PhD-FSTM-2023-003
The Faculty of Science, Technology and Medicine

DISSERTATION

Defence held on 13/01/2023 in Esch-sur-Alzette

to obtain the degree of

DOCTEUR DE L'UNIVERSITÉ DU LUXEMBOURG

EN *BIOLOGIE*

by

Mariana Messias RIBEIRO
Born on 28th of February of 1995 in Belém, Portugal

IDENTIFYING KEY TRANSCRIPTION FACTORS OF CELLULAR MECHANISMS IN SINGLE-CELL ENVIRONMENT FOR REGENERATIVE MEDICINE

Dissertation defence committee

Prof. Dr. Antonio del Sol, Dissertation supervisor
Professor, Université du Luxembourg

Prof. MD. Dr. Ernest Arenas
Professor, Karolinska Institutet

Prof. Dr. Carlos Matute
Professor, Universidad del País Vasco

Prof. Dr. Jens Christian Schwamborn
Professor, Université du Luxembourg

Prof. Dr. Reinhard Schneider, Chairman
Professor, Université du Luxembourg

Affidavit

I hereby confirm that the PhD dissertation entitled “IDENTIFYING KEY TRANSCRIPTION FACTORS OF CELLULAR MECHANISMS IN SINGLE-CELL ENVIRONMENT FOR REGENERATIVE MEDICINE” has been written independently and without any other sources than cited.

Luxembourg, 02/12/2022

Mariana Messias Ribeiro

Acknowledgements

I would like to thank my PhD supervisor Prof. Dr. Antonio del Sol for providing me with the opportunity of doing a PhD in his research group. I appreciate all the help and guidance provided during my studies as well as the insightful scientific discussions we shared. I would also like to thank Prof. Dr. Ernest Arenas for following up on my work throughout these years and for your valuable input on the experimental projects. I am grateful for the time you took to guide me and for all the inspiring discussions we had. Finally, I would like to thank Prof. Dr. Reinhard Schneider for his yearly follow-up of my PhD projects and his valuable feedback. As members of my CET and well-established researchers, you helped me to grow as a scientist, to improve my research projects with our stimulating scientific discussions, and to become independent and critical of my scientific results.

I would also like to take this opportunity to thank the Fond National de la Recherche Luxembourg for funding my research and all my CBG colleagues, especially Dr. Satoshi Okawa for supervising my computational projects and Céline for all the help and support during these years and for becoming one of my closest friends.

To Michi, Izabela, and Kamil for all the relaxing, funny, and memorable moments we shared in Luxembourg. To Ugné for always being there to chat and share our unique moments.

I would also like to thank all my colleagues in Ernest's lab, especially to Dr. Anqi Xiong for taking the time to guide me through my experiments and for her valuable feedback and support. To Emilia and Ka Wai for making my time in Sweden much more fun and for sharing your cooking skills with me.

To Michele, Serge, Cate, Baptiste, and Theresa for keeping up with me for several years now, challenging me to try new things (mostly food), and keeping me motivated to hit the gym (mostly to burn all the food. And challenge Michele).

To my mom and my sister for all their support throughout these years. To my mom for giving 200% every day and shaping me into the person I am today. To my sister for keeping my chair free when I was coming over. I know this is your way to say "I love you".

To Igorko. You know I could not have done this without you. Thank you for supporting me every step of the way and helping me grow as a person. I could not be more thankful to have shared these years of my life with you.

Dedication

This PhD dissertation is dedicated to

My mother, sister, Igorko, and grandmother *for their unconditional support.*

In the memory of ***my grandfather***

Esta dissertação de doutoramento é dedicada

À minha mãe, à minha irmã, ao Igor e à minha avó *pelo vosso apoio incondicional.*

