM to infer the dynamic regulatory map (Schulz et al. 2012). The resulting model file can be loaded into the DREM graphical user interface for visualization purposes. For the computation of the DREM model we utilize iDREM (https://github.com/phoenixding/idrem

#interactive-visualization/) to take advantage of the JSON file containing information about the whole model enabling a command line extraction of top regulators and target genes. dremFLow builds a comprehensive summary in R with a final report as HTML document. The final report includes a GO enrichment analysis on the target gene clusters and top regulators, heatmaps of the top predictor expression profiles, a list of TFs identified as top regulator in multiple gene clusters, number of predicted target genes and the average expression profile of target genes for each top regulator.

Data

To test the pipeline we performed the analysis on a 5-day paired ATAC-seq and RNA-seq time series of human myeloid differentiation (GSE79046) on data of the cellular commitment into macrophages as well as neutrophils and monocytes (Ramirez et al. 2017) and a pseudotime series from CD34+ hematopoietic stem and progenitor cells (HSPCs) that were differentiated to obtain enucleated reticulocytes and FACS sorted to identify eight stages of differentiation (Ludwig et al. 2019).

Application of TimeReg

ATAC-seq samples were aligned with BWA-MEM(H. Li and Durbin 2010) and replicates were merged with samtools(Danecek et al. 2021). RNA-seq alignment was performed with STAR(Dobin et al. 2012), counts were identified with featureCounts(Liao, Smyth, and Shi 2014) and converted to TPM. The required scores for TimeReg were calculated with PECA2 (Duren et al. 2017).

Results Tables

Table B.1.: Number of included genes, differentially expressed genes, peaks, footprints and differentially expressed peaks at each time point.

Time point	DEG	# peaks	# footprints	Diff. peaks
HL60		52505	266489	
3h	1745	31327	153285	1367
6h	2809	25420	134883	3316
12h	4392	34420	163247	5543
24h	5364	51060	233856	22173
48h	5667	42648	226356	25237
96h	6188	40262	202307	26267
120h	6006	66567	282204	35081

Table B.2.: Driver TFs identified by TimeReg for macrophage differentiation

Module	Driver TFs
24h-module1	TOX2
48h-module2	BHLHE41
	ATF3
	HOXA5
	HOXA4
	RXRB
	HOXA2
	SOX4
	HOXA1
	CEBPB
	TEAD3
	NFE2L1
	FOXO1
	JUN
6h-module1	GLI3
	ZBTB7C
	SMAD6
	ZNF354C
	CUX2
	NFATC4

Table B.3.: Comparison between DREMflow and PECA2/TimeReg based on different features.

Features	DREMflow	PECA2/TimeReg
Preprocessing	FastQC, Alignment DESeq2, featureCounts	Not included
Reference genome and gene annotation	User specified version	Specific version only
Module and node assignment	All TG and some TFs by significance	All TFs and all TG Few driver TFs
Output	DREM model for GUI Results HTML Downstream analysis Top regulators GO enrichment TF-TF networks Visualization	Modules TRS heatmaps Driver regulators Ancestor-descendant relationship
Programming languages	Shell, Python, R	Shell, Matlab
Installation requirements	Snakemake, conda and mamba	Homer, Matlab, Bedtools, Samtools
Rerun	From any point of the pipeline	From the beginning
Flexibility	Addition of rules Replacement of rules	No flexibility

Results Figures

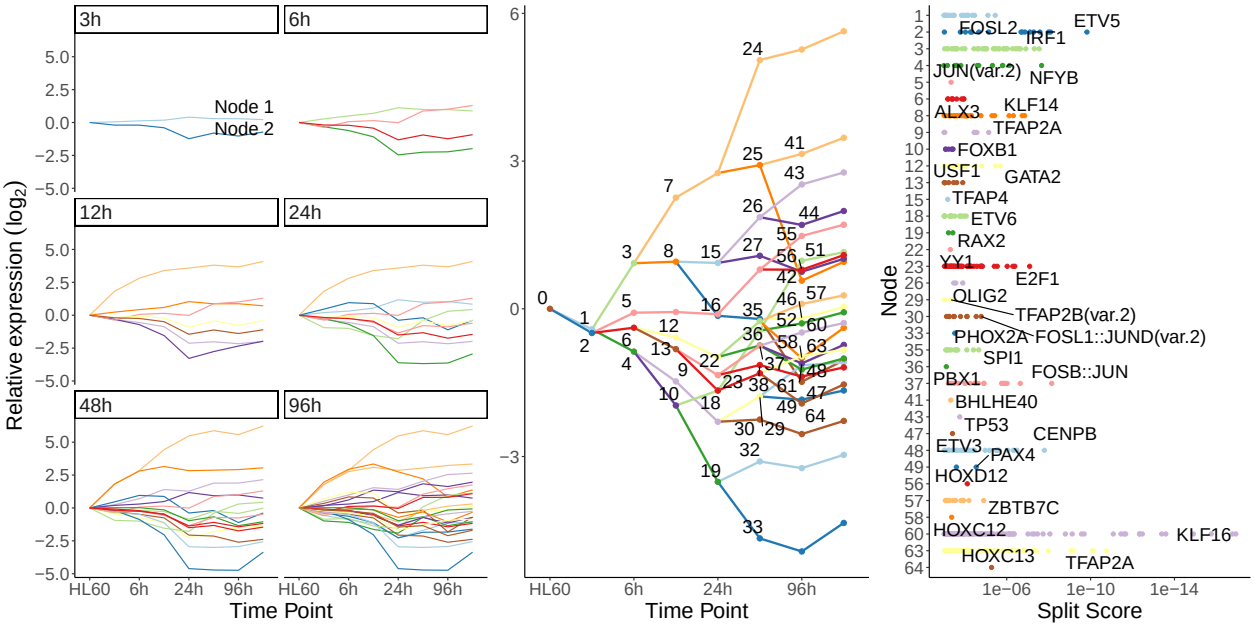


Figure B.1.: (A) Expression profile for the co-expressed gene clusters as estimated by DREMflow for each timepoint. Each color represents an individual co-expressed gene cluster. (B) The computed model showing all split nodes. Nodes after a split are annotated, since they contain unique information such as the assigned genes and TF after the split. (C) TF assigned to the nodes on the y-axis according to their split score. The first and highest ranked TF is labelled.

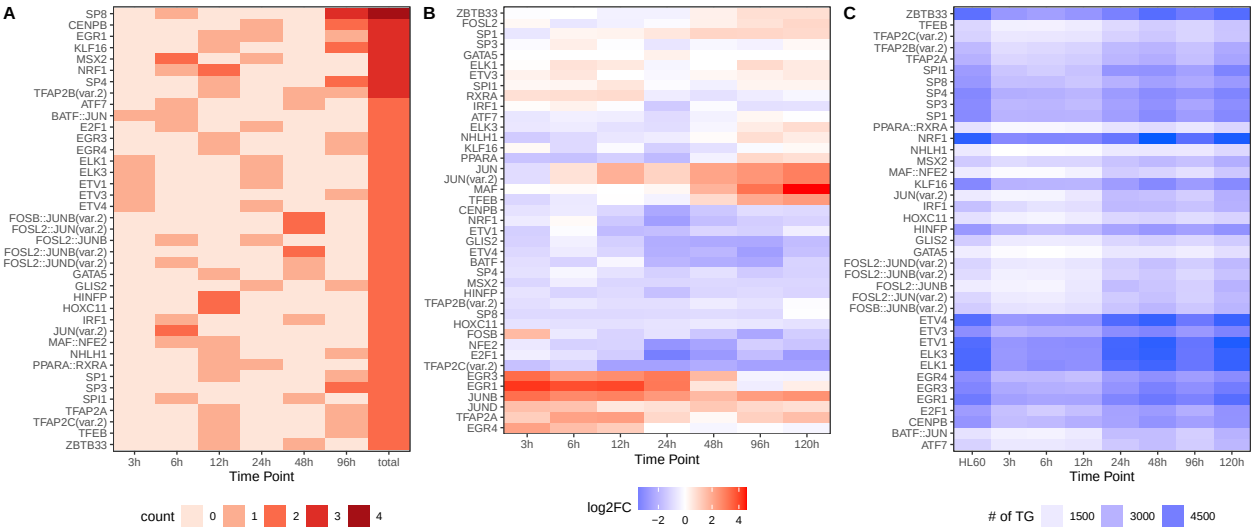


Figure B.2.: (A) Represenation of number of apprearances of reoccurring TF for each time-point and in total in the estimated model. (B) The expression profiles for reoccurring TF as log2FC. (C) Number of estimated target genes for reoccuring TF.

References

Abdolhosseini, Farzad, Behrooz Azarkhalili, Abbas Maazallahi, Aryan Kamal, Seyed Abolfazl Motahari, Ali Sharifi-Zarchi, and Hamidreza Chitsaz. 2019. “Cell Identity Codes: Understanding Cell Identity from Gene Expression Profiles Using Deep Neural Networks.” *Scientific Reports* 9 (1): 2342. https://doi.org/10.1038/s41598-019-38798-y.

Ai, Zhichao, and Irina A. Udalova. 2020. “Transcriptional Regulation of Neutrophil Differentiation and Function During Inflammation.” *Journal of Leukocyte Biology* 107 (3): 419–30. https://doi.org/10.1002/JLB.1RU1219-504RR.

Álvarez-Errico, Damiana, Roser Vento-Tormo, Michael Sieweke, and Esteban Ballestar. 2015. “Epigenetic Control of Myeloid Cell Differentiation, Identity and Function.” *Nature Reviews Immunology* 15 (1): 7–17. https://doi.org/10.1038/nri3777.

Ambrosini, Giovanna, Ilya Vorontsov, Dmitry Penzar, Romain Groux, Oriol Fornes, Daria D. Nikolaeva, Benoit Ballester, et al. 2020. “Insights Gained from a Comprehensive All-Against-All Transcription Factor Binding Motif Benchmarking Study.” *Genome Biology* 21 (1): 114. https://doi.org/10.1186/s13059-020-01996-3.

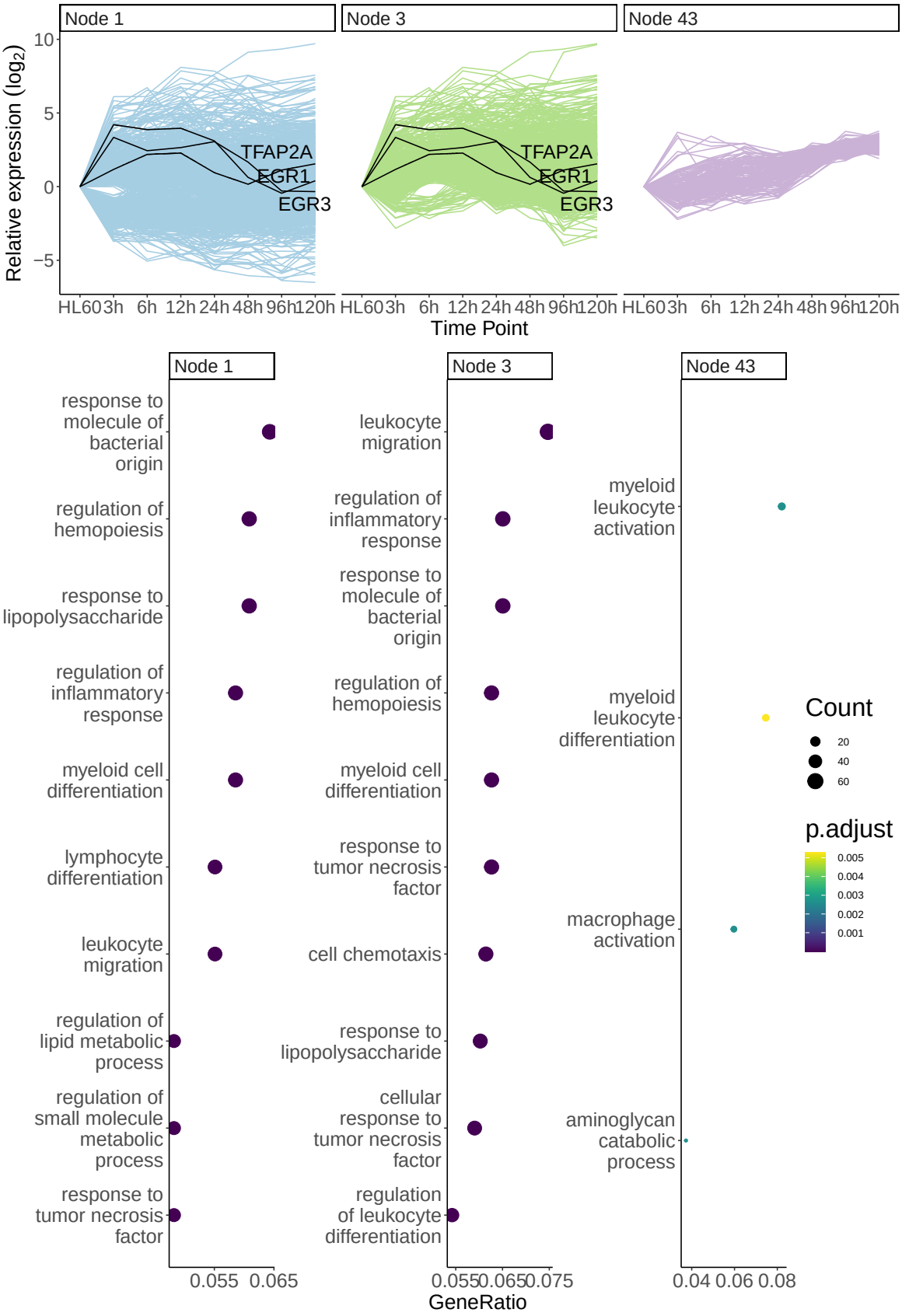


Figure B.3.: (A) Expression profiles of all TG in a co-expressed gene cluster of nodes 1, 3, 14 and 43. Highlighted are expression profiles of top regulators that are TG in this clusters. (B) Gene Ontology enrichment of biological processes for the selected nodes showing hemopoiesis and ERK cascade related terms.

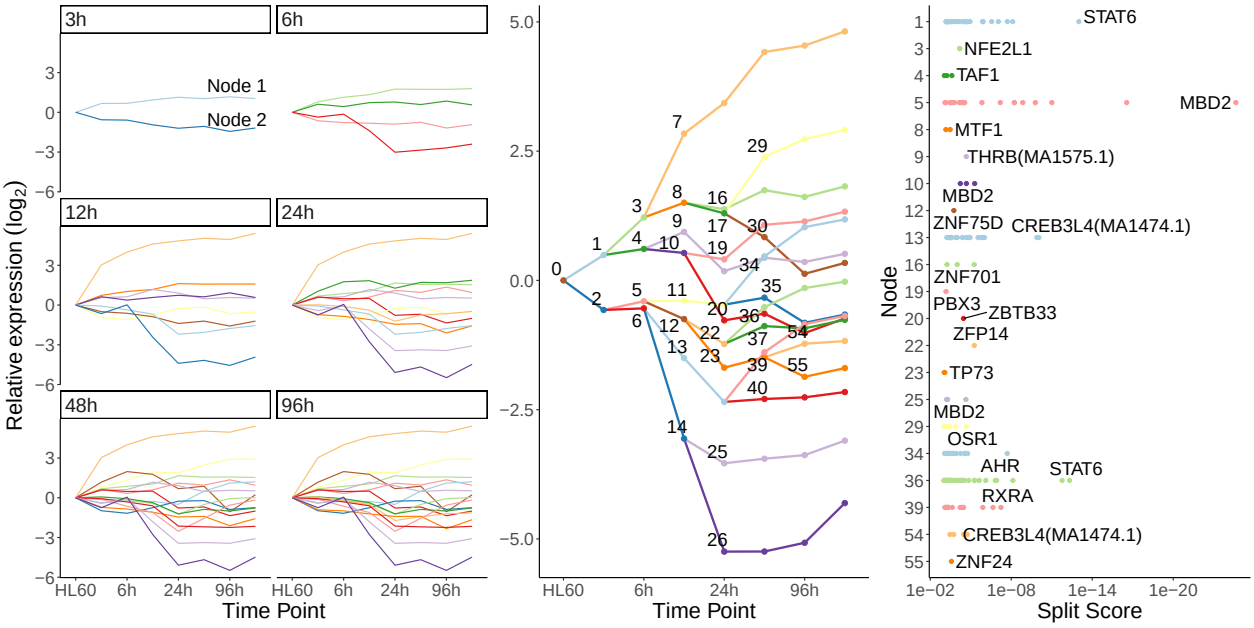


Figure B.4.: (A) Expression profile for the co-expressed gene clusters as estimated by DREMflow for each timepoint. Each color represents an individual co-expressed gene cluster. (B) The computed model showing all split nodes using TF clusters (C) TF clusters assigned to the nodes on the y-axis according to their split score. The first and highest ranked TF cluster is labelled. Each cluster is represented by the first TF in the cluster.

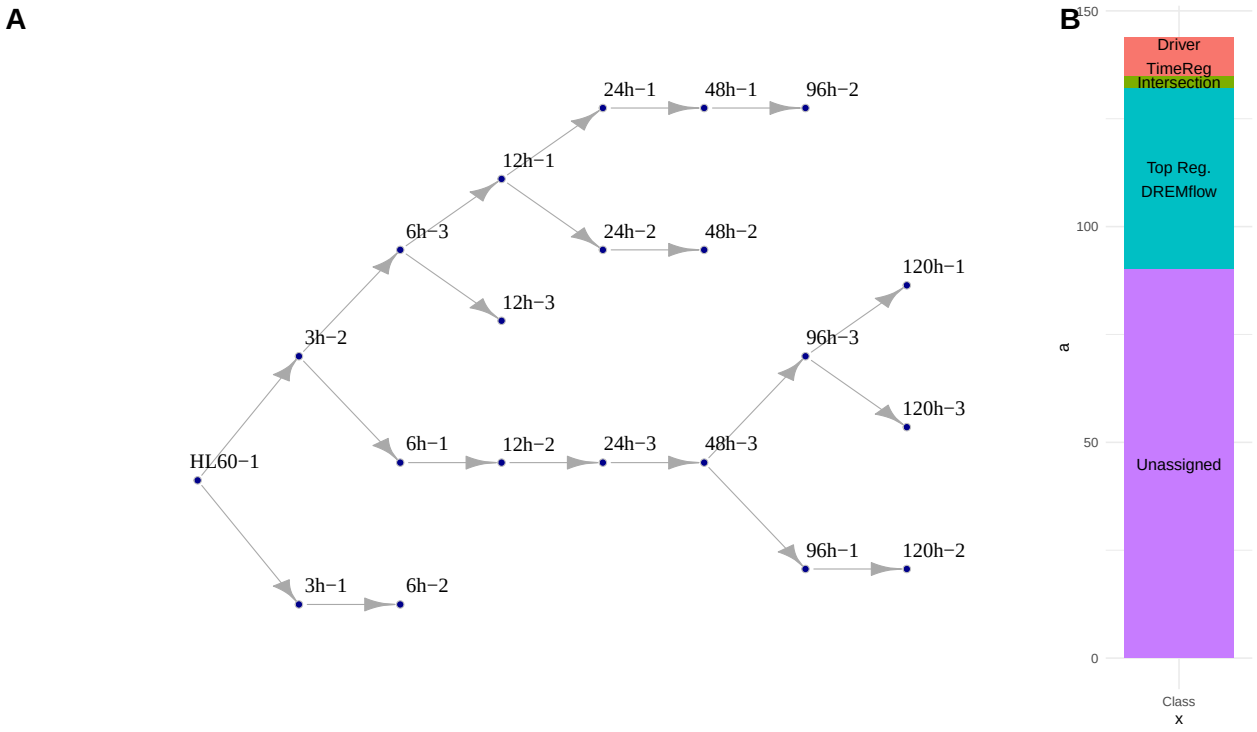


Figure B.5.: TimeReg results

Amemiya, Haley M., Anshul Kundaje, and Alan P. Boyle. 2019. "The ENCODE Blacklist: Identification of Problematic Regions of the Genome." *Scientific Reports* 9 (1): 9354. https://doi.org/10.1038/s41598-019-45839-z.

Andrews, S. 2010. *FastQC: A Quality Control Tool for High Throughput Sequence Data [Online].* http://www.bioinformatics.babraham.ac.uk/projects/fastqc/.

Antunes, Ricardo F, Cláudia Brandão, Margarida Maia, and Fernando A Arosa. 2011. "Red Blood Cells Release Factors with Growth and Survival Bioactivities for Normal and Leukemic T Cells." *Immunology & Cell Biology* 89 (1): 111–21. https://doi.org/10.1038/icb.2010.60.

Arenas, E., M. Denham, and J. C. Villaescusa. 2015. "How to Make a Midbrain Dopaminergic Neuron." *Development* 142 (11): 1918–36. https://doi.org/10.1242/dev.097394.

Arenas, Ernest. 2014. "Wnt Signaling in Midbrain Dopaminergic Neuron Development and Regenerative Medicine for Parkinson's Disease." *Journal of Molecular Cell Biology* 6 (1): 42–53. https://doi.org/10.1093/jmcb/mju001.

Arosa, Fernando, Carlos Pereira, and Ana Fonseca. 2004. "Red Blood Cells as Modula-

tors of T Cell Growth and Survival." *Current Pharmaceutical Design* 10 (2): 191–201. https://doi.org/10.2174/1381612043453432.

Ashburner, Michael, Catherine A. Ball, Judith A. Blake, David Botstein, Heather Butler, J. Michael Cherry, Allan P. Davis, et al. 2000. "Gene Ontology: Tool for the Unification of Biology." *Nature Genetics* 25 (1): 25–29. https://doi.org/10.1038/75556.

Atlasi, Yaser, and Hendrik G. Stunnenberg. 2017. "The Interplay of Epigenetic Marks During Stem Cell Differentiation and Development." *Nature Reviews Genetics* 18 (11): 643–58. https://doi.org/10.1038/nrg.2017.57.

Balwierz, Piotr J., Mikhail Pachkov, Phil Arnold, Andreas J. Gruber, Mihaela Zavolan, and Erik Van Nimwegen. 2014. "ISMARA: Automated Modeling of Genomic Signals as a Democracy of Regulatory Motifs." *Genome Research* 24 (5): 869–84. https://doi.org/10.1101/gr.169508.113.

Bar-Joseph, Ziv, Anthony Gitter, and Itamar Simon. 2012. "Studying and Modelling Dynamic Biological Processes Using Time-Series Gene Expression Data." *Nature Reviews Genetics* 13 (8): 552–64. https://doi.org/10.1038/nrg3244.

Basso, Katia, Adam A Margolin, Gustavo Stolovitzky, Ulf Klein, Riccardo Dalla-Favera, and Andrea Califano. 2005. "Reverse Engineering of Regulatory Networks in Human B Cells." *Nature Genetics* 37 (4): 382–90. https://doi.org/10.1038/ng1532.

Behmoaras, Jacques, Gurjeet Bhangal, Jennifer Smith, Kylie McDonald, Brenda Mutch, Ping Chin Lai, Jan Domin, et al. 2008. "Jund is a determinant of macrophage activation and is associated with glomerulonephritis susceptibility." *Nature Genetics* 40 (5): 553–59. https://doi.org/10.1038/ng.137.

Belton, Jon-Matthew, Rachel Patton McCord, Johan Harmen Gibcus, Natalia Naumova, Ye Zhan, and Job Dekker. 2012. "Hi–C: A Comprehensive Technique to Capture the Conformation of Genomes." *Methods* 58 (3): 268–76. https://doi.org/10.1016/j.ymeth.2012.05.001.

Bencheikh, Laura, M'Boyba Khadija Diop, Julie Rivière, Aygun Imanci, Gerard Pierron, Sylvie Souquere, Audrey Naimo, et al. 2019. "Dynamic Gene Regulation by Nuclear Colony-Stimulating Factor 1 Receptor in Human Monocytes and Macrophages." *Na-*

ture Communications 10 (1): 1935. https://doi.org/10.1038/s41467-019-09970-9.

Berest, Ivan, Christian Arnold, Armando Reyes-Palomares, Giovanni Palla, Kasper Dindler Rasmussen, Holly Giles, Peter-Martin Bruch, et al. 2019. "Quantification of Differential Transcription Factor Activity and Multiomics-Based Classification into Activators and Repressors: diffTF." *Cell Reports* 29 (10): 3147–3159.e12. https://doi.org/10.1016/j.celrep.2019.10.106.

Biedler, J. L., S. Roffler-Tarlov, M. Schachner, and L. S. Freedman. 1978. "Multiple neurotransmitter synthesis by human neuroblastoma cell lines and clones." *Cancer Research* 38 (11 Pt 1): 3751–57.

Bloushtain-Qimron, Noga, Jun Yao, Michail Shipitsin, Reo Maruyama, and Kornelia Polyak. 2009. "Epigenetic Patterns of Embryonic and Adult Stem Cells." *Cell Cycle* 8 (6): 809–17. https://doi.org/10.4161/cc.8.6.7938.

Bohmann, Dirk, Timothy J. Bos, Arie Admon, Tetsuji Nishimura, Peter K. Vogt, and Robert Tjian. 1987. "Human Proto-Oncogene c- *Jun* Encodes a DNA Binding Protein with Structural and Functional Properties Pf Transcription Factor AP-1." *Science* 238 (4832): 1386–92. https://doi.org/10.1126/science.2825349.

Briggs, Robert, and Thomas J. King. 1952. "Transplantation of Living Nuclei from Blastula Cells into Enucleated Frogs' Eggs." *Proceedings of the National Academy of Sciences* 38 (5): 455–63. https://doi.org/10.1073/pnas.38.5.455.

Buccitelli, Christopher, and Matthias Selbach. 2020. "mRNAs, Proteins and the Emerging Principles of Gene Expression Control." *Nature Reviews Genetics* 21 (10): 630–44. https://doi.org/10.1038/s41576-020-0258-4.

Buenrostro, Jason D, Paul G Giresi, Lisa C Zaba, Howard Y Chang, and William J Greenleaf. 2013. "Transposition of Native Chromatin for Fast and Sensitive Epigenomic Profiling of Open Chromatin, DNA-Binding Proteins and Nucleosome Position." *Nature Methods* 10 (12): 1213–18. https://doi.org/10.1038/nmeth.2688.

Cao, Junyue, Darren A. Cusanovich, Vijay Ramani, Delasa Aghamirzaie, Hannah A. Pliner, Andrew J. Hill, Riza M. Daza, et al. 2018. "Joint Profiling of Chromatin Accessibility and Gene Expression in Thousands of Single Cells." *Science* 361 (6409):

1380–85. https://doi.org/10.1126/science.aau0730.

Carbon, Seth, Amelia Ireland, Christopher J. Mungall, ShengQiang Shu, Brad Marshall, Suzanna Lewis, the AmiGO Hub, and the Web Presence Working Group. 2009. "AmiGO: Online Access to Ontology and Annotation Data." *Bioinformatics* 25 (2): 288–89. https://doi.org/10.1093/bioinformatics/btn615.

Catalanotto, Caterina, Carlo Cogoni, and Giuseppe Zardo. 2016. "MicroRNA in Control of Gene Expression: An Overview of Nuclear Functions." *International Journal of Molecular Sciences* 17 (10): 1712. https://doi.org/10.3390/ijms17101712.

Chen, Huidong, Caleb Lareau, Tommaso Andreani, Michael E. Vinyard, Sara P. Garcia, Kendell Clement, Miguel A. Andrade-Navarro, Jason D. Buenrostro, and Luca Pinello. 2019. "Assessment of Computational Methods for the Analysis of Single-Cell ATAC-Seq Data." *Genome Biology* 20 (1): 241. https://doi.org/10.1186/s13059-019-1854-5.

Chen, Siyuan, Jing Yang, Yuquan Wei, and Xiawei Wei. 2020. "Epigenetic regulation of macrophages: from homeostasis maintenance to host defense." *Cellular & Molecular Immunology* 17 (1): 36–49. https://doi.org/10.1038/s41423-019-0315-0.

Chen, Taiping, and Sharon Y. R. Dent. 2014. "Chromatin Modifiers and Remodellers: Regulators of Cellular Differentiation." *Nature Reviews Genetics* 15 (2): 93–106. https://doi.org/10.1038/nrg3607.

Cheng, Yong, Weisheng Wu, Swathi Ashok Kumar, Duonan Yu, Wulan Deng, Tamara Tripic, David C. King, et al. 2009. "Erythroid GATA1 Function Revealed by Genome-Wide Analysis of Transcription Factor Occupancy, Histone Modifications, and mRNA Expression." *Genome Research* 19 (12): 2172–84. https://doi.org/10.1101/gr.098921.109.

Coulon, Antoine, Carson C. Chow, Robert H. Singer, and Daniel R. Larson. 2013. "Eukaryotic Transcriptional Dynamics: From Single Molecules to Cell Populations." *Nature Reviews Genetics* 14 (8): 572–84. https://doi.org/10.1038/nrg3484.

Danecek, Petr, James K Bonfield, Jennifer Liddle, John Marshall, Valeriu Ohan, Martin O Pollard, Andrew Whitwham, et al. 2021. "Twelve Years of SAMtools and BCFtools."

GigaScience 10 (2): giab008. https://doi.org/10.1093/gigascience/giab008.

Dekkers, Koen F., Annette E. Neele, J. Wouter Jukema, Bastiaan T. Heijmans, and Menno P. J. de Winther. 2019. "Human Monocyte-to-Macrophage Differentiation Involves Highly Localized Gain and Loss of DNA Methylation at Transcription Factor Binding Sites." *Epigenetics & Chromatin* 12 (1): 34. https://doi.org/10.1186/s13072-019-0279-4.

Ding, Jun, James S. Hagood, Namasivayam Ambalavanan, Naftali Kaminski, and Ziv Bar-Joseph. 2018. "iDREM: Interactive visualization of dynamic regulatory networks." *PLoS computational biology* 14 (3): e1006019. https://doi.org/10.1371/journal.pcbi.1006019.

Dobin, Alexander, Carrie A. Davis, Felix Schlesinger, Jorg Drenkow, Chris Zaleski, Sonali Jha, Philippe Batut, Mark Chaisson, and Thomas R. Gingeras. 2012. "STAR: Ultrafast Universal RNA-Seq Aligner." *Bioinformatics* 29 (1): 15–21. https://doi.org/10.1093/bioinformatics/bts635.

Dong, Yong, Yimeng Zhang, Yongping Zhang, Xu Pan, Ju Bai, Yijin Chen, Ya Zhou, et al. 2022. "Dissecting the Process of Human Neutrophil Lineage Determination by Using Alpha-Lipoic Acid Inducing Neutrophil Deficiency Model." *Redox Biology* 54 (August): 102392. https://doi.org/10.1016/j.redox.2022.102392.

Duren, Zhana, Xi Chen, Rui Jiang, Yong Wang, and Wing Hung Wong. 2017. "Modeling Gene Regulation from Paired Expression and Chromatin Accessibility Data." *Proceedings of the National Academy of Sciences* 114 (25): E4914–23. https://doi.org/10.1073/pnas.1704553114.

Duren, Zhana, Xi Chen, Jingxue Xin, Yong Wang, and Wing Wong. 2020. "Time Course Regulatory Analysis Based on Paired Expression and Chromatin Accessibility Data." *Genome Research*, March, gr.257063.119. https://doi.org/10.1101/gr.257063.119.

Dzierzak, E., and S. Philipsen. 2013. "Erythropoiesis: Development and Differentiation." *Cold Spring Harbor Perspectives in Medicine* 3 (4): a011601–1. https://doi.org/10.1101/cshperspect.a011601.

Efthymiou, Anastasia G, Guibin Chen, Mahendra Rao, Guokai Chen, and Manfred Boehm.

2014. “Self-Renewal and Cell Lineage Differentiation Strategies in Human Embryonic Stem Cells and Induced Pluripotent Stem Cells.” *Expert Opinion on Biological Therapy* 14 (9): 1333–44. https://doi.org/10.1517/14712598.2014.922533.

Ernst, Jason, Oded Vainas, Christopher T Harbison, Itamar Simon, and Ziv Bar-Joseph. 2007. “Reconstructing Dynamic Regulatory Maps.” *Molecular Systems Biology* 3 (1): 74. https://doi.org/10.1038/msb4100115.

Evans, M. J., and M. H. Kaufman. 1981. “Establishment in Culture of Pluripotential Cells from Mouse Embryos.” *Nature* 292 (5819): 154–56. https://doi.org/10.1038/292154a0.

Ewels, Philip, Måns Magnusson, Sverker Lundin, and Max Käller. 2016. “MultiQC: Summarize Analysis Results for Multiple Tools and Samples in a Single Report.” *Bioinformatics* 32 (19): 3047–48. https://doi.org/10.1093/bioinformatics/btw354.

Feles, Sebastian, Christian Overath, Sina Reichardt, Sebastian Diegeler, Claudia Schmitz, Jessica Kronenberg, Christa Baumstark-Khan, Ruth Hemmersbach, Christine E. Hellweg, and Christian Liemersdorf. 2022. “Streamlining Culture Conditions for the Neuroblastoma Cell Line SH-SY5Y: A Prerequisite for Functional Studies.” *Methods and Protocols* 5 (4): 58. https://doi.org/10.3390/mps5040058.

Fisher, Amanda G. 2002. “Cellular Identity and Lineage Choice.” *Nature Reviews Immunology* 2 (12): 977–82. https://doi.org/10.1038/nri958.

Fong, Bensun C., Imane Chakroun, Mohamed Ariff Iqbal, Smitha Paul, Joseph Bastasic, Daniel O’Neil, Edward Yakubovich, et al. 2022. “The Rb/E2F Axis Is a Key Regulator of the Molecular Signatures Instructing the Quiescent and Activated Adult Neural Stem Cell State.” *Cell Reports* 41 (5): 111578. https://doi.org/10.1016/j.celrep.2022.111578.

Fu, Xing, Ning Fu, Song Guo, Zheng Yan, Ying Xu, Hao Hu, Corinna Menzel, et al. 2009. “Estimating Accuracy of RNA-Seq and Microarrays with Proteomics.” *BMC Genomics* 10 (1): 161. https://doi.org/10.1186/1471-2164-10-161.

Gans, Ian, Ellen I. Hartig, Shusen Zhu, Andrea R. Tilden, Lucie N. Hutchins, Nathaniel J. Maki, Joel H. Graber, and James A. Coffman. 2020. “Klf9 Is a Key Feedforward Reg-

ulator of the Transcriptomic Response to Glucocorticoid Receptor Activity." *Scientific Reports* 10 (1): 11415. https://doi.org/10.1038/s41598-020-68040-z.

Gérard, Deborah, Florian Schmidt, Aurélien Ginolhac, Martine Schmitz, Rashi Halder, Peter Ebert, Marcel H. Schulz, Thomas Sauter, and Lasse Sinkkonen. 2018. "Temporal enhancer profiling of parallel lineages identifies AHR and GLIS1 as regulators of mesenchymal multipotency." *Nucleic Acids Research*, December. https://doi.org/10.1093/nar/gky1240.

Ghanem, N., M. G. Andrusiak, D. Svoboda, S. M. Al Lafi, L. M. Julian, K. A. McClellan, Y. De Repentigny, et al. 2012. "The Rb/E2F Pathway Modulates Neurogenesis Through Direct Regulation of the Dlx1/Dlx2 Bigene Cluster." *Journal of Neuroscience* 32 (24): 8219–30. https://doi.org/10.1523/JNEUROSCI.1344-12.2012.

Grealish, Shane, Elsa Diguet, Agnete Kirkeby, Bengt Mattsson, Andreas Heuer, Yann Bramoulle, Nadja Van Camp, et al. 2014. "Human ESC-Derived Dopamine Neurons Show Similar Preclinical Efficacy and Potency to Fetal Neurons When Grafted in a Rat Model of Parkinson's Disease." *Cell Stem Cell* 15 (5): 653–65. https://doi.org/10.1016/j.stem.2014.09.017.

Grüning, Björn, Ryan Dale, Andreas Sjödin, Brad A. Chapman, Jillian Rowe, Christopher H. Tomkins-Tinch, Renan Valieris, and Johannes Köster. 2018. "Bioconda: Sustainable and Comprehensive Software Distribution for the Life Sciences." *Nature Methods* 15 (7): 475–76. https://doi.org/10.1038/s41592-018-0046-7.

Gusmao, Eduardo G., Christoph Dieterich, Martin Zenke, and Ivan G. Costa. 2014. "Detection of Active Transcription Factor Binding Sites with the Combination of DNase Hypersensitivity and Histone Modifications." *Bioinformatics* 30 (22): 3143–51. https://doi.org/10.1093/bioinformatics/btu519.

Haas, Simon, Andreas Trumpp, and Michael D. Milsom. 2018. "Causes and Consequences of Hematopoietic Stem Cell Heterogeneity." *Cell Stem Cell* 22 (5): 627–38. https://doi.org/10.1016/j.stem.2018.04.003.

Hager, Gordon L., James G. McNally, and Tom Misteli. 2009. "Transcription Dynamics." *Molecular Cell* 35 (6): 741–53. https://doi.org/10.1016/j.molcel.2009.09.005.

Hamada, Michito, Yuki Tsunakawa, Hyojung Jeon, Manoj Kumar Yadav, and Satoru Takahashi. 2020. "Role of MafB in macrophages." *Experimental Animals* 69 (1): 1–10. https://doi.org/10.1538/expanim.19-0076.

Harrison, Stephen C. 1991. "A Structural Taxonomy of DNA-Binding Domains." *Nature* 353 (6346): 715–19. https://doi.org/10.1038/353715a0.

Hermanson, E. 2003. "Nurr1 Regulates Dopamine Synthesis and Storage in MN9D Dopamine Cells." *Experimental Cell Research* 288 (2): 324–34. https://doi.org/10.1016/S0014-4827(03)00216-7.

Hickl, Oskar, Pedro Queirós, Paul Wilmes, Patrick May, and Anna Heintz-Buschart. 2021. "Binny: An Automated Binning Algorithm to Recover High-Quality Genomes from Complex Metagenomic Datasets." https://doi.org/10.1101/2021.12.22.473795.

Hiller, Benjamin M., David J. Marmion, Cayla A. Thompson, Nathaniel A. Elliott, Howard Federoff, Patrik Brundin, Virginia B. Mattis, Christopher W. McMahon, and Jeffrey H. Kordower. 2022. "Optimizing Maturity and Dose of iPSC-Derived Dopamine Progenitor Cell Therapy for Parkinson's Disease." *Npj Regenerative Medicine* 7 (1): 24. https://doi.org/10.1038/s41536-022-00221-y.

Hor, Charlotte N., Jake Yeung, Maxime Jan, Yann Emmenegger, Jeffrey Hubbard, Ioannis Xenarios, Felix Naef, and Paul Franken. 2019. "Sleep–wake-Driven and Circadian Contributions to Daily Rhythms in Gene Expression and Chromatin Accessibility in the Murine Cortex." *Proceedings of the National Academy of Sciences* 116 (51): 25773–83. https://doi.org/10.1073/pnas.1910590116.

Hume, D. A., and S. R. Himes. 2003. "Transcription Factors That Regulate Macrophage Development and Function." In, edited by Siamon Gordon, 158:11–40. Berlin, Heidelberg: Springer Berlin Heidelberg. http://link.springer.com/10.1007/978-3-642-55742-2_2.

Hume, David A., Kim M. Summers, and Michael Rehli. 2016. "Transcriptional Regulation and Macrophage Differentiation." Edited by Siamon Gordon. *Microbiology Spectrum* 4 (3). https://doi.org/10.1128/microbiolspec.MCHD-0024-2015.

Huynh-Thu, Vân Anh, and Pierre Geurts. 2018. "dynGENIE3: Dynamical GENIE3 for the

Inference of Gene Networks from Time Series Expression Data." *Scientific Reports* 8 (1): 3384. https://doi.org/10.1038/s41598-018-21715-0.

Jackson, Michael, Kostas Kavoussanakis, and Edward W. J. Wallace. 2021. "Using Prototyping to Choose a Bioinformatics Workflow Management System." Edited by Francis Ouellette. *PLOS Computational Biology* 17 (2): e1008622. https://doi.org/10.1371/journal.pcbi.1008622.

Jego, G, D Lanneau, A De Thonel, K Berthenet, A Hazoumé, N Droin, A Hamman, et al. 2014. "Dual Regulation of SPI1/PU.1 Transcription Factor by Heat Shock Factor 1 (HSF1) During Macrophage Differentiation of Monocytes." *Leukemia* 28 (8): 1676–86. https://doi.org/10.1038/leu.2014.63.

Jin, Haijing, Ying-Wooi Wan, and Zhandong Liu. 2017. "Comprehensive Evaluation of RNA-Seq Quantification Methods for Linearity." *BMC Bioinformatics* 18 (S4): 117. https://doi.org/10.1186/s12859-017-1526-y.

Johnson, David S., Ali Mortazavi, Richard M. Myers, and Barbara Wold. 2007. "Genome-Wide Mapping of in Vivo Protein-DNA Interactions." *Science* 316 (5830): 1497–1502. https://doi.org/10.1126/science.1141319.

Julian, L M, Y Liu, C A Pakenham, D Dugal-Tessier, V Ruzhynsky, S Bae, S-Y Tsai, G Leone, R S Slack, and A Blais. 2016. "Tissue-Specific Targeting of Cell Fate Regulatory Genes by E2f Factors." *Cell Death & Differentiation* 23 (4): 565–75. https://doi.org/10.1038/cdd.2015.36.

Jung, Sascha, and Antonio Del Sol. 2020. "Multiomics Data Integration Unveils Core Transcriptional Regulatory Networks Governing Cell-Type Identity." *Npj Systems Biology and Applications* 6 (1): 26. https://doi.org/10.1038/s41540-020-00148-4.

Karlebach, Guy, and Ron Shamir. 2008. "Modelling and Analysis of Gene Regulatory Networks." *Nature Reviews Molecular Cell Biology* 9 (10): 770–80. https://doi.org/10.1038/nrm2503.

Kartha, Vinay K., Fabiana M. Duarte, Yan Hu, Sai Ma, Jennifer G. Chew, Caleb A. Lareau, Andrew Earl, et al. 2022. "Functional Inference of Gene Regulation Using Single-Cell Multi-Omics." *Cell Genomics* 2 (9): 100166. https://doi.org/10.1016/j.xgen.2022.1001

66.

Kiani, Karun, Eric M Sanford, Yogesh Goyal, and Arjun Raj. 2022. “Changes in Chromatin Accessibility Are Not Concordant with Transcriptional Changes for Single-Factor Perturbations.” *Molecular Systems Biology* 18 (9). https://doi.org/10.15252/msb.202210979.

Kim, Seongho. 2015. “Ppcor: An R Package for a Fast Calculation to Semi-Partial Correlation Coefficients.” *Communications for Statistical Applications and Methods* 22 (6): 665–74. https://doi.org/10.5351/CSAM.2015.22.6.665.

Klemm, Sandy L., Zohar Shipony, and William J. Greenleaf. 2019. “Chromatin Accessibility and the Regulatory Epigenome.” *Nature Reviews Genetics* 20 (4): 207–20. https://doi.org/10.1038/s41576-018-0089-8.