Em memória do ***meu avô***

Table of Contents

List of Figures

List of Abbreviations

6-OHDA – 6-Hydroxydopamine

AAV – Adeno-associated virus

AD – Alzheimer's Disease

ATAC-seq – Assay for Transposase Accessible Chromatin using sequencing

BBB – Blood brain barrier

BS-seq – Bisulfite sequencing

cDNA – complementary DNA

ChIP – Chromatin Immunoprecipitation

ChIP-seq – ChIP sequencing

CNS – Central Nervous System

CRISPR – Clustered Regularly Interspaced Short Palindromic Repeats

CRISPRa – CRISPR activation

CRT – Cell Replacement Therapy

DAN – Dopaminergic neuron

DE – Differentially Expressed

DNA – Deoxyribonucleic Acid

EPI – Epilepsy

eQTL – expression Quantitative Trait Loci

ESC – Embryonic Stem Cell

GABA – γ-aminobutyric acid

GBM – Glioblastoma

GFAP – Glial Fibrillary Acid Protein

GRN – Gene Regulatory Network

gRNA – guide RNA

GWAS – Genome Wide Association Studies

HVG – High Variable Genes

iDAN – induced DAN

iPSC – induced Pluripotent Stem Cell

JSD – Jensen-Shannon divergence

K – Lysine

mRNA – messenger RNA

MMI – Multivariate Mutual Information

MS – Multiple Sclerosis

NGS – Next Generation Sequencing

PD – Parkinson's Disease

PeriV – Perivascular

PF – Pioneer factor

QC – Quality Control

Rgl – Radial glia

RNA – Ribonucleic Acid

RNA-seq – RNA-sequencing

SCENIC – Single-Cell Regulatory Network Inference and Clustering

scATAC-seq – single-cell ATAC-seq

scRNA-seq – single-cell RNA-sequencing

SN – Substantia Nigra

SNP – Single Nucleotide Polymorphism

SNpc – SN *pars compacta*

TCGA – The Cancer Genome Atlas

TF – Transcription Factor

TH – Tyrosine Hydroxylase

UMI – Unique Molecular Identifier

Summary

Regenerative medicine holds great potential as a therapeutical tool that can be used to tackle the rising incidence of neurodegenerative diseases and other brain disorders. There are still multiple challenges that need to be addressed before successful clinical translation. One of the most promising fields of regenerative medicine is cell replacement therapy. However, current cell conversion protocols have low efficiency and the generated cells lack molecular and functional features described *in vivo*. Another aspect hindering the development of more effective therapies is the lack of knowledge about specific regulatory mechanisms behind disease onset and development. To address these issues, single-cell RNA sequencing data has been used to advance the development of computational and experimental strategies that improve the identification of therapeutical targets and disease-related mechanisms at the cell subpopulation level.

In this PhD dissertation, we focused on developing novel single-cell based computational and experimental methods that advance the development of regenerative medicine approaches. TransSynW is a computational platform that leverages single-cell RNA-sequencing to identify transcription factors that can improve cellular conversion protocols. This method prioritizes pioneer transcription factors, addressing the limitations posed by incomplete epigenetic remodeling observed on the cells generated by current conversion protocols. We applied this method to distinct cellular systems, such as cell types, subtypes and phenotypes, and well-recapitulated known conversion TFs. Moreover, we were able to validate the biological significance of our predictions using a manually curated database of molecular interactions. We applied TransSynW to develop novel direct reprogramming protocols based on endogenous and ectopic regulation of conversion TFs, aiming at generating subpopulations of dopaminergic neurons.

Based on single-cell sequencing data from low-grade glioma and glioblastoma patients, we found that tumor cells derived from the perivascular lineage are uniquely present in glioblastoma patients. To be able to identify this lineage in patients' samples, we identified *PROX1* and *FOXC1* as specifically expressed in radial glia and perivascular derived tumors, respectively. These transcription factors have the potential to become important biomarkers in disease prognosis.

Finally, we delved into disease modelling and developed RNetDys, a multi-omics pipeline that can decipher the impact of disease-associated single nucleotide polymorphisms in the impairment of cell type specific regulatory mechanisms. We applied this pipeline to five diseases with a genetic background and validated the significance of the identified impairments against literature-based evidence.

In summary, this PhD dissertation focuses on overcoming major challenges in cellular conversion protocols, identifying potential prognostic biomarkers, and deciphering the role of disease-associated single nucleotide polymorphisms in the impairment of regulatory mechanisms. The novel findings of this PhD dissertation have potential applications in the different fields of regenerative medicine, such as cell replacement therapy and disease modelling.

1. Introduction

1.1. Central nervous system organization

The nervous system is a complex structure that controls our everyday actions by processing and responding to any perceived stimuli. These responses can be classified in somatic and autonomic functions. Somatic responses encompass the voluntary control of our movements through the connection of the neuronal system to the skeletal muscle while autonomous responses regulate our physiological involuntary processes, such as breathing, heartbeat, and digestion [1,2]. This system can be divided into two main regions: i) the central and ii) the peripheral nervous systems.