Knudsen, Kasper Jermiin, Matilda Rehn, Marie Sigurd Hasemann, Nicolas Rapin, Frederik Otzen Bagger, Ewa Ohlsson, Anton Willer, et al. 2015. “ERG Promotes the Maintenance of Hematopoietic Stem Cells by Restricting Their Differentiation.” *Genes & Development* 29 (18): 1915–29. https://doi.org/10.1101/gad.268409.115.

Koster, J., and S. Rahmann. 2012. “Snakemake–a Scalable Bioinformatics Workflow Engine.” *Bioinformatics* 28 (19): 2520–22. https://doi.org/10.1093/bioinformatics/bts480.

Labzin, Larisa I., Susanne V. Schmidt, Seth L. Masters, Marc Beyer, Wolfgang Krebs, Kathrin Klee, Rainer Stahl, et al. 2015. “ATF3 Is a Key Regulator of Macrophage IFN Responses.” *Journal of Immunology (Baltimore, Md.: 1950)* 195 (9): 4446–55. https://doi.org/10.4049/jimmunol.1500204.

Lachmann, Alexander, Federico M. Giorgi, Gonzalo Lopez, and Andrea Califano. 2016. “ARACNe-AP: Gene Network Reverse Engineering Through Adaptive Partitioning Inference of Mutual Information.” *Bioinformatics* 32 (14): 2233–35. https://doi.org/10.1093/bioinformatics/btw216.

Lambert, Samuel A., Arttu Jolma, Laura F. Campitelli, Pratyush K. Das, Yimeng Yin, Mihai Albu, Xiaoting Chen, Jussi Taipale, Timothy R. Hughes, and Matthew T. Weirauch. 2018. “The Human Transcription Factors.” *Cell* 172 (4): 650–65. https://doi.org/10.1

016/j.cell.2018.01.029.

Langmead, Ben, and Steven L Salzberg. 2012. “Fast Gapped-Read Alignment with Bowtie 2.” *Nature Methods* 9 (4): 357–59. https://doi.org/10.1038/nmeth.1923.

Lara-Astiaso, D., A. Weiner, E. Lorenzo-Vivas, I. Zaretsky, D. A. Jaitin, E. David, H. Keren-Shaul, et al. 2014. “Chromatin State Dynamics During Blood Formation.” *Science* 345 (6199): 943–49. https://doi.org/10.1126/science.1256271.

Larsonneur, Elise, Jonathan Mercier, Nicolas Wiart, Edith Le Floch, Olivier Delhomme, and Vincent Meyer. 2018. “2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM).” In, 2773–75. Madrid, Spain: IEEE. https://doi.org/10.1109/BIBM.2018.8621141.

Lawrence, Toby, and Gioacchino Natoli. 2011. “Transcriptional Regulation of Macrophage Polarization: Enabling Diversity with Identity.” *Nature Reviews Immunology* 11 (11): 750–61. https://doi.org/10.1038/nri3088.

Lee, K.-W., Y. Lee, H.-J. Kwon, and D.-S. Kim. 2005. “Sp1-associated activation of macrophage inflammatory protein-2 promoter by CpG-oligodeoxynucleotide and lipopolysaccharide.” *Cellular and molecular life sciences: CMLS* 62 (2): 188–98. https://doi.org/10.1007/s00018-004-4399-y.

Lee, Tong Ihn, and Richard A. Young. 2013. “Transcriptional Regulation and Its Misregulation in Disease.” *Cell* 152 (6): 1237–51. https://doi.org/10.1016/j.cell.2013.02.014.

Leipzig, Jeremy. 2016. “A Review of Bioinformatic Pipeline Frameworks.” *Briefings in Bioinformatics*, March, bbw020. https://doi.org/10.1093/bib/bbw020.

Leppä, Sirpa, Lila Pirkkala, Helena Saarento, Kevin D. Sarge, and Lea Sistonen. 1997. “Overexpression of HSF2-β Inhibits Hemin-Induced Heat Shock Gene Expression and Erythroid Differentiation in K562 Cells.” *Journal of Biological Chemistry* 272 (24): 15293–98. https://doi.org/10.1074/jbc.272.24.15293.

Li, Guanglan, Wenke Hao, and Wenxue Hu. 2020. “Transcription Factor PU.1 and Immune Cell Differentiation (Review).” *International Journal of Molecular Medicine* 46 (6): 1943–50. https://doi.org/10.3892/ijmm.2020.4763.

Li, Heng, and Richard Durbin. 2010. “Fast and Accurate Long-Read Alignment with

Burrows–Wheeler Transform." *Bioinformatics* 26 (5): 589–95. https://doi.org/10.1093/bioinformatics/btp698.

Li, Heng, Ting Jiang, Meng-Qi Li, Xi-Long Zheng, and Guo-Jun Zhao. 2018. "Transcriptional Regulation of Macrophages Polarization by MicroRNAs." *Frontiers in Immunology* 9 (May): 1175. https://doi.org/10.3389/fimmu.2018.01175.

Li, Ruifang, Sara A. Grimm, and Paul A. Wade. 2021. "A Simple and Robust Method for Simultaneous Dual-Omics Profiling with Limited Numbers of Cells." *Cell Reports Methods* 1 (3): 100041. https://doi.org/10.1016/j.crmeth.2021.100041.

Li, Zhijian, Chao-Chung Kuo, Fabio Ticconi, Mina Shaigan, Julia Gehrmann, Eduardo Gade Gusmao, Manuel Allhoff, Martin Manolov, Martin Zenke, and Ivan G. Costa. 2023. "RGT: A Toolbox for the Integrative Analysis of High Throughput Regulatory Genomics Data." *BMC Bioinformatics* 24 (1): 79. https://doi.org/10.1186/s12859-023-05184-5.

Li, Zhijian, Marcel H. Schulz, Thomas Look, Matthias Begemann, Martin Zenke, and Ivan G. Costa. 2019. "Identification of Transcription Factor Binding Sites Using ATAC-Seq." *Genome Biology* 20 (1). https://doi.org/10.1186/s13059-019-1642-2.

Liao, Yang, Gordon K. Smyth, and Wei Shi. 2014. "featureCounts: An Efficient General Purpose Program for Assigning Sequence Reads to Genomic Features." *Bioinformatics* 30 (7): 923–30. https://doi.org/10.1093/bioinformatics/btt656.

Liao, Yang, Gordon K Smyth, and Wei Shi. 2019. "The r Package Rsubread Is Easier, Faster, Cheaper and Better for Alignment and Quantification of RNA Sequencing Reads." *Nucleic Acids Research* 47 (8): e47–47. https://doi.org/10.1093/nar/gkz114.

Liu, David X, and Lloyd A Greene. 2001. "Regulation of Neuronal Survival and Death by E2F-Dependent Gene Repression and Derepression." *Neuron* 32 (3): 425–38. https://doi.org/10.1016/S0896-6273(01)00495-0.

Liu, Longqi, Lizhi Leng, Chuanyu Liu, Changfu Lu, Yue Yuan, Liang Wu, Fei Gong, et al. 2019. "An Integrated Chromatin Accessibility and Transcriptome Landscape of Human Pre-Implantation Embryos." *Nature Communications* 10 (1): 364. https://doi.org/10.1

038/s41467-018-08244-0.

Lord, Kenneth A, Abbas Abdollahi, Barbara Hoffman-Liebermann, and Dan A Liebermann. 1992. “Proto-Oncogenes of the Fos/Jun Family of Transcription Factors Are Positive Regulators of Myeloid Differentiation.” *Molecular and Cellular Biology*.

Love, Michael I, Wolfgang Huber, and Simon Anders. 2014. “Moderated Estimation of Fold Change and Dispersion for RNA-Seq Data with DESeq2.” *Genome Biology* 15 (12): 550. https://doi.org/10.1186/s13059-014-0550-8.

Ludwig, Leif S., Caleb A. Lareau, Erik L. Bao, Satish K. Nandakumar, Christoph Muus, Jacob C. Ulirsch, Kaitavjeet Chowdhary, et al. 2019. “Transcriptional States and Chromatin Accessibility Underlying Human Erythropoiesis.” *Cell Reports* 27 (11): 3228–3240.e7. https://doi.org/10.1016/j.celrep.2019.05.046.

Luo, Liheng, Michael Gribskov, and Sufang Wang. 2022. “Bibliometric Review of ATAC-Seq and Its Application in Gene Expression.” *Briefings in Bioinformatics* 23 (3): bbac061. https://doi.org/10.1093/bib/bbac061.

Luscombe, Nicholas M., Susan E. Austin, Helen M. Berman, and Janet M. Thornton. 2000. “An Overview of the Structures of Protein-DNA Complexes.” *Genome Biology* 1 (1): reviews001.1. https://doi.org/10.1186/gb-2000-1-1-reviews001.

MacNeil, Lesley T., and Albertha J. M. Walhout. 2011. “Gene Regulatory Networks and the Role of Robustness and Stochasticity in the Control of Gene Expression.” *Genome Research* 21 (5): 645–57. https://doi.org/10.1101/gr.097378.109.

Margolin, Adam A, Ilya Nemenman, Katia Basso, Chris Wiggins, Gustavo Stolovitzky, Riccardo Dalla Favera, and Andrea Califano. 2006. “ARACNE: An Algorithm for the Reconstruction of Gene Regulatory Networks in a Mammalian Cellular Context.” *BMC Bioinformatics* 7 (S1): S7. https://doi.org/10.1186/1471-2105-7-S1-S7.

Martin, G R. 1981. “Isolation of a Pluripotent Cell Line from Early Mouse Embryos Cultured in Medium Conditioned by Teratocarcinoma Stem Cells.” *Proceedings of the National Academy of Sciences* 78 (12): 7634–38. https://doi.org/10.1073/pnas.78.12.7634.

Maston, Glenn A., Sara K. Evans, and Michael R. Green. 2006. “Transcriptional Regulatory Elements in the Human Genome.” *Annual Review of Genomics and Human*

Genetics 7 (1): 29–59. https://doi.org/10.1146/annurev.genom.7.080505.115623.

Mincarelli, Laura, Ashleigh Lister, James Lipscombe, and Iain C. Macaulay. 2018. “Defining Cell Identity with Single-Cell Omics.” *PROTEOMICS* 18 (18): 1700312. https://doi.org/10.1002/pmic.201700312.

Mohr, Jeffrey C., Juan J. De Pablo, and Sean P. Palecek. 2006. “3-D Microwell Culture of Human Embryonic Stem Cells.” *Biomaterials* 27 (36): 6032–42. https://doi.org/10.1016/j.biomaterials.2006.07.012.

Morris, Samantha A. 2019. “The Evolving Concept of Cell Identity in the Single Cell Era.” Edited by Allon Klein and Barbara Treutlein. *Development* 146 (12): dev169748. https://doi.org/10.1242/dev.169748.

Morrison, Sean J., and Judith Kimble. 2006. “Asymmetric and Symmetric Stem-Cell Divisions in Development and Cancer.” *Nature* 441 (7097): 1068–74. https://doi.org/10.1038/nature04956.

Mortazavi, Ali, Brian A Williams, Kenneth McCue, Lorian Schaeffer, and Barbara Wold. 2008. “Mapping and Quantifying Mammalian Transcriptomes by RNA-Seq.” *Nature Methods* 5 (7): 621–28. https://doi.org/10.1038/nmeth.1226.

Nagamura-Inoue, Tokiko, Tomohiko Tamura, and Keiko Ozato. 2001. “Transcription Factors That Regulate Growth and Differentiation of Myeloid Cells.” *International Reviews of Immunology* 20 (1): 83–105. https://doi.org/10.3109/08830180109056724.

Narayanasamy, Shaman, Yohan Jarosz, Emilie E. L. Muller, Anna Heintz-Buschart, Malte Herold, Anne Kaysen, Cédric C. Laczny, Nicolás Pinel, Patrick May, and Paul Wilmes. 2016. “IMP: A Pipeline for Reproducible Reference-Independent Integrated Metagenomic and Metatranscriptomic Analyses.” *Genome Biology* 17 (1): 260. https://doi.org/10.1186/s13059-016-1116-8.

Nishida, Keishin, Martin C. Frith, and Kenta Nakai. 2009. “Pseudocounts for Transcription Factor Binding Sites.” *Nucleic Acids Research* 37 (3): 939–44. https://doi.org/10.1093/nar/gkn1019.

Notta, Faiyaz, Sasan Zandi, Naoya Takayama, Stephanie Dobson, Olga I. Gan, Gavin Wilson, Kerstin B. Kaufmann, et al. 2016. “Distinct routes of lineage development

reshape the human blood hierarchy across ontogeny." *Science (New York, N.Y.)* 351 (6269): aab2116. https://doi.org/10.1126/science.aab2116.

Novak, J. P., and C. C. Stewart. 1991. "Stochastic Versus Deterministic in Haemopoiesis: What Is What?" *British Journal of Haematology* 78 (2): 149–54. https://doi.org/10.1111/j.1365-2141.1991.tb04409.x.

O'Brien, Jacob, Heyam Hayder, Yara Zayed, and Chun Peng. 2018. "Overview of MicroRNA Biogenesis, Mechanisms of Actions, and Circulation." *Frontiers in Endocrinology* 9 (August): 402. https://doi.org/10.3389/fendo.2018.00402.

Orecchioni, Marco, Yanal Ghosheh, Akula Bala Pramod, and Klaus Ley. 2019. "Macrophage Polarization: Different Gene Signatures in M1(LPS+) Vs. Classically and M2(LPS–) Vs. Alternatively Activated Macrophages." *Frontiers in Immunology* 10 (May): 1084. https://doi.org/10.3389/fimmu.2019.01084.

Pang, Zhiping P., Nan Yang, Thomas Vierbuchen, Austin Ostermeier, Daniel R. Fuentes, Troy Q. Yang, Ami Citri, et al. 2011. "Induction of Human Neuronal Cells by Defined Transcription Factors." *Nature* 476 (7359): 220–23. https://doi.org/10.1038/nature10202.

Pape, Utz J., Sven Rahmann, and Martin Vingron. 2008. "Natural Similarity Measures Between Position Frequency Matrices with an Application to Clustering." *Bioinformatics* 24 (3): 350–57. https://doi.org/10.1093/bioinformatics/btm610.

Park, Peter J. 2009. "ChIP–seq: Advantages and Challenges of a Maturing Technology." *Nature Reviews Genetics* 10 (10): 669–80. https://doi.org/10.1038/nrg2641.

Perrin, Hannah J., Kevin W. Currin, Swarooparani Vadlamudi, Gautam K. Pandey, Kenneth K. Ng, Martin Wabitsch, Markku Laakso, Michael I. Love, and Karen L. Mohlke. 2021. "Chromatin Accessibility and Gene Expression During Adipocyte Differentiation Identify Context-Dependent Effects at Cardiometabolic GWAS Loci." Edited by Chris Cotsapas. *PLOS Genetics* 17 (10): e1009865. https://doi.org/10.1371/journal.pgen.1009865.

Pfisterer, Ulrich, Agnete Kirkeby, Olof Torper, James Wood, Jenny Nelander, Audrey Dufour, Anders Björklund, Olle Lindvall, Johan Jakobsson, and Malin Parmar. 2011. "Di-

rect Conversion of Human Fibroblasts to Dopaminergic Neurons." *Proceedings of the National Academy of Sciences* 108 (25): 10343–48. https://doi.org/10.1073/pnas.1105135108.

Pinho, Sandra, and Paul S. Frenette. 2019. "Haematopoietic Stem Cell Activity and Interactions with the Niche." *Nature Reviews Molecular Cell Biology* 20 (5): 303–20. https://doi.org/10.1038/s41580-019-0103-9.

Portales-Casamar, Elodie, Supat Thongjuea, Andrew T. Kwon, David Arenillas, Xiaobei Zhao, Eivind Valen, Dimas Yusuf, Boris Lenhard, Wyeth W. Wasserman, and Albin Sandelin. 2010. "JASPAR 2010: The Greatly Expanded Open-Access Database of Transcription Factor Binding Profiles." *Nucleic Acids Research* 38 (suppl_1): D105–10. https://doi.org/10.1093/nar/gkp950.

Pundhir, Sachin, Felicia Kathrine Bratt Lauridsen, Mikkel Bruhn Schuster, Janus Schou Jakobsen, Ying Ge, Erwin Marten Schoof, Nicolas Rapin, Johannes Waage, Marie Sigurd Hasemann, and Bo Torben Porse. 2018. "Enhancer and Transcription Factor Dynamics During Myeloid Differentiation Reveal an Early Differentiation Block in Cebpa Null Progenitors." *Cell Reports* 23 (9): 2744–57. https://doi.org/10.1016/j.celrep.2018.05.012.

Quinlan, Aaron R., and Ira M. Hall. 2010. "BEDTools: A Flexible Suite of Utilities for Comparing Genomic Features." *Bioinformatics* 26 (6): 841–42. https://doi.org/10.1093/bioinformatics/btq033.

Ramirez, Ricardo N., Nicole C. El-Ali, Mikayla Anne Mager, Dana Wyman, Ana Conesa, and Ali Mortazavi. 2017. "Dynamic Gene Regulatory Networks of Human Myeloid Differentiation." *Cell Systems* 4 (4): 416–429.e3. https://doi.org/10.1016/j.cels.2017.03.005.

Ramos, Borja Gomez, Jochen Ohnmacht, Nikola de Lange, Aurélien Ginolhac, Elena Valceschini, Aleksandar Rakovic, Rashi Halder, et al. 2023. "Multi-Omics Analysis Identifies LBX1 and NHLH1 as Central Regulators of Human Midbrain Dopaminergic Neuron Differentiation." https://doi.org/10.1101/2023.01.27.525898.

Ranzoni, Anna Maria, Andrea Tangherloni, Ivan Berest, Simone Giovanni Riva, Brynelle

Myers, Paulina M. Strzelecka, Jiarui Xu, et al. 2021. “Integrative Single-Cell RNA-Seq and ATAC-Seq Analysis of Human Developmental Hematopoiesis.” *Cell Stem Cell* 28 (3): 472–487.e7. https://doi.org/10.1016/j.stem.2020.11.015.

Rauschmeier, René, Charlotte Gustafsson, Annika Reinhardt, Noelia A-Gonzalez, Luigi Tortola, Dilay Cansever, Sethuraman Subramanian, et al. 2019. “Bhlhe40 and Bhlhe41 Transcription Factors Regulate Alveolar Macrophage Self-Renewal and Identity.” *The EMBO Journal* 38 (19). https://doi.org/10.15252/embj.2018101233.

Richardson, Edward T., Supriya Shukla, Nancy Nagy, W. Henry Boom, Rose C. Beck, Lan Zhou, Gary E. Landreth, and Clifford V. Harding. 2015. “ERK Signaling Is Essential for Macrophage Development.” Edited by Kevin D Bunting. *PLOS ONE* 10 (10): e0140064. https://doi.org/10.1371/journal.pone.0140064.

Rosenbauer, Frank, and Daniel G. Tenen. 2007. “Transcription Factors in Myeloid Development: Balancing Differentiation with Transformation.” *Nature Reviews Immunology* 7 (2): 105–17. https://doi.org/10.1038/nri2024.

Sánchez Alvarado, Alejandro, and Shinya Yamanaka. 2014. “Rethinking Differentiation: Stem Cells, Regeneration, and Plasticity.” *Cell* 157 (1): 110–19. https://doi.org/10.1016/j.cell.2014.02.041.

Sandelin, A. 2004. “JASPAR: An Open-Access Database for Eukaryotic Transcription Factor Binding Profiles.” *Nucleic Acids Research* 32 (90001): 91D–94. https://doi.org/10.1093/nar/gkh012.

Santoni de Sio, F. R., L. Passerini, M. M. Valente, F. Russo, L. Naldini, M. G. Roncarolo, and R. Bacchetta. 2017. “Ectopic FOXP3 Expression Preserves Primitive Features Of Human Hematopoietic Stem Cells While Impairing Functional T Cell Differentiation.” *Scientific Reports* 7 (1): 15820. https://doi.org/10.1038/s41598-017-15689-8.

Schliep, A., I. G. Costa, C. Steinhoff, and A. Schonhuth. 2005. “Analyzing Gene Expression Time-Courses.” *IEEE/ACM Transactions on Computational Biology and Bioinformatics* 2 (3): 179–93. https://doi.org/10.1109/TCBB.2005.31.

Schmidt, Florian, Nina Gasparoni, Gilles Gasparoni, Kathrin Gianmoena, Cristina Cadenas, Julia K. Polansky, Peter Ebert, et al. 2017. “Combining Transcription Factor

Binding Affinities with Open-Chromatin Data for Accurate Gene Expression Prediction." *Nucleic Acids Research* 45 (1): 54–66. https://doi.org/10.1093/nar/gkw1061.

Schmidt, Florian, Fabian Kern, Peter Ebert, Nina Baumgarten, and Marcel H Schulz. 2019. "TEPIC 2—an Extended Framework for Transcription Factor Binding Prediction and Integrative Epigenomic Analysis." Edited by Bonnie Berger. *Bioinformatics* 35 (9): 1608–9. https://doi.org/10.1093/bioinformatics/bty856.

Schuettengruber, Bernd, Henri-Marc Bourbon, Luciano Di Croce, and Giacomo Cavalli. 2017. "Genome Regulation by Polycomb and Trithorax: 70 Years and Counting." *Cell* 171 (1): 34–57. https://doi.org/10.1016/j.cell.2017.08.002.

Schulz, Marcel H, William E Devanny, Anthony Gitter, Shan Zhong, Jason Ernst, and Ziv Bar-Joseph. 2012. "DREM 2.0: Improved Reconstruction of Dynamic Regulatory Networks from Time-Series Expression Data." *BMC Systems Biology* 6 (1): 104. https://doi.org/10.1186/1752-0509-6-104.

Schweitzer, Jeffrey S., Bin Song, Todd M. Herrington, Tae-Yoon Park, Nayeon Lee, Sanghyeok Ko, Jeha Jeon, et al. 2020. "Personalized iPSC-Derived Dopamine Progenitor Cells for Parkinson's Disease." *New England Journal of Medicine* 382 (20): 1926–32. https://doi.org/10.1056/NEJMoa1915872.

Sherf, Orna, Limor Nashelsky Zolotov, Keren Liser, Hadas Tilleman, Vukasin M. Jovanovic, Ksenija Zega, Marin M. Jukic, and Claude Brodski. 2015. "Otx2 Requires Lmx1b to Control the Development of Mesodiencephalic Dopaminergic Neurons." Edited by Renping Zhou. *PLOS ONE* 10 (10): e0139697. https://doi.org/10.1371/journal.pone.0139697.

Silvennoinen, Katri, Nikola de Lange, Sara Zagaglia, Simona Balestrini, Ganna Androsova, Merel Wassenaar, Pauls Auce, et al. 2019. "Comparative Effectiveness of Antiepileptic Drugs in Juvenile Myoclonic Epilepsy." *Epilepsia Open* 4 (3): 420–30. https://doi.org/10.1002/epi4.12349.

Song, Lingyun, and Gregory E. Crawford. 2010. "DNase-Seq: A High-Resolution Technique for Mapping Active Gene Regulatory Elements Across the Genome from Mammalian Cells." *Cold Spring Harbor Protocols* 2010 (2): pdb.prot5384. https://doi.org/

10.1101/pdb.prot5384.

Song, Qiao, Yuli Hou, Yiyin Zhang, Jing Liu, Yaqi Wang, Jingxuan Fu, Chi Zhang, et al. 2022. “Integrated Multi-Omics Approach Revealed Cellular Senescence Landscape.” *Nucleic Acids Research* 50 (19): 10947–63. https://doi.org/10.1093/nar/gkac885.

Specht, Alicia T, and Jun Li. 2017. “LEAP: Constructing Gene Co-Expression Networks for Single-Cell RNA-Sequencing Data Using Pseudotime Ordering.” Edited by Ziv Bar-Joseph. *Bioinformatics* 33 (5): 764–66. https://doi.org/10.1093/bioinformatics/btw729.

Spitz, François, and Eileen E. M. Furlong. 2012. “Transcription Factors: From Enhancer Binding to Developmental Control.” *Nature Reviews Genetics* 13 (9): 613–26. https://doi.org/10.1038/nrg3207.

T’Jonck, Wouter, Martin Guilliams, and Johnny Bonnardel. 2018. “Niche Signals and Transcription Factors Involved in Tissue-Resident Macrophage Development.” *Cellular Immunology* 330 (August): 43–53. https://doi.org/10.1016/j.cellimm.2018.02.005.

Tabata, Tetsuya. 2001. “Genetics of Morphogen Gradients.” *Nature Reviews Genetics* 2 (8): 620–30. https://doi.org/10.1038/35084577.

Takahashi, Kazutoshi, and Shinya Yamanaka. 2006. “Induction of Pluripotent Stem Cells from Mouse Embryonic and Adult Fibroblast Cultures by Defined Factors.” *Cell* 126 (4): 663–76. https://doi.org/10.1016/j.cell.2006.07.024.

———. 2013. “Induced Pluripotent Stem Cells in Medicine and Biology.” *Development* 140 (12): 2457–61. https://doi.org/10.1242/dev.092551.

The FANTOM Consortium, Owen J L Rackham, Jaber Firas, Hai Fang, Matt E Oates, Melissa L Holmes, Anja S Knaupp, et al. 2016. “A Predictive Computational Framework for Direct Reprogramming Between Human Cell Types.” *Nature Genetics* 48 (3): 331–35. https://doi.org/10.1038/ng.3487.

Tsompana, Maria, and Michael J Buck. 2014. “Chromatin Accessibility: A Window into the Genome.” *Epigenetics & Chromatin* 7 (1): 33. https://doi.org/10.1186/1756-8935-7-33.

van der Raadt, Jori, Sebastianus H C van Gestel, Nael Nadif Kasri, and Cornelis A Albers.

2019. “ONECUT Transcription Factors Induce Neuronal Characteristics and Remodel Chromatin Accessibility.” *Nucleic Acids Research* 47 (11): 5587–5602. https://doi.org/10.1093/nar/gkz273.

Vasimuddin, Md., Sanchit Misra, Heng Li, and Srinivas Aluru. 2019. “2019 IEEE International Parallel and Distributed Processing Symposium (IPDPS).” In, 314–24. Rio de Janeiro, Brazil: IEEE. https://doi.org/10.1109/IPDPS.2019.00041.

Veenvliet, Jesse V., Maria T. M. Alves dos Santos, Willemieke M. Kouwenhoven, Lars von Oerthel, Jamie L. Lim, Annemarie J. A. van der Linden, Marian J. A. Groot Koerkamp, Frank C. P. Holstege, and Marten P. Smidt. 2013. “Specification of Dopaminergic Subsets Involves Interplay of En1 and Pitx3.” *Development* 140 (16): 3373–84. https://doi.org/10.1242/dev.094565.

Veremeyko, Tatyana, Amanda W. Y. Yung, Daniel C. Anthony, Tatyana Strekalova, and Eugene D. Ponomarev. 2018. “Early Growth Response Gene-2 Is Essential for M1 and M2 Macrophage Activation and Plasticity by Modulation of the Transcription Factor CEBPβ.” *Frontiers in Immunology* 9 (November): 2515. https://doi.org/10.3389/fimmu.2018.02515.

Vierbuchen, Thomas, Austin Ostermeier, Zhiping P. Pang, Yuko Kokubu, Thomas C. Südhof, and Marius Wernig. 2010. “Direct Conversion of Fibroblasts to Functional Neurons by Defined Factors.” *Nature* 463 (7284): 1035–41. https://doi.org/10.1038/nature08797.

Villaescusa, J Carlos, Bingsi Li, Enrique M Toledo, Pia Rivetti di Val Cervo, Shanzheng Yang, Simon RW Stott, Karol Kaiser, et al. 2016. “A PBX1 Transcriptional Network Controls Dopaminergic Neuron Development and Is Impaired in Parkinson’s Disease.” *The EMBO Journal* 35 (18): 1963–78. https://doi.org/10.15252/embj.201593725.

Wagner, Günter P., Koryu Kin, and Vincent J. Lynch. 2012. “Measurement of mRNA Abundance Using RNA-Seq Data: RPKM Measure Is Inconsistent Among Samples.” *Theory in Biosciences* 131 (4): 281–85. https://doi.org/10.1007/s12064-012-0162-3.

Wang, Mengmeng, King-Hwa Ling, Jun Tan, and Cheng-Biao Lu. 2020. “Development and Differentiation of Midbrain Dopaminergic Neuron: From Bench to Bedside.” *Cells*

9 (6): 1489. https://doi.org/10.3390/cells9061489.

Wang, Zhong, Mark Gerstein, and Michael Snyder. 2009. “RNA-Seq: A Revolutionary Tool for Transcriptomics.” *Nature Reviews Genetics* 10 (1): 57–63. https://doi.org/10.1038/nrg2484.

Weber, Andreas P. M. 2015. “Discovering New Biology Through RNA-Seq.” *Plant Physiology*, September, pp.01081.2015. https://doi.org/10.1104/pp.15.01081.

Weiskopf, Kipp, Peter J. Schnorr, Wendy W. Pang, Mark P. Chao, Akanksha Chhabra, Jun Seita, Mingye Feng, and Irving L. Weissman. 2016. “Myeloid Cell Origins, Differentiation, and Clinical Implications.” *Microbiology Spectrum* 4 (5). https://doi.org/10.1128/microbiolspec.MCHD-0031-2016.

Wingender, Edgar, Torsten Schoeps, and Jürgen Dönitz. 2013. “TFClass: An Expandable Hierarchical Classification of Human Transcription Factors.” *Nucleic Acids Research* 41 (D1): D165–70. https://doi.org/10.1093/nar/gks1123.

Wingender, Edgar, Torsten Schoeps, Martin Haubrock, Mathias Krull, and Jürgen Dönitz. 2018. “TFClass: Expanding the Classification of Human Transcription Factors to Their Mammalian Orthologs.” *Nucleic Acids Research* 46 (D1): D343–47. https://doi.org/10.1093/nar/gkx987.

Wu, Tianzhi, Erqiang Hu, Shuangbin Xu, Meijun Chen, Pingfan Guo, Zehan Dai, Tingze Feng, et al. 2021. “clusterProfiler 4.0: A Universal Enrichment Tool for Interpreting Omics Data.” *The Innovation* 2 (3): 100141. https://doi.org/10.1016/j.xinn.2021.100141.

Xiang, Yangfei, Yoshiaki Tanaka, Benjamin Patterson, Young-Jin Kang, Gubbi Govindaiah, Naomi Roselaar, Bilal Cakir, et al. 2017. “Fusion of Regionally Specified hPSC-Derived Organoids Models Human Brain Development and Interneuron Migration.” *Cell Stem Cell* 21 (3): 383–398.e7. https://doi.org/10.1016/j.stem.2017.07.007.

Xu, Yang, Edmon Begoli, and Rachel Patton McCord. 2022. “sciCAN: Single-Cell Chromatin Accessibility and Gene Expression Data Integration via Cycle-Consistent Adversarial Network.” *Npj Systems Biology and Applications* 8 (1): 33. https://doi.org/10.1038/s41540-022-00245-6.

Yan, Kai, Tian-Tian Da, Zhen-Hua Bian, Yi He, Meng-Chu Liu, Qing-Zhi Liu, Jie Long, et al. 2020. "Multi-Omics Analysis Identifies FoxO1 as a Regulator of Macrophage Function Through Metabolic Reprogramming." *Cell Death & Disease* 11 (9): 800. https://doi.org/10.1038/s41419-020-02982-0.

Yang, B-H, S Hagemann, P Mamareli, U Lauer, U Hoffmann, M Beckstette, L Föhse, et al. 2016. "Foxp3+ T Cells Expressing RORγt Represent a Stable Regulatory T-Cell Effector Lineage with Enhanced Suppressive Capacity During Intestinal Inflammation." *Mucosal Immunology* 9 (2): 444–57. https://doi.org/10.1038/mi.2015.74.

Yang, Bi-Huei, Ke Wang, Shuo Wan, Yan Liang, Xiaomei Yuan, Yi Dong, Sunglim Cho, et al. 2019. "TCF1 and LEF1 Control Treg Competitive Survival and Tfr Development to Prevent Autoimmune Diseases." *Cell Reports* 27 (12): 3629–3645.e6. https://doi.org/10.1016/j.celrep.2019.05.061.

Yesudhas, Dhanusha, Maria Batool, Muhammad Anwar, Suresh Panneerselvam, and Sangdun Choi. 2017. "Proteins Recognizing DNA: Structural Uniqueness and Versatility of DNA-Binding Domains in Stem Cell Transcription Factors." *Genes* 8 (8): 192. https://doi.org/10.3390/genes8080192.

Yoo, Andy B., Morris A. Jette, and Mark Grondona. 2003. "SLURM: Simple Linux Utility for Resource Management." In, edited by Dror Feitelson, Larry Rudolph, and Uwe Schwiegelshohn, 2862:44–60. Berlin, Heidelberg: Springer Berlin Heidelberg. http://link.springer.com/10.1007/10968987_3.

Yu, Guangchuang, Li-Gen Wang, Yanyan Han, and Qing-Yu He. 2012. "clusterProfiler: An R Package for Comparing Biological Themes Among Gene Clusters." *OMICS: A Journal of Integrative Biology* 16 (5): 284–87. https://doi.org/10.1089/omi.2011.0118.

Zakrzewski, Wojciech, Maciej Dobrzyński, Maria Szymonowicz, and Zbigniew Rybak. 2019. "Stem Cells: Past, Present, and Future." *Stem Cell Research & Therapy* 10 (1): 68. https://doi.org/10.1186/s13287-019-1165-5.

Zhang, Hong, Meritxell Alberich-Jorda, Giovanni Amabile, Henry Yang, Philipp B. Staber, Annalisa Di Ruscio, Robert S. Welner, et al. 2013. "Sox4 Is a Key Oncogenic Target

in C/EBPα Mutant Acute Myeloid Leukemia." *Cancer Cell* 24 (5): 575–88. https://doi.org/10.1016/j.ccr.2013.09.018.

Zhang, Kai, Mengchi Wang, Ying Zhao, and Wei Wang. 2019. "Taiji: System-Level Identification of Key Transcription Factors Reveals Transcriptional Waves in Mouse Embryonic Development." *Science Advances* 5 (3): eaav3262. https://doi.org/10.1126/sciadv.aav3262.

Zhang, Yong, Tao Liu, Clifford A Meyer, Jérôme Eeckhoute, David S Johnson, Bradley E Bernstein, Chad Nusbaum, et al. 2008. "Model-Based Analysis of ChIP-Seq (MACS)." *Genome Biology* 9 (9): R137. https://doi.org/10.1186/gb-2008-9-9-r137.

Zhang, Ziqi, Chengkai Yang, and Xiuwei Zhang. 2022. "scDART: Integrating Unmatched scRNA-Seq and scATAC-Seq Data and Learning Cross-Modality Relationship Simultaneously." *Genome Biology* 23 (1): 139. https://doi.org/10.1186/s13059-022-02706-x.

Zhao, Shanrong, Wai-Ping Fung-Leung, Anton Bittner, Karen Ngo, and Xuejun Liu. 2014. "Comparison of RNA-Seq and Microarray in Transcriptome Profiling of Activated T Cells." Edited by Shu-Dong Zhang. *PLoS ONE* 9 (1): e78644. https://doi.org/10.1371/journal.pone.0078644.

Zhao, Shanrong, Zhan Ye, and Robert Stanton. 2020. "Misuse of RPKM or TPM normalization when comparing across samples and sequencing protocols." *RNA (New York, N.Y.)* 26 (8): 903–9. https://doi.org/10.1261/rna.074922.120.

Zhao, Shanrong, and Baohong Zhang. 2015. "A Comprehensive Evaluation of Ensembl, RefSeq, and UCSC Annotations in the Context of RNA-Seq Read Mapping and Gene Quantification." *BMC Genomics* 16 (1): 97. https://doi.org/10.1186/s12864-015-1308-8.

Supplemental Material manuscript

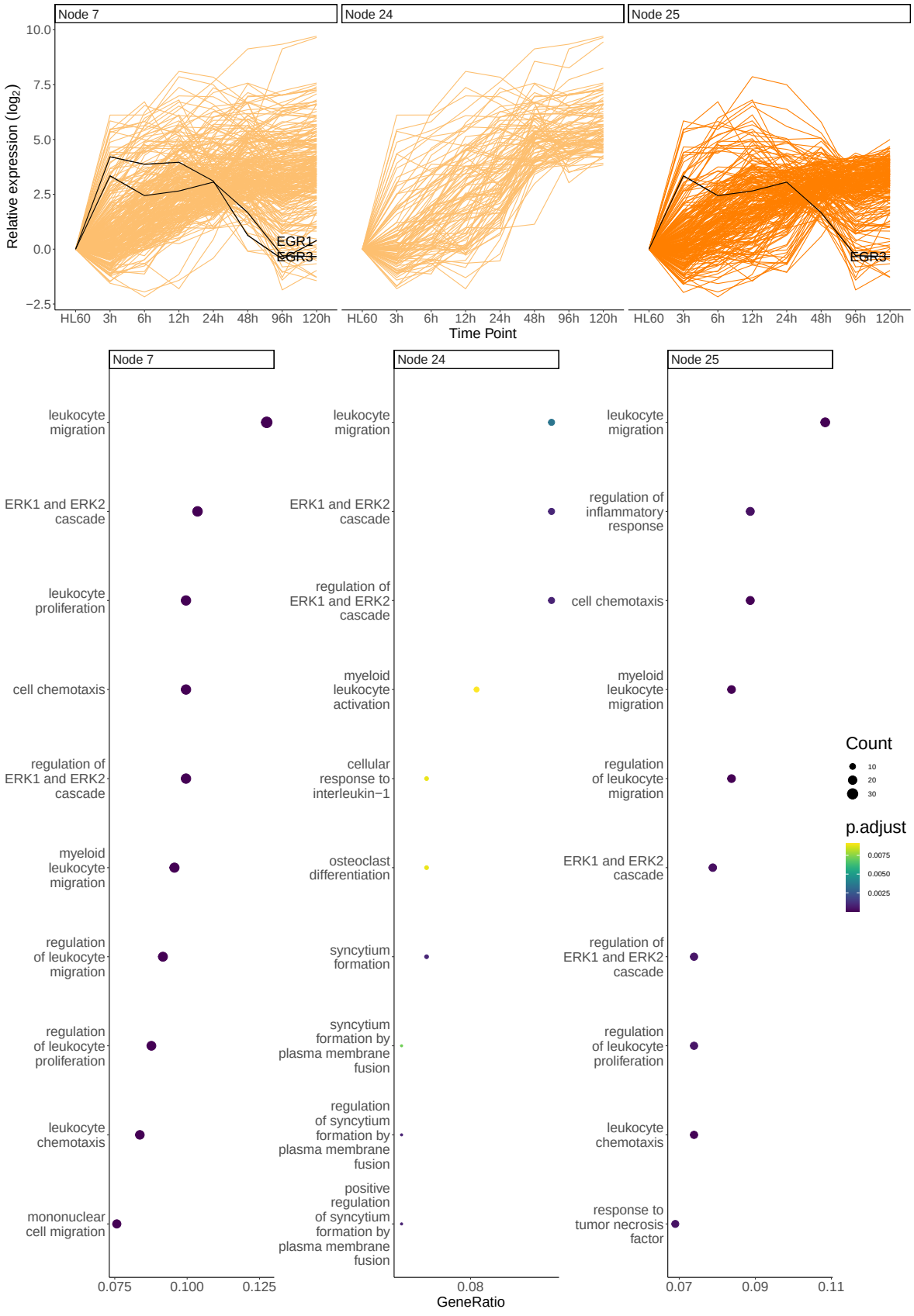


Figure B.6.: Gene Ontology enrichment of biological processes for the selected nodes showing enrichment for ERK cascade related terms.

Table B.4.: Best ranking for selected TF at each time point they are ranked 10 or higher.

TF	n	3h	6h	12h	24h	48h	96h
NRF1	3	44	2	3	59	3	14
MSX2	3	175	8	138	10	18	59
SP8	4	28	9	36	114	24	2
EGR1	3	124	16	5	7	12	5
KLF16	3	208	27	6	66	35	1
SP4	3	30	23	7	41	40	2
TFAP2B(var.2)	3	40	80	2	117	1	5
CENPB	3	169	83	76	6	65	1

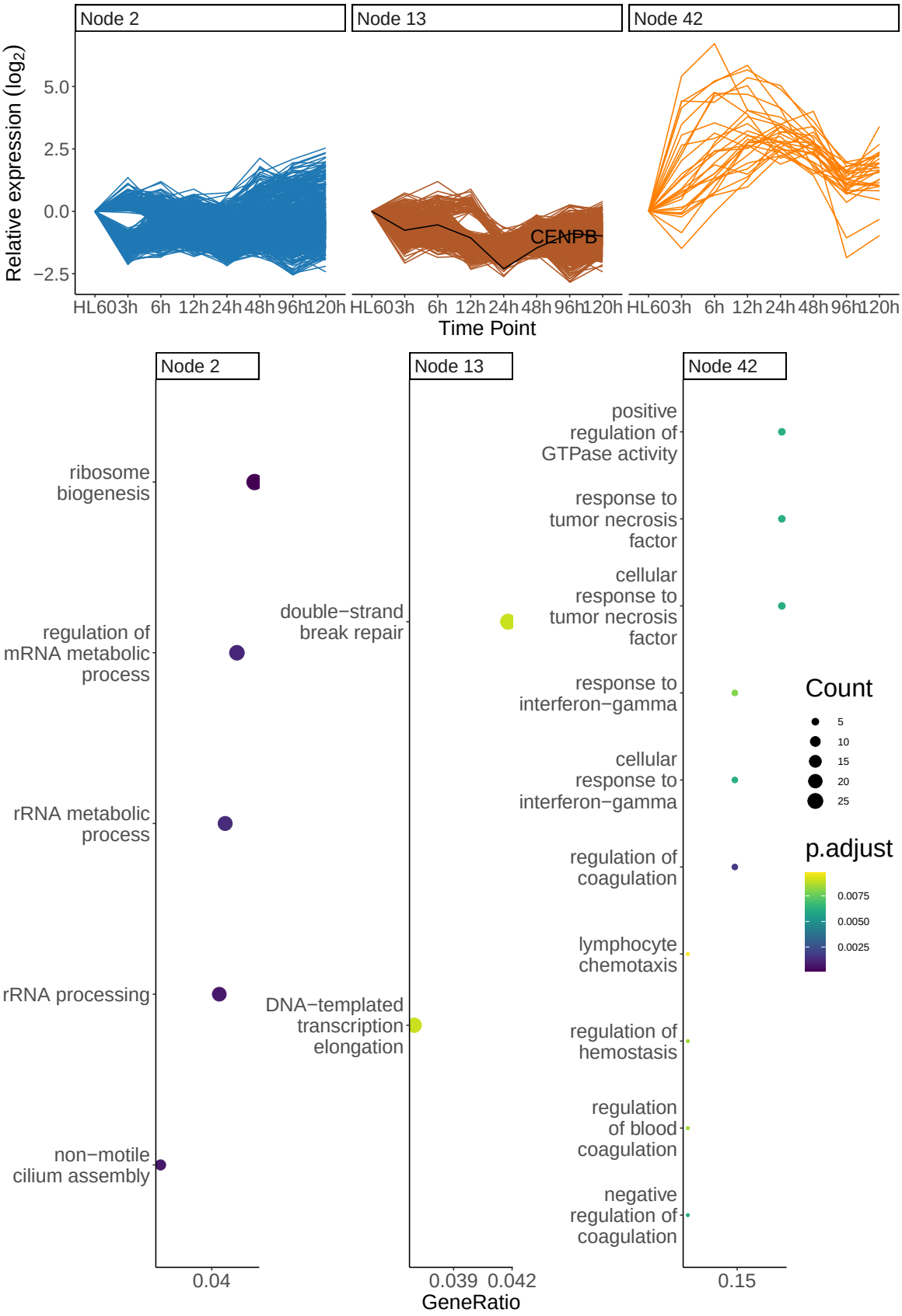


Figure B.7.: Gene Ontology enrichment of biological processes for the selected nodes showing general terms.

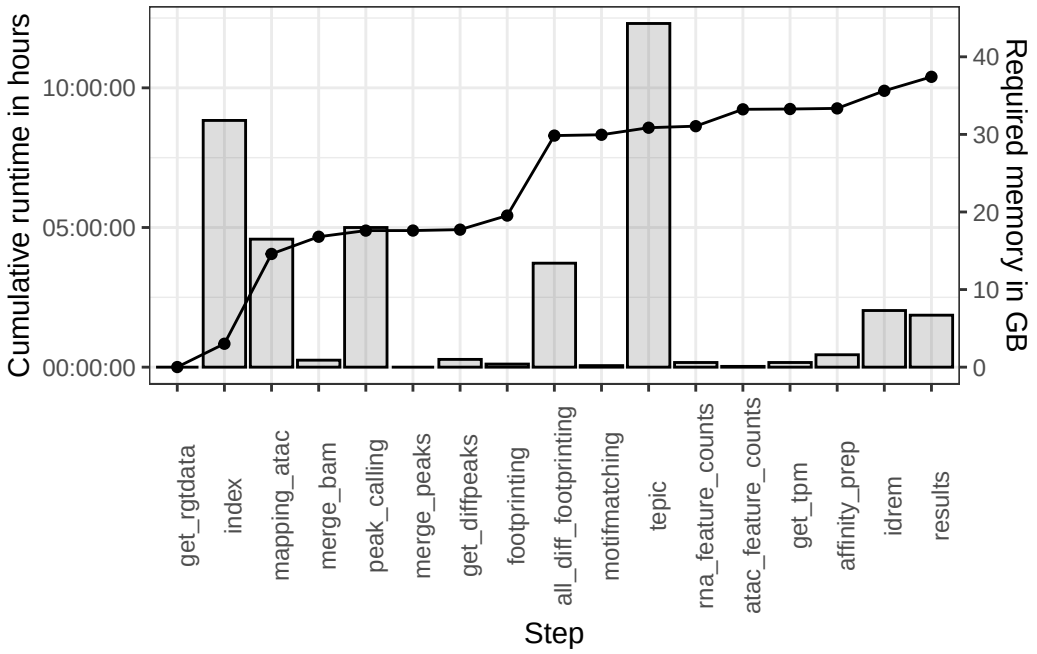


Figure B.8.: Runtime in hours and the memory requirement for each step

B.2. Manuscript - Multi-omics analysis identifies LBX1 and NHLH1 as central regulators of human midbrain dopaminergic neuron differentiation

For the following manuscript I established the EPIC-DREM pipeline for ATAC-seq based on the work of Gérard et al. (2018) and performed the analysis with EPIC-DREM as described in Section 3.1. It is currently available on bioRxiv (Ramos et al. (2023)) and the revised manuscript for the submission to EMBO Reports will include the results from Section 4.1.4.

bioRxiv preprint doi: https://doi.org/10.1101/2023.01.27.525898; this version posted January 27, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Multi-omics analysis identifies LBX1 and NHLH1 as central regulators of human midbrain dopaminergic neuron differentiation

Borja Gomez Ramos[1,2], Jochen Ohnmacht[1,2], Nikola de Lange[2], Aurélien Ginolhac[1], Elena Valceschini[1], Aleksandar Rakovic[3], Rashi Halder[2], François Massart[2], Christine Klein[3], Roland Krause[2], Marcel H. Schulz[4-6], Thomas Sauter[1], Rejko Krüger[2,7,8] and Lasse Sinkkonen[1,9]

[1] Department of Life Sciences and Medicine (DLSM), University of Luxembourg, L-4362 Belvaux, Luxembourg

[2] Luxembourg Centre for Systems Biomedicine (LCSB), University of Luxembourg, L-4362 Belvaux, Luxembourg

[3] Institute of Neurogenetics, University of Lübeck, D-23538, Lübeck, Germany

[4] Institute for Cardiovascular Regeneration, Goethe University, D-60590, Frankfurt, Germany

[5] German Centre for Cardiovascular Research, Partner site Rhein-Main, 60590 Frankfurt am Main, Germany.

[6] Cardio-Pulmonary Institute, Goethe University, Frankfurt am Main, Germany

[7] Centre Hospitalier de Luxembourg (CHL), L-1210, Luxembourg, Luxembourg

[8] Luxembourg Institute of Health (LIH), L-1445, Luxembourg, Luxembourg

[9] Corresponding author: Dr. Lasse Sinkkonen (lasse.sinkkonen@uni.lu)

bioRxiv preprint doi: https://doi.org/10.1101/2023.01.27.525898; this version posted January 27, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Abstract

Midbrain dopaminergic neurons (mDANs) control voluntary movement, cognition, and reward behavior under physiological conditions and are implicated in human diseases such as Parkinson's disease (PD). Many transcription factors (TFs) controlling human mDAN differentiation during development have been described, but much of the regulatory landscape remains undefined. Using a tyrosine hydroxylase (TH) iPSC reporter line, we have generated time series transcriptomic and epigenomic profiles of purified mDANs during differentiation. Integrative analysis predicted novel central regulators of mDAN differentiation and super-enhancers were used to prioritize key TFs. We find LBX1, NHLH1 and NR2F1/2 to be necessary for mDAN differentiation and show that overexpression of either LBX1 or NHLH1 can also improve mDAN specification. NHLH1 is necessary for the induction of neuronal miR-124, while LBX1 regulates cholesterol biosynthesis, possibly through mTOR signaling. Consistently, rapamycin treatment led to an inhibition of mDAN differentiation. Thus, our work reveals novel regulators of human mDAN differentiation.

Keywords

Dopaminergic neurons; induced pluripotent stem cells; multi-omics data integration; chromatin; enhancers; transcription factors

bioRxiv preprint doi: https://doi.org/10.1101/2023.01.27.525898; this version posted January 27, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Introduction

Induced pluripotent stem cell (iPSC) technology presents a unique system to study how transcription factors (TFs) control cell differentiation and specification in humans. TFs bind to genomic regulatory regions such as enhancers and promoters to mediate their action. The TFs occupying an enhancer region collectively control transcriptional initiation of their target genes through different mechanisms (1, 2). In particular super-enhancers (SEs), dense clusters of enhancers under high regulatory load (3), have been associated with cell identity genes, master TFs, and are enriched with disease-associated genetic variants (4, 5). Identifying TF binding profiles and active enhancer regions across the genome requires laborious and indirect methods such as chromatin immunoprecipitation sequencing (ChIP-seq). The presence of regulatory proteins alone does not necessarily indicate active gene regulation. To overcome this limitation, these methods can be combined with gene expression analysis in an integrative multi-omics approach to determine the role of specific TFs or enhancers in gene regulation. Moreover, a time course analysis combining transcriptomic and epigenomic data has shown great potential for studying processes like cell differentiation and revealing the regulatory events' hierarchy. Such approaches have been able to generate gene regulatory networks (GRN) that also take into account the regulatory landscape of the cells, facilitating the identification of key TFs (6–8).

Midbrain dopaminergic neurons (mDANs) are widely used in biomedical research due to their involvement in different psychiatric and neurological disorders such as schizophrenia, drug addiction and PD (9–11). The current protocols for generating mDANs from iPSCs produce heterogeneous populations and incompletely specified cells (12, 13). The genetic program underlying mDAN development has been extensively investigated (reviewed in 14), with most of the studies relying on transcriptomic data and mice as the model organism. With the emergence of single-cell technologies, improved insights into human and mouse midbrain development have revealed differences in the temporal dynamics, cell composition, and expression of TFs (13, 15).

All these findings highlight our limited understanding of human development, hampering our ability to apply the developmental knowledge to improve iPSC differentiation protocols. Only a few studies have considered the epigenetic landscape of mDANs during development (16–18). Furthermore, cellular heterogeneity present in the iPSC-derived cultures obscure physiological or biological insights from the cell type of interest (19, 20).

In this study, we profiled differentiating human iPSC-derived mDANs at the transcriptomic and epigenomic levels. Integrative analysis using our EPIC-DREM pipeline (8, 21–23) generated time-point-specific gene regulatory interactions. Together with mapping of cell type-specific SEs, this allowed the identification of putative key TFs controlling mDANs. We show that LBX1, NHLH1 and NR2F1/2 are necessary for mDAN differentiation, with LBX1 and NHLH1 also able to increase the number of mDANs. Further characterization of these TFs revealed the control of cholesterol

bioRxiv preprint doi: https://doi.org/10.1101/2023.01.27.525898; this version posted January 27, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

biosynthesis by LBX1 and induction of miR-124 by NHLH1 as a few of the mechanisms contributing to mDAN specification. In summary, this study provides novel profiling of differentiating mDANs. Our data can be exploited for further purposes such as studies on disease-associated regulatory genetic variation.

Materials and Methods

Cell lines

The human iPSC line GM17602 (Coriell) was used in this study as a control and for the generation of a tyrosine hydroxylase (TH) reporter cell line. This iPSC line was previously characterized in (24) (called HFF) and used in (19) for the generation of the reporter line. Briefly, in the reporter line, the T2A coding sequence was fused with the mCherry open reading frame, and it was biallelically inserted in place of the stop codon in the endogenous TH locus using CRISPR/Cas9 editing (Figure 1A).

Cell culture and differentiation

hiPSC maintenance, generation of small molecule neural precursor cells (smNPC) and differentiation towards mDANs are described in (25). In Figure 1A, a schematic representation of the mDAN differentiation protocol can be observed with the different media used and for how long the cells were kept in culture. This differentiation protocol is composed of three different media. smNPCs are incubated in Differentiation medium one containing 100 ng/ml of FGF8b, 1 µM of purmorphamine (PMA), and 200 µM of ascorbic acid which starts with smNPC and is used during the first eight days of differentiation. From day 8 until day 10 of differentiation cells are kept in Differentiation medium two, composed of 0.5 µM PMA and 200 µM of ascorbic acid. Lastly, maturation medium containing 200 µM of ascorbic acid, 10 ng/ml of brain-derived neurotrophic factor (BDNF), 10 ng/ml of glial cell-derived neurotrophic factor (GDNF), 500 µM of dcAMP, and 1 ng/ml of TGFβ3 which is used from day 10 until the desired time point. The molecules used in the three different media were mixed in N2B27 medium (1:1 of Dulbecco's modified Eagle medium/Nutrient Mixture F-12 [DMEM/F12] and Neurobasal medium supplemented with 1% Pen/Strep, 1% GlutaMAX, 1% B27 supplement minus vitamin A, and 0.5% N2 supplement, all from Gibco).

Astrocyte differentiation was induced based on the protocol from (26), with small changes. Briefly, smNPC were seeded in N2B27 medium complemented with 3 µM CHIR99021, 0.5 µM PMA and 150 µM ascorbic acid. After two days, fresh medium plus 20 ng/ml of FGF-2 was added to the smNPC culture. On day four, the cells were split into a new plate to start neural stem cell (NSC) generation. The medium used for the generation and maintenance of NSC contained DMEM/F-12, 1% Pen/Strep, 1% GlutaMAX, 1% B27 supplement serum-free (with vitamin A), 1% N2 supplement, 40 ng/ml EGF,

bioRxiv preprint doi: https://doi.org/10.1101/2023.01.27.525898; this version posted January 27, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

40n g/ml FGF-2, and 1.5 ng/ul hLIF. The cells were kept in this medium for 3-4 passages. Then, astrocyte differentiation was started in DMEM/F-12 supplemented with 1% Pen/Strep, 1% GlutaMAX, and 1% FBS.

All the plates used were previously coated with Geltrex matrix from Gibco.

Flow cytometry and FACS

On the day of analysis, medium was removed from the cells and Accutase (Gibco) was added to the well. Cells were incubated in Accutase at 37°C until detachment (10-30 min). Then, 2 volumes of DMEM/F-12 were added to the well. Cells were gently pipetted up and down until dissociation and collected in a 15 ml Falcon tube after passing them through a 50 µm cell strainer to obtain a single-cell suspension. Falcon tubes were centrifuged 3 min at 300xg and room temperature (RT). Pellets were resuspended in PBS and cells were transferred to a 1.5 ml Eppendorf tube. 4',6-diamidino-2-phenylindole (DAPI) was added to the cells at 5 µg/ml and incubated for 5 min at 4°C in a tube rotator. After incubation, the cells were washed twice with 2% (w/v) BSA in PBS. Then, cells were ready for either flow cytometry analysis or fluorescence-assisted cell sorting (FACS). For flow cytometry analysis, the BD LSRFortessa™ cell analyzer was used. For FACS, the BD FACSAria™ III sorter was used. FlowJo version 10 software was used for data processing and generation of plots.

Total RNA extraction, cDNA synthesis and RT-qPCR

For total RNA extraction, Quick-RNA miniprep kit from Zymo Research (R1055) was used as per manufacturer's instructions. When total RNA had to be extracted from FACS sorted cells, Quick-RNA microprep kit from Zymo Research (R1050) was used. Briefly, to avoid RNA degradation due to FACS, cells were collected in batches for no longer than 10 min. Then, sorted cells were pelleted by centrifugation for 3min at 500 g and 4°C. Lysis buffer from the kit was added to the pellet until 150 000 – 300 000 cells were collected. RNA extraction was finalized following manufacturer's instruction.

For cDNA synthesis, amounts from 200 ng to 1 ug of total RNA were used, depending on sample availability. To perform the reaction, dNTPs (0.5 mM, ThermoFisher, R0181), oligo dT-primer (2.5 µM), 1µl RevertAid reverse transcriptase (200 U/µl, ThermoFisher, EP0441), and 1µl Ribolock RNase inhibitor (40 U/µl, ThermoFisher, EO0381) were mixed with the proper amount of total RNA in a final volume of 40 µl. cDNA synthesis was performed at 42°C for 1 hr. Reaction was terminated by incubating the reaction at 70°C for 10 min. cDNA was diluted 1:10, 1:5 or 1:2 in DNase/RNase free water, depending on the amount of total RNA used (1 µg, 500 ng or 200 ng, respectively). Diluted cDNA was stored at -20°C.

RT-qPCR was performed in an Applied Biosystems 7500 Fast Real-Time PCR system. For the reaction, 5 µl of diluted cDNA was mixed with 1xAbsolute Blue qPCR SYBR green low ROX mix (ThermoFisher, AB4322B) and 500 nM primer concentration, in a final volume of 20 µl per well using

bioRxiv preprint doi: https://doi.org/10.1101/2023.01.27.525898; this version posted January 27, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

AmpliStart 96well plate (Westburg, WB1900). PCR reaction had the following settings: 95°C for 15min and then 40 cycles of 95°C for 15sec, 55°C for 15sec and 72°C for 30sec. The $2^{-(\Delta\Delta Ct)}$ method was used to calculate gene expression levels. ΔΔCt was calculated using the following formula: $(\Delta Ct_{(target\ gene)}-\Delta Ct_{(housekeeping\ gene)}) - (\Delta Ct_{(target\ gene)}- \Delta Ct_{(housekeeping\ gene)})_{reference\ condition}$. ACTB was used as the housekeeping gene. GraphPad Prism 9 was used to create plots and perform statistics. Data was always normalized to the reference condition and one sample t test was performed to determine significance.

Omni-ATAC

The Omni-ATAC protocol was performed with 50 000 cells (or 25 000 cells for day 50 neurons) with minor modifications from (27). Using TDE1 Tagment DNA Enzyme and TD Buffer from Illumina. Cells were either derived from FACS or directly from cell culture plates after detachment with Accutase. After amplification with primer sets as described in (28), libraries underwent size selection using AMPure XP beads (Beckman Coulter) to remove large fragments. Libraries were stored at -20°C. For library quality control, the Agilent High Sensitivity DNA kit (5067-4626) was used in a 2100 Bioanalyzer instrument. Libraries used for sequencing presented a fragment distribution starting from ~200 bp until around 1000 bp, with nucleosomal pattern peaks.

Low-input ChIP-seq and calling of SEs

To perform low-input chromatin immunoprecipitation (ChIP) for H3K27ac, the Low Cell ChIP-Seq Kit from Active Motif was used (53084). For smNPC and sorted neurons, a total of 150 000 and 200 000 cells were used, respectively. The protocol was performed according to manufacturer's instructions, with minor changes. Sonication of the cells was performed using a Bioruptor® Pico sonication device from Diagenode. The settings used for sonication were 40 cycles of 30 sec off and 30 sec on at 8°C. After sonication, 20% of the sonicated sample volume was saved as input. For the immunoprecipitation (IP) reaction, a total of 4 µg of H3K27ac antibody was used per reaction. Details about the antibody can be found in the Antibodies section. Input samples that were collected after sonication were processed together with IP reactions starting from the step where reversal of cross-links and DNA purification was performed. After this point, samples were processed in parallel. Libraries were stored at -20°C until quality control and sequencing. For library quality control, the same procedure as in the Omni-ATAC protocol was used. For low-input ChIP, good libraries presented an average of 600 bp fragment distribution. The low-input ChIP-seq protocol was validated by comparing the identified SEs to those previously detected in smNPC via regular high-input ChIP-seq methods (29). SEs were considered as regions larger than 10 kb.

Immunocytochemistry

PhenoPlate 96-well (PerkinElmer, 6055308) previously coated with geltrex was used for immunocytochemistry. Cells were fixed in 4% paraformaldehyde for 15 min at RT. Then, cells were

bioRxiv preprint doi: https://doi.org/10.1101/2023.01.27.525898; this version posted January 27, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

washed three times with PBS. For permeabilization, cells were incubated 1 hr at RT in PBS, 0.4% Triton-X, 10% goat serum and 2% BSA. After 1 hr, cells were washed twice with PBS. Primary antibody was diluted in PBS, 0.1% Triton-X, 1% goat serum and 0.2% BSA. Cells were incubated with primary antibody shaking overnight at 4°C. Next day, cells were washed three times with PBS. Then, secondary antibody was diluted in the same buffer as the primary and incubated for 2-3 hrs at RT. Finally, cells were washed three times with PBS. In the first wash, DAPI was added to the PBS (5 µg/ml) and incubated for 15 min at RT. Images were taken using a Zeiss spinning disk confocal microscope. Image processing was done in ImageJ.

Bacterial culture, plasmid extraction and lentivirus production

Glycerol stocks of bacteria containing the plasmid of interest were taken from -80°C. Without letting the glycerol stocks thaw and with the help of a P10 pipette tip, bacteria were added to a polypropylene graduated culture tube (Roth, EC01.1) containing 5 ml of LB Broth medium (20.6 g/l, Roth, X968.2) supplemented with Ampicillin (100 µg/ml). Bacteria were incubated at 37°C and 120 rpm shaking for 5 hrs. Then, bacteria were transferred to an Erlenmeyer containing 150 ml of LB Broth medium supplemented with ampicillin. Erlenmeyer was incubated overnight at 37°C and shaken at 120 rpm. After bacterial expansion, plasmid extraction was performed using NucleoBond Xtra Midi EF (740420.50) as per manufacturer's instructions.

For lentivirus production, 8 million HEK293T cells were seeded in a T75 flask using 15 ml of DMEM (Gibco) supplemented with 1% Pen/Strep, and 10% FBS and transfected the next day for lentiviral production. Third-generation lentiviral particles were produced. Briefly, 4 µg of pMDG, 2 µg of pMDL, 2 µg of pREV, and 8 µg of the plasmid of interest were mixed with 200 µl of CaCl2 (1 M, Sigma, 21115-100ML). Volume was completed with sterile water up to 800 µl. The 800 µl of the plasmid mixture was mixed with 800 µl of HEPES buffered saline (Sigma, 51558-50ML) by making bubbles slowly in a dropwise manner. This transfection mixture was incubated for 20 min at RT. In the meanwhile, 16 µl of 25mM chloroquine was added to the T75 flask containing the HEK293T cells and incubated for a minimum of 5 min to facilitate transfection. Next, the transfection mixture was added to the HEK293T cells. After 4-6hr, medium was removed from HEK293T cells and 14 ml of fresh medium was added. After 48 hrs, lentiviral particles were ready for collection. HEK293T medium from the flask was collected in a 15ml Falcon tube and centrifuged for 10 min at 2000 rpm and 4°C. The supernatant was cleared by filtering through a 0.45 µm filter (Sartorious, 16537). Filtered lentiviral particles were aliquoted in cryovials (1 ml aliquots) and stored at -80°C.

Transduction

Two different approaches for transduction of differentiating neurons were used in this study: early and late transductions. All the lentiviral particles used contained a GFP reporter that helped control for transduction efficiency. For overexpression constructs a codon-optimized cDNA sequence of the TF

bioRxiv preprint doi: https://doi.org/10.1101/2023.01.27.525898; this version posted January 27, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

was used. Lentiviral particles were previously tested to adjust transduction efficiencies to ~80%, determined by GFP positive cells using flow cytometry.

For early transduction, smNPC were seeded in a 6-well plate with a density of 1-2 million cells per well and differentiation was started by seeding them directly in a differentiation medium. Next day, medium was removed and lentiviral particles were added to the cells in a final volume of 1 ml. Plate was sealed with parafilm and centrifuged for 10 min at 250 g and RT for spinfection. After spinfection, 1 ml of differentiation medium was added to the cells. Lentiviral particles were incubated overnight. Next day, lentiviral particles were removed, and fresh differentiation medium was added to the cells. Differentiation continued until the day of analysis.

For late transduction, differentiating cells were split on day 8 of differentiation into a 6 well plate at a density of 3 million cells per well. The next day, the medium was removed and transduction was performed as described for early transduction. Differentiation continued until the day of analysis.

GW3965, ABX464, and Rapamycin treatments

GW3965-HCl (Selleckchem, S2630), ABX464 (Selleckchem, S0076), and Rapamycin (Selleckchem, S1039) molecules were tested to determine their working concentration in our cultures. For GW3965-HCl, 0.1 µM was sufficient to induce SREBF1 mRNA levels. For ABX464, 10 µM was sufficient to induce miR-124-3p. For Rapamycin, 10 nM was sufficient to downregulate SREBF1 mRNA levels.

The effect of the molecules was tested during normal differentiation and under LBX1 or NHLH1 knock-down (KD) conditions. Briefly, neuronal differentiations were started and on day 8 of differentiation, cells were split into a 6-well plate with a density of 3 million cells per well. Next day, late transduction with shLBX1, shNHLH1 or shScramble was performed as described above. On day 10 of differentiation, transduced cultures were treated either with GW3965, ABX464, or rapamycin. Cells transduced with shLBX1 were treated with GW3965 or Rapamycin, while cells transduced with shNHLH1 were treated with ABX464. Treatments continued until day 15, when cells were analyzed.

TaqMan assay

TaqMan assay was performed to determine miR-124-3p levels using TaqMan™ MicroRNA Reverse Transcription Kit (ThermoFisher, 4366596), TaqMan™ MicroRNA Assay hsa-miR-124-3p (ThermoFisher, AssayID 003188_mat, 4440886), TaqMan™ MicroRNA Control Assay U6 snRNA (ThermoFisher, AssayID 001973, 4427975), TaqMan™ MicroRNA Assay hsa-miR-423-5p (ThermoFisher, AssayID 002340, 4427975), and TaqMan™ Fast Advanced Master Mix (ThermoFisher, 4444556). First, reverse transcription reaction was performed using 0.15 µl of 100 mM dNTPs, 1 µl of MultiScribe™ Reverse Transcriptase (50 U/µl), 1.5 µl of 10X Reverse Transcription buffer, 0.19 µl of RNase inhibitor (20 U/µl), 10 ng of total RNA, and 3 µl of RT primer (either U6 or miR-124-3p primer) in a final volume of 15 µl. Reactions were incubated for 30 min at 16°C and then

bioRxiv preprint doi: https://doi.org/10.1101/2023.01.27.525898; this version posted January 27, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

30 min at 42°C. Reaction was terminated by incubating the tubes for 5 min at 85°C. Then, PCR was performed using the same plates and machine as for RT-qPCR. For the PCR reaction, 1 µl of 20X TaqMan MicroRNA Assay (either from U6 or miR-124-3p), 1.33 µl from RT reaction, and 10 µl of TaqMan™ Fast Advanced Master Mix were mixed in a final volume of 20 µl. The PCR reaction settings were: 10 min at 95°C followed by 40 cycles of 15 sec at 95°C and 1 min at 60°C. To calculate miR-124-3p levels, the $2^{-(\Delta\Delta Ct)}$ method was used again, where U6 snRNA represented the housekeeping gene. GraphPad Prism 9 was used as described for RT-qPCR analysis.

Sequencing

Prior RNA-seq, RNA quality was determined by using the Agilent RNA 6000 Nano kit (5064-1511) in an Agilent 2100 Bioanalyzer machine. Samples selected for sequencing had a RIN value > 7. RNA-seq from time course data, including smNPC, mDANs and astrocytes samples, was done using the TruSeq Stranded mRNA library prep kit, single-end 75 bp read length, and a NextSeq500 machine.

RNA-seq from KD samples was done using an Illumina stranded mRNA library prep. Kit, paired-end 50 bp read length, and a NovaSeq6000 machine.

ATAC-seq samples were sequenced on a NextSeq500 machine using paired-end 75 bp read length.

ChIP-seq samples were sequenced on a NextSeq500 machine using single-end 75 bp read length.

RNA-seq analysis

For the time course RNA-seq, including smNPC, mDANs at days 15, 30 and 50 of differentiation, non-mDAN at days 15 and 50 of differentiation, and astrocytes at day 65 of differentiation the following tools were used. Raw fastq files were assessed for quality using FastQC (https://www.bioinformatics.babraham.ac.uk/projects/fastqc/). Summary of sample quality controls was obtained using MultiQC (30). Next, the Paleomix pipeline was used for trimming sequencing adapaters using AdapterRemoval (31, 32). After adapter removal, ribosomal RNA was filtered from the data using SortMeRNA (33). Alignment to the reference genome was done using STAR (34). BAM files were validated using Picard (https://broadinstitute.github.io/picard/). Quality reads (≥Q30) were filtered using SAMtools (35). Gene counts were obtained using FeatureCounts from the Rsubread package (36). Differential expression analysis was done using the R package DESeq2 (37). For more information about the specific versions and settings used for the different tools, please refer to our repository (RNA-seq folder, RNA-seq_DataAnalysis_TimeCourse.rmd script). Genome version and annotation were GRCh38 patch 12 and Gencode human release 31, respectively.

For the RNA-seq data from the different TF knock-down experiments, including LBX1 KD, NHLH1 KD and NR2F1/2 KD, a snakemake pipeline was used (38). This pipeline includes the tools STAR, SAMtools, FastQC, FastQ Screen, AdapterRemoval, Rsubread, DESeq2, ggplot2 and apeglm (39–41).

bioRxiv preprint doi: https://doi.org/10.1101/2023.01.27.525898; this version posted January 27, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

For more details about the pipeline, please refer to our repository (RNA-seq folder, RNA-seq_DataAnalysis_TF_KDs.rmd script). Genome version and annotation were GRCh38 release 102.

ATAC-seq analysis

For the time course ATAC-seq, including smNPC, mDANs on days 15, 30, and 50 of differentiation, non-mDAN on days 15 and 50 of differentiation, and astrocytes on day 65 of differentiation the following tools were used. Raw fastq files were assessed for quality using FastQC. A summary of sample quality control was obtained using MultiQC. Using the Paleomix pipeline, trimming of sequencing adapters was done using AdapterRemoval and mapping to the reference genome with BWA (42). BAM files were validated using Picard. Quality reads (≥Q30) were filtered using SAMtools. Peak calling was performed using Genrich (https://github.com/jsh58/Genrich). For more information about the specific versions and settings used for the different tools, please refer to our repository (README.md file inside the ATAC-seq folder). Genome version was GRCh38 patch 1.

ChIP-seq analysis

For H3K27ac ChIP-seq, including smNPC, mDANs at days 30 and 50, and non-mDANs at day 50, the following tools were used. After merging R1 and R2 raw fastq files as described by Active Motif, the new fastq files were assessed for quality using FastQC. Using the Paleomix pipeline, trimming of sequencing adapters was done using AdapterRemoval and mapping to the reference genome with BWA. BAM files were validated using Picard. Quality reads (≥Q30) were filtered using SAMtools. New BAM files were sorted according to mapping position using SAMtools prior the molecular identifier de-duping step. For de-duping, a perl script provided by Active Motif was used (not provided, it should be requested from the manufacturer. Script name rmDupByMids.pl.txt, version from 2019). For calling enhancers and super-enhancers HOMER was used (43). For more information about the specific versions and settings used for the different tools, please refer to our repository (ChIP-seq folder, Lowinput_ChIP-seq_analysis.rmd script). Genome version was GRCh38 patch 12.

EPIC-DREM analysis

EPIC-DREM was applied as a snakemake pipeline. As an input, pre-processed BAM files from the ATAC-seq analysis and the gene counts from the previously described RNA-seq analysis were used, including smNPC, and mDANs at days 15, 30 and 50. ATAC-seq peak calling was performed with Genrich over replicates. The Regulatory Genomics Toolbox was used to identify footprints in called peaks (21) and subsequently, TF-gene affinities were calculated using TEPIC (22). The resulting time-point-specific lists of TF-gene links were merged and filtered according to expression, removing links from unexpressed TF. TF was considered unexpressed if it presented with a transcripts per million (TPM) value < 1 in all analyzed time points. Time-point-specific GRN were identified with interactive Dynamics Regulatory Events Miner (iDREM) (44). Results from iDREM were further processed in R

bioRxiv preprint doi: https://doi.org/10.1101/2023.01.27.525898; this version posted January 27, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

for visualization and GO enrichment was performed using clusterProfiler (45). For more details about the different parameters used, please refer to our repository (EPIC-DREM folder).

Primers

Primer	sequence (5' → 3')
hACTB_F	AAACTGGAACGGTGAAGGTG
hACTB_R	AGAGAAGTGGGGTGGCTTTT
hNR2F1_F	GAGCAGGTGGAGAAGCTCAA
hNR2F1_R	CAGGCGTCTGACGTGAACAG
hNR2F2_F	AGGCGCTGCACGTTGAC
hNR2F2_R	AGGCATCTGAGGTGAACAGGACTA
hNHLH1_F	ACGCTACCCCTGAGAGTCTAGAAA
hNHLH1_R	TCTGGGTGCTCAAGGCTCAT
hLBX1_F	AAGGCCGCGACGGTATG
hLBX1_R	GCGACTTTCGCCGCTTCTTA
hSOX4_F	CCTAATTTCTCCATGTTTACACTTCAAT
hSOX4_R	GTGGACACTGGTGGCAGGTT
hHOXB2_F	CCGAGGAAGAGCTGGATTTTT
hHOXB2_R	GTTAGGGAAACTGCAGGTCGAT
hLMX1B_F	GCCGAAAGGTCCGAGAGA
hLMX1B_R	CTTCTTCATCTTTGCTCTTTGGTT
hNHLH1_OE_F	CCGACAAGAAGCTCTCCAAGA
hNHLH1_OE_R	TGGTTCAGGTAGGAGATATAGCAGATG
hSREBF1_F	GCTCCTCCATCAATGACAAAATC
hSREBF1_R	TGCAGAAAGCGAATGTAGTCGAT
hSREBF2_F	CCTGTCATTCGAGTCAGGTTCTG
hSREBF2_R	CAATCACACCATTTACCAGCCATA
hSCARB2_F	CTACAGGGAACTCAGAAACAAAGCA
hSCARB2_R	CCAACAGATTGGTCTCGTTCAA
hGAPDH_F	GCATCCTGGGCTACACTGAG
hGAPDH_R	GGTGGTCCAGGGGTCTTACT

bioRxiv preprint doi: https://doi.org/10.1101/2023.01.27.525898; this version posted January 27, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Antibodies

Antigen	Specie	Dilution	Cat No
TH	Rabbit	1:250	Sigma T8700-1VL
NeuN Alexa fluor 488 conjugated	Mouse	1:250	Sigma MAB377X
Histone H3 acetyl K27 (H3K27ac)	Rabbit	**	Abcam Ab4729
Rabbit IgG Alexa fluor 647	Goat	1:1000	Invitrogen A27040

Bacterial glycerol stocks

All bacterial glycerol stocks are from GeneCopoeia. The different shRNA constructs used were:

Name	Vector	shRNA sequence	Cat No
Scramble shRNA (shScramble)	psi-LVRU6GP	GCTTCGCGCCGTAGTCTTA	CSHCTR001-LVRU6GP
shRNA targeting HOXB2 (shHOXB2)	psi-LVRU6GP	GGTATTACTGAATTAGCGTTT	HSH008988-LVRU6GP-b
shRNA targeting LMX1B (shLMX1B)	psi-LVRU6GP	GGGTGACTACGAGAAGGAGAA	CS-HSH101934-LVRU6GP-01-c
shRNA targeting NHLH1 (shNHLH1)	psi-LVRU6GP	GCTATATCTCCTACCTGAACC	HSH011829-LVRU6GP-a
shRNA targeting NR2F2 (shNR2F2)	psi-LVRU6GP	GGAGGAACCACATATAACACT	CS-HSH110558-LVRU6GP-01-c
shRNA targeting NR2F1 (shNR2F1)	psi-LVRU6GP	CCGCAGGAACTTAACTTACAC	HSH110557-LVRU6GP-b
shRNA targeting LBX1 (shLBX1)	psi-LVRU6GP	GACGTAGAGTCCGCCAAGAAA	HSH000806-LVRU6GP-d

The different overexpression constructs used were:

Name	Vector	Accession No	Cat No
Control empty vector (GFP OE)	pEZ-Lv228	**	EX-NEG-Lv228
NHLH1 overexpression (NHLH1 OE)	pEZ-Lv228	NM_005598	EX-F0569-Lv228
LBX1 overexpression (LBX1 OE)	pEZ-Lv228	NM_006562	EX-T1288-Lv228
NR2F2 overexpression (NR2F2 OE)	pEZ-Lv228	NM_021005	EX-C0221-Lv228

bioRxiv preprint doi: https://doi.org/10.1101/2023.01.27.525898; this version posted January 27, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Results

Generation of paired transcriptomic and chromatin accessibility profiles of mDAN differentiation

mDANs were enriched for transcriptomic and epigenomic profiling based on the expression of P2A-mCherry reporter, stably expressed under the control of the endogenous promoter of the TH gene, coding for tyrosine hydroxylase, the rate-limiting enzyme for dopamine biosynthesis (Figure 1A, (19)). The presence of the mCherry reporter allowed the purification of mDANs from heterogeneous iPSC-derived neuronal cultures by FACS at multiple time points of differentiation (Figure 1A-B). Reporter expression correlated with TH expression, as validated by immunocytochemistry (Figure 1B). Using the reporter cell line, paired transcriptomic and chromatin accessibility profiles were generated from neuronal progenitor cells (smNPCs) and mDANs after 15, 30, and 50 days of differentiation, as shown in Figure 1A. Time points were selected based on culture features reported by (46): mDANs appear in the culture by day 15, after the cells are incubated in maturation medium (Figure 1A); they start to show electrophysiological activity after 30 days of differentiation (46). Finally, by day 50 mDANs begin to resemble more mature neurons. Gene expression and chromatin accessibility profiles were generated from iPSC-derived astrocytes differentiated from the same reporter cell line for 65 days (Figure 1, (26)) as additional reference data.

The data generated from mDANs showed cell-type-specific gene expression and chromatin accessibility profiles based on established markers of these cells (TH, EN1, DDC, and LMX1B) when compared to non-mDANs, and in contrast to either pluripotency/early neuroectoderm markers (PAX6) or glial markers (GFAP) (Figure 1C, (14)(47)(48)). The transcriptome analysis of the mCherry population across differentiation confirmed a downregulation of pluripotency genes in parallel with induction of pan-neuronal markers and, importantly, mDAN-specific marker genes (Supplementary Figure 1, (13, 49)). Moreover, some of the genes selective for either A9 and A10 mDAN subtypes became upregulated, while others remained undetected, suggesting that the cells have not adopted a subtype-specific identity. Principal component analysis (PCA) confirmed a clear separation of smNPCs, mDANs and astrocytes at the level of both transcriptome and chromatin accessibility (Figure 1D). Taken together, our reporter cell line and the obtained genome-wide profiles can be used for further characterization of the regulatory landscape of mDAN differentiation.

Integrative analysis predicts key regulators of human mDAN differentiation

To integrate the paired transcriptomic and epigenomic profiles to identify key regulators of mDAN differentiation, our previously published EPIC-DREM pipeline was adapted for use with ATAC-seq data and applied (Figure 1E, (8)). In short, genome-wide accessible regions were determined per cell type and time point, and footprinting analysis was performed to predict TF binding sites (TFBS) in these regions using HINT-ATAC (21). Computed TF-gene scores based on accessibility signal and TF

bioRxiv preprint doi: https://doi.org/10.1101/2023.01.27.525898; this version posted January 27, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

binding strength were integrated to finally define regulators across time in previously described statistical framework (22, 50). The time point-specific predictions were combined with the time series gene expression changes by DREM (23) to build a time point-specific gene regulatory network of mDAN differentiation. DREM identifies gene expression bifurcation points across the time points analyzed, corresponding to groups of co-expressed genes. For each bifurcation point (also called split node), TFs are ranked according to the highest number of target genes within the group based on accessibility data predictions. Therefore, DREM highlights the TFs controlling most of the observed transcriptional changes across time.

Figure 2A shows the result of EPIC-DREM on mDAN time course data. There are a total of 26 split nodes with a total of 327 TFs ranked differently across them (Supplementary Table 1). Highlighted in red are different top-ranked TFs known to be involved in the regionalization, differentiation, and specification of mDANs. For example, well-described TFs controlling mDAN differentiation, such as NR4A2 (also known as NURR1), LMX1A/B and EN1, were identified among the main regulators (51–54). Pioneer TFs, critical for cell reprogramming due to their ability to alter chromatin structure (55), such as ASCL1 and NEUROG2, also appeared as top-ranked TFs (56, 57). Other factors more involved in the regionalization and differentiation of the midbrain such as GBX2, NKX6-1, PBX1, SREBF1, NR1H2 (also known as LXRβ), and MSX1 were captured by EPIC-DREM (58–62). Thus, our EPIC-DREM predictions are in line with the existing literature.

Moreover, pathway enrichment analysis of the genes included in the first four split nodes revealed enrichment for neuronal functions such as axon development and synapsis formation among the upregulated genes (Figure 2B, nodes 1 and 2) and enrichment for ribosomal RNA production and cell division for downregulated genes (Figure 2B, nodes 3 and 4, respectively), consistent with switching from the multipotent state to neuronal differentiation. Thus, EPIC-DREM can highlight the main biological processes governing mDAN differentiation.

Identification of key TFs controlled by super-enhancers

While EPIC-DREM allowed us to predict the TFs controlling the highest number of target genes during mDAN differentiation, we set out to further prioritize these TFs by identifying those that are themselves under high regulatory load in a cell type-selective manner (4, 63). To do this we performed low input ChIP-seq for H3K27ac, a histone mark associated with active enhancers, for mDANs at days 30 and 50 of differentiation to determine which TFs display association with SEs at the respective time point. For the association, SE regions were overlapped with the genomic coordinates of the genes encoding for TFs and expressed in a time point-specific manner to obtain a list of 49 TFs controlled by SEs across both time points (Figure 3A, Supplementary Table 2). Finally, the list of all SE-associated TFs at either time point was compared with a list of TFs combining the top 20 ranked TFs from each of the split

bioRxiv preprint doi: https://doi.org/10.1101/2023.01.27.525898; this version posted January 27, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

nodes from EPIC-DREM (Figure 3B, Supplementary Table 2). Among the 49 SE-associated TFs, 17 were also among the top 20 ranked TFs from EPIC-DREM (Figure 3B, Supplementary Table 2).

The list of 17 TFs was further explored to select the most promising candidates for functional analysis. For this, a literature search was performed to see whether these TFs had already been associated with mDAN function or development. For example, as TCF4 and MEIS1 have been previously related to mDAN subset specification and striatal dopaminergic system formation, respectively, and consequently were not included in follow-up experiments (64, 65). From the remaining TFs, HOXB2, LBX1, NHLH1, NR2F1 (also known as COUP-TFI), NR2F2 (also known as COUP-TFII) and SOX4 were found to present the strongest SE signals and most dynamic gene expression profiles and, therefore, were selected for functional analysis as novel candidate regulators of mDAN differentiation (Figure 3C).

Among the candidates, HOXB2 has been mainly associated with neural crest and hindbrain patterning (66). No relation of HOXB2 with mDANs has been described before. HOXB2 is clustered with other HOXB genes that are also proximal to the same SE. Most of the chromatin accessibility is observed within the HOXB2 gene body (Figure 3C). HOXB2 was abundantly expressed in smNPCs and early time points of differentiation but decreased in day 50 mDANs, in keeping with its predicted role as a regulator of downregulated genes (Figure 2A, Supplementary Table 1).

LBX1 has been implicated in the development of different types of neurons such as retrotrapezoid nucleus (RTN) neurons, GABAergic neurons, and somatosensory interneurons in the spinal cord (67–69), highlighting a role for this TF in cell fate decisions. LBX1 was the only TF among the selected candidates to present an mDAN-specific SE (Figure 3C). Consistently, LBX1 expression was induced >10-fold upon early mDAN differentiation.

NHLH1 is mainly characterized as an early pan-neuronal marker that determines many neuronal cell fate decisions (70–73). NHLH1 expression was continuously increased during differentiation with almost 70 RPKM detected in day 50 mDANs (Figure 3C).

Temporal expression dynamics of NR2F1 and NR2F2 in the developing brain are important for the specification and balance of different neuronal lineages as well as neuron-to-glia transitions (74–77). Recently, NR2F1 was found to be deregulated in iPSCs from PD patients carrying the LRRK2-G2019S mutation (29). During early mDAN differentiation NR2F1 expression increased to particularly high levels (>200 RPKM), while NR2F2 expression decreased in later time points compared to smNPCs (Figure 3C).

SOX4, like many SOX genes, has been described to be implicated in neurogenesis and maintenance of progenitor cells during development. Interestingly, the role of this TF in generating TH-expressing cells from sympathetic or enteric nervous systems has been described (78, 79). Also, SOX4 expression

bioRxiv preprint doi: https://doi.org/10.1101/2023.01.27.525898; this version posted January 27, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

increased to >100 RPKM during mDAN differentiation and was accompanied by an increased enhancer signal at the locus (Figure 3C).