The central nervous system (CNS) comprises two major organs of the human body: the brain and the spinal cord. These structures are enveloped by the meninges and are protected by the cranial cavity in the skull and by the spinal canal in the vertebral column, respectively [3]. Likewise, the CNS encompasses the white matter, composed of oligodendrocytes and axons ensheathed by myelin, and the gray matter which consists mostly of neuronal somas (cell bodies). In the brain, the gray matter is located in the outer part while the white matter can be found in the inner portion. In the spinal cord, the location of these regions is inversed.

The brain is our most intricate and enigmatic organ (Figure 1). It has the capability to process and interpret information transmitted by multiple signals throughout the body to accordingly determine the most suitable response [4]. This structure can be divided in three main sections: i) the cerebrum, ii) the cerebellum, and iii) the brainstem. The cerebrum is the major component of the brain (around eighty percent), and it can be subdivided into cerebral cortex, hippocampus, and amygdala [4]. Together, this structure is responsible for the highest neural abilities, such as cognition, visual processing, language, and memory, as well as for further specialized faculties, for instance consciousness and perception [4,5]. The main role of the cerebellum is to control motor functions by establishing the communication between cerebral cortex, spinal cord, and muscle [4]. It regulates coordination, timing, and precision of our everyday movements.

The brainstem is positioned at the bottom of the brain, and it links the cerebrum to cerebellum and spinal cord [6]. Unlike the cerebrum, this structure is small, only accounting

for around three percent of the brain's weight [7]. The brainstem can be divided into three major parts: i) the medulla oblongata, ii) the pons, and iii) the midbrain (Figure 1) [6]. The nerve tracts connecting the brain to the spinal cord, in addition to the ones linked to several organs, all pass through the medulla oblongata, making this part of the brainstem responsible for autonomic functions, such as breathing and blood pressure. The pons bridges the connection between the medulla oblongata and the midbrain [4,6]. Here, the nerves that go through it connect to the cerebellum, regulating muscle coordination. In addition, the pons also relays nerves that connect to the head and face, contributing for the control of eyeball movement, facial expression, and salivation.

The midbrain serves as a connection center that transmits motor and sensory signals between spinal cord, pons, and cerebral cortex (Figure 1) [6–8]. Two of its main regions are the cerebral peduncles and the tegmentum. Each of the cerebral peduncles creates a lobe on the lower part of the tegmentum. The connecting parts consist mostly of substantia nigra (SN), a nucleus of dopaminergic neurons (DANs) that has a crucial function in regulating reward and motor movement [9]. The SN is the main input into the circuitry of the basal ganglia, a group of subcortical structures that regulate, among other functions, cognition, emotions, and voluntary movement. The SN is subdivided into *pars reticulata* and *compacta* (SNpc), the latter presenting a darker coloration due to the elevated levels of neuromelanin derived from dopamine synthesis [9,10]. The SNpc is the primary region of the brain responsible for dopamine production, a crucial neurotransmitter that impacts movement, cognitive functions, and emotional responses. In particular, the SNpc is connected to the striatum, one of the main components of the basal ganglia, by DANs forming the nigrostriatal pathway (Figure 1). Degeneration of the DANs forming the nigrostriatal pathway in the SNpc is profoundly involved in the onset of the motor deficits seen in Parkinson's Disease (PD) [11].