Overall, these candidates presented great potential as cell fate and differentiation determinants with no previous relation to mDANs.

LBX1, NHLH1, and NR2F1/2 are necessary for mDAN neurogenesis

To functionally validate the data-driven predictions *in vitro*, TFs were knocked down (KD) during differentiation to determine their impact on the mDAN number in mixed cultures. A lentiviral vector containing a shRNA targeting the different candidate TFs and a GFP reporter was used to control for transduction efficiency. To assess the effect of the TFs, two different experimental designs for transduction were used: early and late transduction (Figure 4A). For early transduction, cells were transduced on day one of differentiation while for late transduction, cells were transduced on day nine, just before they entered the maturation phase. In both cases, the cells were analyzed on day 15 (Figure 4B shows the results from the KD experiments). LMX1B served as a positive control as its role in mDAN differentiation is well characterized (53, 80). SOX4 was omitted as we were unable to identify an shRNA with a KD efficiency of >25%. As expected, LMX1B KD reduced mDAN numbers in the cultures as assessed by the mCherry reporter signal (Figure 4B). Upon early transduction, LMX1B KD reduced the numbers of mDAN in the culture by ~60%. However, the KD efficiency for LMX1B, as measured by RT-qPCR, was very variable, which can be explained by the low levels of GFP positive cells (~40%) upon analysis on day 15 of differentiation (Supplementary Figure 2A), possibly masking the KD of this TF by the GFP negative cells in the culture. On the other hand, with late transduction, good transduction efficiency and strong KDs were observed on the day of analysis for LMX1B. Interestingly, late KD of this TF reduced the mDAN numbers by only 30% but also decreased the overall cell density in the cultures. This highlights the dual role of LMX1B during differentiation and correlates with previous studies (80), emphasizing the biological relevance of our cellular system.

LBX1 KD by early transduction was found to severely decrease the cell numbers, with very few cells remaining by day 5-6 post transduction, preventing a more detailed analysis of the effect of this TF on mDANs (Supplementary Figure 2A). Although LBX1 KD by late transduction also resulted in some decrease in cell numbers, the surviving population displayed good transduction efficiencies and allowed us to study the impact of LBX1 on mDANs. After late transduction, LBX1 KD efficiency on the day of analysis was ~60% and mDAN numbers were reduced by 50% (Figure 4B).

NHLH1 KD had the strongest effect on mDAN numbers among the candidates tested. KD efficiencies in both early and late transduction were around 80% and mDAN numbers were reduced by roughly 80% and 60% in early and late transduction, respectively. Although the total number of cells was not affected by the NHLH1 KD in early transduction, GFP-positive cells represented only half of the cells on the day of analysis, similar to what was observed for LMX1B KD (Supplementary Figure 2A).

bioRxiv preprint doi: https://doi.org/10.1101/2023.01.27.525898; this version posted January 27, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Since NR2F1 and NR2F2 are known to act redundantly (74), lentiviral particles expressing the respective shRNAs were combined to target both TFs to avoid possible compensatory mechanisms. While individual KD of NR2F1 or NR2F2 was successful (data not shown), in dual KD conditions only NR2F1 levels were significantly reduced (Figure 4B). Nevertheless, the combined NR2F1/2 KDs markedly affected mDAN differentiation, reducing mDAN numbers in the culture by close to 80% and 50% in early and late transduction, respectively. In addition, when NR2F1/2 were KD, cells could keep good transduction efficiencies until the day of analysis despite the negative effect on mDAN numbers, in contrast to what was observed for LMX1B and NHLH1 KDs (Supplementary Figure 2A).

Lastly, for HOXB2 good transduction efficiencies were observed on the day of analysis, with >60% KD efficiency in both early and late transduction. However, although HOXB2 KD reduced the numbers of mDANs in the culture, the effect was the weakest among the tested TFs.

In summary, LBX1, NHLH1 and NR2F1/2 were all found to be necessary for mDAN differentiation. Their depletion during the specification of these neurons exhibited a more robust phenotype than that of LMX1B, a well-established regulator of mDAN differentiation, thereby demonstrating an important regulatory role for LBX1, NHLH1, and NR2F1/2. On the other hand, loss of HOXB2 only had a limited effect on mDAN levels.

Elevated expression of LBX1 or NHLH1 can increase mDAN numbers

To further determine the role of the identified TFs in mDANs, LBX1, NHLH1 and NR2F1/2 were overexpressed during differentiation using the same lentiviral transduction approach as in the earlier KD experiments. Figure 4C shows the results from the overexpression experiment for the three different TFs. For LBX1, strong overexpression (9-20-fold compared to control vector) and high levels of transduced cells were observed on the day of analysis (Supplementary Figure 2B) for both early and late transduction. However, the induction of LBX1 had opposite effects depending on the time point of differentiation. While LBX1 overexpression early during differentiation had a negative impact on mDAN numbers, late overexpression resulted in significantly increased mDAN numbers (Figure 4C).

Overexpression of NHLH1 during early differentiation resulted in an almost 2-fold increase in mDANs, despite a modest increase of gene expression levels by 50% and reduced number of transduced cells based on the GFP signal on the day of analysis. With late induction of NHLH1 expression, again, high levels of transduced cells, and stable overexpression could be observed, leading to a significant increase in mDAN numbers.

We were unable to achieve any meaningful overexpression of NR2F1, most likely due to its already very high endogenous expression levels (Figure 3C and data not shown). Therefore, only NR2F2 overexpression was tested. While the expression fold change upon NR2F2 overexpression was comparable to that achieved for NHLH1, this had almost no effect on the cultures. Early transduction

bioRxiv preprint doi: https://doi.org/10.1101/2023.01.27.525898; this version posted January 27, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

showed a very minor, albeit significant increase in mDAN levels, while late transduction for NR2F2 overexpression had no observable effect on mDAN numbers. Interestingly, NR2F2 overexpression could not induce NR2F1 expression (Figure 4C).

Taken together, LBX1 and NHLH1 were able to increase the number of mDAN, with the timing of overexpression being the key to producing this effect for LBX1. NHLH1 showed again the strongest effect on mDAN numbers, as also observed for its KD, and the effect was independent of the developmental timing. Lastly, NR2F2 overexpression did not present any benefit regarding mDAN neurogenesis. This is likely due to the considerable levels of expression by the endogenous and functionally redundant NR2F1.

NHLH1 controls miR-124-3p expression in mDANs

To further characterize the role of NHLH1 in mDAN differentiation, RNA-seq was performed using the samples from the late transduction KD experiments (Figure 4A). NHLH1 KD led to a total of 491 differentially expressed genes (DEGs) (absolute $\log_2$-fold change > 1, FDR < 0.05) (Figure 5A, Supplementary Table 3). Using an Ingenuity Pathway Analysis (IPA) for upstream regulators explaining the transcriptional changes produced by NHLH1 KD (Figure 5B, top 10 upstream regulators based on z-score, p-value < 0.05) predicted miR-124-3p to be downregulated and to belong to the key regulators driving the expression changes (81). Consistently, g:Profiler (82) analysis of upregulated genes from NHLH1 KD predicted miR-124-3p to control many of the genes (Supplementary Table 3). Moreover, miR-124-3p was the only regulator predicted by both methods, IPA and g:Profiler. miR-124 is the most abundant microRNA in the brain with neurogenic properties and has been associated with dopaminergic neurodegeneration in PD (83). Indeed, further exploration of our epigenomic data from mDANs confirmed the three loci coding for miR-124 (MIR124-1 to -3) to be highly accessible with large regions occupied by H3K27ac (Supplementary Figure 3A). Moreover, the primary transcripts of miR-124-1 and miR-124-2 presented an increased expression specifically in mDANs (Supplementary Figure 3A). To validate the predicted reduction in miR-124 expression upon NHLH1 depletion, a TaqMan microRNA assay was used. Importantly, a strong and significant downregulation of mature miR-124-3p could be confirmed (Figure 5C).

Taking advantage of TargetScan (https://www.targetscan.org), a database containing predicted microRNA targets (84), a list of predicted targets for miR-124-3p was filtered for the genes also expressed in our RNA-seq data (Supplementary Table 3). The obtained gene list was compared with the list of upregulated DEGs from NHLH1 KD to determine how many of the genes could be affected by the downregulation of miR-124 (Figure 5D). Strikingly, over 12% of all upregulated genes belonged to primary targets of miR-124, a significantly larger proportion than expected by chance (hypergeometric test, p-value = 3.409e-14). Plotting the 122 targets of miR-124-3p in the KD RNA-seq data confirmed a strong upregulation, as expected (Figure 5E). When the expression of the same targets

bioRxiv preprint doi: https://doi.org/10.1101/2023.01.27.525898; this version posted January 27, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

in smNPCs, across mDAN time course, and astrocytes was plotted, it became clear that most of the microRNA targets are enriched in astrocytes or smNPCs (Figure 5E). Overall, the results highlight miR-124 as a likely mediator of NHLH1-controlled gene regulation, contributing to mDAN specification.

To directly test the contribution of miR-124, a small molecule treatment was used to stimulate miR-124 expression during differentiation and determine whether there would be any benefit for mDAN differentiation. Recent studies have described a new role for ABX464, a quinoline with antiviral properties, in the selective and specific induction of miR-124 (85–87). This molecule binds to the Cap binding complex (CBC) at the 5' end of the primary transcript and promotes the selective splicing of LINC00599, the host gene of miR-124-1. We found that ABX464 could induce miR-124-3p by around two-fold, but the required concentration had a negative effect on cell density in culture (Supplementary Figure 3B). Nevertheless, as small molecules can be powerful tools for their use in biomedicine, and their optimization is possible by chemical modifications, we proceed with the testing ABX464 during mDAN differentiation. ABX464 was added to the cells when they entered the maturation medium on day 9 of differentiation. The molecule was tested under normal differentiation and in cells transduced by shNHLH1.

Interestingly, ABX464 treatment strongly affected both mRNA and microRNA expression (Figure 5F, for original Ct values, please see Supplementary Figure 3D). Moreover, ABX464 also increased U6 snRNA, preventing its use for normalization, while another endogenous microRNA (miR-423-5p) and SCARB2 mRNA were also affected. Hence, we normalized all mRNAs and microRNAs to ACTB expression. Although not significant, ABX464 treatment appeared to increase miR-124-3p levels, while it was downregulated upon NHLH1 KD, as expected. NHLH1 expression was also upregulated upon ABX464 treatment that even reversed NHLH1 repression in NHLH1 KD cells. The rescue of NHLH1 and the increase of miR-124-3p in the cells treated with both ABX464 and shNHLH1 was not enough to rescue the mDAN loss (Figure 5G). However, ABX464 treatment alone significantly increased mDAN numbers in the culture, possibly due to increased NHLH1 and miR-124-3p expression (Figure 5F-G and Supplementary Figure 3D).

LBX1 regulates cholesterol metabolism

LBX1 KD resulted in a total of 2241 DEGs as assessed by RNA-seq (Figure 6A, Supplementary Table 4). Next, pathway enrichment analysis of the identified DEGs was performed using IPA to determine the main processes altered by the LBX1 KD (81). Figure 6B shows the top 10 pathways affected by the KD based on the significance of the enrichment. Cholesterol biosynthesis appeared as one of the main pathways affected by LBX1 and was predicted to be reduced upon LBX1 depletion. Indeed, most genes involved in cholesterol biosynthesis according to the human metabolic reconstruction (RECON) (88), were significantly downregulated upon the KD (Figure 6C). TFs involved in cholesterol metabolism have been previously associated with mDAN development and neurogenesis (60, 62). Consistently, we

bioRxiv preprint doi: https://doi.org/10.1101/2023.01.27.525898; this version posted January 27, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

found the upstream transcriptional regulators of cholesterol metabolism, SREBF1 and SREBF2 (89), to be downregulated by LBX1 KD), although only SREBF1 was reduced more than 2-fold (Figure 6D). The decreased levels were also confirmed by RT-qPCR (Figure 6E). Moreover, upstream regulators of SREBF1, NR1H3 and NR1H2 (LXRα/β) were also found to be deregulated (LXRα/NR1H3 log_2-fold change = -0.515 and LXRβ/NR1H2 log_2-fold change = 0.87). This is consistent with the observed pathway enrichment for LXR activation (Figure 6B). Finally, SREBF1 and NR1H2 appeared as top regulators in our EPIC-DREM analysis (Figure 2A).

These results suggested a novel role of LBX1 in controlling mDAN differentiation via cholesterol metabolism. Therefore, stimulating the cholesterol metabolism pathway should promote mDAN differentiation and could rescue the LBX1 KD effect. GW3965 is a potent and well-described synthetic agonist of NR1H3/2 (LXRα/β). LXR activation can lead to SREBF1 induction and improved mDAN differentiation (60). Hence, neurons were treated with GW3965 under normal differentiation or upon LBX1 KD, starting from day 9 of differentiation, and analyzed on day 15 of differentiation. Although GW3965 treatment induced a high expression of SREBF1 in comparison with control conditions, the treatment did not increase mDAN numbers and did not rescue the LBX1 KD effect (Figure 6F-G and Supplementary Figure 4). Thereby, suggesting that the critical role of LBX1 in mDAN differentiation lies upstream of cholesterol metabolism.

Further exploration of the DEGs altered by LBX1 KD highlighted the downregulation of eIF2 signaling and alterations in the regulation of eIF4 and p70S6K signaling (Figure 6B). Consistently, we found mTOR signaling to be predicted as downregulated upon LBX1 KD (mTOR signaling, -log(p-value) = 3.97, z-score = -0.707, Supplementary Table 4). It has been previously shown that mTOR signaling can control lipid metabolism through SREBF1, and mTOR inhibition leads to eIF4 sequestering by the eukaryotic initiation factor 4E-binding proteins (4E-BPs), impeding translation (90, 91). Therefore, downregulation of mTOR signaling due to the LBX1 KD could explain the observed downregulation in translation and cholesterol biosynthesis. Indeed, we found sirolimus (rapamycin), an mTOR inhibitor, as a chemical drug showing increased predicted activity among upstream regulators in our IPA analysis (activation z-score = 3.031, p-value of overlap = 1.21e-14, Supplementary Table 4). Therefore, the role of mTOR signaling during mDAN differentiation was tested using rapamycin. Neurons were treated with rapamycin as they were treated with GW3965, under normal differentiation and under LBX1 KD. Rapamycin treatment downregulates LBX1 in cells only transduced with shScramble (Figure 6F). In addition, rapamycin treatment was able to downregulate SREBF1 and SREBF2 in a similar way as observed before for LBX1 KD alone. Lastly, rapamycin treatment was also able to decrease the number of mDANs present in cultures (Figure 6G). However, rapamycin did not induce a similar loss of overall cell numbers as observed upon LBX1 KD (Supplementary Figure 4).

bioRxiv preprint doi: https://doi.org/10.1101/2023.01.27.525898; this version posted January 27, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

In summary, LBX1 KD results in an extensive deregulation of the neuronal transcriptome. Cholesterol biosynthesis appeared as one of the main affected pathways regulated by LBX1, possibly through the regulation of SREBF1/2 and NR1H3/2. Stimulating this pathway using GW3965 did not result in increased numbers of mDANs, suggesting that misregulation of the TFs involved in lipid metabolism alone is insufficient to explain the observed impact on mDANs. Perturbation of mTOR signaling provides a possible explanation of the results observed upon LBX1 KD. Indeed, mTOR inhibition was able to reproduce many of the changes induced by LBX1 KD.

Discussion

Here we present a multi-omics data integration approach to discover novel TFs controlling mDAN differentiation. Although previous studies have used similar transcriptomic and epigenomic profiling of mDANs (16, 18), our assessment of human mDAN differentiation is cell-type-specific and time-resolved. Together with the *de novo* identification of key regulators of mDANs, we also provide a functional validation and characterization of the newly identified TFs.

The EPIC-DREM pipeline here offers an unbiased and data-driven tool for identifying TFs controlling processes across time based on transcriptomic and epigenomic data. The workflow can also be used for processes other than differentiation such as drug treatments and disease progression. This study used EPIC-DREM for the first time in conjunction with ATAC-seq data. Previously, EPIC-DREM operated on ChIP-seq datasets (8) and some of the tools integrated into this pipeline have been optimized for the use of other methods to detect accessible regions such as DNaseI-seq and NOMe-seq (22). EPIC-DREM can be applied to an ample range of epigenomic datasets that together with transcriptomic data can help to uncover key regulators and biological insights.

For the first time, we identified TFs controlled by SEs during mDAN differentiation. A total of 49 TFs were found to be under the control of SEs between day 30 and day 50 of differentiation (Supplementary Table 2). Of these, 17 were among the top TFs predicted by EPIC-DREM (Figure 3B). Among the remaining 32 TFs that did not overlap with EPIC-DREM, we found additional promising candidates. For example, ZFHX4 and CSRNP3 are highly expressed in iPSC-derived mDANs (data not shown) and in developing human mDANs *in vivo* (http://linnarssonlab.org/ventralmidbrain/, (13). These TFs were not selected in our analysis because their motifs are not known, namely, their TFBS in the accessible regions cannot be determined through footprinting and the consequent prediction as regulators was not possible. This highlights a limitation of the approach used here, but simultaneously opens new research avenues for previously less studied TFs that seem to play an important role in neurogenesis.

bioRxiv preprint doi: https://doi.org/10.1101/2023.01.27.525898; this version posted January 27, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Gene perturbation experiments were performed to validate the identified candidates: HOXB2, LBX1, NHLH1, NR2F1, and NR2F2. After TF depletion during differentiation, it was found that LBX1, NHLH1 and NR2F1/2 presented a stronger phenotype than a well-established and characterized TF involved in mDAN neurogenesis: LMX1B. Further characterization of the role of these TFs revealed that only LBX1 and NHLH1 inductions during differentiation were able to increase mDAN numbers, while induction time was critical for LBX1.

Overall, our results are in line with previous findings. Loss of function in the LBX1 protein, specifically in the protein-protein interaction domain, in addition to being lethal during development due to cardiac abnormalities in mice, has been shown to impair neurogenesis during the development of the neural tube (92). This finding can be correlated with what was observed during LBX1 KD in mDANs, where differentiating cells always struggle to survive. Moreover, the findings on the development of spinal cord interneurons, RTN, and dorsal neurons from the hindbrain have linked LBX1's function to cell fate decisions (67, 68, 92). Here we have added a role for LBX1 in mDAN differentiation.

On the other hand, NHLH1 loss of function has been shown to be lethal in mice, but only in adult life. However, if accompanied by loss of NHLH2, mice die already at birth (72, 93). Furthermore, NHLH1 has been shown to control neurogenesis by regulating the expression of neuronal-specific genes during *Xenopus* development (94). Our results support an important role for NHLH1 in controlling mDAN differentiation, consistent with previous findings.

After seeing the relevance of LBX1 and NHLH1 during differentiation, transcriptomic analysis after TF depletion was performed to determine the main processes controlled by these TFs. LBX1 KD showed a clear downregulation of cholesterol biosynthesis-related genes together with SREBF1 and SREBF2. As SREBF1 is at the same time regulated by LXRα/β, and all of them have been previously related to mDAN neurogenesis, it was decided to stimulate this pathway using GW3965. However, the treatment with this molecule was not sufficient to rescue the LBX1 KD effect or increase mDAN numbers in the differentiation protocol used for this study. It is possible that for GW3965 treatment to be effective, it should be either added from the beginning of differentiation or cells should be treated for a more extended period. We refrained from these experiments because the main effects of LBX1 KD were studied late in differentiation during the first 5 days of maturation.

On the other hand, as stimulation of TFs involved in cholesterol metabolism did not rescue the LBX1 KD effect, we focused on another enriched pathway, eIF signaling. This pathway is involved in translation, a process also downregulated by LBX1 KD. Exploration of the data for a common origin of cholesterol biosynthesis and translation downregulation revealed mTOR signaling inhibition as a possible explanation and in line with previous findings (90, 91, 95–97). mTOR signaling was recently shown to play different roles in mDAN biology, controlling their morphology, dopamine release, and

bioRxiv preprint doi: https://doi.org/10.1101/2023.01.27.525898; this version posted January 27, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

electrophysiology (98). Indeed, inhibiting mTOR signaling using rapamycin recapitulated most of the features observed by LBX1 KD.

Transcriptional profiling and more detailed analysis of neurons upon NHLH1 KD revealed a downregulation of the mature form of miR-124-3p. This relationship between NHLH1 and miR-124 has not been described before. miR-124 is a potent neurogenic microRNA that has been found to be downregulated in PD patients (99) and shown to be protective in PD animal models (100, 101). After determining the capacity of ABX464 in inducing miR-124 levels in neurons, the molecule was used to determine its effect during mDAN differentiation and as an attempt to rescue the NHLH1 KD. It was found that for inducing at least a 2-fold increase of miR-124-3p, the concentration needed was causing neurotoxicity, possibly due to the already high expression of miR-124 in the neuronal lineage. Independently of that, ABX464 treatment was able to increase mDAN numbers and to induce NHLH1 expression. Therefore, NHLH1 and miR-124 seem to be part of a positive feedback loop. To unveil potential applications of ABX464, its neurotoxicity should be reduced for example by chemical modification of the molecule. ABX464 is already in clinical trials for the treatment of rheumatoid arthritis and ulcerative colitis due to its anti-inflammatory potential (87, 102). Namely, ABX464 safety and efficacy has been already evaluated and the repurposing of its use for PD could be taken into consideration to increase miR-124-3p levels as a neuroprotective treatment.

Data availability

The source RNA-seq, ATAC-seq and ChIP-seq fastq files have been deposited at https://ega-archive.org/, under the accession number EGAD00001009288. Additional intermediate files such as RNA-seq counts, ATAC-seq peaks and bigwig files, and ChIP-seq H3K27ac SEs and bigwig files, can be provided upon request.

Analysis code can be found in the following repository:
https://gitlab.lcsb.uni.lu/borja.gomezramos/gomezramos_et_al_2023/-/tree/main/

Funding

This work was supported by the Luxembourg National Research Fund within the National Centre of Excellence in Research on Parkinson's Disease (NCER-PD; FNR/NCER13/BM/11264123), and the PEARL program (FNR/P13/6682797 to R.K.), as well as the PARK-QC DTU (PRIDE17/12244779/PARK-QC).; L.S., B.G.R. and J.O. have received funding from Fondation du Pélican de Mie et Pierre Hippert-Faber and Luxembourg Rotary Foundation. The genome-editing platform in Lübeck is supported by the DFG (FOR 2488 to A.R.).

bioRxiv preprint doi: https://doi.org/10.1101/2023.01.27.525898; this version posted January 27, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Conflict of Interest

The authors declare no conflict of interest

Acknowledgments

The authors would like to thank Marie Catillon for support with experiments. The computational analysis presented in this paper were carried out using the HPC facilities of the University of Luxembourg.

Author Contributions

B.G.R., J.O., M.S., T.S., R.K. and L.S. conceived the project and designed the experiments. B.G.R. performed most of the experiments and bioinformatic analysis. J.O. established and performed ATAC-seq protocol and performed bioinformatic analysis. N.d.L., Ro.K., and M.S. established the EPIC-DREM pipeline and performed the analysis. A.G. designed the RNA-seq pipeline and performed bioinformatic analysis. A.R. designed and generated the iPSC reporter line in consultation with C.K.. R.H. prepared RNA-seq libraries and performed the sequencing. F.M. performed rapamycin treatments. Ro.K., M.S., T.S., R.K., and L.S. provided input on data interpretation and revision of the study. L.S. supervised the project. B.G.R. and L.S. wrote the manuscript. All authors revised and approved the final manuscript.

Supplementary Information

The Supplementary Information contains four supplementary figures and four supplementary tables.

bioRxiv preprint doi: https://doi.org/10.1101/2023.01.27.525898; this version posted January 27, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

References

1. Levine, M. (2010) Transcriptional Enhancers in Animal Development and Evolution. *Curr. Biol.*, **20**, R754–R763.
2. Spitz, F. and Furlong, E.E.M. (2012) Transcription factors: From enhancer binding to developmental control. *Nat. Rev. Genet.*, **13**, 613–626.
3. Whyte, W.A., Orlando, D.A., Hnisz, D., Abraham, B.J., Lin, C.Y., Kagey, M.H., Rahl, P.B., Lee, T.I. and Young, R.A. (2013) Master transcription factors and mediator establish super-enhancers at key cell identity genes. *Cell*, **153**, 307–319.
4. Hnisz, D., Abraham, B.J., Lee, T.I., Lau, A., Saint-André, V., Sigova, A. a, Hoke, H. a and Young, R. a (2013) Super-enhancers in the control of cell identity and disease. *Cell*, **155**, 934–47.
5. Parker, S.C.J., Stitzel, M.L., Taylor, D.L., Orozco, J.M., Erdos, M.R., Akiyama, J.A., van Bueren, K.L., Chines, P.S., Narisu, N., Black, B.L., et al. (2013) Chromatin stretch enhancer states drive cell-specific gene regulation and harbor human disease risk variants. *Proc. Natl. Acad. Sci. U. S. A.*, **110**, 17921–6.
6. Duren, Z., Chen, X., Xin, J., Wang, Y. and Wong, W.H. (2020) Time course regulatory analysis based on paired expression and chromatin accessibility data. *Genome Res.*, **30**, 622–634.
7. Rauch, A., Haakonsson, A.K., Madsen, J.G.S., Larsen, M., Forss, I., Madsen, M.R., Van Hauwaert, E.L., Wiwie, C., Jespersen, N.Z., Tencerova, M., et al. (2019) Osteogenesis depends on commissioning of a network of stem cell transcription factors that act as repressors of adipogenesis. *Nat. Genet.*, **51**, 716–727.
8. Gérard, D., Schmidt, F., Ginolhac, A., Schmitz, M., Halder, R., Ebert, P., Schulz, M.H., Sauter, T. and Sinkkonen, L. (2018) Temporal enhancer profiling of parallel lineages identifies AHR and GLIS1 as regulators of mesenchymal multipotency. *Nucleic Acids Res.*, 10.1093/nar/gky1240.
9. Volkow, N.D., Wise, R.A. and Baler, R. (2017) The dopamine motive system: implications for drug and food addiction. *Nat. Rev. Neurosci.*, **18**, 741–752.
10. Grace, A.A. and Gomes, F. V. (2019) The Circuitry of Dopamine System Regulation and its Disruption in Schizophrenia: Insights Into Treatment and Prevention. *Schizophr. Bull.*, **45**, 148–157.
11. Okano, H. and Morimoto, S. (2022) iPSC-based disease modeling and drug discovery in cardinal neurodegenerative disorders. *Cell Stem Cell*, **29**, 189–208.
12. Fernandes, H.J.R., Patikas, N., Foskolou, S., Field, S.F., Park, J.E., Byrne, M.L., Bassett, A.R. and Metzakopian, E. (2020) Single-Cell Transcriptomics of Parkinson's Disease Human In Vitro Models Reveals Dopamine Neuron-Specific Stress Responses. *Cell Rep.*, **33**, 108263.
13. La Manno, G., Gyllborg, D., Codeluppi, S., Nishimura, K., Salto, C., Zeisel, A., Borm, L.E., Stott, S.R.W., Toledo, E.M., Villaescusa, J.C., et al. (2016) Molecular Diversity of Midbrain Development in Mouse, Human, and Stem Cells. *Cell*, **167**, 566-580.e19.

bioRxiv preprint doi: https://doi.org/10.1101/2023.01.27.525898; this version posted January 27, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

14. Arenas, E., Denham, M. and Villaescusa, J.C. (2015) How to make a midbrain dopaminergic neuron. *Development*, **142**, 1918–1936.

15. Ásgrímsdóttir, E.S. and Arenas, E. (2020) Midbrain Dopaminergic Neuron Development at the Single Cell Level: In vivo and in Stem Cells. *Front. Cell Dev. Biol.*, **8**, 1–20.

16. Xia, N., Fang, F., Zhang, P., Cui, J., Tep-Cullison, C., Hamerley, T., Lee, H.J., Palmer, T., Bothner, B., Lee, J.H., et al. (2017) A Knockin Reporter Allows Purification and Characterization of mDA Neurons from Heterogeneous Populations. *Cell Rep.*, **18**, 2533–2546.

17. Fernández-Santiago, R., Carballo-Carbajal, I., Castellano, G., Torrent, R., Richaud, Y., Sánchez-Danés, A., Vilarrasa-Blasi, R., Sánchez-Pla, A., Mosquera, J.L., Soriano, J., et al. (2015) Aberrant epigenome in iPSC-derived dopaminergic neurons from Parkinson's disease patients. *EMBO Mol. Med.*, **7**, 1529–1546.

18. Meléndez-Ramírez, C., Cuevas-Diaz Duran, R., Barrios-García, T., Giacoman-Lozano, M., López-Ornelas, A., Herrera-Gamboa, J., Estudillo, E., Soto-Reyes, E., Velasco, I. and Treviño, V. (2021) Dynamic landscape of chromatin accessibility and transcriptomic changes during differentiation of human embryonic stem cells into dopaminergic neurons. *Sci. Reports 2021 111*, **11**, 1–18.

19. Rakovic, A., Voß, D., Vulinovic, F., Meier, B., Hellberg, A.K., Nau, C., Klein, C. and Leipold, E. (2022) Electrophysiological Properties of Induced Pluripotent Stem Cell-Derived Midbrain Dopaminergic Neurons Correlate With Expression of Tyrosine Hydroxylase. *Front. Cell. Neurosci.*, **16**, 121.

20. Sandor, C., Robertson, P., Lang, C., Heger, A., Booth, H., Vowles, J., Witty, L., Bowden, R., Hu, M., Cowley, S.A., et al. (2017) Transcriptomic profiling of purified patient-derived dopamine neurons identifies convergent perturbations and therapeutics for Parkinson's disease. *Hum. Mol. Genet.*, **26**, 552–566.

21. Li, Z., Schulz, M.H., Zenke, M. and Costa, I.G. (2018) Identification of Transcription Factor Binding Sites using ATAC-seq. *Biorxiv*, 10.1101/362863.

22. Schmidt, F., Gasparoni, N., Gasparoni, G., Gianmoena, K., Cadenas, C., Polansky, J.K., Ebert, P., Nordstrom, K., Barann, M., Sinha, A., et al. (2017) Combining transcription factor binding affinities with open-chromatin data for accurate gene expression prediction. *Nucleic Acids Res.*, **45**, 54–66.

23. Schulz, M.H., Devanny, W.E., Gitter, A., Zhong, S., Ernst, J. and Bar-Joseph, Z. (2012) DREM 2.0: Improved reconstruction of dynamic regulatory networks from time-series expression data. *BMC Syst. Biol.*, **6**.

24. Zanon, A., Kalvakuri, S., Rakovic, A., Foco, L., Guida, M., Schwienbacher, C., Serafin, A., Rudolph, F., Trilck, M., Grünewald, A., et al. (2019) Corrigendum: SLP-2 interacts with Parkin in mitochondria and prevents mitochondrial dysfunction in Parkin-deficient human iPSC-derived neurons and Drosophila. *Hum. Mol. Genet.*, **28**, 1225.

bioRxiv preprint doi: https://doi.org/10.1101/2023.01.27.525898; this version posted January 27, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

25. Hanss, Z., Larsen, S.B., Antony, P., Mencke, P., Massart, F., Jarazo, J., Schwamborn, J.C., Barbuti, P.A., Mellick, G.D. and Krüger, R. (2021) Mitochondrial and Clearance Impairment in p.D620N VPS35 Patient-Derived Neurons. *Mov. Disord.*, **36**, 704.
26. Palm, T., Bolognin, S., Meiser, J., Nickels, S., Träger, C., Meilenbrock, R.L., Brockhaus, J., Schreitmüller, M., Missler, M. and Schwamborn, J.C. (2015) Rapid and robust generation of long-term self-renewing human neural stem cells with the ability to generate mature astroglia. *Sci. Rep.*, **5**, 1–16.
27. Corces, M.R., Trevino, A.E., Hamilton, E.G., Greenside, P.G., Sinnott-Armstrong, N.A., Vesuna, S., Satpathy, A.T., Rubin, A.J., Montine, K.S., Wu, B., et al. (2017) An improved ATAC-seq protocol reduces background and enables interrogation of frozen tissues. *Nat. Methods*, **14**, 959–962.
28. Buenrostro, J.D., Giresi, P.G., Zaba, L.C., Chang, H.Y. and Greenleaf, W.J. (2013) Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nat. Methods*, **10**, 1213–1218.
29. Walter, J., Bolognin, S., Poovathingal, S.K., Magni, S., Gérard, D., Antony, P.M.A., Nickels, S.L., Salamanca, L., Berger, E., Smits, L.M., et al. (2021) The Parkinson's-disease-associated mutation LRRK2-G2019S alters dopaminergic differentiation dynamics via NR2F1. *Cell Rep.*, **37**.
30. Ewels, P., Magnusson, M., Lundin, S. and Käller, M. (2016) MultiQC: summarize analysis results for multiple tools and samples in a single report. *Bioinformatics*, **32**, 3047–3048.
31. Schubert, M., Lindgreen, S. and Orlando, L. (2016) AdapterRemoval v2: rapid adapter trimming, identification, and read merging. *BMC Res. Notes*, **9**.
32. Schubert, M., Ermini, L., Sarkissian, C. Der, Jónsson, H., Ginolhac, A., Schaefer, R., Martin, M.D., Fernández, R., Kircher, M., McCue, M., et al. (2014) Characterization of ancient and modern genomes by SNP detection and phylogenomic and metagenomic analysis using PALEOMIX. *Nat. Protoc.*, **9**, 1056–1082.
33. Kopylova, E., Noé, L. and Touzet, H. (2012) SortMeRNA: fast and accurate filtering of ribosomal RNAs in metatranscriptomic data. *Bioinformatics*, **28**, 3211–3217.
34. Dobin, A., Davis, C.A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M. and Gingeras, T.R. (2013) STAR: ultrafast universal RNA-seq aligner. *Bioinformatics*, **29**, 15–21.
35. Danecek, P., Bonfield, J.K., Liddle, J., Marshall, J., Ohan, V., Pollard, M.O., Whitwham, A., Keane, T., McCarthy, S.A., Davies, R.M., et al. (2021) Twelve years of SAMtools and BCFtools. *Gigascience*, **10**, 1–4.
36. Liao, Y., Smyth, G.K. and Shi, W. (2019) The R package Rsubread is easier, faster, cheaper and better for alignment and quantification of RNA sequencing reads. *Nucleic Acids Res.*, **47**, e47–e47.

bioRxiv preprint doi: https://doi.org/10.1101/2023.01.27.525898; this version posted January 27, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

37. Love, M.I., Huber, W. and Anders, S. (2014) Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.*, **15**, 550.
38. Köster, J., Mölder, F., Jablonski, K.P., Letcher, B., Hall, M.B., Tomkins-Tinch, C.H., Sochat, V., Forster, J., Lee, S., Twardziok, S.O., et al. (2021) Sustainable data analysis with Snakemake. *F1000Research*, **10**.
39. Wickham, H. (2016) ggplot2. 10.1007/978-3-319-24277-4.
40. Wingett, S.W. and Andrews, S. (2018) FastQ Screen: A tool for multi-genome mapping and quality control. *F1000Research*, **7**, 1338.
41. Zhu, A., Ibrahim, J.G. and Love, M.I. (2019) Heavy-tailed prior distributions for sequence count data: removing the noise and preserving large differences. *Bioinformatics*, **35**, 2084–2092.
42. Li, H. and Durbin, R. (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*, **25**, 1754–1760.
43. Heinz, S., Benner, C., Spann, N., Bertolino, E., Lin, Y.C., Laslo, P., Cheng, J.X., Murre, C., Singh, H. and Glass, C.K. (2010) Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. *Mol. Cell*, **38**, 576–589.
44. Ding, J., Hagood, J.S., Ambalavanan, N., Kaminski, N. and Bar-Joseph, Z. (2018) iDREM: Interactive visualization of dynamic regulatory networks. *PLOS Comput. Biol.*, **14**, e1006019.
45. Wu, T., Hu, E., Xu, S., Chen, M., Guo, P., Dai, Z., Feng, T., Zhou, L., Tang, W., Zhan, L., et al. (2021) clusterProfiler 4.0: A universal enrichment tool for interpreting omics data. *Innov. (Cambridge*, **2**.
46. Reinhardt, P., Glatza, M., Hemmer, K., Tsytsyura, Y., Thiel, C.S., Höing, S., Moritz, S., Parga, J.A., Wagner, L., Bruder, J.M., et al. (2013) Derivation and Expansion Using Only Small Molecules of Human Neural Progenitors for Neurodegenerative Disease Modeling. *PLoS One*, **8**.
47. Nakamura, H. and Watanabe, Y. (2005) Isthmus organizer and regionalization of the mesencephalon and metencephalon. *Int. J. Dev. Biol.*, **49**, 231–235.
48. Yang, Z. and Wang, K.K.W. (2015) Glial fibrillary acidic protein: from intermediate filament assembly and gliosis to neurobiomarker. *Trends Neurosci.*, **38**, 364–374.
49. Anderegg, A., Poulin, J.-F. and Awatramani, R. (2015) Molecular heterogeneity of midbrain dopaminergic neurons – Moving toward single cell resolution. *FEBS Lett.*, **589**, 3714–3726.
50. Schmidt, F., Kern, F., Ebert, P., Baumgarten, N. and Schulz, M.H. (2019) TEPIC 2—an extended framework for transcription factor binding prediction and integrative epigenomic analysis. *Bioinformatics*, **35**, 1608–1609.
51. Hermanson, E., Joseph, B., Castro, D., Lindqvist, E., Aarnisalo, P., Wallén, Å., Benoit, G., Hengerer, B., Olson, L. and Perlmann, T. (2003) Nurr1 regulates dopamine synthesis and storage in MN9D dopamine cells. *Exp. Cell Res.*, **288**, 324–334.

bioRxiv preprint doi: https://doi.org/10.1101/2023.01.27.525898; this version posted January 27, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

52. Veenvliet, J. V., dos Santos, M.T.M.A., Kouwenhoven, W.M., Von Oerthel, L., Lim, J.L., van der Linden, A.J.A., Koerkamp, M.J.A.G., Holstege, F.C.P. and Smidt, M.P. (2013) Specification of dopaminergic subsets involves interplay of En1 and Pitx3. *Development*, **140**, 3373–3384.

53. Sherf, O., Zolotov, L.N., Liser, K., Tilleman, H., Jovanovic, V.M., Zega, K., Jukic, M.M. and Brodski, C. (2015) Otx2 Requires Lmx1b to Control the Development of Mesodiencephalic Dopaminergic Neurons. *PLoS One*, **10**.

54. Hoekstra, E.J., von Oerthel, L., van der Linden, A.J.A., Schellevis, R.D., Scheppink, G., Holstege, F.C.P., Groot-Koerkamp, M.J., van der Heide, L.P. and Smidt, M.P. (2013) Lmx1a is an activator of Rgs4 and Grb10 and is responsible for the correct specification of rostral and medial mdDA neurons. *Eur. J. Neurosci.*, **37**, 23–32.

55. Iwafuchi-Doi, M. and Zaret, K.S. (2016) Cell fate control by pioneer transcription factors. *Development*, **143**, 1833–1837.

56. Herdy, J., Schafer, S., Kim, Y., Ansari, Z., Zangwill, D., Ku, M., Paquola, A., Lee, H., Mertens, J. and Gage, F.H. (2019) Chemical modulation of transcriptionally enriched signaling pathways to optimize the conversion of fibroblasts into neurons. *Elife*, **8**.

57. Smith, D.K., Yang, J., Liu, M.L. and Zhang, C.L. (2016) Small Molecules Modulate Chromatin Accessibility to Promote NEUROG2-Mediated Fibroblast-to-Neuron Reprogramming. *Stem cell reports*, **7**, 955–969.

58. Andersson, E., Tryggvason, U., Deng, Q., Friling, S., Alekseenko, Z., Robert, B., Perlmann, T. and Ericson, J. (2006) Identification of intrinsic determinants of midbrain dopamine neurons. *Cell*, **124**, 393–405.

59. Martinez-Barbera, J.P., Signore, M., Pilo Boyl, P., Puelles, E., Acampora, D., Gogoi, R., Schubert, F., Lumsden, A. and Simeone, A. (2001) Regionalisation of anterior neuroectoderm and its competence in responding to forebrain and midbrain inducing activities depend on mutual antagonism between OTX2 and GBX2. *Development*, **128**, 4789–4800.

60. Toledo, E.M., Yang, S., Gyllborg, D., van Wijk, K.E., Sinha, I., Varas-Godoy, M., Grigsby, C.L., Lönnerberg, P., Islam, S., Steffensen, K.R., et al. (2020) Srebf1 Controls Midbrain Dopaminergic Neurogenesis. *Cell Rep.*, **31**.

61. Villaescusa, J.C., Li, B., Toledo, E.M., Rivetti di Val Cervo, P., Yang, S., Stott, S.R., Kaiser, K., Islam, S., Gyllborg, D., Laguna-Goya, R., et al. (2016) A PBX1 transcriptional network controls dopaminergic neuron development and is impaired in Parkinson's disease. *EMBO J.*, **35**, 1963–1978.

62. Sacchetti, P., Sousa, K.M., Hall, A.C., Liste, I., Steffensen, K.R., Theofilopoulos, S., Parish, C.L., Hazenberg, C., Richter, L.Ä., Hovatta, O., et al. (2009) Liver X Receptors and Oxysterols Promote Ventral Midbrain Neurogenesis In Vivo and in Human Embryonic Stem Cells. *Cell Stem Cell*, **5**, 409–419.

63. Galhardo, M., Berninger, P., Nguyen, T.P., Sauter, T. and Sinkkonen, L. (2015) Cell type-

bioRxiv preprint doi: https://doi.org/10.1101/2023.01.27.525898; this version posted January 27, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

selective disease-association of genes under high regulatory load. *Nucleic Acids Res.*, **43**, 8839–8855.

64. Lyu, S., Xing, H., Liu, Y., Girdhar, P., Zhang, K., Yokoi, F., Xiao, R. and Li, Y. (2020) Deficiency of Meis1, a transcriptional regulator, in mice and worms: Neurochemical and behavioral characterizations with implications in the restless legs syndrome. *J. Neurochem.*, **155**, 522–537.
65. Mesman, S., Wever, I. and Smidt, M.P. (2021) Tcf4 Is Involved in Subset Specification of Mesodiencephalic Dopaminergic Neurons. *Biomedicines*, **9**.
66. Parker, H.J., De Kumar, B., Green, S.A., Prummel, K.D., Hess, C., Kaufman, C.K., Mosimann, C., Wiedemann, L.M., Bronner, M.E. and Krumlauf, R. (2019) A Hox-TALE regulatory circuit for neural crest patterning is conserved across vertebrates. *Nat. Commun. 2019 101*, **10**, 1–15.
67. Gross, M.K., Dottori, M. and Goulding, M. (2002) Lbx1 specifies somatosensory association interneurons in the dorsal spinal cord. *Neuron*, **34**, 535–549.
68. Hernandez-Miranda, L.R., Ibrahim, D.M., Ruffault, P.L., Larrosa, M., Balueva, K., Müller, T., De Weerd, W., Stolte-Dijkstra, I., Hostra, R.M.W., Brunet, J.F., et al. (2018) Mutation in LBX1/Lbx1 precludes transcription factor cooperativity and causes congenital hypoventilation in humans and mice. *Proc. Natl. Acad. Sci. U. S. A.*, **115**, 13021–13026.
69. Huang, M., Huang, T., Xiang, Y., Xie, Z., Chen, Y., Yan, R., Xu, J. and Cheng, L. (2008) Ptf1a, Lbx1 and Pax2 coordinate glycinergic and peptidergic transmitter phenotypes in dorsal spinal inhibitory neurons. *Dev. Biol.*, **322**, 394–405.
70. Ratié, L., Ware, M., Jagline, H., David, V. and Dupé, V. (2014) Dynamic expression of Notch-dependent neurogenic markers in the chick embryonic nervous system. *Front. Neuroanat.*, **8**, 158.
71. De Smaele, E., Fragomeli, C., Ferretti, E., Pelloni, M., Po, A., Canettieri, G., Coni, S., Di Marcotullio, L., Greco, A., Moretti, M., et al. (2008) An Integrated Approach Identifies Nhlh1 and Insm1 as Sonic Hedgehog-regulated Genes in Developing Cerebellum and Medulloblastoma. *Neoplasia*, **10**, 89-IN36.
72. Krüger, M., Ruschke, K. and Braun, T. (2004) NSCL-1 and NSCL-2 synergistically determine the fate of GnRH-1 neurons and control necdin gene expression. *EMBO J.*, **23**, 4353–4364.
73. Schmid, T., Krüger, M. and Braun, T. (2007) NSCL-1 and -2 control the formation of precerebellar nuclei by orchestrating the migration of neuronal precursor cells. *J. Neurochem.*, **102**, 2061–2072.
74. Naka, H., Nakamura, S., Shimazaki, T. and Okano, H. (2008) Requirement for COUP-TFI and II in the temporal specification of neural stem cells in CNS development. *Nat. Neurosci. 2008 119*, **11**, 1014–1023.
75. Bonzano, S., Crisci, I., Podlesny-Drabiniok, A., Rolando, C., Krezel, W., Studer, M. and De Marchis, S. (2018) Neuron-Astroglia Cell Fate Decision in the Adult Mouse Hippocampal

bioRxiv preprint doi: https://doi.org/10.1101/2023.01.27.525898; this version posted January 27, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Neurogenic Niche Is Cell-Intrinsically Controlled by COUP-TFI In Vivo. *Cell Rep.*, **24**, 329–341.

76. Teratani-Ota, Y., Yamamizu, K., Piao, Y., Sharova, L., Amano, M., Yu, H., Schlessinger, D., Ko, M.S.H. and Sharov, A.A. (2016) Induction of specific neuron types by overexpression of single transcription factors. *Vitr. Cell. Dev. Biol. - Anim.*, **52**, 961–973.
77. Zhang, K., Yu, F., Zhu, J., Han, S., Chen, J., Wu, X., Chen, Y., Shen, T., Liao, J., Guo, W., et al. (2020) Imbalance of Excitatory/Inhibitory Neuron Differentiation in Neurodevelopmental Disorders with an NR2F1 Point Mutation. *Cell Rep.*, **31**, 107521.
78. Memic, F., Knoflach, V., Morarach, K., Sadler, R., Laranjeira, C., Hjerling-Leffler, J., Sundström, E., Pachnis, V. and Marklund, U. (2018) Transcription and Signaling Regulators in Developing Neuronal Subtypes of Mouse and Human Enteric Nervous System. *Gastroenterology*, **154**, 624–636.
79. Potzner, M.R., Tsarovina, K., Binder, E., Penzo-Méndez, A., Lefebvre, V., Rohrer, H., Wegner, M. and Sock, E. (2010) Sequential requirement of Sox4 and Sox11 during development of the sympathetic nervous system. *Development*, **137**, 775–784.
80. Wever, I., Largo-Barrientos, P., Hoekstra, E.J. and Smidt, M.P. (2019) Lmx1b Influences Correct Post-mitotic Coding of Mesodiencephalic Dopaminergic Neurons. *Front. Mol. Neurosci.*, **12**.
81. Krämer, A., Green, J., Pollard Jr, J. and Tugendreich, S. (2014) Causal analysis approaches in Ingenuity Pathway Analysis. *Bioinformatics*, **30**, 523–530.
82. Raudvere, U., Kolberg, L., Kuzmin, I., Arak, T., Adler, P., Peterson, H. and Vilo, J. (2019) g:Profiler: a web server for functional enrichment analysis and conversions of gene lists (2019 update). *Nucleic Acids Res.*, **47**, W191–W198.
83. Angelopoulou, E., Paudel, Y.N. and Piperi, C. (2019) miR-124 and Parkinson's disease: A biomarker with therapeutic potential. *Pharmacol. Res.*, **150**.
84. Agarwal, V., Bell, G.W., Nam, J.-W. and Bartel, D.P. (2015) Predicting effective microRNA target sites in mammalian mRNAs. *Elife*, **4**, e05005.
85. Tazi, J., Begon-Pescia, C., Campos, N., Apolit, C., Garcel, A. and Scherrer, D. (2021) Specific and selective induction of miR-124 in immune cells by the quinoline ABX464: a transformative therapy for inflammatory diseases. *Drug Discov. Today*, **26**, 1030–1039.
86. Vautrin, A., Manchon, L., Garcel, A., Campos, N., Lapasset, L., Laaref, A.M., Bruno, R., Gislard, M., Dubois, E., Scherrer, D., et al. (2019) Both anti-inflammatory and antiviral properties of novel drug candidate ABX464 are mediated by modulation of RNA splicing. *Sci. Rep.*, **9**.
87. Daien, C., Krogulec, M., Gineste, P., Steens, J.-M., Desroys du Roure, L., Biguenet, S., Scherrer, D., Santo, J., Ehrlich, H. and Durez, P. (2022) Safety and efficacy of the miR-124 upregulator ABX464 (obefazimod, 50 and 100 mg per day) in patients with active rheumatoid arthritis and inadequate response to methotrexate and/or anti-TNFα therapy: a placebo-controlled phase II study. *Ann. Rheum. Dis.*, **81**, annrheumdis-2022-222228.

bioRxiv preprint doi: https://doi.org/10.1101/2023.01.27.525898; this version posted January 27, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

88. Brunk, E., Sahoo, S., Zielinski, D.C., Altunkaya, A., Dräger, A., Mih, N., Gatto, F., Nilsson, A., Preciat Gonzalez, G.A., Aurich, M.K., et al. (2018) Recon3D enables a three-dimensional view of gene variation in human metabolism. *Nat. Biotechnol.*, **36**, 272–281.

89. Horton, J.D., Goldstein, J.L. and Brown, M.S. (2002) SREBPs: activators of the complete program of cholesterol and fatty acid synthesis in the liver. *J. Clin. Invest.*, **109**, 1125–1131.

90. Wang, H., Liu, Y., Ding, J., Huang, Y., Liu, J., Liu, N., Ao, Y., Hong, Y., Wang, L., Zhang, L., et al. (2019) Targeting mTOR suppressed colon cancer growth through 4EBP1/eIF4E/PUMA pathway. *Cancer Gene Ther. 2019 276*, **27**, 448–460.

91. Liu, G.Y. and Sabatini, D.M. (2020) mTOR at the nexus of nutrition, growth, ageing and disease. *Nat. Rev. Mol. Cell Biol. 2020 214*, **21**, 183–203.

92. Decourtye, L., McCallum-Loudeac, J.A., Zellhuber-McMillan, S., Young, E., Sircombe, K.J. and Wilson, M.J. (2022) Characterization of a novel Lbx1 mouse loss of function strain. *Differentiation.*, **123**, 30–41.

93. Cogliati, T., Good, D.J., Haigney, M., Delgado-Romero, P., Eckhaus, M.A., Koch, W.J. and Kirsch, I.R. (2002) Predisposition to Arrhythmia and Autonomic Dysfunction in Nhlh1-Deficient Mice. *Mol. Cell. Biol.*, **22**, 4977.

94. Bao, J., Talmage, D.A., Role, L.W. and Gautier, J. (2000) Regulation of neurogenesis by interactions between HEN1 and neuronal LMO proteins. *Development*, **127**, 425–435.

95. Wang, B.T., Ducker, G.S., Barczak, A.J., Barbeau, R., Erle, D.J. and Shokat, K.M. (2011) The mammalian target of rapamycin regulates cholesterol biosynthetic gene expression and exhibits a rapamycin-resistant transcriptional profile. *Proc. Natl. Acad. Sci. U. S. A.*, **108**, 15201–15206.

96. Fonseca, B.D., Smith, E.M., Yelle, N., Alain, T., Bushell, M. and Pause, A. (2014) The ever-evolving role of mTOR in translation. *Semin. Cell Dev. Biol.*, **36**, 102–112.

97. Yecies, J.L., Zhang, H.H., Menon, S., Liu, S., Yecies, D., Lipovsky, A.I., Gorgun, C., Kwiatkowski, D.J., Hotamisligil, G.S., Lee, C.H., et al. (2011) Akt stimulates hepatic SREBP1c and lipogenesis through parallel mTORC1-dependent and independent pathways. *Cell Metab.*, **14**, 21–32.

98. Kosillo, P., Ahmed, K.M., Aisenberg, E.E., Karalis, V., Roberts, B.M., Cragg, S.J. and Bateup, H.S. (2022) Dopamine neuron morphology and output are differentially controlled by mTORC1 and mTORC2. *Elife*, **11**.

99. Yang, Y., Li, Y., Yang, H., Guo, J. and Li, N. (2021) Circulating MicroRNAs and Long Non-coding RNAs as Potential Diagnostic Biomarkers for Parkinson's Disease. *Front. Mol. Neurosci.*, **14**, 28.

100. Zhang, F., Yao, Y., Miao, N., Wang, N., Xu, X. and Yang, C. (2022) Neuroprotective effects of microRNA 124 in Parkinson's disease mice. *Arch. Gerontol. Geriatr.*, **99**.

101. Saraiva, C., Paiva, J., Santos, T., Ferreira, L. and Bernardino, L. (2016) MicroRNA-124 loaded nanoparticles enhance brain repair in Parkinson's disease. *J. Control. Release*, **235**, 291–305.

bioRxiv preprint doi: https://doi.org/10.1101/2023.01.27.525898; this version posted January 27, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

102. Vermeire, S., Hébuterne, X., Tilg, H., De Hertogh, G., Gineste, P. and Steens, J.M. (2021) Induction and Long-term Follow-up With ABX464 for Moderate-to-severe Ulcerative Colitis: Results of Phase IIa Trial. *Gastroenterology*, **160**, 2595-2598.e3.

bioRxiv preprint doi: https://doi.org/10.1101/2023.01.27.525898; this version posted January 27, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Figures

Figure 1: Epigenomic and transcriptomic analysis of human mDAN differentiation. A) Gene editing scheme of the human iPSC reporter line, a representative flow cytometry comparison of iPSC-derived neurons from the lines with and without the reporter, and the differentiation protocol used with the different time points analyzed. **B)** Validation of the reporter line by immunocytochemistry. TH staining correlated with mCherry signal. NeuN was used as a neuronal marker while DAPI stained all nuclei. Scale bar = 100 µm. **C)** Gene expression and chromatin accessibility profiles from the different samples analyzed showed specific enrichments for cell type specific markers of mDANs (TH, EN1, DDC, LMX1B), astrocytes (GFAP), and non-mDANs (PAX6). ATAC-seq tracks are plotted under the same scale for comparison purposes. N = 3 for all datasets and samples. **D)** PCA plots of the RNA-seq and ATAC-seq data reveal the differentiation dynamics. **E)** EPIC-DREM pipeline scheme showing the different steps in data integration for the generation of time point-specific gene regulatory networks across differentiation. ATAC-seq was used to define gene regulatory regions and the TFs associated to them were identified through footprinting analysis. Integration with the transcriptional changes defined by RNA-seq allowed the creation of sets of co-expressed genes associated with a ranking of TFs controlling them across time.

Figure 2: Integrative analysis of time series transcriptomic and chromatin accessibility data predicts key regulators of mDAN differentiation. A) EPIC-DREM results from the integration of time course data on gene expression and chromatin accessibility profiles of mDAN differentiation from Figure 1. X-axis represents the time points analyzed and y-axis represents expression $\log_2$-fold changes across time. EPIC-DREM result contained a total of 26 split nodes with an associated list of TFs ranked according to their regulatory importance for the genes contained within the node. See also Supplementary Table 1. Highlighted in red are TFs ranked as top regulators and previously associated to control of mDAN differentiation. Highlighted in blue are the novel identified TFs that were selected for functional analysis. **B)** Pathway enrichment analysis for the genes contained in the first 4 nodes created by EPIC-DREM captures the main biological processes regulated at early stages of neuronal differentiation.

Figure 3: Identification of key TFs controlled by super-enhancers. A) Schematic representation of the selection of TFs controlled by super-enhancers. First, TF has to be expressed at the specific time point of analysis and second, its gene body had to be located under a super-enhancer region in order to be included. **B)** Venn analysis of the top 20 TFs across all split nodes from EPIC-DREM with the list of TFs controlled by super-enhancers across the analyzed time points. **C)** H3K27ac signal and chromatin accessibility profiles at the loci of the novel candidate TFs, highlighting the identified super-enhancer regions in mDANs, together with the expression dynamics during mDAN differentiation. ATAC-seq and ChIP-seq tracks are plotted under the same scale per dataset for comparison purposes.

bioRxiv preprint doi: https://doi.org/10.1101/2023.01.27.525898; this version posted January 27, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Figure 4. LBX1, NHLH1, and NR2F1/2 are necessary for mDAN differentiation and LBX1 and NHLH1 can also improve mDAN neurogenesis. A) Graphic representation of the two different transduction approaches (day 1 and day 9) used in this study and the day of analysis (day 15) for all samples. **B)** TF knock down results for early and late transduction of shRNA lentiviral particles. mRNA levels were always normalized to shScramble per replicate. Relative mDAN numbers were calculated based on mCherry signal and normalized to the mCherry population of the shScramble condition per replicate. One sample t-test was used for statistical analysis, taking 100 as the theoretical mean. **C)** TF overexpression results for early and late transduction of lentiviral particles containing codon optimized cDNA. Expression was normalized to GFP overexpression per replicate. Relative mDAN numbers were also calculated according to mCherry signal but normalized to GFP overexpression per replicate. One sample t-test was used for statistical analysis, taking 1 and 100 as the theoretical mean for gene expression and for mDAN numbers, respectively. For knock-down and overexpression experiments, N $\geq$ 3. * = p-value < 0.05, ** = p-value < 0.01, *** = p-value < 0.001, **** = p-value < 0.0001, and ns = not significant.

Figure 5: NHLH1 controls mDAN differentiation by regulating miR-124-3p expression. A) Volcano plot showing DEGs from RNA-seq analysis of neurons at day 15 following a late transduction with shScramble or shNHLH1. N = 3. Black lines represent cut-off according to FDR < 0.05 and absolute $\log_2$-fold change > 1. **B)** Ingenuity Pathway Analysis (IPA) predicted top 10 upstream regulators based on absolute Z-score value using DEGs from panel A. P-value < 0.05 for all regulators. **C)** Bar plots showing results from TaqMan assay determining miR-124-3p levels upon NHLH1 KD. Expression was normalized to shScramble samples. One sample t-test was used for statistical analysis, taking 1 as the theoretical mean. N = 3 **D)** Venn analysis of predicted and expressed miR-124-3p targets from TargetScan with upregulated genes upon NHLH1 KD. Two overlaps were done, one using the upregulated genes with a p-value < 0.05 and the second with upregulated genes with a $\log_2$-fold change > 1 and p-value < 0.05. For overlaps, a hypergeometric test was used to determine statistical significance. RF = representation factor. RF > 1 indicates more overlap than expected by chance. **E)** Heatmaps showing the expression of the 122 predicted miR-124-3p targets upregulated upon NHLH1 KD. Genes were plotted in the RNA-seq data from the KD experiment and the time course data containing smNPC, mDANs and astrocytes. **F)** RT-qPCR and TaqMan assays were used to determine the expression of different genes and microRNAs upon ABX464 treatment and NHLH1 KD conditions. Expression values were all normalized to ACTB. Normalization between groups was done using shScramble as the reference condition. One sample t-test was used for statistical analysis, taking 1 as the theoretical mean. N = 3. **G)** Effect of ABX464 treatment and NHLH1 KD conditions on mDAN numbers based on mCherry signal. mDAN numbers were normalized to shScramble. One sample t-test was used for statistical analysis, taking 100 as the theoretical mean. N = 3. * = p-value < 0.05, ** = p-value < 0.01, and ns = not significant.

bioRxiv preprint doi: https://doi.org/10.1101/2023.01.27.525898; this version posted January 27, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Figure 6: LBX1 controls cholesterol metabolism. A) Volcano plot showing DEGs from RNA-seq analysis of neurons at day 15 following a late transduction with shScramble or shLBX1. N = 3. Black lines represent cut-off according to FDR < 0.05 and absolute log_2-fold change > 1. **B)** Top 10 pathways from Ingenuity Pathway Analysis (IPA) based on significance of enrichment using DEGs from panel A. X-axis represents the -log_{10} p-value of the enrichment. **C)** Heatmap showing the differences in expression between shScramble and shLBX1 of the genes involved in cholesterol biosynthesis. **D)** Expression levels (FPKM) of LBX1, SREBF1 and SREBF2 in the RNA-seq analysis. **E)** RT-qPCR validation of SREBF1/2 downregulation due to LBX1 KD in late transduced neurons at day 15 of differentiation. N = 5. mRNA levels were normalized to shScramble per replicate. One sample t-test was used for statistical analysis, taking 1 as the theoretical mean. **F)** Impact of GW3965 and rapamycin treatments on gene expression in differentiating mDANs and upon LBX1 KD. Expression of LBX1, SREBF1 and SREBF2 was normalized to shScramble per replicate. N = 4. One sample t-test was used for statistical analysis, taking 1 as the theoretical mean **G)** Effect of GW3965 and rapamycin treatments on mDAN numbers based on mCherry signal and normalized to the mCherry population of the shScramble per replicate. N = 4. One sample t-test was used for statistical analysis, taking 100 as the theoretical mean. * = p-value < 0.05, ** = p-value < 0.01, *** = p-value < 0.001, and ns = not significant.

bioRxiv preprint doi: https://doi.org/10.1101/2023.01.27.525898; this version posted January 27, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Figure 1

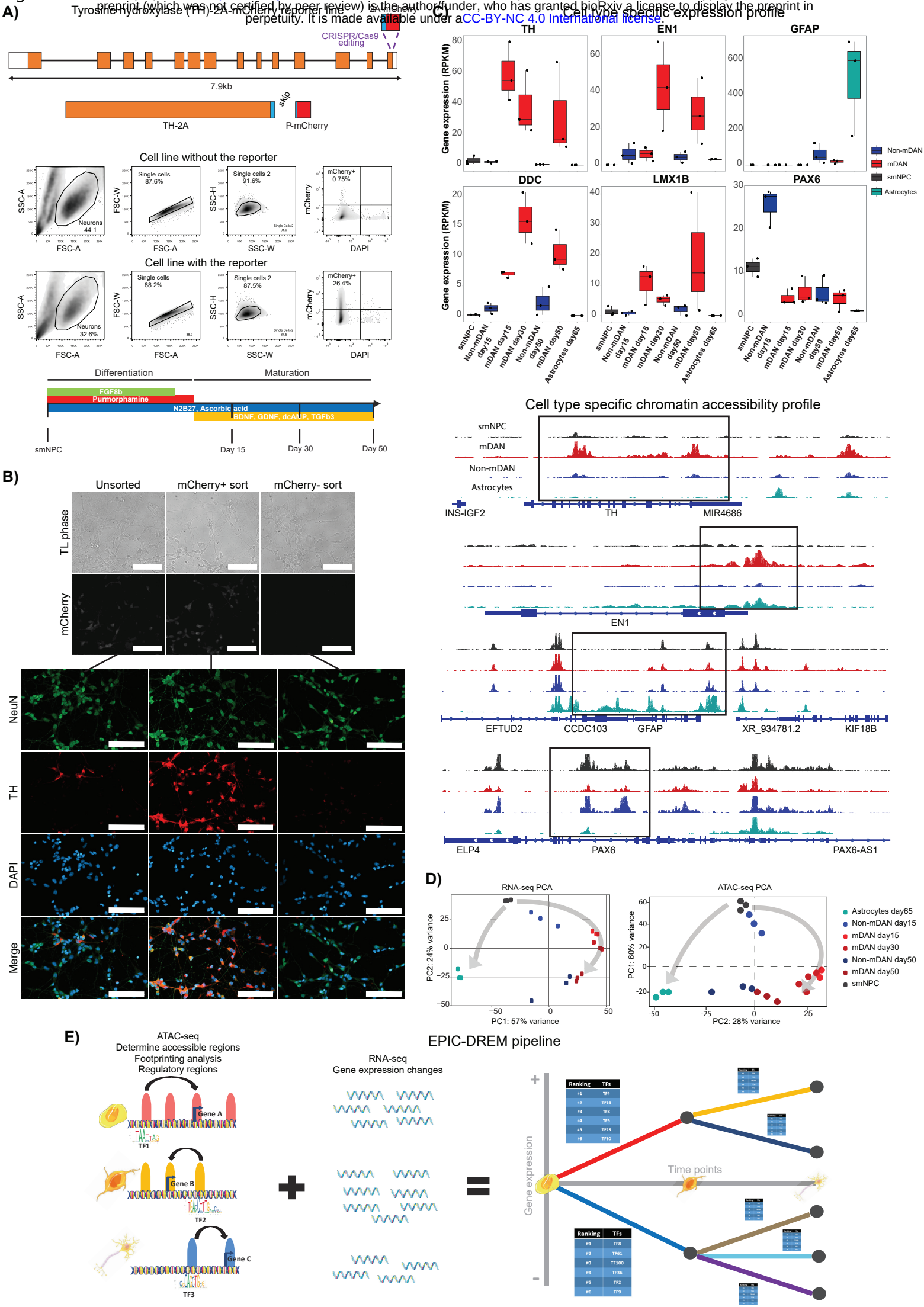
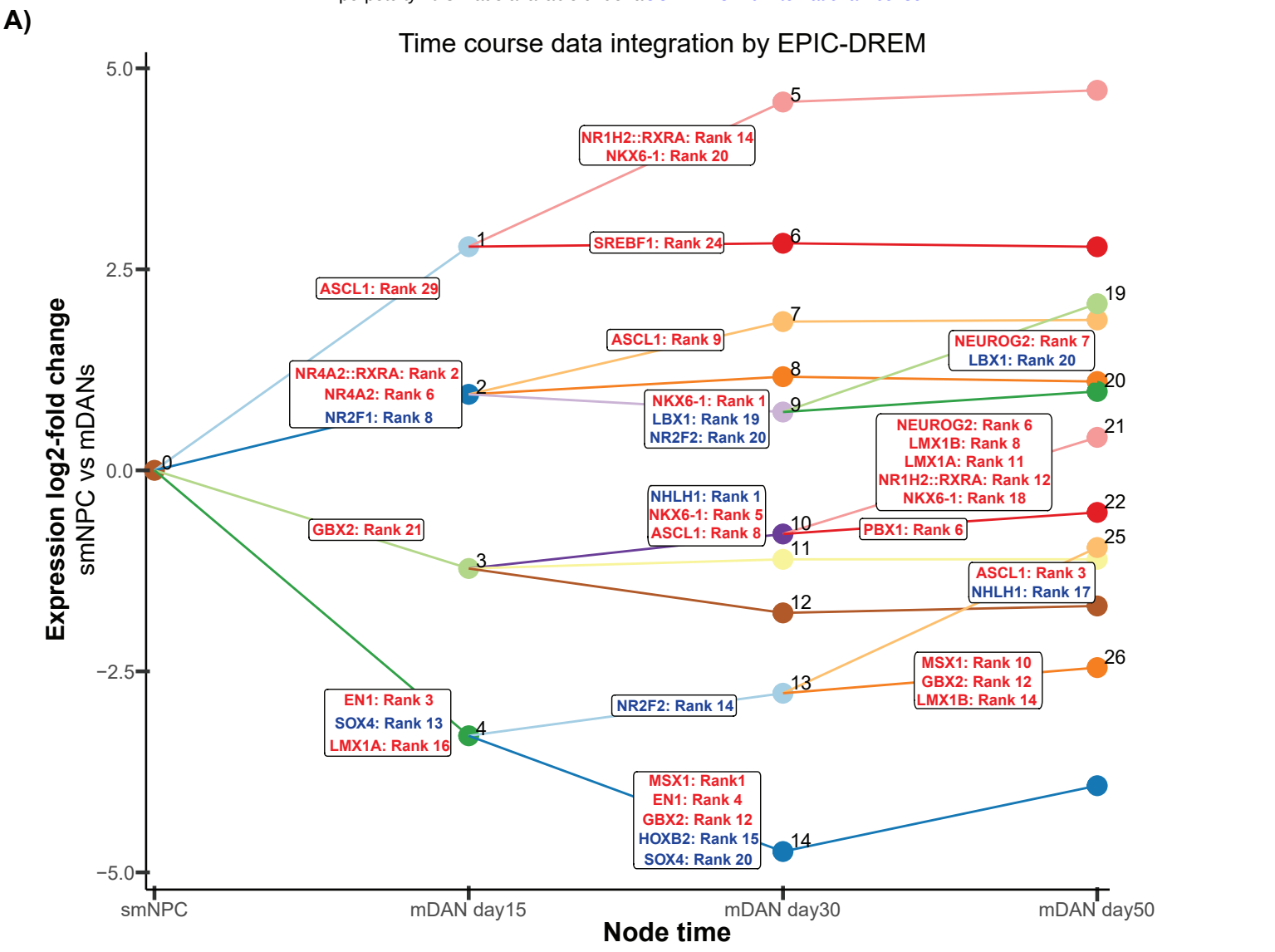


Figure 2

bioRxiv preprint doi: https://doi.org/10.1101/2023.01.27.525898; this version posted January 27, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.



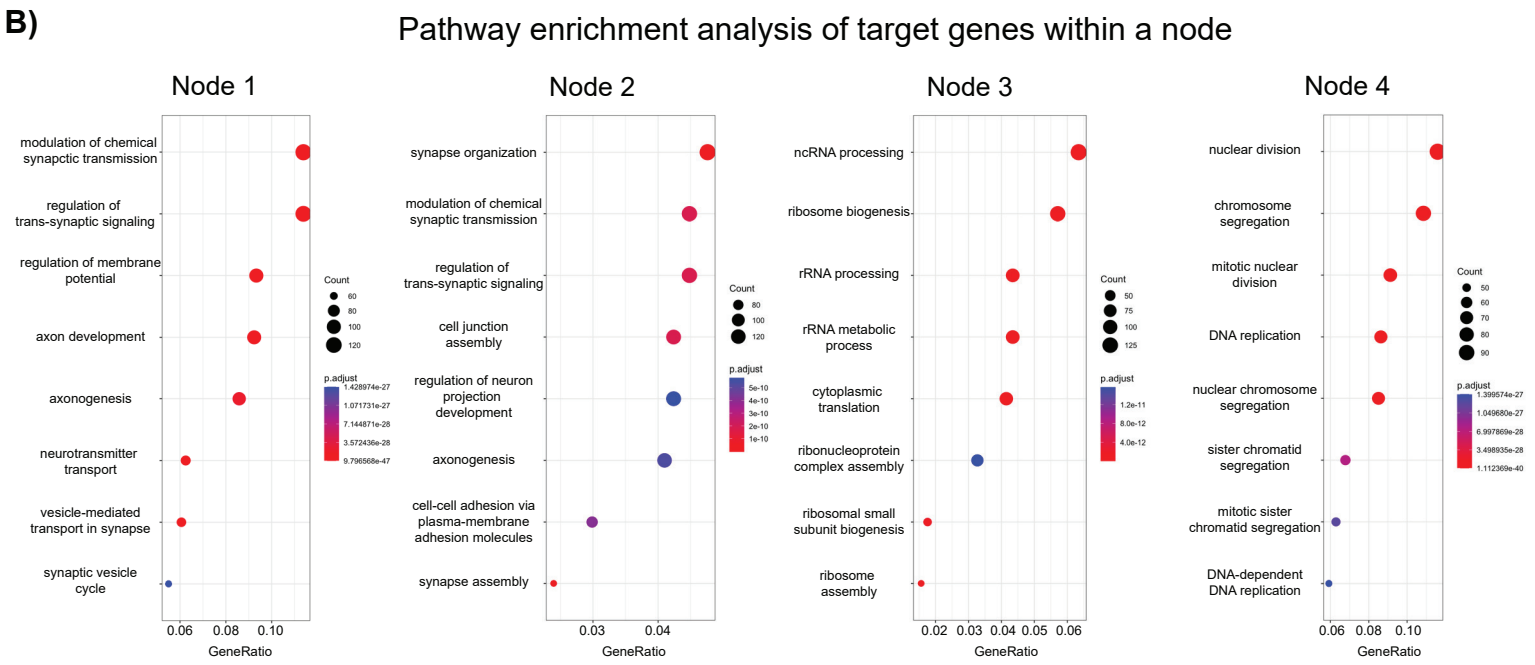


Figure 3

bioRxiv preprint doi: https://doi.org/10.1101/2023.01.27.525898; this version posted January 27, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

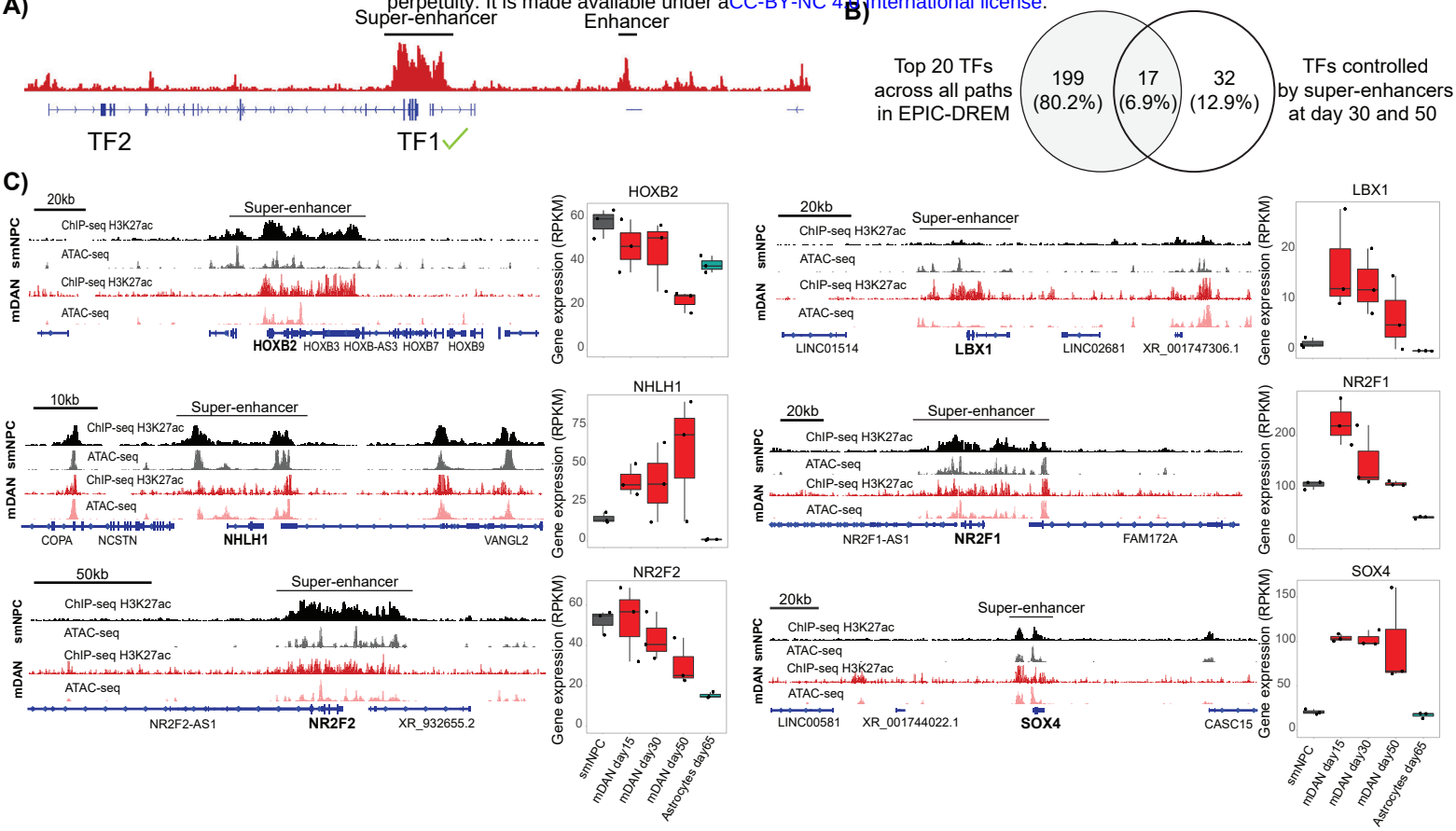


Figure 4

bioRxiv preprint doi: https://doi.org/10.1101/2023.01.27.525898; this version posted January 27, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

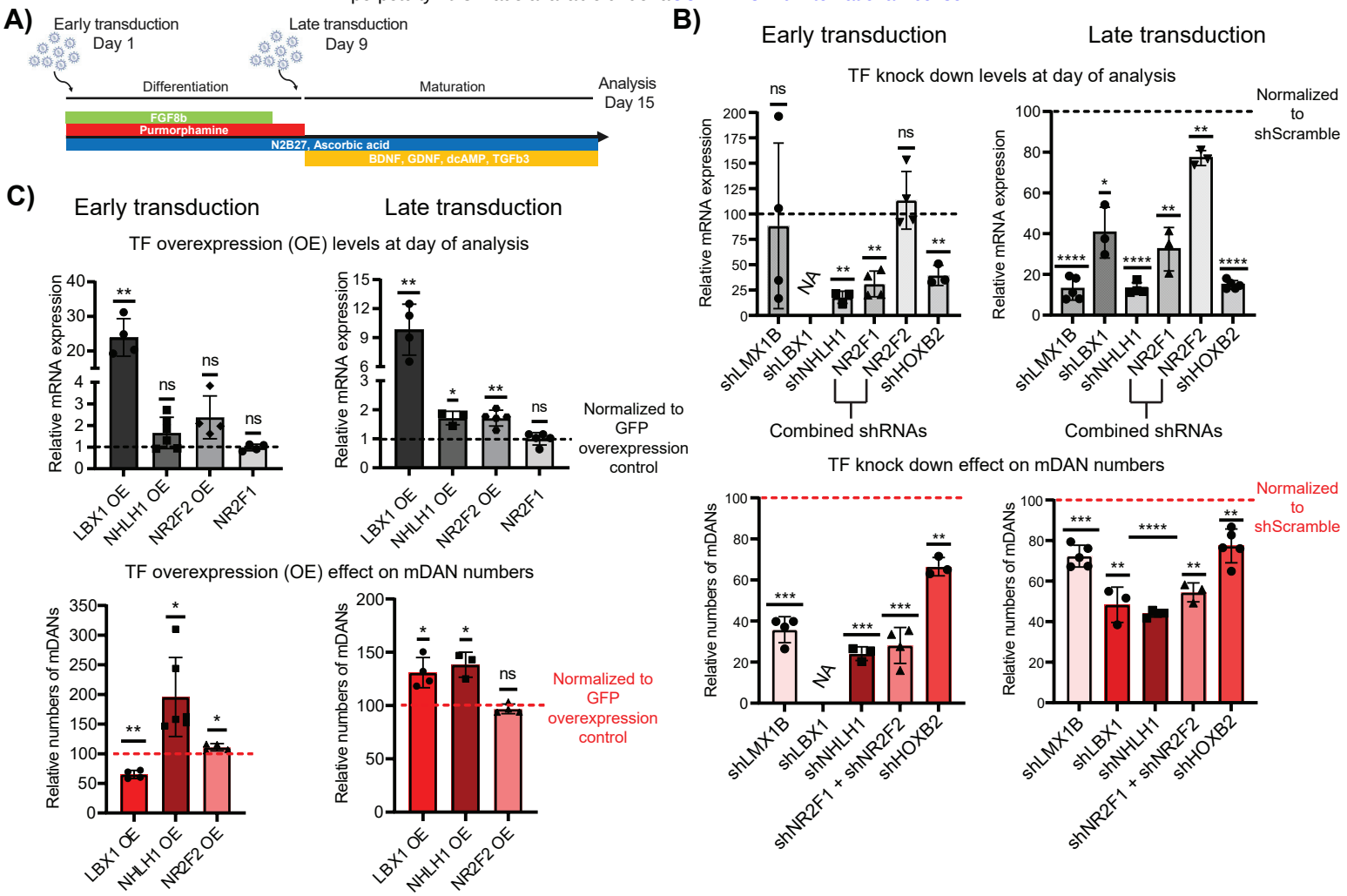


Figure 5

bioRxiv preprint doi: https://doi.org/10.1101/2023.01.27.525898; this version posted January 27, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

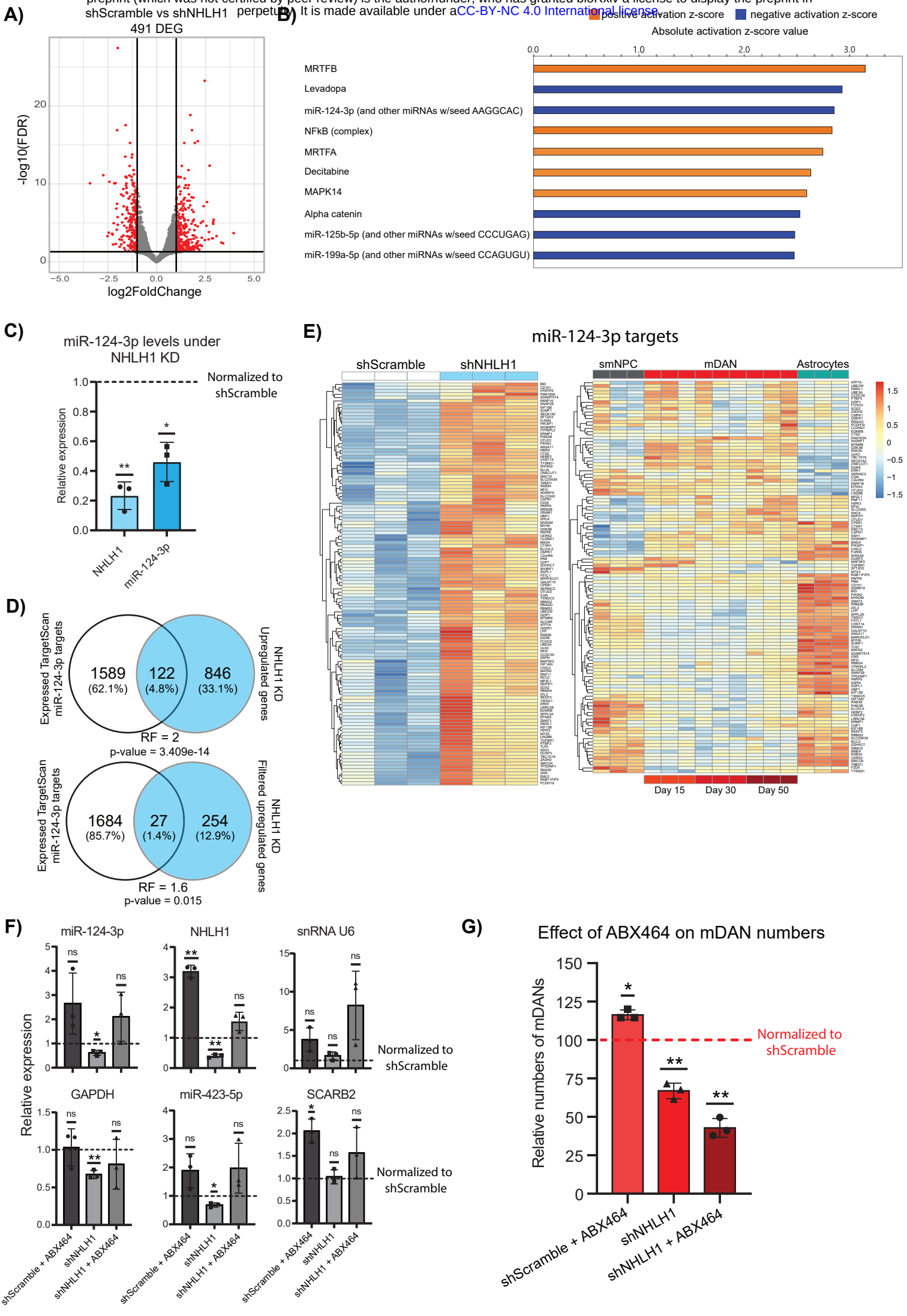
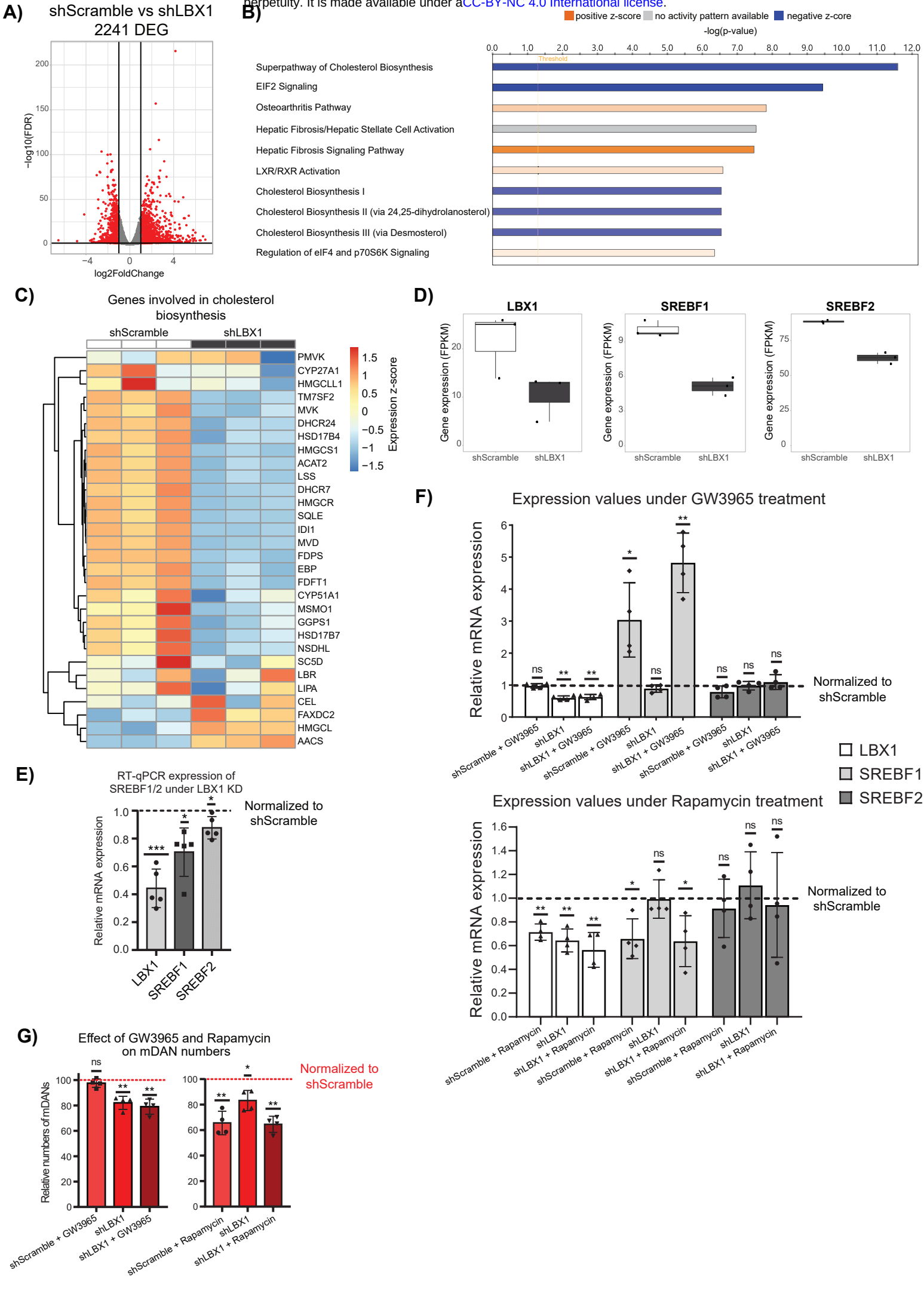


Figure 6

bioRxiv preprint doi: https://doi.org/10.1101/2023.01.27.525898; this version posted January 27, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.