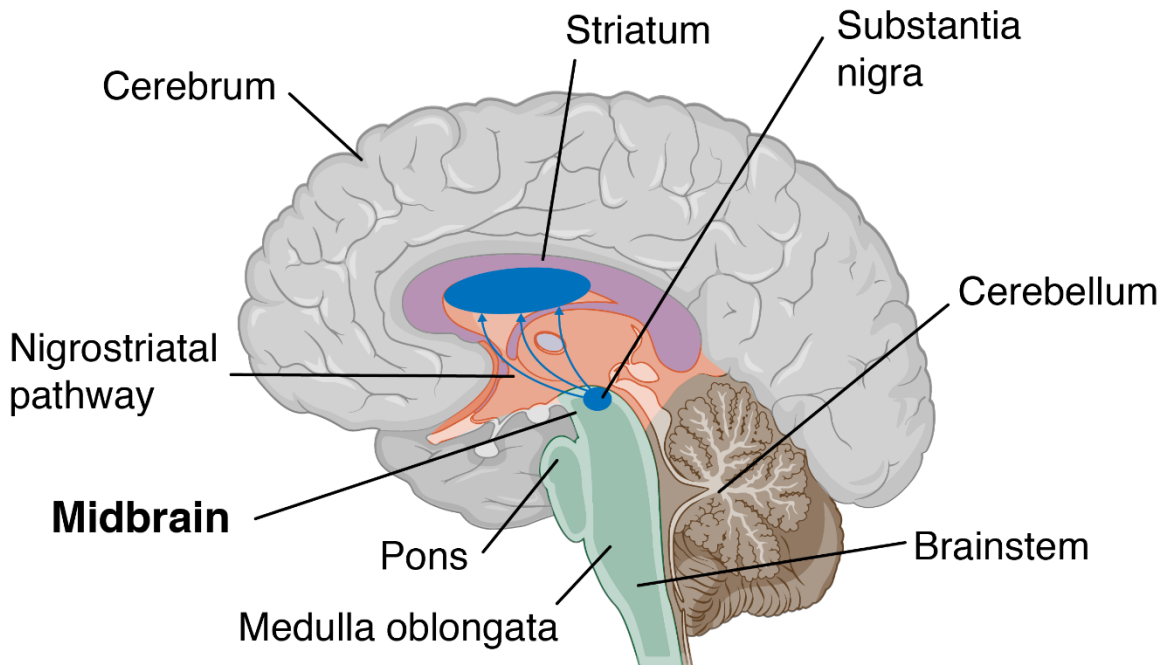


Figure 1: Schematic representation of the brain. This structure can be divided in three main sections: cerebrum, cerebellum, and brainstem. The brainstem can be divided into four major parts, including medulla oblongata, pons, and midbrain. Within the midbrain, we can find the substantia nigra (SN), a nucleus of dopaminergic neurons (DANs) that has a crucial function in regulating reward and motor movement. In particular, the SN *pars compacta* is connected to the striatum by DANs, forming the nigrostriatal pathway.

1.1.1. Classification and function of the major brain cell types

Our brain is composed of two main types of cells: i) neurons and ii) glial cells. Neurons are the building foundations of the brain, being responsible for the entire network of communication of the nervous system. Glial cells are non-neuronal cells whose main functions are to provide support to the metabolically demanding neuronal functions, structural maintenance, and to facilitate synapse transmission [12]. Thus, the crosstalk between neurons and glial cells is of the upmost importance for the correct functioning and regulation of brain homeostasis.

1.1.1.1. Neurons

Neurons are elongated, asymmetric cells that have the capacity of being electrically excitable [13]. Due to this ability, neurons' primary function is to respond to external stimuli by generating electrical impulses, named action potentials, and conduct them throughout the body's neuronal network. The morphology of these cells is very distinct, with a round cell body (soma) and branching cellular processes that extend into opposite

directions, forming dendrites and axons. The neuronal morphology is held together by a complex network of microtubules [14]. One of its main structural components is beta tubulin III (TUBB3). This protein is expressed specifically in neurons during early development and is commonly used as a positive marker [14,15].

Dendrite arborization allows for multiple signals from different cells to be received and processed by an individual neuron. If the sum of these electrical impulses exceeds a threshold, the neuron will fire and transmit an action potential down its myelinated axon (Figure 2) [13,16]. Once the pulse reaches the axon terminal, the electrical signal is converted into a chemical one that can then be transmitted to the target neuron or effector cell. This chemical signal is sent along the axon terminal of the presynaptic neuron to the dendritic branches of the postsynaptic neuron through the space in between these neurons, also called a synapse [16,17]. In these structures, the chemical pulse is transmitted from the presynaptic to the postsynaptic neurons by the release of neurotransmitters. These molecules can propagate two types of messages: i) excitatory and ii) inhibitory. Excitatory neurotransmitters induce the depolarization of the cellular membrane, resulting in the generation and conduction of the action potential to the postsynaptic neuron. On the other hand, inhibitory pulses promote hyperpolarization of the postsynaptic membrane, leading to the blockage of the propagation of the action potential. There are two neurotransmitters worth mentioning, glutamate and γ-aminobutyric acid (GABA), as they are responsible for excitatory and inhibitory synaptic function, respectively.