B.3. Manuscript - Risk factors for cognitive disorders after surgery and anesthesia

I supported the transcriptomics analysis, the data management and data ingestion. Only a minor part of the analysis was included in the final manuscript. The manuscript is currently in re-submission to The Lancet.

The Lancet

Risk factors for cognitive disorders after surgery and anesthesia

--Manuscript Draft--

Manuscript Number:	THELANCET-D-21-08716
Article Type:	Article
Keywords:	postoperative delirium; cognitive dysfunction; Postoperative complications; anesthesia; Cohort study; neuroimaging; risk factors; transcriptome; blood specimen collection
Corresponding Author:	Georg Winterer, MD, PhD Charite Universitatsmedizin Berlin GERMANY
First Author:	Georg Winterer, MD, PhD
Order of Authors:	Georg Winterer, MD, PhD
	Levent Akyuez, PhD
	Anna Androsova, PhD
	Diana Boraschi, PhD
	Friedrich Borchers, MD
	Jeroen de Bresser, MD, PhD
	Sreyoshi Chatterjee, MSc
	Marta M. Correia
	Nikola M. de Lange
	Thomas Bernd Dschietzig, MD, PhD
	Soumyabrata Ghosh, PhD
	Insa Feinkohl, PhD
	Jürgen Gallinat, MD, PhD
	Daniel Hadzidiakos, MD
	Sven Hädel, MSc
	John-Dylan Haynes, PhD
	Stefanie Heilmann-Heimbach, PhD
	Maria Heinrich, MD
	Jeroen Hendrikse, MD, PhD
	Per Hoffmann, PhD
	Jürgen Janke, PhD
	Ilse M. J. Kant, PhD
	Angelie Kraft, BSc
	Roland Krause, PhD
	Jochen Kruppa, PhD
	Simone Kühn, PhD
	Gunnar Lachmann, MD, PhD
	Florian Lammers-Lietz, MD
	Markus Laubach, MD

Powered by Editorial Manager® and ProduXion Manager® from Aries Systems Corporation

	Christoph Lippert, PhD
	David K. Menon
	Simone J. T. van Montfort, PhD
	Rudolf Mörgeli, MD
	Anika Müller, MD
	Henk-Jan Mutsaerts, MD, PhD
	Markus Nöthen, MD
	Peter Nürnberg, PhD
	Kwaku Ofosu, MD
	Malte Pietzsch, PhD
	Sophie K. Piper, PhD
	Tobias Pischon, MPH, PhD
	Jacobus Preller, MBChB
	Konstanze Scheurer, PhD
	Reinhard Schneider, PhD
	Kathrin Scholtz, PhD
	Peter H. Schreier, PhD
	Arjen J. C. Slooter, MD, PhD
	Emmanuel A. Stamatakis, PhD
	Clarissa von Häfen, PhD
	Edwin van Dellen, MD, PhD
	Hans-Dieter Volk, MD, PhD
	Simon Weber
	Janine Wiebach
	Anton Wiehe, BSc
	Jeanne M. Winterer, MSc
	Stefan Winzeck, PhD
	Alissa Wolf, MD
	Fatima Yürek, MD
	Norman Zacharias, PhD
	Claudia D. Spies, MD, PhD
Manuscript Region of Origin:	GERMANY
Abstract:	BACKGROUND: Postoperative delirium (POD) is a multietiological condition and affects 20% of older surgical patients frequently followed by postoperative cognitive dysfunction (POCD). It is associated with poor clinical outcome and mortality. We aimed to establish a multimodal biomarker database to identify risk factors and to develop a multivariate risk prediction algorithm. METHODS: BioCog is a multicentric prospective cohort study. Patients aged ≥65 years were enrolled before elective major surgery. Clinical, neuropsychological, neuroimaging and blood-based including transcriptomic data were collected pre- and postoperatively. POD was assessed for up to seven postoperative days and POCD after three months. All preoperative and precipitating perioperative factors underwent univariate analyses followed by multivariate analyses with logistic regression (LogReg) and gradient boosted trees (GBT).

Powered by Editorial Manager® and ProduXion Manager® from Aries Systems Corporation

	RESULTS: 184 of 929 (20%) patients experienced POD. POCD was found in 66 of 578 (11%). POD/POCD risk was associated with preoperative comorbidities, cognitive decline, brain atrophy, duration of surgery/anesthesia, age-associated loss of functional/physical reserve and several blood-based parameters. In multivariate analyses, highest average prediction (AP) and area-under-the-curve (AUC) values were obtained with a combination of clinical, blood-based and neuroimaging parameters: LogReg: AP=54 · 5±8 · 6 (AUC: 79 · 3±5 · 0); GBT: AP=53 · 8±7 · 3 (AUC: 78 · 9±4 · 6). These analyses also showed that preoperatively pathological factors strongly interact with precipitant conditions, in particular the duration of anesthesia and surgery. CONCLUSION: Several predisposing and precipitating risk markers were identified. Models combining clinical, laboratory and imaging parameters predict POD best. This study constitutes the basis for hypothesis-driven analyses and a currently implemented multivariate prediction expert system for clinical practice.

Powered by Editorial Manager® and ProduXion Manager® from Aries Systems Corporation

Manuscript Click here to view linked References

Risk factors for cognitive disorders after surgery and anesthesia

Georg Winterer, MD, PhD[1,3,4], Levent Akyuez, PhD[5], Anna Androsova, PhD[6,7], Diana Boraschi, PhD[8], Friedrich Borchers, MD[1], Jeroen de Bresser, MD, PhD[9], Sreyoshi Chatterjee, MSc[6], Marta M. Correia[10], Nikola M. de Lange[6], Thomas Bernd Dschietzig, MD, PhD[11,12], Soumyabrata Ghosh, PhD[6], Insa Feinkohl, PhD[13], Paul Fletcher, MD, PhD[16], Jürgen Gallinat, MD, PhD[17], Daniel Hadzidiakos, MD[1], Sven Hädel, MSc[3], John-Dylan Haynes, PhD[18], Stefanie Heilmann-Heimbach, PhD[19], Maria Heinrich, MD[1,20], Jeroen Hendrikse, MD, PhD[21], Per Hoffmann, PhD[19,22,23], Jürgen Janke, PhD[13,14], Ilse M. J. Kant, PhD[24], Angelie Kraft, BSc[3,25], Roland Krause, PhD[6], Jochen Kruppa, PhD[20,26], Simone Kühn, PhD[1,17], Gunnar Lachmann, MD, PhD[1,20], Florian Lammers-Lietz, MD[1,3], Markus Laubach, MD[1,3], Christoph Lippert, PhD[27], David K. Menon[28], Simone J. T. van Montfort, PhD[24], Rudolf Mörgeli, MD[1], Anika Müller, MD[1], Henk-Jan Mutsaerts, MD, PhD[21], Markus Nöthen, MD[19], Peter Nürnberg, PhD[30,31], Kwaku Ofosu, MD[1], Malte Pietzsch, PhD[33], Sophie K. Piper, PhD[20,26], Tobias Pischon, MPH, PhD[2,13,14,15], Jacobus Preller, MBChB[28], Konstanze Scheurer, PhD[1], Reinhard Schneider, PhD[6], Kathrin Scholtz, PhD[1], Peter H. Schreier, PhD[3,31], Arjen J. C. Slooter, MD, PhD[24], Emmanuel A. Stamatakis, PhD[29,34], Clarissa von Häfen, PhD[1], Edwin van Dellen, MD, PhD[24,35], Hans-Dieter Volk, MD, PhD[5], Simon Weber[33], Janine Wiebach[20,26], Anton Wiehe, BSc[3,25], Jeanne M. Winterer, MSc[3,36,] Stefan Winzeck, PhD[29,37], Alissa Wolf, MD[1], Fatima Yürek, MD[1], Norman Zacharias, PhD[1,3] and Claudia D. Spies, MD, PhD[1] on behalf of the BioCog consortium

Affiliations

1 Charité–Universitätsmedizin Berlin, corporate member of Freie Universität Berlin and Humboldt-Universität zu Berlin, Department of Anesthesiology and Operative Intensive Care Medicine (CCM, CVK), Berlin, Germany

2 Charité–Universitätsmedizin Berlin, corporate member of Freie Universität Berlin and Humboldt-Universität zu Berlin

3 Pharmaimage Biomarker Solutions GmbH, Berlin, Germany

4 PI Health Solutions GmbH, Berlin Germany

5 Berlin Institute of Health at Charité –Universitätsmedizin Berlin, BIH Center for Regenerative therapies (BCRT), Berlin, Germany

6 Bioinformatics core, Luxembourg Center for Systems Biomedicine (LCSB), University of Luxembourg, Belvaux, Luxembourg

7 Swiss Data Science Center EPFL & ETH Zurich, Zurich, Switzerland

8 Institute of Biochemistry, Consiglio Nazionale delle Ricerche (CNR) di Pisa, Pisa, Italy

9 Department of Radiology, Leiden University Medical Center, Leiden, Netherlands

10 MRC Cognition and Brain Sciences Unit, University of Cambridge, Cambridge, United Kingdom

11 Immundiagnostik AG, Bensheim, Germany

12 MHB Medizinische Hochschule Brandenburg, Neuruppin, Germany

13 Molecular Epidemiology Research Group, Max-Delbrück-Center for Molecular Medicine in the Helmholtz Association (MDC), Berlin, Germany

14 Biobank Technology Platform, Max-Delbrück-Center for Molecular Medicine in the Helmholtz Association (MDC), Berlin, Germany

15 Berlin Institute of Health at Charité – Universitätsmedizin Berlin, Core Facility Biobank, Berlin, Germany

16 Department of Psychiatry, University of Cambridge, Cambridge, United Kingdom

17 Department of Psychiatry, University Medical-Center Hamburg-Eppendorf, Hamburg, Germany

18 Berlin Center for Advanced Neuroimaging (BCAN), Charité–Universitätsmedizin Berlin, corporate member of Freie Universität Berlin and Humboldt-Universität zu Berlin, Berlin, Germany

19 Institute of Human Genetics, University of Bonn, Bonn, Germany

20 Berlin Institute of Health at Charité –Universitätsmedizin Berlin, Berlin, Germany

21 Department of Radiology and Brain Center Rudolf Magnus, University Medical Center Utrecht (UMC), Utrecht, Netherlands

22 Division of Medical Genetics, University Hospital, Basel, Switzerland

23 Human Genetics Research Group, Department of Biomedicine, University of Basel, Basel, Switzerland

24 Department of Intensive Care Medicine and Brain Center, University Medical Center Utrecht (UMC), Utrecht University, Utrecht, the Netherlands

25 AdaLab UG, Hamburg, Germany

26 Charité–Universitätsmedizin Berlin, corporate member of Freie Universität Berlin and Humboldt-Universität zu Berlin, Institute of Biometry and Clinical Epidemiology

27 Hasso-Plattner Institute, University of Potsdam, Potsdam, Germany

28 John Farman ICU, Cambridge University Hospitals NHS Foundation Trust, Cambridge, United Kingdom

29 Division of Anaesthesia, Department of Medicine, University Of Cambridge, Cambridge, United Kingdom

30 Cologne Center for Genomics, University of Cologne, Cologne, Germany

31 Institute for Genetics of the University of Cologne, Cologne, Germany

32 Atlas Biolabs GmbH, Berlin, Germany

33 Cellogic GmbH (Cellogic), Berlin, Germany

34 Department of Clinical Neurosciences, School of Clinical Medicine, University of Cambridge, Cambridge, United Kingdom

35 Department of Psychiatry and UMC Utrecht Brain Center, University Medical Center Utrecht (UMC), Utrecht, Netherlands

36 Charité–Universitätsmedizin Berlin, corporate member of Freie Universität Berlin and Humboldt-Universität zu Berlin, Department of Psychiatry (CCM), Berlin, Germany

37 BioMedIA, Department of Computing, Imperial College London, London, United Kingdom

Corresponding author:

Georg Winterer, MD, PhD – Associate Professor of Psychiatry

Clinical Neuroscience Research Group

Charité–Universitätsmedizin Berlin, corporate member of Freie Universität Berlin and Humboldt-Universität zu Berlin, Department of Anesthesiology and Operative Intensive Care Medicine (CCM, CVK)

Lindenberger Weg 80, D-13125 Berlin, Germany

Phone: +49 30 450 540 105

Fax: +49 30 450 540 767

E-mail: georg.winterer@charite.de

BACKGROUND: Postoperative delirium (POD) is a multietiological condition and affects 20% of older surgical patients frequently followed by postoperative cognitive dysfunction (POCD). It is associated with poor clinical outcome and mortality. We aimed to establish a multimodal biomarker database to identify risk factors and to develop a multivariate risk prediction algorithm.

METHODS: BioCog is a multicentric prospective cohort study. Patients aged ≥65 years were enrolled before elective major surgery. Clinical, neuropsychological, neuroimaging and blood-based including transcriptomic data were collected pre- and postoperatively. POD was assessed for up to seven postoperative days and POCD after three months. All preoperative and precipitating perioperative factors underwent univariate analyses followed by multivariate analyses with logistic regression (LogReg) and gradient boosted trees (GBT).

RESULTS: 184 of 929 (20%) patients experienced POD. POCD was found in 66 of 578 (11%). POD/POCD risk was associated with preoperative comorbidities, cognitive decline, brain atrophy, duration of surgery/anesthesia, age-associated loss of functional/physical reserve and several blood-based parameters. In multivariate analyses, highest average prediction (AP) and area-under-the-curve (AUC) values were obtained with a combination of clinical, blood-based and neuroimaging parameters: LogReg: AP=54·5±8·6 (AUC: 79·3±5·0); GBT: AP=53·8±7·3 (AUC: 78·9±4·6). These analyses also showed that preoperatively pathological factors strongly interact with precipitant conditions, in particular the duration of anesthesia and surgery.

CONCLUSION: Several predisposing and precipitating risk markers were identified. Models combining clinical, laboratory and imaging parameters predict POD best. This study constitutes the basis for hypothesis-driven analyses and a currently implemented multivariate prediction expert system for clinical practice.

The research leading to these results has received funding from the European Union Seventh Framework Program [FP7/2007-2013] under grant agreement n° 602461. The study was registered at Clinicaltrials.gov under no. NCT02265263.

KEYWORDS: *postoperative delirium, cognitive dysfunction, postoperative complications, anesthesia, cohort study, neuroimaging, risk factors, transcriptome, blood specimen collection*

1 Introduction

Delirium is defined by disturbances in attention, awareness, cognition, psychomotor behavior and emotional state.[1] According to the 5th edition of the Diagnostic and Statistical Manual of Mental Disorders (DSM-V), delirium is a consequence of another medical condition, intoxication or withdrawal, or multietiological. Delirium incidence after surgery ranges from 5-50% for different patient cohorts and surgical procedures.[2] Postoperative delirium (POD) occurs most frequently in older patients (20% incidence in >65y and >30% in >80y old).[3] POD incidence is assumed to rise with increasing life expectancy. For Europe, the WHO has predicted life expectancy of 80y in 2050 compared to 60y in 1950 and in 2010, mean life expectancy at the age of 65y was still 19y.[4,5] This challenges healthcare systems since POD is associated with hospitalization, treatment costs, re-institutionalization and mortality risk.[6] Cognitive dysfunction may persist for months after surgery.[7] This condition, referred to as postoperative cognitive dysfunction (POCD), is associated with early labor market leave, social transfer payment dependence and increased mortality.[8]

As preventable measures have already been described, early detection of patients at risk for postoperative neurocognitive disorders is essential, especially as prehabilitation programs can be time consuming and should be carefully weighted against a delay of surgical intervention in some patients.[9]

Previous literature reports POD/POCD risk to be related to patient characteristics, e.g. age, comorbidities, medication, cognitive status, functionality, type of surgery.[3,10] A variety of serological, molecular and imaging biomarkers have been discussed on the pathogenesis, including biomarkers of inflammation and immune reaction, metabolism and brain damage.[2]

The aim of this study was to create a clinical database and a biorepository to establish a valid set of clinical, blood-based and neuroimaging biomarkers for multivariate risk and clinical outcome prediction of POD/POCD.

2 Methods

2.1 Study design

The BioCog (Biomarker Development for Postoperative Cognitive Impairment in the Elderly, www.biocog.eu) study is an EU-funded multicenter prospective observational cohort study with the aim to identify risk factors for POD and POCD.

The study was conducted in line with the Helsinki declaration (clinicaltrials.gov: NCT02265263). All procedures were approved by the local ethics committees in Berlin, Germany (EA2/092/14) and Utrecht, Netherlands (14-469). All participants gave written informed consent prior to inclusion.

2.2 Participants

Patients were enrolled at the Charité – Universitätsmedizin Berlin, Germany, and the University Medical Center Utrecht, Netherlands. Consenting patients of European (Caucasian) descent aged ≥65 years presenting for elective non-cardiac surgery with an expected duration >60min were included. Patients were excluded from study according to the following criteria:

- positive screening for pre-existing major neurocognitive disorder defined as a Mini-Mental Status Examination (MMSE) score ≤23 points
- any condition interfering with neurocognitive assessment (severe sensory impairment, neuropsychiatric illness, intracranial surgery)
- unavailability for follow-up assessment
- accommodation in an institution due to official or judicial order
- inability to give informed consent (including detection of preoperative delirium assessed at baseline)

2.3 Study procedures

Data were collected one day before and at multiple time points after surgery including medical history and clinical assessments, neuropsychological testing, blood collection and neuroimaging. Follow-up assessments were scheduled for up to five years, but data from the follow-up three months after surgery were analyzed here only. Figure 1 displays the study procedure.

2.4 Outcome

Co-primary endpoints were POD during the first seven days after surgery and POCD at three months follow-up.

Independently of the routine hospital procedures, POD assessments were conducted twice daily during the first seven postoperative days by a clinical research team, which was trained and supervised by delirium experts. POD was defined according to DSM-5 criteria.[1] Since it is known that even ICU physician tend to underdiagnose delirium, four results from four additional assessment tools were recorded at each visits.[11] Patients were considered delirious in case of:

- ≥2 cumulative points on the Nursing Delirium Screening Scale (Nu-DESC), and/or[12]
- a positive Confusion Assessment Method (CAM) score on a general ward, and/or[13]
- a positive CAM for the Intensive Care Unit (CAM-ICU) score on an intensive care unit, and/or[14]
- patient´s chart review showing descriptions of delirium (e.g. confused, agitated, drowsy, disorientated, delirious, received antipsychotic therapy).

Neuropsychological testing took place before surgery and at hospital discharge. Patients were invited for follow-up assessment three months postoperatively.

To adjust for natural variability in cognitive performance and learning effects in repeated cognitive testing, POCD was defined according to the Reliable Change Index (RCI) with the ISPOCD (International Study of Postoperative Cognitive Dysfunction) criteria by Rasmussen implemented in the POCD package for R (https://github.com/Wiebachj/POCDr).[15,16] The RCI is the patient's performance change in a cognitive test (pre- vs. postoperatively) in relation to the mean change in a non-surgical control group. POCD was diagnosed in patients with an RCI score below -1·96 in at least two out of seven single cognitive test parameters and/or a compound RCI averaged over all single RCIs (see supplement 1). Demographic and cognitive properties of the control cohort and test stability have been previously described.[17]

2.5 Mortality and post-hoc analysis of loss to follow-up

Mortality data until the 90th postoperative day were only collected in Berlin. Post-hoc analysis of loss to follow-up (missing neuropsychological testing and death before follow-up) was conducted (supplement 4).

2.6 Potential predictor parameters

An explicit parameter list and assessment details are given in supplement 1.

2.7 Clinical assessments

Sociodemographic data, preoperative diagnoses and medication, health-related quality of life, nutritional status, tobacco and alcohol use were recorded before surgery. A functional and physical assessment battery including frailty and functional impairment was conducted. Potentially precipitating factors (duration of surgery and of anesthesia, postoperative anticholinergic medication and pain, duration of ICU and hospital stay) were recorded.

2.8 Neuropsychological data

Patients underwent cognitive testing with the Cambridge Neuropsychological Test Automated Battery (CANTAB) battery and additional paper and pencil tests. Initial Mini-Mental-State- Examination (MMSE) score, battery test scores and overall preoperative cognitive impairment (PreCI) in the test battery were analyzed as risk factors. PreCI was defined through comparison of cognitive test scores with those of a control group analogously to the definition of POCD.[15,16] Readers should note that all patients, including those with PreCI, obtained MMSE>23 which was part of our inclusion criteria. Details on the CANTAB assessment are given in supplement 1.

2.9 Laboratory parameters

Routine clinical and experimental parameters were assessed. Transcriptomic analyses were conducted for consenting patients. Details on parameters and methods used in transcriptomic analyses are given in supplement 1.

2.10 Neuroimaging

The imaging protocol comprised whole brain T1-weighted structural magnetic resonance imaging (MRI) sequences, T2-weighted high-resolution hippocampus imaging and diffusion tensor imaging (DTI). We calculated global and regional brain volumes including hippocampal subregions, cortical thickness and curvature, mean diffusivity, kurtosis and fractional anisotropy. MRI data analyzed here were acquired within two weeks before surgery (figure 1).

2.11 Statistics

2.11.1 Estimation of sample size

Sample size was based on the following considerations: With a 20-30% incidence of POD, a 10% drop-out rate and an assumed effect size of Hedge's g=0·5, a sample size of 198-222 yields 80% power for a two-sided t-test at α=0·05.[18] Due to lower POCD incidence (8-12%) and higher drop-out after three

months (20%), N=360-540 are needed (GPower software).[19] To allow a 1:1 split into test and validation sample, a sample size of n=1200 patients was planned.

2.11.2 Univariate statistical analyses

For descriptive purposes, associations of pre-/perioperative parameters with POD/POCD were analyzed using simple logistic regression. We report odds ratios (OR) with 95% confidence intervals (CI). Cut-offs were set for continuous variables associated with POD/POCD from the literature (supplement 1).

P-values in the mRNA microarray analysis were adjusted for multiple testing and $p<0{\cdot}05$ was considered significant.

Postoperative survival was assessed using Kaplan-Meier curves stratified for POD, PreCI and a combination thereof. Differences in survival were exploratorily tested using the log-rank test.

2.11.3 Multivariate statistical analyses

For multivariate POD risk prediction, gradient boosted trees (GBT) were used as a machine learning approach and compared to logistic regression. The hyperparameter tuning considered different regularization strategies for logistic regression, like LASSO and Elastic Net. Analyses were conducted separately for different parameter domains: clinical/neuropsychological, blood-based, neuroimaging. Models were first trained and evaluated per domain and then on aggregation of all domains.

For the neuroimaging domain, relevant features (hippocampus volume, brain volume, Nucleus basalis Meynert volume) were preselected based on expert knowledge. For clinical, neuropsychological and blood-based parameter domains, no a priori feature selection was conducted. For comparison, an additional prediction model was trained considering potentially peri- and postoperative precipitating factors (precipitants), e.g. duration of anesthesia, site of surgery, uncontrolled post-operative pain, post-operative anticholinergic medication. Precision recall curves and receiver operating curves (ROC) with average precision (AP) values and area-under-the-curve (AUC) values with 95% CI are provided.

Models were validated using nested cross validation to ensure generalization despite hyperparameter optimization (for details see supplement 1). We did not perform prediction modeling of POCD due to its low incidence.

2.12 Role of funding source

The funding source had no involvement in study design, data collection, analysis or interpretation, writing or submitting the manuscript.

3 Results

We recruited 933 patients between 2014/11 and 2017/04. Table 1 characterizes the sample. The patient flow chart is given in figure 2a.

Preoperative cognitive assessment was completed in 928 patients and PreCI was found in 122/928 (13%) patients. POD assessments were available for 929 patients. 184/929 (20%) patients developed POD. Figure S1 (supplement 2) gives an overview of daily POD incidence. 664/933 (71%) patients provided follow-up data after three months, but only 641/933 (69%) patients underwent complete neuropsychological assessment. POCD was found in 66/641 (10%, figure 2b).

3.1 POD and POCD risk factors (univariate analyses)

Figure 3 displays OR with 95% CIs for parameters with CIs excluding unity. Sample and effect sizes for all parameters are given in supplements 2 (POD) and 3 (POCD).

Preoperative comorbidity, advanced age and related geriatric deficits (functional and motor impairment, malnutrition, tumor) and preoperative longterm benzodiazepine intake were all associated increasing risk for POD/POCD (supplement 2, tables S2-S4; supplement 3, S18-S20). PreCI (OR: 2·57 [1·69; 3·88]), but especially Verbal Recognition Memory impairment (OR: 3·12 [1.41; 6·92]) and the MMSE score (OR: 3·10 [1·96; 4·85]) increased risk for POD. PreCI was unrelated to POCD (supplement 3, tables S21-22).

There were several perioperative precipitating factors which increased the risk to develop POD including duration of surgery and anesthesia, surgery with opening of body cavities, postoperative pain and anticholinergic medication (supplement 2, tables S8-10, figure S3-4).

Several preoperative blood-based metabolic markers were associated with POD (tables S11). Laboratory signs of inflammation were associated with POD/POCD risk, and high OR were especially found for interleukin 8 >200pg/mL and POD (OR: 13·90 [1·83; 105·79], supplement 2, tables S12, S13) and immature granulocytes >0·9% for POCD (OR: 5·21 [1·31; 18·22], supplement 3, tables S25, S29). A higher ratio of β-amyloid 42 and 40 concentrations was associated with POD (OR: 0·74 [0·56; 0·93], supplement 2, table S12).

For neuroimaging data, a general pattern emerged showing that lower regional brain volumes were associated with POD/POCD risk. In figure 3, an excerpt from all neuroimaging parameters is given including whole brain and hippocampal volume (see also supplement 2, figures S5, tables S14-15; supplement 3 S27-28, figure S7).

mRNA microarray analysis showed several genes differentially expressed in POD patients at $p<0·05$, but with a very low fold-change (supplement 2, figure S6). We cannot infer any significant call of differentially expressed genes (supplement 2, tables S16, S17).

3.2 Multivariate prediction of POD

Figure 4 (top) shows the AP recall curves of the fine-tuned GBT models compared with logistic regression as a benchmark (further details in supplement 4). Comparing the models across domains (clinical, blood-based and neuroimaging), the model based on clinical parameters performs best. Combining the clinical parameters with biomarkers improves model performance slightly (supplement 4). Furthermore, the model is improved when adding precipitants to the combined models with clinical, blood-based and neuroimaging parameters, and in particular the parameters duration of surgery and anesthesia >4h boost the performance. Overall, the following factors are

some of the most important POD predictors in this work (figure 4; supplement 4, figure S11): preoperative cognitive impairment, frailty, low hippocampus volume, functional and physical impairment, site of surgery, long duration of surgery/anesthesia, anemia and inflammation markers.

3.3 Post-hoc analysis of loss to follow-up and mortality

Patients with POD had prolonged hospital stay (median, IQR: 11d, 7-23d vs. 5d, 3-8d), were admitted to the ICU more often (112 of 745 [15%] vs. 133 of 184 [72%]) and needed prolonged intensive care (median, IQR: 1d, 0-1d vs. 0d, 0-0d) compared to those without POD.

90-day mortality was assessed for 648 of 651 patients enrolled in Berlin. Figure 5 shows the respective Kaplan-Meier curves. POD (but not PreCI) patients show a statistically significant poorer survival rate. A post-hoc analysis of factors associated with mortality and loss to follow-up is given in supplement 5.

4 Discussion

We present results from the most extensive cohort study to date on POD/POCD that makes use of a sample of Dutch and German over 65 years who were all cognitively unimpaired at enrollment. We provide a multimodal database encompassing a wide range of markers across different domains: clinical/neuropsychological, blood-based and neuroimaging.

We found that several previously reported clinical and biomarker risk factors predict POD/POCD including anemia, metabolic changes, preoperative cognitive impairment, poor physical status, long-term benzodiazepine intake or the duration of surgical intervention and anesthesia.[20–22] This once again supports the notion of a multietiological character of these two clinical conditions.

For instance, our findings are in line with results obtained in the SAGES project. The SAGES project is a POD cohort study in 566 patients (age mean±SD: 77±5y) with major non-cardiac surgery.[23] Their neuroimaging data from 146 patients suggest that Alzheimer-like brain changes can increase POD

risk.[24] We found that a low preoperative MMSE score and a presurgical pattern of low hippocampal and brain volume is associated with POD. SAGES also reported preoperative changes of blood-based brain-injury- and Alzheimer-related biomarkers (neurofilament light).[25] In our study, β-amyloid contribute to POD risk prediction in multivariate analyses. This may suggest pathobiological commonalities of POD with Alzheimer's dementia.

In contrast, we also obtained findings deviating from what has been reported by others. Among others, a large observational study found alcohol-related predictors to be associated with POD in 774 patients with head/neck carcinoma.[26] However, we found no association of the Alcohol Use Disorder Identification Test (AUDIT) score with POD, which might be due to the low prevalence of hazardous alcohol consumption (<7%). Also, self-report in the AUDIT alone might have poor predictive quality compared to a medical history of alcohol abuse or combination with additional markers, e.g., in our multivariate model γ-glutamyltransferase contributed to risk prediction.[27]

Despite previous reports, we found no association of POD and POCD.[7] We need to acknowledge the low POCD prevalence in our study, which might be caused by the low follow-up (69% for the cognitive testing), although the ISPOCD study reported a similar low prevalence (10%) with a comparable follow-up rate (78%). POD was associated with loss-to-follow-up, which might have masked a statistically significant association of POD with subsequent POCD in our study. This is further substantiated by the finding that duration of hospital and ICU stay predict POD risk, but not POCD, which might be explained by higher early mortality after complicated surgery. Loss to follow-up could have masked a fraction of POCD diagnoses in patients who died early after discharge or found follow-up assessments to stressful.[15]

Our multivariate analyses identified risk factors. In the basic multivariate preoperative prediction model, clinical markers and biomarkers separately predicted POD. However, biomarkers did not substantially improve POD risk prediction as compared to clinical parameters. An obvious explanation

for this is that clinical conditions and biomarkers can be closely associated with each other. For instance, low hippocampal volume is a hallmark of dementia as reflected by a cognitive decline in MMSE. Low serum proteins can be related to frailty and malnutrition.[28] Similarly, the effects of inflammation, which is considered a key feature in POD development, might rather be mediated by chronic disease burden in older patients driven by chronic inflammation (inflammageing).[2,29] Thus, inflammation-related organ damage might be more relevant for POD than inflammation itself. Even though, this does not exclude the possibility that inflammation constitutes a pathological relevant final common path of aging in POD development.

Apart from canonical inflammatory markers we found a remarkable association of interleukin 8 and POD substantiating findings from our own work.[16] IL8 is considered a chemoattractant with a half-life of days to weeks. High plasma levels have been discussed to misguide immune cells from the site of inflammation rather than orchestrating a locally limited response to tissue damage, carrying peripheral inflammation over to the central nervous system in delirium.[30]

When perioperative precipitants such as duration of anesthesia and surgery were included in the prediction models, a combined clinical and biomarker model was clearly superior as opposed to models that rely only on clinical parameters. This indicates that preoperative risk status and perioperative precipitants interact such that preoperative serological changes and/or brain neurodegeneration become meaningful for POD prediction in patients who undergo long surgical procedures. These analyses also show that POD risk is not fully captured by preoperative information alone.

Some of our multivariate GBT models integrate information from biomarkers with clinical data better than logistic regression models. However, to fully exploit the potential of machine learning algorithms and to further improve prediction algorithms, a big data approach with sample sizes that are at least one order of magnitude higher would be needed. This could be achieved with a future multicentric

collaboration study or data pooling. This way, relatively rare pathological risk factors also could be identified and patient populations could be covered that were underrepresented in our cohort such as patients receiving cardiovascular surgery or patients with advanced dementia.

We anticipate that our results will facilitate further sophisticated hypothesis-driven analyses and subgrouping of patients for better understanding of pathophysiological processes and conception of clinical studies. The combination of clinical parameters and biomarker might be particularly helpful for the overall risk assessment in those patients who are scheduled for long surgical procedures. In addition, the preoperative clinical information together with biomarker data can guide prevention and treatment strategies to reduce POD risk ahead of elective surgeries (modifiable risk factors) and to constantly re-evaluate the patient's POD risk during postoperative care.

5 Acknowledgments

The authors made the following contributions to the work presented here: Georg Winterer: conceptualization, funding acquisition, investigation, methodology, resources, software, supervision, visualization, writing original draft; Levent Akyuez: investigation; Anna Androsova: resources, investigation; Diana Boraschi: investigation, resources, writing original draft; Friedrich Borchers: data curation, investigation, resources, validation, writing original draft; Jeroen de Bresser: formal analysis, investigation, writing original draft; Sreyoshi Chatterjee: formal analysis, investigation; Marta M. Correia: formal analysis; Nikola M. de Lange: formal analysis, investigation; Thomas Bernd Dschietzig: investigation, methodology, resources, writing original draft; Soumyabrata Ghosh: formal analysis, investigation; Insa Feinkohl: data curation, investigation, resources, writing original draft; Jürgen Gallinat: methodology; Daniel Hadzidiakos: data curation, investigation, resources, validation; Sven Hädel: data curation, software; John-Dylan Haynes: methodology, resources; Stefanie Heilmann-Heimbach: investigation; Maria Heinrich: data curation, investigation, methodology, resources, validation, writing original draft; Jeroen Hendrikse: formal analysis, investigation; Per Hoffmann:

investigation; Jürgen Janke: investigation; Ilse M. J. Kant: formal analysis, investigation, resources, validation; Angelie Kraft: formal analysis, resources, writing original draft; Roland Krause: formal analysis, investigation, resources; Jochen Kruppa: data curation, formal analysis, methodology, software, writing original draft; Simone Kühn: formal analysis, investigation, methodology, resources, software; Gunnar Lachmann: investigation, resources; Florian Lammers-Lietz: data curation, formal analysis, investigation, visualization, writing original draft; Markus Laubach: investigation; Christoph Lippert: formal analysis, investigation; David K. Menon: formal analysis, investigation; Simone J. T. van Montfort, investigation, resources, validation; Rudolf Mörgeli: formal analysis, investigation, resources; Anika Müller: investigation, resources; Henk-Jan Mutsaerts: formal analysis, investigation; Markus Nöthen: formal analysis, investigation; Peter Nürnberg: formal analysis, investigation, methodology; Kwaku Ofosu: investigation, resources; Malte Pietzsch: conceptualization, formal analysis; Sophie K. Piper: formal analysis, investigation, methodology, software, writing original draft; Tobias Pischon: conceptualization, funding acquisition, investigation, supervision, methodology, resources, revision of manuscript; Jacobus Preller: formal analysis, methodology, writing – review & editing; Konstanze Scheurer: project administration; Reinhard Schneider: methodology, resources; Kathrin Scholtz: data curation, project administration, writing original draft; Peter H. Schreier: investigation, resources; Arjen J. C. Slooter: conceptulatization, methodology, resources, supervision, writing original draft; Emmanuel A. Stamatakis: formal analysis, methodology, investigation, writing original draft; Clarissa von Häfen: investigation, project administration, resources; Edwin van Dellen: investigation, formal analysis, supervision, methodology; Hans-Dieter Volk: methodology, resources; Simon Weber: conceptualization, formal analysis; Janine Wiebach: formal analysis, methodology; Anton Wiehe: formal analysis, software, visualization, writing original draft; Jeanne M. Winterer: data curation; Stefan Winzeck: formal analysis, investigation; Alissa Wolf: investigation, resources; Fatima Yürek: investigation, resources, writing original draft; Norman Zacharias: data curation, formal analysis, investigation, methodology, validation; Claudia D. Spies: conceptualization, funding

acquisition, investigation, methodology, resources, software, supervision, visualization, writing original draft.

Special thanks to Judy Veldhuijzen (UMC Utrecht) who supplied neuropsychological expertise. Clinical data management was provided by Olaf Bender and Alexander Krannich at Koordininierungszentrum für Klinische Studien (KKS Berlin). We thank our team of MD students and study nurses: Tuba Gülmez, Felix Müller, Emmanuel Keller, Eleftheria Papadaki, Saya Speidel, Bennet Borak, Steffi Herferth, Johannes Lange, Mario Lamping, Helene Michler, Juliane Dörfler, Anton Jacobshagen, Petra Kozma, Marinus Fislage, Wolf Rüdiger Brockhaus, Luisa Rothe, Pola Neuling, Ken-Dieter Michel, Zdravka Bosancic, Firas Nosirat, Maryam Kurpanik, Sophia Kuenz, Lukas-Sebastian Roediger, Irene Mergele, Anja Nottbrock, Leopold Rupp, Marie Graunke, Victoria Windmann. The authors further wish to thank the team of the student apprentices / interners of the Department of Anesthesiology at the Charité Universitätsmedizin Berlin. Magnet resonance imaging has been supported by the Berlin Center for Advanced Neuroimaging core staff, including Andrea Hassenpflug, Yvonne Kamm, Karl Bormann, Stefan Hetzer and Christian Labadie. Henning Krampe supported the study by recruiting and supervising students for neuropsychological testing. From the team of UMC Utrecht we thank our team of study nurses: Ada van Kampen, Gea, Sandra Numan and our team of students: Emily Tegnell, Lieke Hermans, Lara Mentink, Ellen Aarts, Rutger van de Leur, Rianne Tessers, Beatrijs Gelderblom, Carla Kraan, Ilona Bader, Dorian Brouwer, Jolien, Marielle de Vreede, Willem-jan Wreesman, Susan Haidari, Corinne Eertink, Prescilla Uijtewaal, Rebecca Hekking, Joyce van Loon, Michel Boon, Raoul Lieben, Yarit Wiggerts, Daan Kuppens, Aletta van den Bosch, Myriam Jaarsma-Coes, Rosa Smoor, Fienke Ditzel. Special thanks to Eline de Graaff (UMC Utrecht) for general research coordination. Special thanks to Niels Blanken for support in neuroimaging (UMC Utrecht). Michaela Renzulli at ALTA (Siena, Italy) provided additional administrative and coordinating services.

6 Funding

The research leading to these results has received funding from the European Union Seventh Framework Program [FP7/2007-2013] under grant agreement n° 602461.

7 Data availability

Individual participant data that underlie the results reported in this article, after de-identification (text, tables, figures, and appendices) may be made available upon request following publication to researchers who provide a methodologically sound proposal. Proposals should be directed to claudia.spies@charite.de and georg.winterer@charite.de. To gain access, requesting researchers will need to sign a data access agreement. Analyses will be limited to those approved in appropriate ethics and governance arrangements. All study documents which do not identify individuals (e.g. study protocol, informed consent form) will be freely available on request.

8 Conflict of interest

Georg Winterer and Claudia Spies are currently licensing a Class IIa medical device (web-based software tool for multivariate risk prediction of POD and POCD in clinical practice). This diagnostic software includes a pending patent application. Claudia Spies is director of the Department of Anesthesiology and Operative Intensive Care Medicine at the Charité with a long-standing research experience in postoperative delirium and neurocognitive disorders. Georg Winterer, the coordinator of the BioCog project, leadingly drafted the grant application and study protocol. In this regard, Claudia was critically involved in the clinical part of the study protocol. Winterer is the head of the Clinical Neuroscience Research Group at the Department of Anesthesiology and Operative Intensive Care Medicine at the Charité. He has extensive experience in biomarker-based risk prediction in schizophrenia, addiction and dementia research and development. He is also founder and CEO of PharmaImage Biomarker Solutions GmbH Berlin (Germany) and President of its subsidary

Pharmaimage Biomarkers Incl. (Cambridge, MA, USA) and PI Health Solutions GmbH Berlin (Germany).

Prof. Spies, Prof. Winterer, Dr. Boraschi, Dr. de Bresser, Morgado Correia, Prof. Dschietzig, Dr. Heinrich, Dr. Hoffmann, Dr. Janke, Dr. Kant, Dr. Kruppa, Prof. Kühn, Dr. Menon, Dr. Mörgeli, Dr. Mutsaerts, Dr. Pietzsch, Dr. Scholtz, Dr. Weber, Dr. Wiebach report grants from the European Commission during the conduct of the study. Prof. Winterer reports grants from the Deutsche Forschungsgemeinschaft/German Research Society and from the German Ministry of Health. Prof. Spies reports grants from the European Commission, from Drägerwerk AG & Co. KgaA, Deutsche Forschungsgemeinschaft/German Research Society, Deutsches Zentrum für Luft- und Raumfahrt e.V. (DLR)/German Aerospace Center, Einstein Stiftung Berlin/Einstein Foundation Berlin, Gemeinsamer Bundesausschuss/Federal Joint Committee (G-BA), Inneruniversitäre Forschungsförderung/Inner University Grants, Projektträger im DLR/Project Management Agency, Stifterverband für die deutsche Wissenschaft e.V./Non-Profit-Society Promoting Science and Education, WHOCC, Baxter Deutschland GmbH, Cytosorbents Europe GmbH, Edwards Lifesciences Germany GmbH, Fresenius Medical Care, Grünenthal GmbH, Masimo Europe Ltd., Pfizer Pharma PFE GmbH, personal fees from Georg Thieme Verlag, grants from Dr. F. Köhler Chemie GmbH, Sintetica GmbH, AGUETTANT Deutschland GmbH, AbbVie Deutschland GmbH & Co. KG, Amomed Pharma GmbH, InTouch Health, Copra System GmbH, Correvio GmbH, Max-Planck-Gesellschaft zur Förderung der Wissenschaften e.V., Medtronic, Philips Electronics Nederland BV, BMG and BMBF outside the study conduction. She has patents 10 2014 215 211.9, 10 2018 115364.8, 10 2018 110 275.5, 50 2015 010 534.8, 50 2015 010 347.7 and 10 2014 215 212.7 licensed. Gunnar Lachmann and Maria Heinrich report grants from the BIH Charité Clinician Scientist Program during the conduct of the study. Prof. Dschietzig reports personal fees from Immundiagnostik AG during the conduct of the study. Dr. Lammers-Lietz reports personal fees from PharmaImage GmbH during the conduct of the study. Dr. Hoffmann reports personal fees from Life&Brain GmbH outside the submitted work. Dr. Lachmann reports personal fees from Sobi outside

the submitted work. Dr. Mutsaerts reports grants from Dutch Heart Foundation, Horizon 2020/Eurostars, EU JPND during the conduct of the study. Dr. Heilmann-Heimbach is an employee of Life&Brain GmbH. Dr. Borchers reports grants from the European Commission, German Aerospace Center, Federal Joint Committee, Project Management Agency, Dr. F. Köhler Chemie GmbH, Sintetica GmbH and personal fees from the Charité BeST Simulation & Training Center. None of the other authors have a conflict of interest to declare.

9 References

1. American Psychiatric Association. Diagnostic and Statistical Manual. 5th ed. Washington DC: APA Press; 2013.

2. Androsova G, Krause R, Winterer G, Schneider R. Biomarkers of postoperative delirium and cognitive dysfunction. Front Aging Neurosci 2015;7:112.

3. Kotfis K, Szylińska A, Listewnik M, et al. Early delirium after cardiac surgery: an analysis of incidence and risk factors in elderly (≥65 years) and very elderly (≥80 years) patients. Clin Interv Aging 2018;13:1061–70.

4. Robine J-M, Cambois E. Healthy life expectancy in Europe. Population Societies 2013;No 499(4):1–4.

5. World Health Organization. World report on Ageing and Health [Internet]. Geneva, Switzerland: WHO Press; 2015. Available from: https://apps.who.int/iris/bitstream/handle/10665/186463/9789240694811_eng.pdf;jsessionid=B1F84CF274FDAEC798FA94D46F57C8B5?sequence=1

6. Robinson TN, Raeburn CD, Tran ZV, Angles EM, Brenner LA, Moss M. Postoperative delirium in the elderly: risk factors and outcomes. Ann Surg 2009;249(1):173–8.

7. Rudolph JL, Marcantonio ER, Culley DJ, et al. Delirium is associated with early postoperative cognitive dysfunction. Anaesthesia 2008;63(9):941–7.

8. Steinmetz J, Christensen KB, Lund T, Lohse N, Rasmussen LS, ISPOCD Group. Long-term consequences of postoperative cognitive dysfunction. Anesthesiology 2009;110(3):548–55.

9. Humeidan ML, Reyes J-PC, Mavarez-Martinez A, et al. Effect of Cognitive Prehabilitation on the Incidence of Postoperative Delirium Among Older Adults Undergoing Major Noncardiac Surgery: The Neurobics Randomized Clinical Trial. JAMA Surgery 2021;156(2):148–56.

10. Paredes S, Cortínez L, Contreras V, Silbert B. Post-operative cognitive dysfunction at 3 months in adults after non-cardiac surgery: a qualitative systematic review. Acta Anaesthesiol Scand 2016;60(8):1043–58.

11. van Eijk MMJ, van Marum RJ, Klijn IAM, de Wit N, Kesecioglu J, Slooter AJC. Comparison of delirium assessment tools in a mixed intensive care unit*. Critical Care Medicine 2009;37(6):1881.

12. Gaudreau J-D, Gagnon P, Harel F, Tremblay A, Roy M-A. Fast, systematic, and continuous delirium assessment in hospitalized patients: the nursing delirium screening scale. J Pain Symptom Manage 2005;29(4):368–75.

13. Inouye SK, van Dyck CH, Alessi CA, Balkin S, Siegal AP, Horwitz RI. Clarifying confusion: the confusion assessment method. A new method for detection of delirium. Ann Intern Med 1990;113(12):941–8.

14. Ely EW, Margolin R, Francis J, et al. Evaluation of delirium in critically ill patients: validation of the Confusion Assessment Method for the Intensive Care Unit (CAM-ICU). Crit Care Med 2001;29(7):1370–9.

15. Rasmussen LS, Larsen K, Houx P, et al. The assessment of postoperative cognitive function. Acta Anaesthesiol Scand 2001;45(3):275–89.

16. Spies CD, Knaak C, Mertens M, et al. Physostigmine for prevention of postoperative delirium and long-term cognitive dysfunction in liver surgery: A double-blinded randomised controlled trial. Eur J Anaesthesiol 2021;

17. Feinkohl I, Borchers F, Burkhardt S, et al. Stability of neuropsychological test performance in older adults serving as normative controls for a study on postoperative cognitive dysfunction. BMC Res Notes 2020;13(1):55.

18. Clerx L, Visser PJ, Verhey F, Aalten P. New MRI markers for Alzheimer's disease: a meta-analysis of diffusion tensor imaging and a comparison with medial temporal lobe measurements. J Alzheimers Dis 2012;29(2):405–29.

19. Moller JT, Cluitmans P, Rasmussen LS, et al. Long-term postoperative cognitive dysfunction in the elderly ISPOCD1 study. ISPOCD investigators. International Study of Post-Operative Cognitive Dysfunction. Lancet 1998;351(9106):857–61.

20. Rudolph JL, Jones RN, Rasmussen LS, Silverstein JH, Inouye SK, Marcantonio ER. Independent vascular and cognitive risk factors for postoperative delirium. Am J Med 2007;120(9):807–13.

21. Katlic MR, Coleman J, Khan K, Wozniak SE, Abraham JH. Sinai Abbreviated Geriatric Evaluation: Development and Validation of a Practical Test. Ann Surg 2019;269(1):177–83.

22. Kazmierski J, Kowman M, Banach M, et al. Incidence and predictors of delirium after cardiac surgery: Results from The IPDACS Study. J Psychosom Res 2010;69(2):179–85.

23. Schmitt EM, Saczynski JS, Kosar CM, et al. The Successful Aging After Elective Surgery Study: Cohort Description and Data Quality Procedures. Journal of the American Geriatrics Society 2015;63(12):2463–71.

24. Racine AM, Fong TG, Travison TG, et al. Alzheimer's-related cortical atrophy is associated with postoperative delirium severity in persons without dementia. Neurobiol Aging 2017;59:55–63.

25. Fong TG, Vasunilashorn SM, Ngo L, et al. Association of Plasma Neurofilament Light with Postoperative Delirium. Annals of Neurology [Internet] [cited 2020 Oct 18];n/a(n/a). Available from: https://onlinelibrary.wiley.com/doi/abs/10.1002/ana.25889

26. Shah S, Weed HG, He X, Agrawal A, Ozer E, Schuller DE. Alcohol-related predictors of delirium after major head and neck cancer surgery. Arch Otolaryngol Head Neck Surg 2012;138(3):266–71.

27. Heil T, Spies C, Bullmann C, et al. [The relevance of CDT (carbohydrate-deficient transferrin). Preoperative diagnosis of chronic alcohol abuse in intensive care patients following elective tumor resection]. Anaesthesist 1994;43(7):447–53.

28. Jayanama K, Theou O, Blodgett JM, Cahill L, Rockwood K. Frailty, nutrition-related parameters, and mortality across the adult age spectrum. BMC Med 2018;16(1):188.

29. Ferrucci L, Fabbri E. Inflammageing: chronic inflammation in ageing, cardiovascular disease, and frailty. Nature Reviews Cardiology 2018;15(9):505–22.

30. Remick DG. Interleukin-8. Critical Care Medicine 2005;33(12):S466.

Table 1: Sample description of the entire sample (N=933)

		Median (IQR)	Range (min.-max.)
Age (years)		72 (68-76)	65-91
MMSE score (points)		29 (28-30)	24-30
GDS		1 (0-3)	0-12.86
Duration of anesthesia (min)		204 (122-308)	10-1669
Duration of surgery (min)		102 (55-191)	3-594
Duration of hospital stay (days)		6 (3-10)	0-131
Duration of ICU stay (days)		0 (0-0.2)	0-55
		Absolute n	Relative frequency
PreCI		122/928	13%
POD		184/929	20%
POCD at 3 months		66/641	10%
Site of surgery	Intracranial*	10/912	1%
	Intrathoracic, -abdominal, -pelvic	397/912	44%
	peripheral	505/912	55%
Type of anesthesia	general	689/913	76%
	regional	56/913	6%
	combined	168/913	18%
ASA-PS	I	36/933	4%
	II	559/933	60%
	III	337/933	36%
	IV	1/933	<1%
Women		395/933	42%
ISCED level	1+2	152/843	18%
	3+4	343/843	41%
	5+6	348/843	41%

IQR: interquartile range; MMSE: Mini-Mental Status Examination; GDS: geriatric depression scale; ICU: intensive care unit; PreCI: preoperative cognitive impairment; POD: postoperative delirium; POCD: postoperative cognitive dysfunction; ASA-PS: American Society of Anesthesiologists Physical Status; ISCED: International Standard Classification for Education

*intracranial surgery not affecting brain parenchyma (e.g. meningioma).

Figure 1: BioCog study procedure. Patients were recruited at the anesthesiology outpatient clinics of the study centers. Within two weeks before scheduled surgery, baseline clinical, cognitive, laboratory and neuroimaging data as well as blood samples were collected. After surgery, patients were screened for POD twice daily for up to a maximum of seven days or until discharge from hospital. Patients were invited for follow-up assessment three months after surgery. During this visit participants underwent neurocognitive testing, blood sampling neuroimaging and clinical assessments. To maximize the number of patients returning for a follow-up investigation, assessments were allowed to take place two months after surgery at the earliest and six months after surgery at the latest.

Figure 2: Patient flow chart (A) and absolute prevalences of preoperative cognitive impairment (PreCI), delirium postoperative cognitive dysfunction (POCD) with missing values (B). Arrows with numbers indicate the number of patients with PreCI and POD who developed POD or POCD, respectively, or were lost to follow-up.

Figure 3: Summary of parameters that were significantly associated with POD (left box, green) and POCD (right box, red). Odds ratios (OR) with 95% confidence interval (95% CI) are shown (only parameters are depicted with CI excluding unity). The diameter of the circle corresponds to the number of available datasets. For details on sample size and numeric values, see supplementary materials 3 and 4.

* Tumor includes diagnoses of leukemia and lymphoma.

** The upper limits for the 95% CI have been truncated. The true upper limits are 105.8 for the association of IL8 and POD and 18·2 for Immature granulocytes and POCD.

Abbreviations:

adj.: adjusted for assessment in different study centers, p: points

age & comorbidity: ASA: American Society of Anesthesiologists Physical Status, CAD: coronary artery disease CCI: Charlson comorbidity index

inflammation: CRP: C-reactive protein, imm. granuloc.: immature granulocytes

cognition: GPT: Grooved Pegboard Test (completion time), MMSE: Mini-mental status examination, preop. cogn. impairment: preoperative cognitive impairment, VRM: Verbal Recognition Memory

functionality & geriatric assessment: GDS: Geriatric depression scale, MNA-SF: Mini-nutritional assessment short form, TUG: Timed up-and-go test, frailty refers to Fried's frailty phenotype

metabolic: BMI: body mass index, HDL: high density lipoprotein, LDL: low density lipoprotein

drugs & medication: GGT: γ-glutamyltransferase, BDZ: preoperative longterm prescription of benzodiazepines

neuroimaging: CA: cornu ammonis, dia.: diameter (cortical thickness), FA: fractional anisotropy, HPC: hippocampus, vol.: volume

Figure 4: Top: The AP recall curves for the different fine-tuned GBT POD prediction models as well as logistic regressions (LogReg) as a benchmark are shown. Different models are compared (clinical, blood, sparse imaging domains and domains combined). Clinical refers to all clinical assessments, including sociodemographic data, preoperative diagnoses and medication as well as neuropsychological testing. "Blood" refers to all blood-based laboratory parameters including experimental and routine clinical parameters. "Sparse imaging" is a set of neuroimaging markers (hippocampus volume, brain volume, volume of nucleus Basalis of Meynert). Additionally, models without (left) and with (right) precipitating factors (precipitants) are compared, i.e., basic models and models including precipitants. Precipitants are perioperative precipitating factors: site of surgery, duration of anesthesia, uncontrolled post-operative pain, post-operative anticholinergic medication. For the different data types, the models were trained and evaluated separately. The binarization

threshold selected per model to calculate precision, recall, and specificity are those that bring the model closest to a sensitivity of 95%. The AP metrics are averaged across a tenfold test split and presented with their respective standard errors (SE). The shaded regions display the SE derived from ten trials. The dashed lines correspond to the logistic regressions and the full lines to the GBT. Colors encode the used data types. In supplement 4, we also present receiver operating characteristics (ROC) with are under the curves (AUC) metric as well as details on the AP metrics. As mentioned in the methods section (supplement 2), sensitivity and specificity of the ROC are less informative compared to the AP recall curves when given imbalanced data as in our case.

Bottom: Feature importances of the GBT model trained on clinical, laboratory, sparse neuroimaging data without precipitants, i.e. basic model (left) and with precipitants (right) averaged over a tenfold test split. To reduce clutter, only features were shown that in sum explain 90% of the importance. Note: In all models, age, sex, weight, height and BMI are included as static parameters. It is obvious from the model comparisons that GBT can help to improve the models in certain cases. Precipitating factors obviously play a major role, in particular duration of surgery and duration of anesthesia, but also uncontrolled post-operative pain. Note: post-operative anticholinergic medication does not appear to be a major factor for POD risk.

Abbreviations: vol: volume, IL: interleukin, IL8, EDTA: interleukin 8 from EDTA-anticoagulated whole blood, MMSE: Mini-Mental-Status-Examination-Test score, ASA-PS: Physical status classification system of the American Society of Anesthesiologists (ASA), BDZ: benzodiazepines, S100A12: S100 calcium-binding protein A12, CCI: Charlson comorbidity index, MDA: malondialdehyde, EQ5D: Life-Quality, HDL: high-density lipoprotein, leptinRec: leptin receptor, TP4240: total plasma Aβ42/40 ratio, GGT: γ-glutamyltransferase, CRP: C-reactive protein, MNA-SF: Mini-Nutritional Assessment (short form), Triglyc: triglycerides, ISCED: International Standard Classification of Education score, HbA1c: hemoglobin A1c, ALAT: alanine aminotransferase, IADL: Instrumental Activities of Daily Living, ADMA:

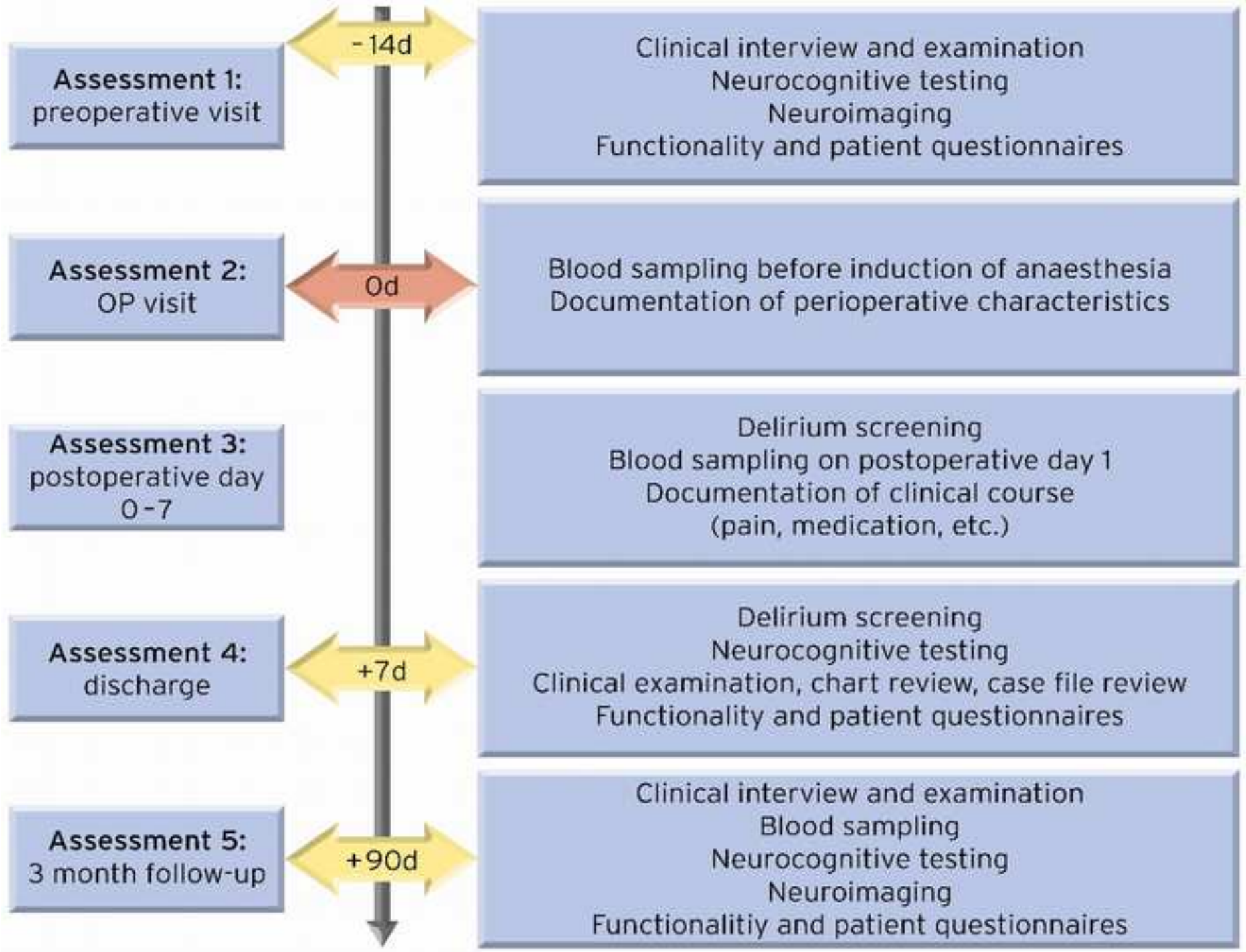
Asymmetric dimethylarginine, NIDDM: non insulin dependent diabetes mellitus, leptinSLR: soluble leptin receptor.

*Tumor includes preoperative leukemia and lymphoma.

Figure 5: Survival curves for preoperative cognitive impairment (N=678, top), POD (postoperative delirium) (N=683, middle) and the interaction of both (N=678, bottom). The respective inlay is a magnification of the larger plot. The log-rank test indicates a statistically significant difference in survival for patients with POD ($X^2(1)=38{\cdot}5$, $p<0{\cdot}001$), but not for patients with preoperative impairment ($X^2(1)=1{\cdot}9$, $p=0{\cdot}2$). Compared to all other groups, patients with both preoperative cognitive impairment and POD have a significantly lower survival rate ($X^2(1)=15{\cdot}4$, $p<0{\cdot}001$), but compared to other patients with POD, those with previous cognitive impairment have no higher mortality ($X^2(1)=0{\cdot}3$, $p=0{\cdot}6$).

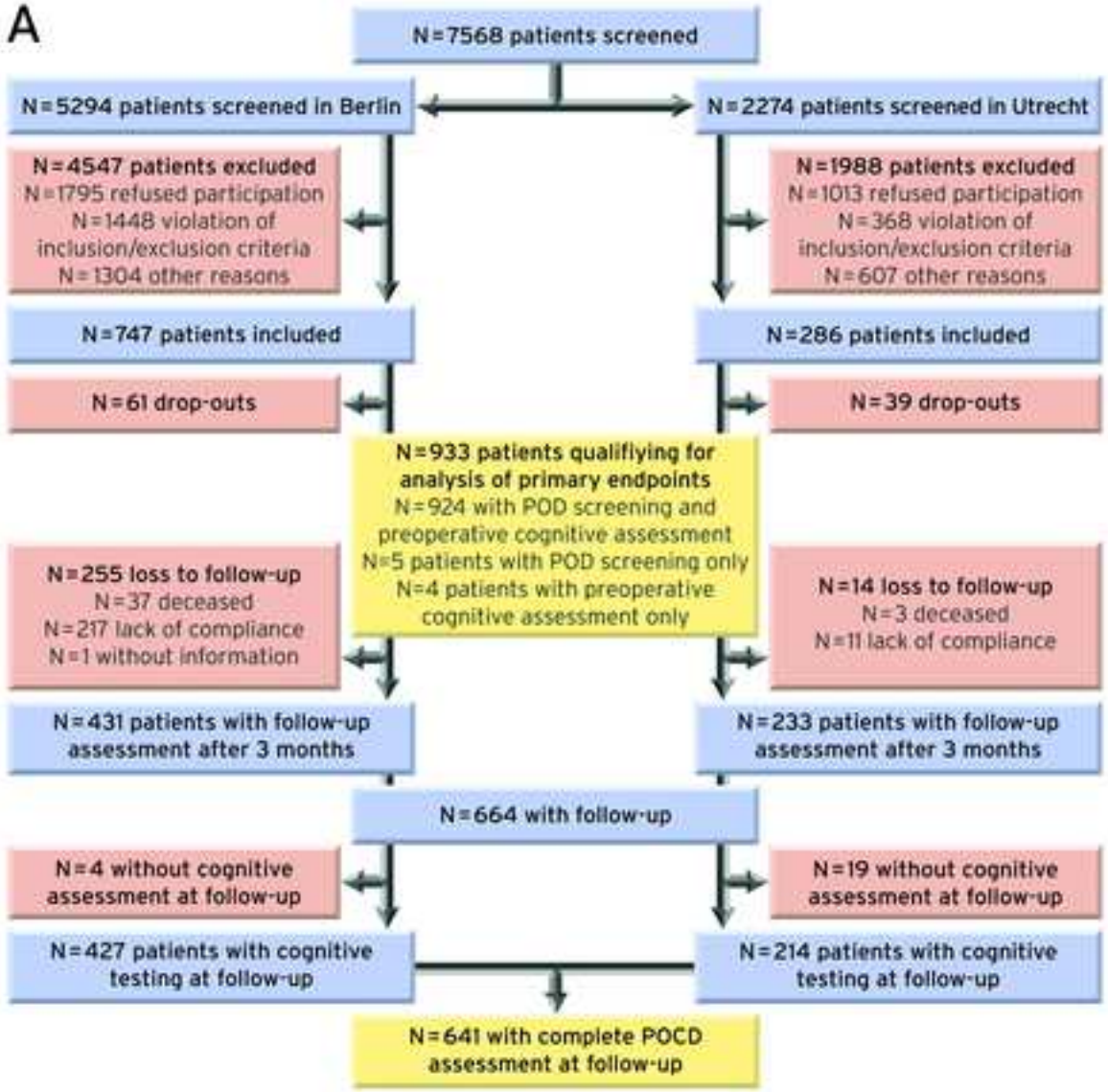
Figure

Click here to access/download;Figure;Figure 1.jpg



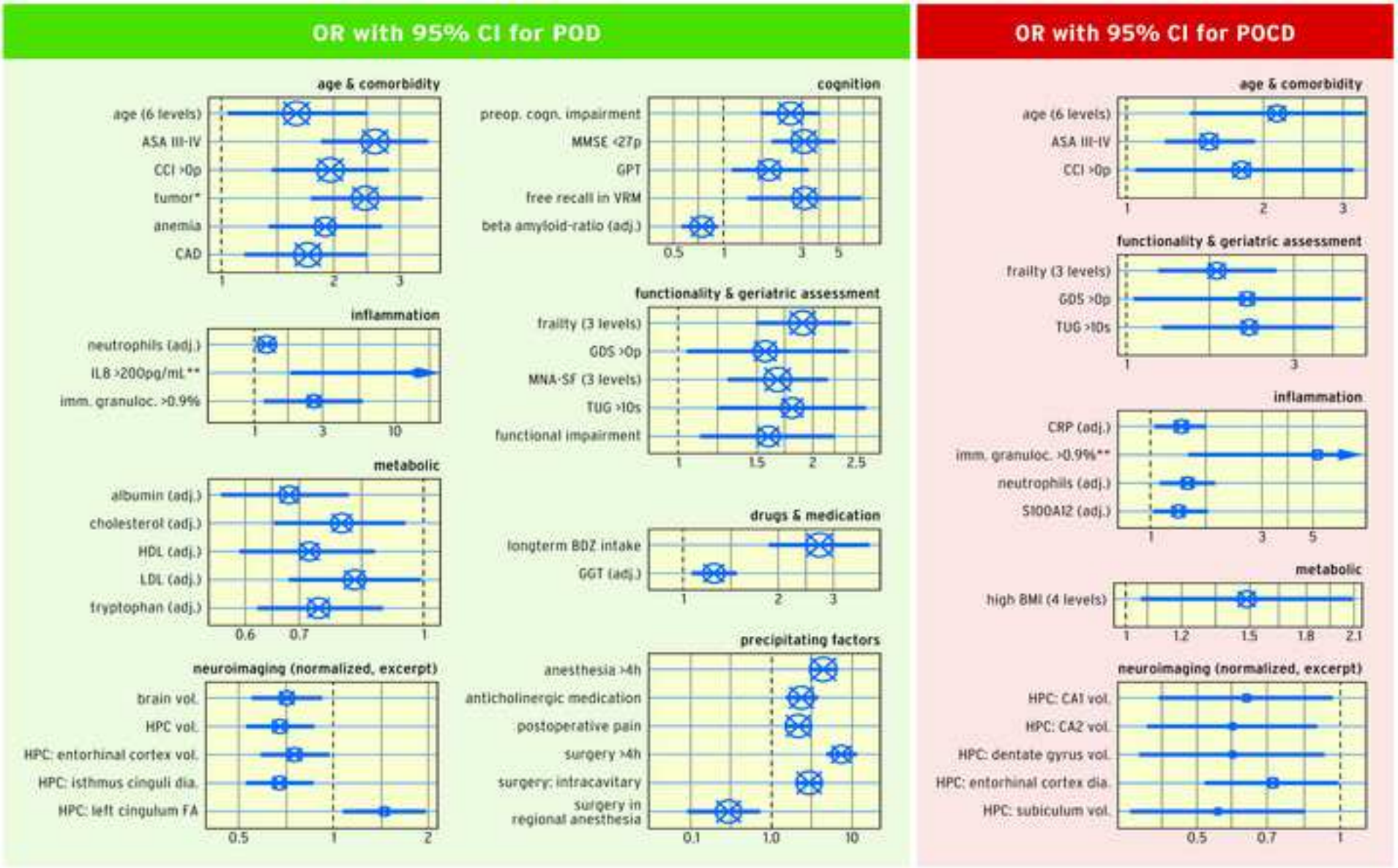
Figure

Click here to access/download;Figure;Figure 2-fin_750dpi.jpg



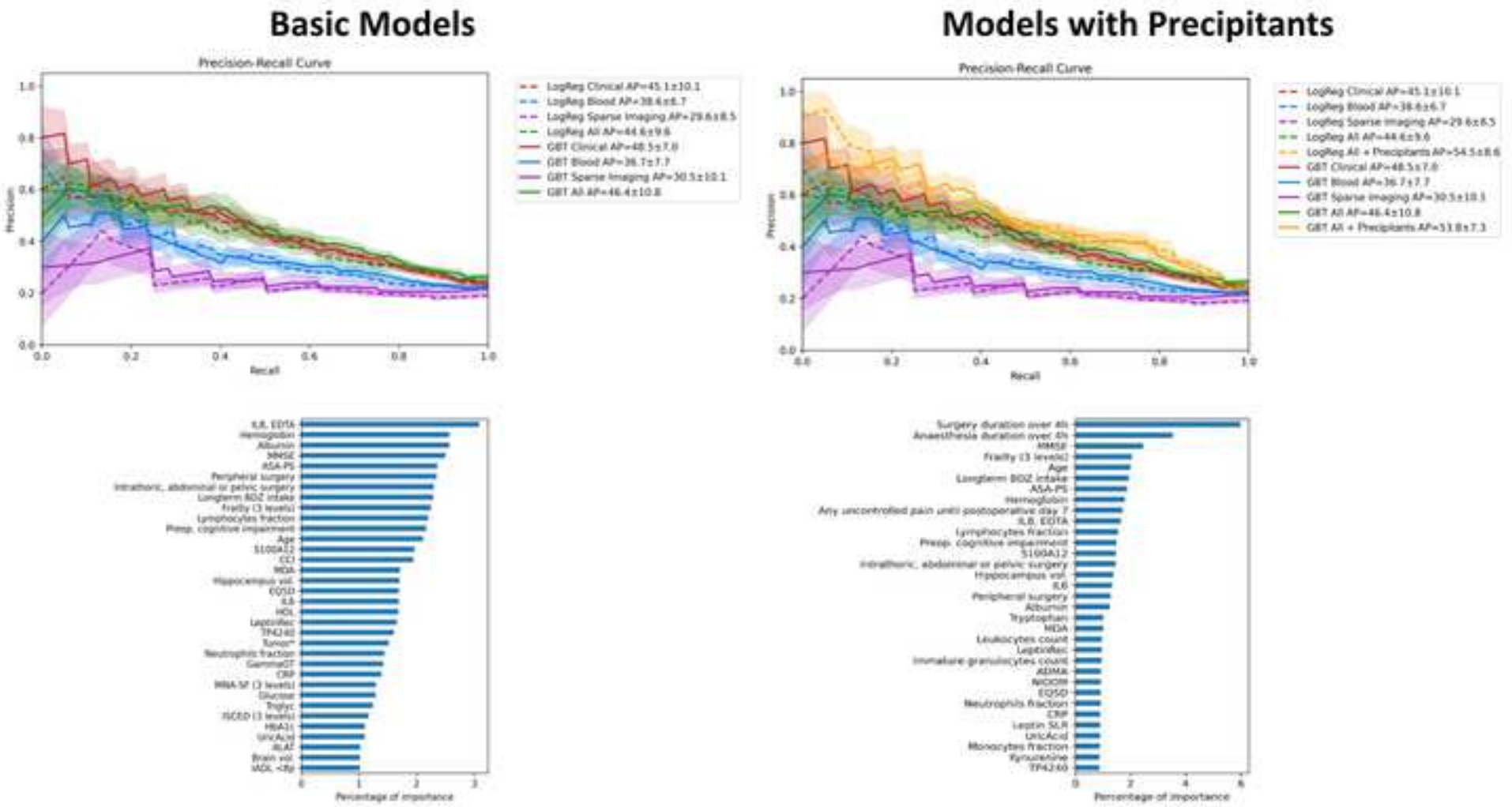
Figure

Click here to access/download;Figure;Figure 3_final 5-19052021_1000dpi.jpg



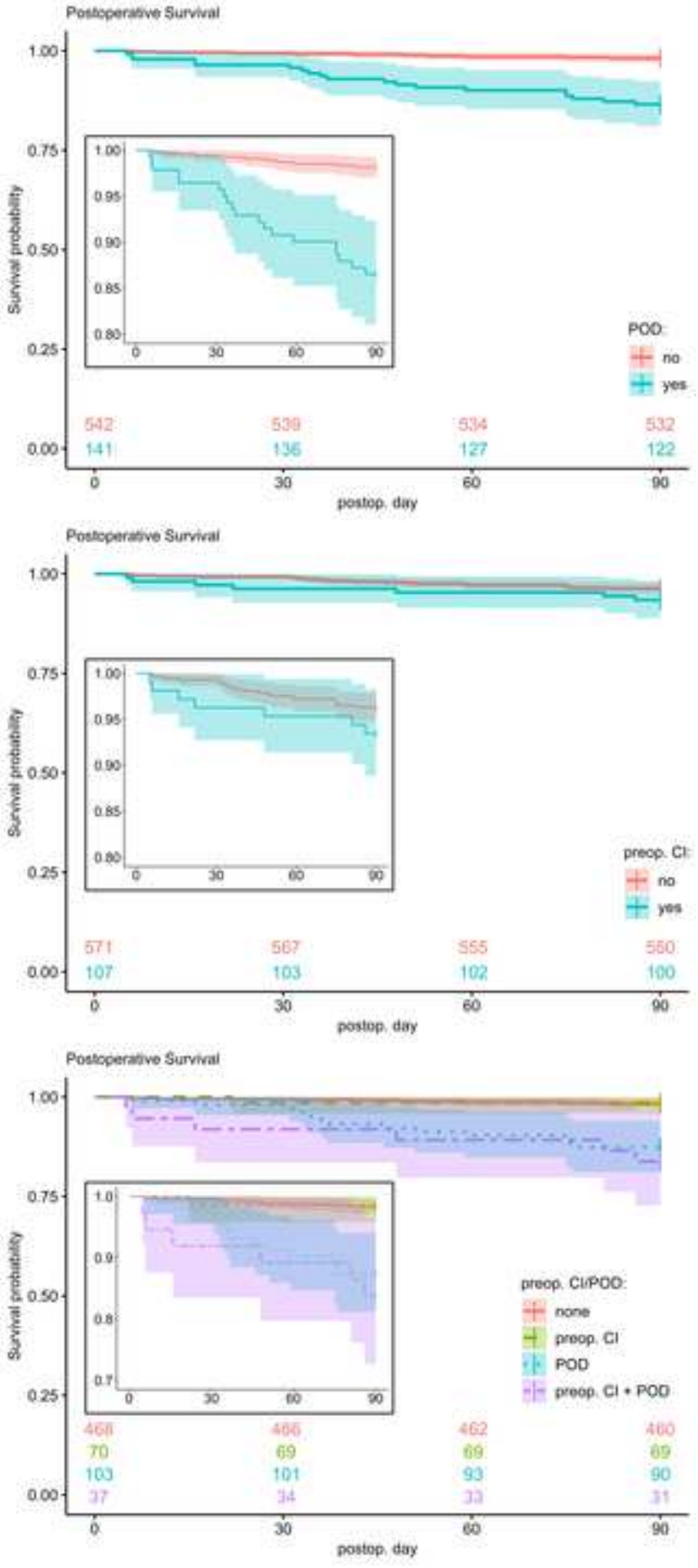
Figure

Click here to access/download;Figure;Figure 4_adalab_update_fixed.jpg



Figure

Click here to access/download;Figure;Figure 5-fin_1000dpi.jpg



B.4. Manuscript - Comparative effectiveness of antiepileptic drug combination therapy based on mode of action

I supported earlier works on the data set in my Master thesis. I discontinued to work on the manuscript for separation of work for the Master (Silvennoinen et al. 2019) and the PhD. This manuscript is in preparation.

Comparative effectiveness of antiepileptic drug combination therapy based on mode of action

Milena Zizovic, Emadeldin Hassanin, Nikola de Lange, The EpiPGX consortium, Sanjay Sisodiya, Chantal Depondt, Roland Krause

2023-02-28

Introduction

Antiepileptic drugs (AED) are frequently administered as combination therapy. Combinations of AEDs with different modes of action (MOA) are sometimes considered to deliver the best balance of reduction in seizure frequency with minimal adverse drug responses (ADR). Certain AED combinations have been described in animal experiments as particularly effective, notably the combination of levetiracetam and topiramate (Schidlitzki et al. 2020). In the clinical application, combination therapies are more complex to study much insight is provided from analysis of retrospective data[citations]. A previous study (Margolis et al. 2014) relied on electronic health records of 90,000 cases and showed that AED combinations using different MOAs are more effective. Hospital records often do not provide adequate evidence of the success of an AED. Therefore, the comparative effectiveness between AED combinations was assessed by the persistence of the AEDs in Margolis et al. (2014), the best proxy variable available to those authors. Here we describe the effectiveness of AEDS combinations in the EpiPGX cohort of over 12,000 cases of European descent. We evaluate the differences in outcomes comparing different-MOA combination therapy with same-MOA combination therapy of people with epilepsy, based on seizure reduction and adequacy of each AED trial.

Methods

AED categorization

AEDs reduce seizures through different MOA which can be categorized into: (1) Modulation of voltage-gated channels either by blocking such as of sodium, and calcium channels, or enhancing opening such as for potassium channels; (2) Enhancement of GABA inhibition either acting at the level of GABA A receptors, GAT1 (GABA transporter), or GABA transaminase; (3) Modulation of synaptic release through SV2A; (4) Multiple modes of action.

The aim of these MOA is to control abnormal inhibitory or excitatory neurotransmission on different levels (Brodie 2010; Johannessen and Landmark 2010; Rogawski and Lo 2016). The summary of AEDs below shows the drugs included in the analyses of the EpiPGX cohort for each of the four categories (Table 1).

Table 1: Antiepileptic drugs and their mode of action classification.

Mode of action (MOA)	Antiepileptic drug	AED
Sodium channel blocker (SCB)	Carbamazepine	CBZ
	Eslicarbazepine	ESL
	Ethotoin	ETN
	Oxcarbazepine	OXC
	Phenytion	PHT
	Lamotrigine	LTG
	Lacosamide	LCM
Calcium channel blocker (CCB)	Ethosuximide	ESM
GABA analogs (GA)	Benzodiazepines	BDZ
	Gabapentin	GBP
	Phenobarbital	PB
	Pregabalin	PGB
	Primidon	PRM
	Tiagabine	TGB
	Vigabatrin	VGB
SV2A binding (SV)	Levetiracetam	LEV
Multiple mechanisms (MM)	Felbamate	FBM
	Valproate	VPA
	Topiramate	TPM
	Zonisamide	ZNS

Participants

The EpiPGX study cohort consists of 12822 patients and more than 39,000 AEDs trials. Patients were recruited in specialized epilepsy clinics in Belgium, Germany, Ireland, Italy, the Netherlands, the United Kingdom, and Australia. There were 10146 patients using AEDs. Patients who were receiving monotherapy only were excluded:
only those who were receiving two or more AEDs were selected. The earliest recorded drug treatment in the data set was dated June 2, 1993, the latest July 2, 2016. We selected 2481

patients with concomitant use of two or more AEDs for more than 90 days, accounting for 6087 combination trials. The cut-off of 90 days was based on our previous studies, including Margolis 2014. Concomitance was defined by a number of days overlapping with continuous use of the second AED. AED combinations with fewer than 5 trials were excluded from analyses. AEDs thus not included are shown in Supplementary (Supplementary Table 1). Furthermore, focal epilepsy cases and generalized epilepsy cases were analyzed separately. The focal group included 1650 people with 4273 combination trials and the group with generalized epilepsy had 473 patients and 912 combination trials (Figure 1). Thirty patients diagnosed with epilepsy with both localized and generalized features which were not included in the analyses.

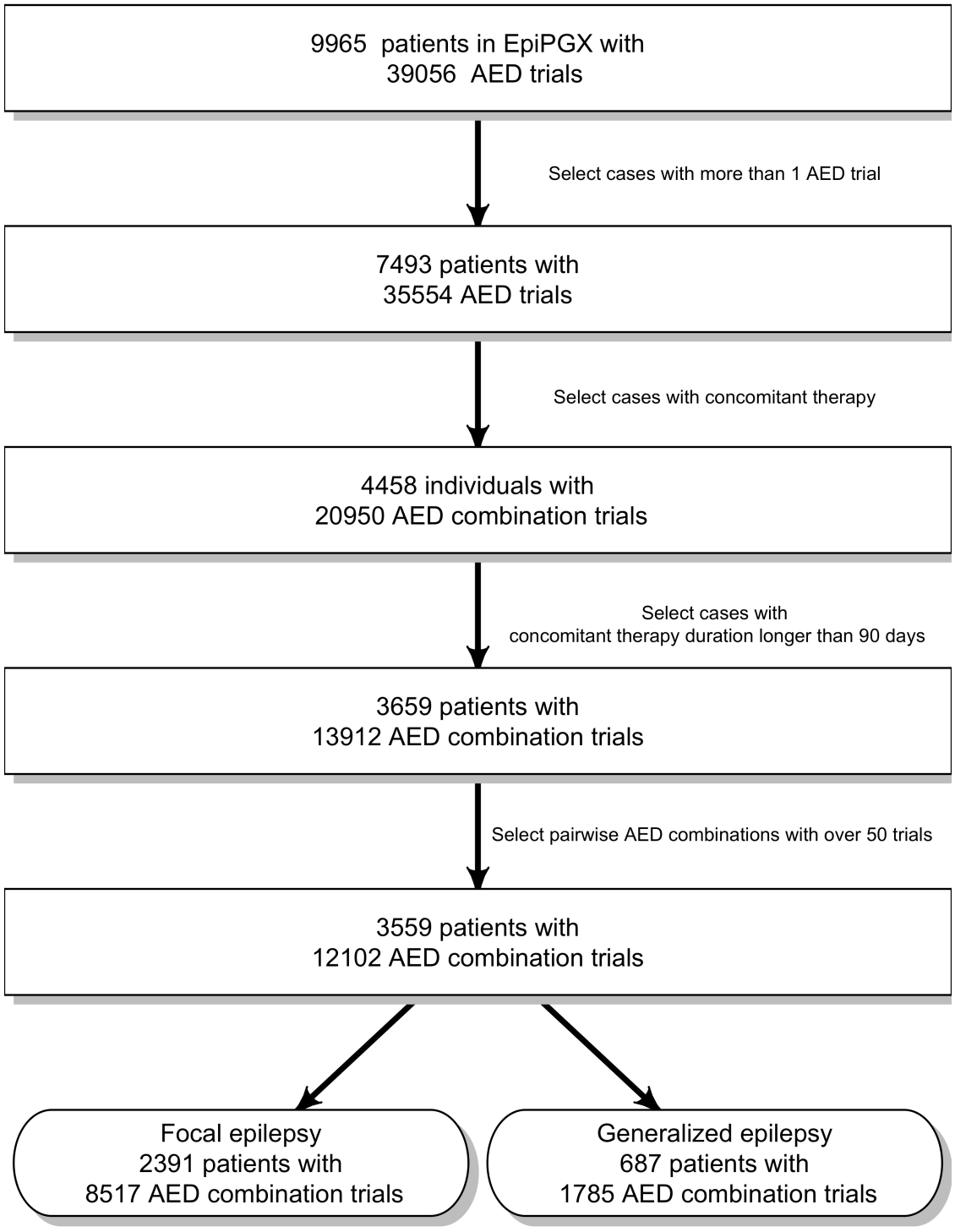


Figure 1: **Flowchart for the inclusion of patients and AED trials.** The number of patients, AED trials and pairwise combination trials for which defined data were available is indicated for each step. Abbreviations: AED - antiepileptic drug.

Analytical workflow

Concomitant therapy of more than two drugs was decomposed and analyzed in pairwise combinations. The analyses consider frequency of AED or MOA application and frequency of AED or MOA application in pairwise combinations applied in cohorts. For example, in estimation of VPA application frequency VPA would be counted once while in frequencies of VPA in pairwise combinations it would be counted three times (Figure 2).