Glutamate has been described to have a central role in the plasticity of the nervous system, namely in the synapse connections involved in memory storage [17,18]. This neurotransmitter is involved in the pathogenesis of multiple neurodegenerative diseases, namely Alzheimer's Disease (AD) and PD [19]. Glutamate excitotoxicity is triggered when this neurotransmitter is not cleared from the synaptic cleft by the glial cells, specifically astrocytes, increasing its extracellular concentration [19,20]. This effect is a major contributor to neuronal cell death of glutamatergic neurons observed in AD patients [19,21]. Moreover, mutations in PD-associated genes can cause a dysregulation in glutamate mediated synapses, which leads to glutamate excitotoxicity [22,23]. This excitotoxic process has been associated to DAN degeneration in several parkinsonian mouse models [21,24,25].

On the other hand, GABA is the neurotransmitter targeted for the treatment of neurological disorders such as epilepsy (EPI) [26]. Reduction of GABA mediated inhibitory

pulses in the cerebral cortex leads to electrical patterns like the ones observed upon an epileptic episode [26,27]. Additionally, chemical agonists of GABA have been shown to be antiepileptic while GABA antagonists promote epileptic seizures [28,29].

Finally, dopamine is another major neurotransmitter that has a vital role in multiple brain functions, such as motor movement and reward pathways. The synthesis of this neurotransmitter requires a specific amino acid, phenylalanine, and is dependent on the action of an enzyme only present in DANs, the L-tyrosine hydroxylase (TH) [30]. The depletion of dopamine releasing neurons is closely related to the onset of PD [31]. Indeed, the most administered treatment for the symptoms of PD is the prescription of levodopa, a small molecule that is converted to dopamine in the brain [32].

Together, all these neurotransmitters are released in the more than 10^{14} synapses present in our brain, reflecting the capacity of neurons to transmit millions of impulses in a split second [33]. To maintain this unparallel processing power, glial cells play an essential role in meeting the high metabolic demands arising from neuronal functions and sustaining a microenvironment suitable for an efficient transmission of this massive amount of information.

1.1.1.2. Astrocytes

Astrocytes are specialized, star-shaped cells that provide essential support to the neuronal tissue. These glial cells extend long cellular processes from their soma that branch out into the processes called endfeet [34]. Astrocytes are the most abundant cell type in the brain, and they can be identified by the presence of glial fibrillary acid protein (GFAP), a uniquely expressed filament protein [34,35]. Depending on their anatomical position and morphology, this class of glial cells can be divided into two subtypes: i) fibrous or ii) protoplasmic. Fibrous astrocytes are located in the white matter and display thin and elongated processes that connect to nodes of Ranvier, the short myelin gaps in neuronal axons [36,37]. In contrast, protoplasmic astrocytes reside in the gray matter and display a globoid morphology composed of numerous irregularly ramified processes which enclose the synapses [36,37]. This coupled organization allows astrocytes to establish a unique crosstalk with neurons, vital for the maintenance of the neuronal microenvironment, namely through shuttling and recycling of metabolites and neurotransmitters.

In the synapses, electrical impulses transmitted by neurons are transformed into chemical ones, in a process mediated by neurotransmitters. Clearance of these chemicals,

especially glutamate, is essential for correct transmission of countless signals and to prevent excitotoxicity. Astrocytes use glutamate transporter 1 and the glutamate-aspartate transporter to clear glutamate from the synaptic cleft (Figure 2). Glutamate is metabolized into glutamine by an astrocyte-specific enzyme named glutamine synthetase [38]. Additionally, pyruvate carboxylase, an enzyme essential to the replenishment of glutamate, is uniquely expressed in astrocytes, allowing the synthesis of glutamine from glucose [39]. Astrocytes then shuttle the glutamine back to the neurons, where it is converted again into glutamate in an energy-dependent process mediated by the enzyme glutaminase [38,40].

Astrocytes are able to effectively recycle neurotransmitters and uptake glucose due to their ideal position: their processes surround synapses while their endfeet connect with vascular cells, such as pericytes [41]. Pericytes are mural cells embedded within the capillaries that are known by their stem cell-like attributes and by their role in regulating the brain's blood flow and the blood brain barrier (BBB).

Due to their connection to the circulatory system, astrocytes uptake glucose when glutamate is released into the synaptic cleft (Figure 2) [42]. Several studies have shown that, due to the differences in the metabolic rate, astrocytes and neurons use complementary metabolic pathways to oxidize glucose [43–45]. Astrocytes have a lower metabolic rate and direct access to glucose therefore they preferentially choose the glycolytic pathway which leads to the production of lactate that is shuttled to neurons though the monocarboxylate transporter 1 [42,46,47]. Neurons have a high metabolic rate therefore lactate enters as a substrate in the tricyclic acid cycle to produce adenosine 5′-triphosphate [42,43,47]. Together with regulating oxidative stress and contributing to synaptogenesis, astrocytes have a pivotal role in sensing brain homeostasis and their unique position makes them the ideal cell type to provide metabolic support to neurons and to maintain neuronal excitability and synaptic transmission.