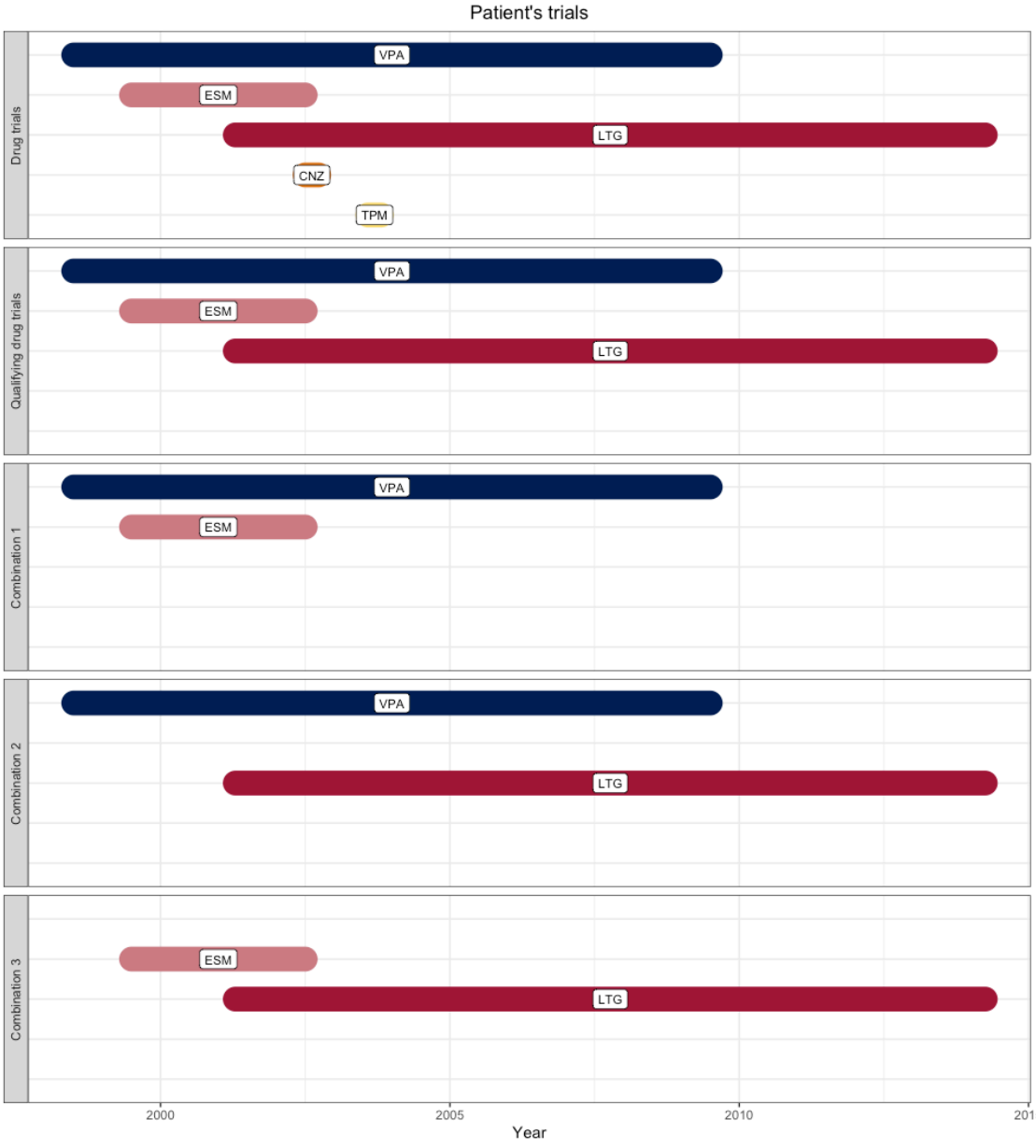


Figure 2: ***Drug trials from 1998 until 2014 in an individual with epilepsy showing therapy concomitance as an example of the analytical flow.*** *Concomitant therapy of more than two drugs was decomposed and analyzed in pairwise combinations. AEDs with short durations of treatment, (CNZ and TPM in the above example) were excluded. Abbreviations: CNZ- Clonazepam, ESM- Ethosuximide, LTG- Lamotrigine, TPM- Topiramate, VPA- Valproate.*

Data distribution and drug persistence

Drug persistence is defined as the duration of time from initiation to discontinuation of therapy. Previous studies reported that most of the survival data were skewed: our data are also skewed (Supplementary Figure 1). Moreover, a simple comparison of therapy duration does not consider if a trial is still ongoing or interrupted. This led to conducting survival analysis such as Kaplan-Meyer estimation, widely used as a nonparametric estimator, and Cox proportional hazard regression model which considers ongoing trials.

Survival analysis

Kaplan Meyer and Cox proportional hazard regression model, two models were made – GA combination model comparing all combinations containing a GA AED among themselves, and SCB model comparing all combinations containing a SCB drug. GA+GA and SCB+SCB were used as a reference respectively, to compare same MOA based combinations to different MOA based combinations in each model. Kaplan Meyer survival analysis generates survival curves of persistence probability over time, using median persistence time to compare combinations by 50% probability of therapy continuance. Cox proportional hazard regression models generate hazard ratios to compare combinations by probability to be discontinued. The risk of the therapy discontinuation is assessed comparing hazard ratios of all GA combinations with the reference HR and the same is done for SCB combinations. Combination with a hazard ratio less than 1 is considered as less likely to be discontinued. In both tests, the p-value is used to estimate statistical significance of the shown difference. The survival analysis was performed in R.

Results

Study sample

Table 2: Frequencies of MOA-based trials among 9965 people in EpiPGX, 6109 diagnosed with focal epilepsy and 2097 patients diagnosed with generalized epilepsy.

MOA	All cases trials	All cases frequency	Focal cases trials	Focal cases frequency	Generalized cases trials	Generalized cases frequency
CCB	564	0.014	147	0.006	341	0.051
GA	8045	0.206	5866	0.227	948	0.141
MM	9819	0.252	5615	0.217	2563	0.382
SCB	15811	0.406	11059	0.427	2049	0.306
SV	4743	0.122	3187	0.123	805	0.120

Table 3: Frequencies of AED trials among 9965 people in EpiPGX, 6109 diagnosed with focal epilepsy and 2097 patients diagnosed with generalized epilepsy.

MOA	AED	All cases trials	All cases frequency	Focal cases trials	Focal cases frequency	Generalized cases trials	Generalized cases frequency
CCB	Ethosuximide	564	0.014	147	0.006	341	0.051
GA	Clobazam	2109	0.054	1423	0.055	281	0.042
GA	Clonazepam	644	0.017	351	0.014	128	0.019
GA	Diazepam	171	0.004	100	0.004	19	0.003
GA	Gabapentin	1067	0.027	897	0.035	46	0.007
GA	Lorazepam	37	0.001	27	0.001	8	0.001
GA	Nitrazepam	12	0.000	9	0.000	1	0.000
GA	Phenobarbital	1572	0.040	1076	0.042	288	0.043
GA	Pregabalin	716	0.018	597	0.023	22	0.003
GA	Primidone	429	0.011	311	0.012	74	0.011
GA	Tiagabine	269	0.007	247	0.010	12	0.002
GA	Vigabatrin	1019	0.026	828	0.032	69	0.010
MM	Felbamate	74	0.002	51	0.002	13	0.002
MM	Topiramate	2687	0.069	1888	0.073	446	0.067
MM	Valproate	6200	0.159	3100	0.120	1930	0.288
MM	Zonisamide	858	0.022	576	0.022	174	0.026
SCB	Carbamazepine	5664	0.145	4201	0.162	491	0.073
SCB	Eslicarbazepine	131	0.003	113	0.004	13	0.002
SCB	Ethotoin	1	0.000	1	0.000	0	0
SCB	Lacosamide	903	0.023	678	0.026	58	0.009
SCB	Lamotrigine	5285	0.136	3171	0.123	1165	0.174
SCB	Oxcarbazepine	1433	0.037	1089	0.042	93	0.014
SCB	Phenytoin	2394	0.061	1806	0.070	229	0.034
SV	Levetiracetam	4743	0.122	3187	0.123	805	0.120

AED trial distribution by person-years

Some drugs are indicated for only generalized but not focal epilepsies such as Ethosuximide (BROWNE et al. 1975).

Number of AED prescriptions per year from 1970-2015

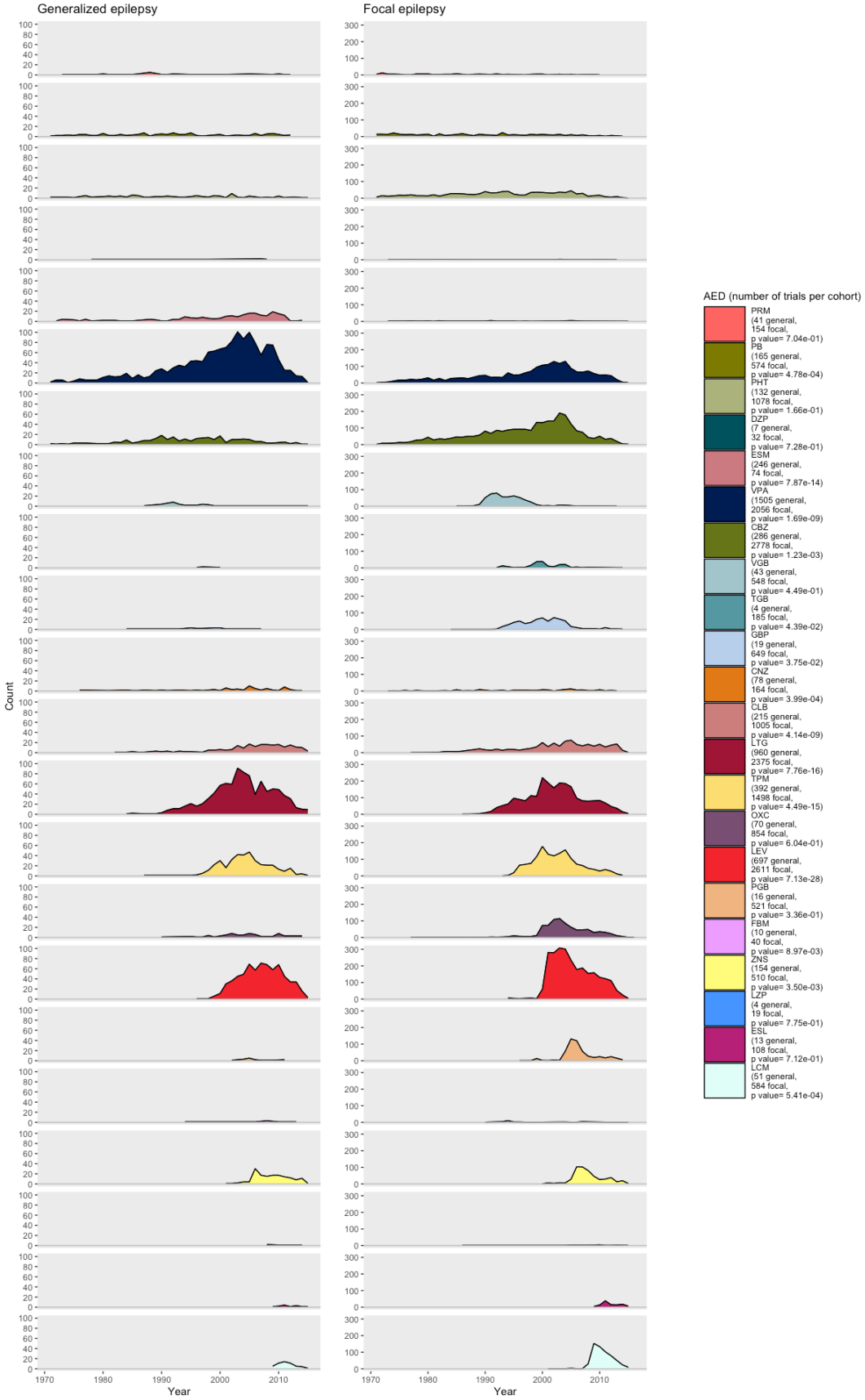


Figure 3: Comparison of AED use in focal and generalized epilepsy patients by the treatment start year

The plots show count of trials started per year for each AED in focal and generalized epilepsy in the EpiPGX cohort. The AEDs normally recommended for treatment of focal epilepsy have a lower count in generalized epilepsy cohort than in focal over time.

To further verify the statistical significance we performed a Mann Whitney U non-parametric test since the data is non-uniform. The difference in the number of an AED trial started per year was statistically significant between the two cohorts for 16 out of 22 drugs (p values < 0.05) (Figure 3).

Survival analysis

Survival analysis comparing two groups - same-MOA versus different-MOA based combinations

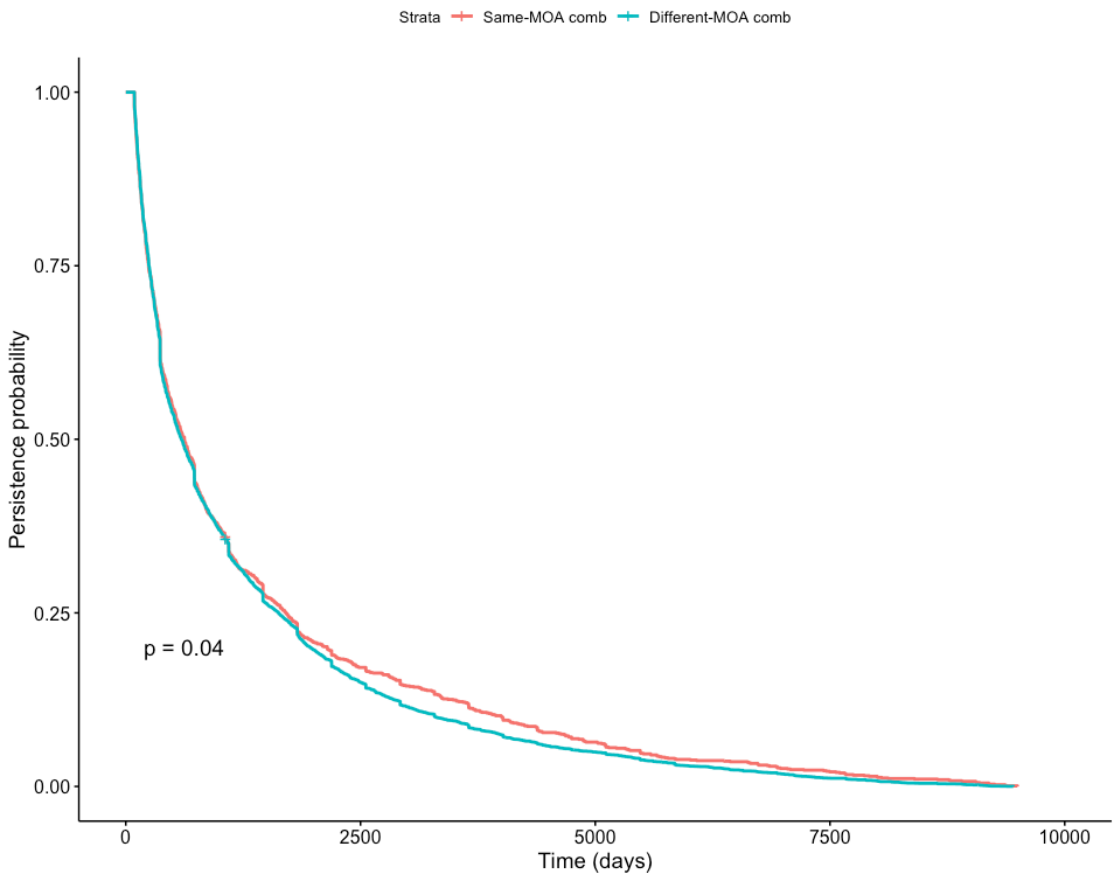


Figure 4: Kaplan Meyer survival curves of persistence comparisons among same- and different-MOA based combinations in all epilepsy cases.

Table 4: Kaplan Meyer summary for same- and different-MOA based combinations in all epilepsy cases. The table includes the number of trials, number of events and median persistence time.

MOA combination	Records	Events	Median	0.95LCL	0.95UCL
Same-MOA combinations	1240	1239	623	548	704
Different-MOA combinations	4760	4758	595	554	630

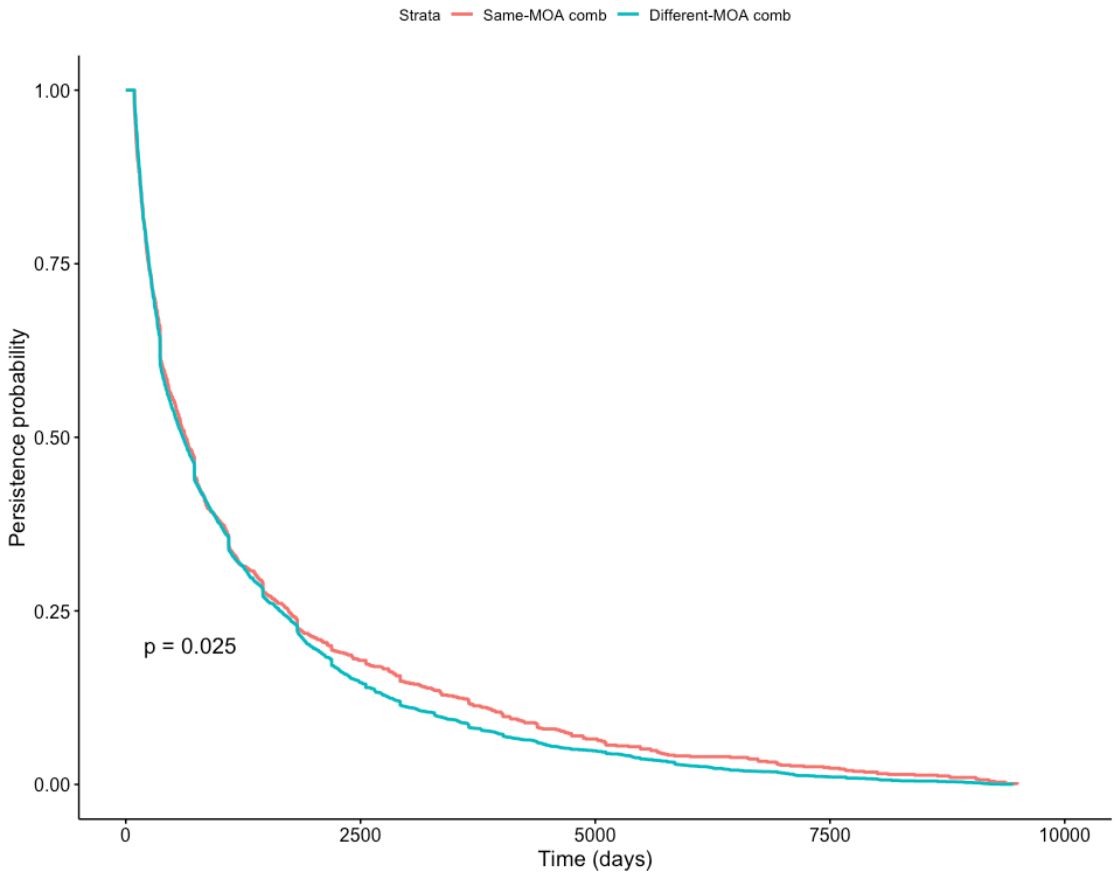


Figure 5: Kaplan Meyer survival curves of persistence comparisons among same- and different-MOA based combinations in focal epilepsy cases.

Table 5: Kaplan Meyer summary for same- and different-MOA based combinations in focal epilepsy cases. The table includes the number of trials, number of events and median persistence time.

MOA combination	Records	Events	Median	0.95LCL	0.95UCL
Same-MOA combinations	901	901	639	564	730
Different-MOA combinations	3305	3305	609	558	657

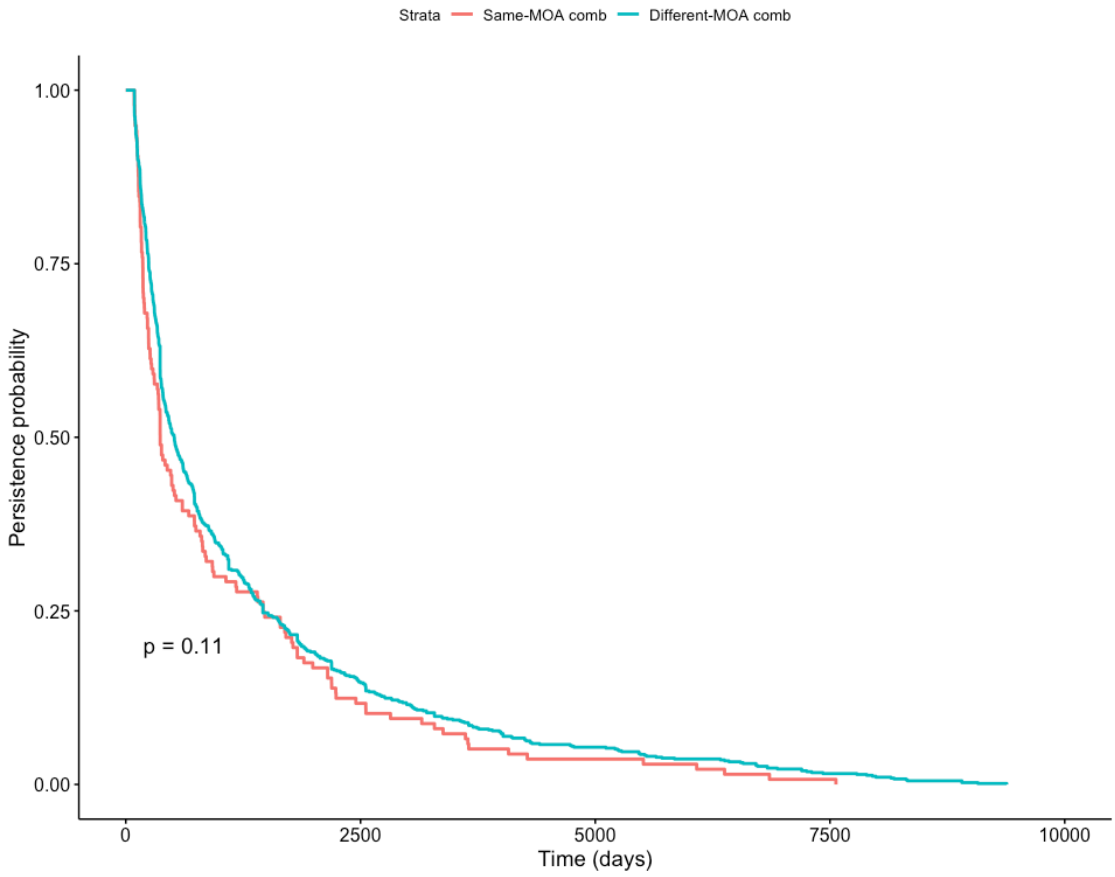


Figure 6: Kaplan Meyer survival curves of persistence comparisons among same- and different-MOA based combinations in generalised epilepsy cases.

Table 6: Kaplan Meyer summary for same- and different-MOA based combinations in generalised epilepsy cases. The table includes the number of trials, number of events and median persistence time.

MOA combination	Records	Events	Median	0.95LCL	0.95UCL
Same-MOA combinations	137	137	366	304	534
Different-MOA combinations	765	765	516	442	603

Group of all combinations containing same mode of action drugs and group of all different-MOA based drug combinations were compared in all cases with qualifying treatments, focal and generalized groups using Kaplan Meyer survival analysis. As expected, the Kaplan Meyer survival curves comparing treatment persistence of AED combinations over time (median in days) show that different MOA-based combinations have higher persistence than same-MOA based combinations. All findings are statistically significant ($p < 0.05$). (Figure 4, Figure 5, Figure 6, Table 4, Table 5, Table 6). Cox proportional hazard regression was not performed since it is a comparison of only two groups.

Investigation of interesting AEDs combinations

Retention of levetiracetam and topiramate combination in focal epilepsy patients

Kaplan Meyer survival analysis

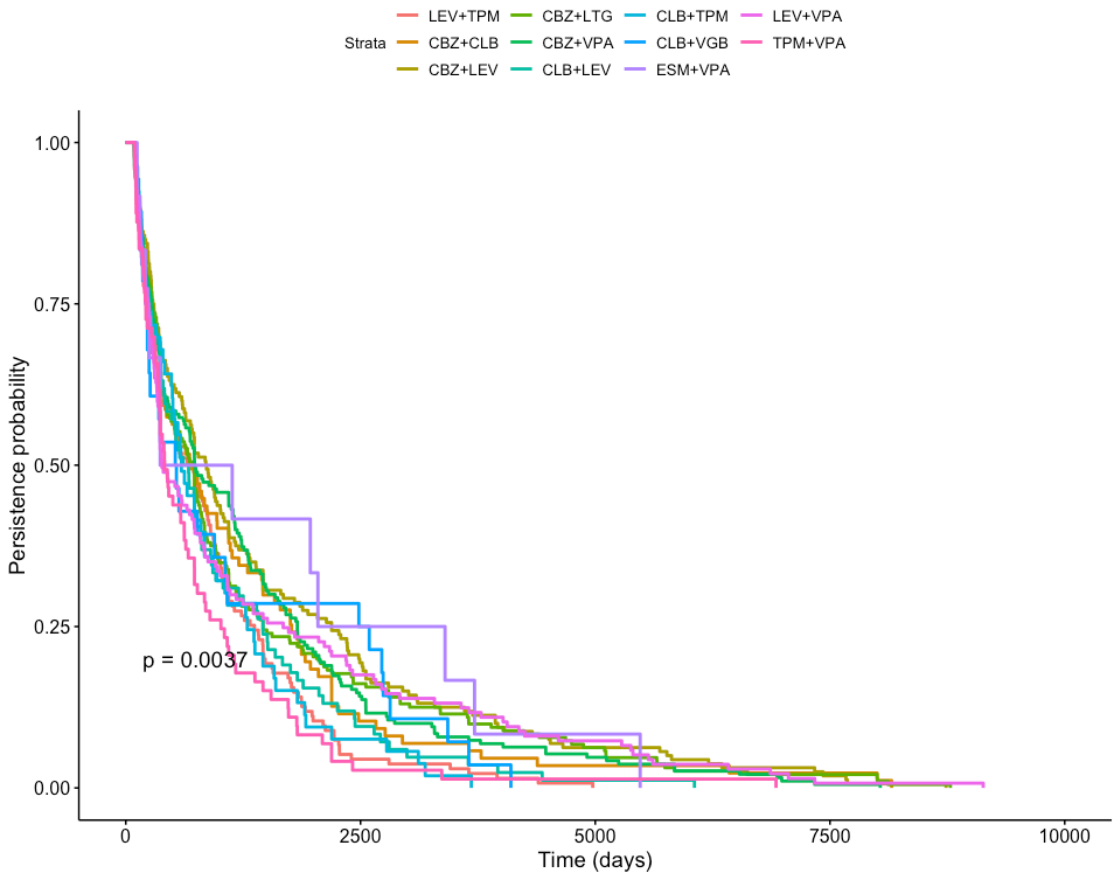


Figure 7: Kaplan Meyer survival curves of persistence comparisons of Levetiracetam-Topiramate to other AED combinations used focal epilepsy patients therapy.

Table 7: Kaplan Meyer summary for Levetiracetam-Topiramate combination comparison to other AED combinations used in focal epilepsy cases. The table includes the number of trials, number of events and median persistence time.

AED combination	Records	Events	Median	0.95LCL	0.95UCL
LEV+TPM	135	135	700	487	883
CBZ+CLB	87	87	685	419	1096
CBZ+LEV	160	160	856	639	1049
CBZ+LTG	192	192	730	516	805
CBZ+VPA	190	190	736	623	1141

AED combination	Records	Events	Median	0.95LCL	0.95UCL
CLB+LEV	84	84	640	416	808
CLB+TPM	53	53	592	495	928
CLB+VGB	28	28	533	248	2482
ESM+VPA	12	12	750	246	0
LEV+VPA	137	137	385	365	727
TPM+VPA	73	73	413	365	639

Cox proportional hazard regression

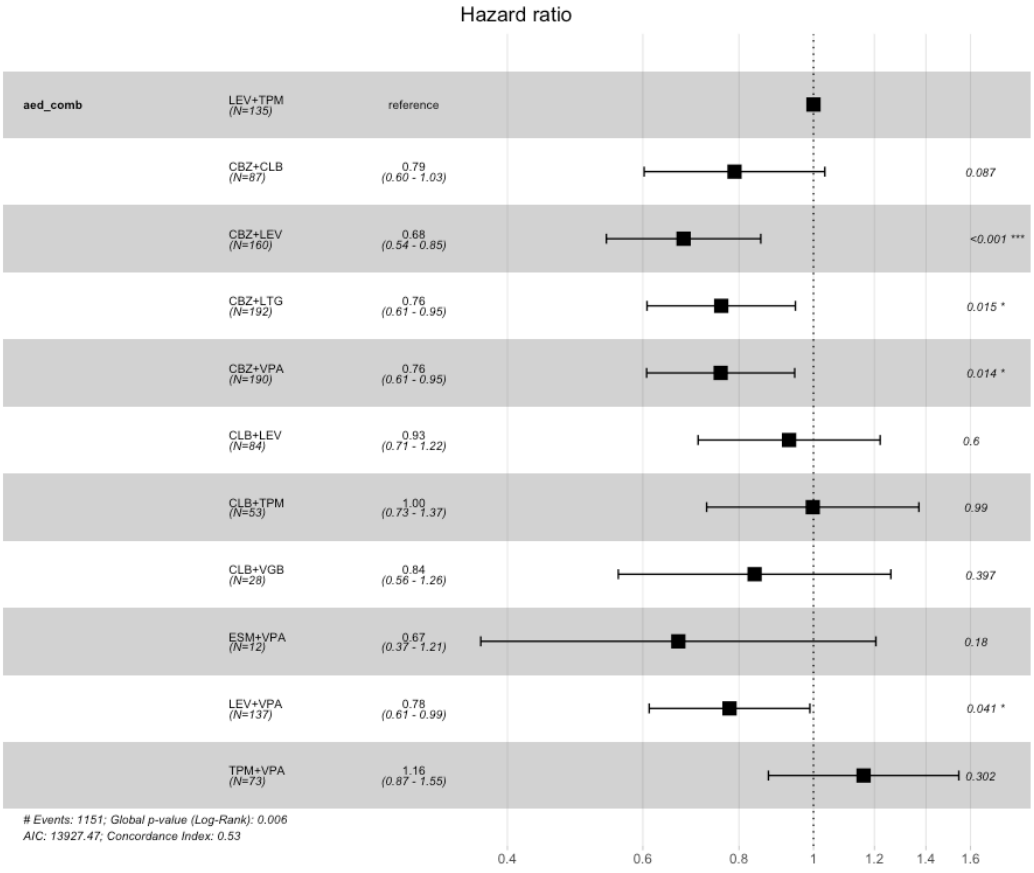


Figure 8: Multivariate Cox proportional hazards regression model for risk of treatment non-persistence for levetiracetam-topiramate combination compared to other AED combinations in focal epilepsy patients.

Table 8: Hazard Ratios of levetiracetam-topiramate combination therapy discontinuation compared to other AED combinations in focal epilepsy cases.

AED combination	Reference	Hazard ratio	Lower CI	Higher CI	P value
CBZ+CLB	LEV+TPM	0.789	0.602	1.035	0.087
CBZ+LEV	LEV+TPM	0.678	0.538	0.854	< 0.001
CBZ+LTG	LEV+TPM	0.759	0.608	0.947	0.015

AED combination	Reference	Hazard ratio	Lower CI	Higher CI	P value
CBZ+VPA	LEV+TPM	0.757	0.607	0.945	0.014
CLB+LEV	LEV+TPM	0.930	0.708	1.221	0.600
CLB+TPM	LEV+TPM	0.998	0.726	1.371	0.990
CLB+VGB	LEV+TPM	0.838	0.557	1.261	0.397
ESM+VPA	LEV+TPM	0.667	0.369	1.206	0.180
LEV+VPA	LEV+TPM	0.778	0.611	0.989	0.041
TPM+VPA	LEV+TPM	1.162	0.874	1.545	0.302

The levetiracetam-topiramate combination was investigated in focal epilepsy patients because of the significant antiepileptogenic and disease-modifying effect found in a temporal lobe epilepsy mouse model in the study of Löscher 2020. Its retention was compared to the most frequent AED combinations in all MOA-based combination group in all cases of focal epilepsy with qualifying treatments. The LEV+TPM was not the most frequent but second-ranked in MM+SV group, and therefore, after selection of all the first ranked AED combinations within each MOA based combination group LEV+TPM had to be added for the comparison. In addition, ESM+LTG from CCB+SCB group was excluded since it was the most frequently used with its 4 trials which is not enough data for Cox proportional hazard analysis, leaving this group without a representative for the comparison. LEV+TPM was compared to the most frequent AED combinations from each remaining MOA based combinations groups including MM+SV first ranked AED combination. (Figure 7, Figure 8, Table 7, Table 8)

Retention of lamotrigine and valproate combination in generalized epilepsy patients

Kaplan Meyer survival analysis

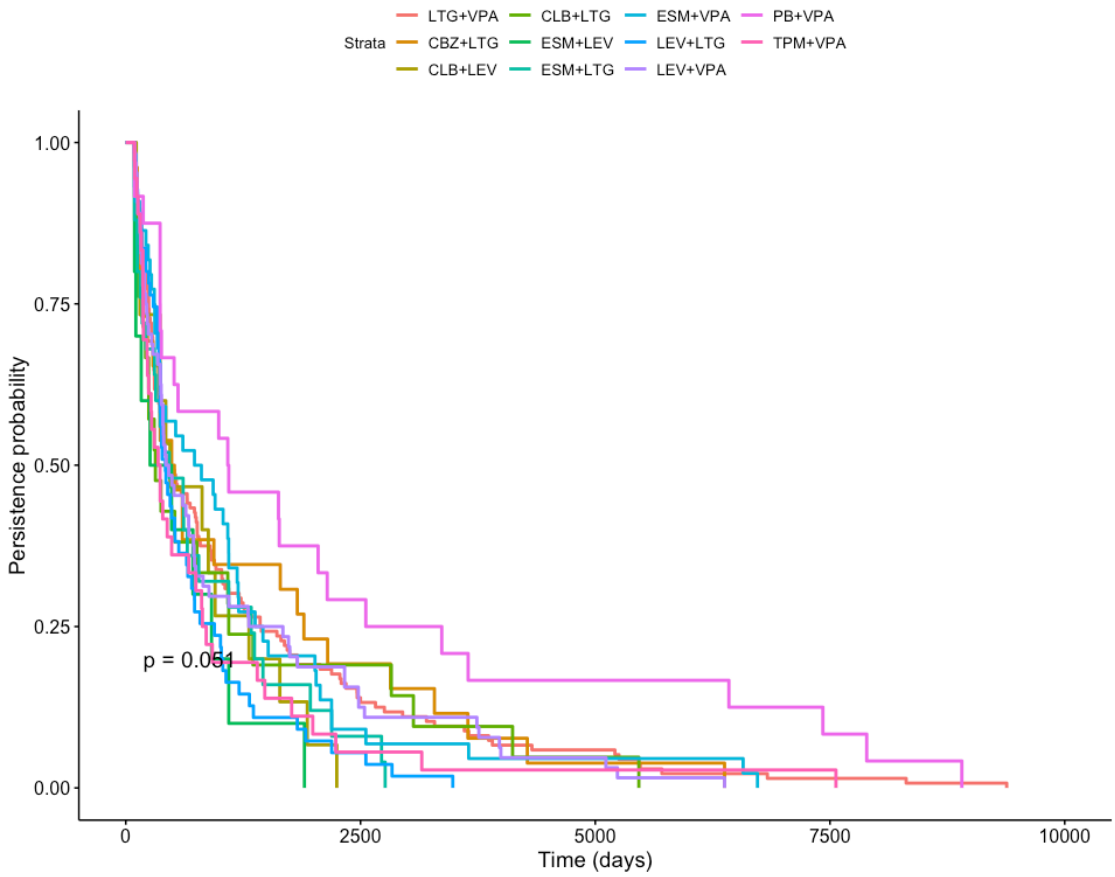


Figure 9: Kaplan Meyer survival curves of persistence comparisons of lamotrigine-valproate combination to other most frequent AED combinations used generalized epilepsy patients therapy.

Table 9: Kaplan Meyer summary for lamotrigine-valproate combination comparison to other most frequent AED combinations used in generalized epilepsy cases. The table includes the number of trials, number of events and median persistence time.

AED combination	Records	Events	Median	0.95LCL	0.95UCL
LTG+VPA	136	136	464	365	760
CBZ+LTG	26	26	503	289	1897
CLB+LEV	15	15	470	365	1640
CLB+LTG	21	21	315	210	1354
ESM+LEV	10	10	372	107	0
ESM+LTG	25	25	462	306	1336
ESM+VPA	44	44	768	365	1186

AED combination	Records	Events	Median	0.95LCL	0.95UCL
LEV+LTG	55	55	421	365	647
LEV+VPA	64	64	442	380	726
PB+VPA	24	24	1091	514	2556
TPM+VPA	36	36	354	231	746

Cox proportional hazard regression

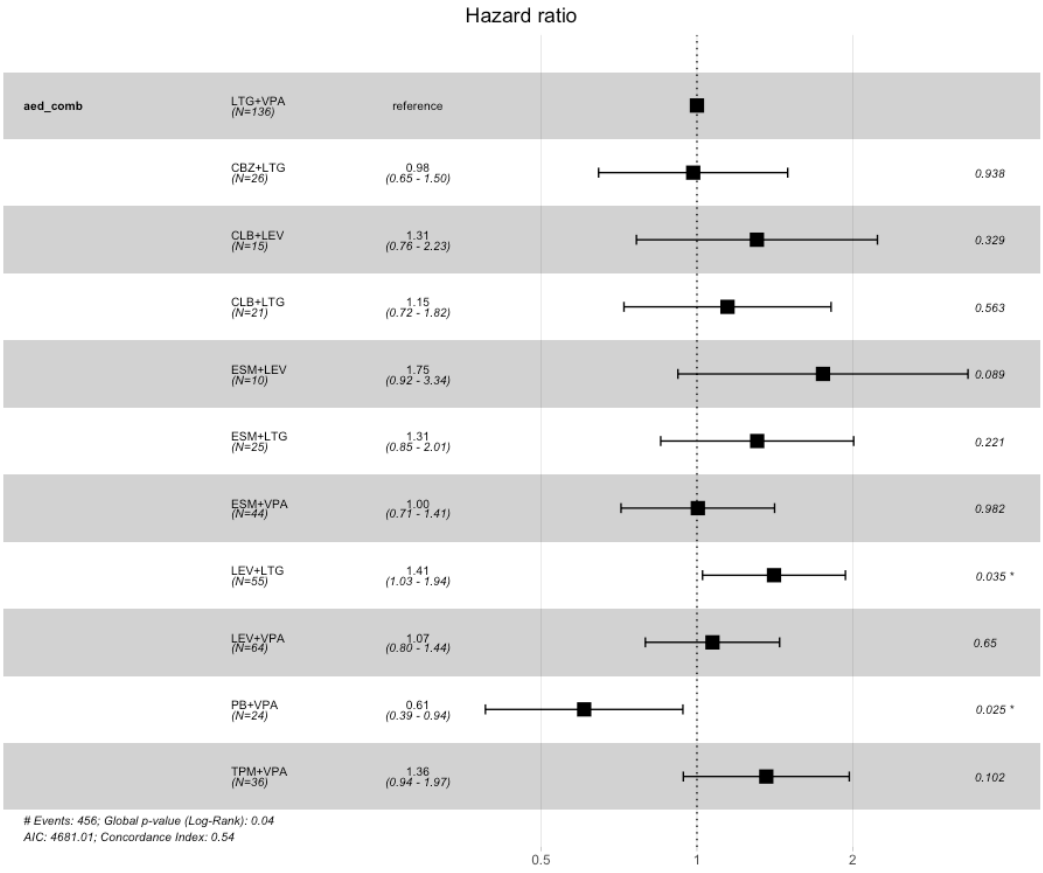


Figure 10: Multivariate Cox proportional hazards regression model for risk of treatment non-persistence for lamotrigine-valproate combinations compared to other most frequent AED combinations used in generalized epilepsy patients.

Table 10: Hazard Ratios of lamotrigine-valproate combination in therapy discontinuation compared to other most frequent AED combinations in generalized epilepsy cases.

AED combination	Reference	Hazard ratio	Lower CI	Higher CI	P value
CBZ+LTG	LTG+VPA	0.983	0.646	1.498	0.938
CLB+LEV	LTG+VPA	1.306	0.764	2.232	0.329
CLB+LTG	LTG+VPA	1.146	0.723	1.816	0.563
ESM+LEV	LTG+VPA	1.752	0.919	3.340	0.089
ESM+LTG	LTG+VPA	1.307	0.851	2.008	0.221

AED combination	Reference	Hazard ratio	Lower CI	Higher CI	P value
ESM+VPA	LTG+VPA	1.004	0.713	1.413	0.982
LEV+LTG	LTG+VPA	1.409	1.025	1.936	0.035
LEV+VPA	LTG+VPA	1.072	0.795	1.445	0.650
PB+VPA	LTG+VPA	0.605	0.390	0.939	0.025
TPM+VPA	LTG+VPA	1.361	0.941	1.970	0.102

Lamotrigin-Valproate combination was investigated in generalised epilepsy patients as the most frequently used AED combination. Its retention was compared to retention of most frequently used AED combinations in other MOA-based combination groups. Additionally, CLB+GBP from GA+GA group was excluded since it was the most frequently used with its 3 trials which is not enough data for Cox proportional hazard analysis, leaving this group without a representative for the comparison. (Figure 9, Figure 10, Table 9, Table 10)

Discussion

Same MOA-based combinations are less retained than different MOA-based combinations in all cohorts. Literature comparison, support the results by other studies, others that do not aggree

*The animal model TLE is not good representation of human TLE, more particularly the combination Lev+TPM

*Top ranked AED combinations are not significantly different in application in generalized epilepsy.

References

Brodie, Martin J. 2010. "Antiepileptic drug therapy the story so far." *Seizure: European Journal of Epilepsy* 19 (10): 650–55. https://doi.org/10.1016/j.seizure.2010.10.027.

BROWNE, THOMAS R., FRITZ E. DREIFUSS, PAUL R. DYKEN, DAVID J. GOODE, J. KIFFIN PENRY, ROGER J. PORTER, BILLY G. WHITE, and PHILIP T. WHITE. 1975. "Ethosuximide in the Treatment of Absence (Petit Mal) Seizures." *Neurology* 25 (6): 515–15. https://doi.org/10.1212/WNL.25.6.515.

Johannessen, Svein I, and Cecilie Johannessen Landmark. 2010. "Antiepileptic Drug Interactions - Principles and Clinical Implications," 254–67.

Margolis, Jay M., Bong-Chul Chu, Zhixiao J. Wang, Ronda Copher, and Jose E. Cavazos. 2014. "Effectiveness of Antiepileptic Drug Combination Therapy for Partial-Onset Seizures Based on Mechanisms of Action." *JAMA Neurology* 71 (8): 985. https://doi.org/10.1001/jamaneurol.2014.808.

Rogawski, Michael A, and Wolfgang Lo. 2016. "Mechanisms of Action of Antiseizure Drugs and the Ketogenic Diet," no. 1: 1–28.

Schidlitzki, Alina, Pablo Bascuñana, Prashant K. Srivastava, Lisa Welzel, Friederike Twele, Kathrin Töllner, Christopher Käufer, et al. 2020. “Proof-of-Concept That Network Pharmacology Is Effective to Modify Development of Acquired Temporal Lobe Epilepsy.” *Neurobiology of Disease* 134 (February): 104664. https://doi.org/10.1016/j.nbd.2019.104664.