Astrocytes also play a role in neuroinflammation. Upon brain insult, astrocytes can become reactive through a process named astrogliosis and release a variety of effector molecules, such as cytokines, growth factors and signaling molecules [48,49]. The major hallmarks of this process are the upregulation of GFAP, cellular hypertrophy and proliferation [49]. Astrogliosis can be triggered by mild occurrences, such as mild trauma, viral and bacterial infections, or by more severe events such as neurodegeneration and ischemia. If the brain injury is very severe (e.g., stroke, neurodegenerative disease or acute infection),

astrocytic scars start forming. The scar formation consists of the aggregation and intertwining of the processes of the recently generated and elongated reactive astrocytes, forming compact borders that highlight the areas of severe tissue damage or necrosis [50].

Similar to the macrophage nomenclature, reactive astrocytes have been classified as A1 and A2 types, depending on their functions [51,52]. A1 astrocytes lose most of their physiological functions and promote neuronal death and toxicity. Specifically, A1 astrocytes have been shown to promote formation of weaker synapses when compared to healthy astrocytes. Notably, A1 astrocytes have been detected in several neurodegenerative diseases [52,53]. It was reported that in AD, the majority of the astrocytes located in the prefrontal cortex express high levels of GFAP and an A1 type marker, C3 [52]. Furthermore, it has been shown that inhibiting the formation of A1 astrocytes reduces the degeneration of DANs and improves motor function in a PD model [53]. On the other hand, A2 astrocytes seem to have a neuroprotective role as they have been shown to promote tissue repair and neuronal survival [52]. Specifically, A2 astrocytes release TGFβ, an anti-inflammatory cytokine, that has been reported to play a neuroprotective role and promote synaptogenesis [54].

Astrogliosis is a complex multifactorial mechanism, triggered as a first line of defense to repair brain damage. However, upon severe injury, this process can get exacerbated and lead to the formation of A1 astrocytes which have been described to be involved in the development of neurodegenerative diseases [51,52]. Notably, it has been described that A1 astrocytes can be converted into the A2 type, healthy astrocytes, and even neurons [55–57]. Inducing cellular conversion of A1 astrocytes is a strategy that could be applied to regenerative medicine therapies for neurodegenerative diseases.

1.1.1.3.Microglia

Microglia represent the main immune cell type in the brain and play an essential role in containing neuroinflammation and infection. These immune cells reside in the brain and have a remarkable morphological plasticity, which allows them to navigate and access injury sites easily and phagocytose any harmful elements [58,59]. Under normal conditions, microglia have highly branched and fine processes, which allows them to effectively monitor the parenchyma of the brain for signs of cellular injury or pathogens. Once these cells find damage-associated molecular patterns, molecules derived from invaders or cellular damage, microglia become activated and change their morphology [60]. Their cell bodies increase in size and their processes are completely retracted, acquiring an amoeboid configuration [61].

The conventional classification system consists of two microglial states of activation: i) the M1 or classic and ii) the M2 or alternative [62]. These phenotypes have opposite functions. In the M1 state, microglia releases proinflammatory cytokines and reactive nitrogen and oxygen species, while M2 is an anti-inflammatory state where microglia release trophic factors instead [62]. Recently, it has been described that these two phenotypes are just part of a spectrum of activation profiles, some even specific to certain neurological diseases [63,64].

The main functions of microglia are elimination of pathogens, apoptotic cells and toxins, synaptic pruning, myelin maintenance, and tissue surveillance (Figure 2) [58]. However, this brain cell type has also been shown to have a role in the development of several neurodegenerative diseases. In AD for instance, the rate of clearance of amyloid beta plaques by the microglia is not fast enough to prevent their overactivation. This leads to a sustained release of inflammatory cytokines that promote the persistent neuroinflammation observed in AD [65,66]. In PD, the role of microglia remains elusive, but some studies show that the extensive death of DANs associated with the disease can lead to an increase in the release of proinflammatory cytokines by activated microglia, leading to an exacerbated neuroinflammatory response [67–69].

Under physiological conditions, microglia's spectrum of activated states greatly contributes to the maintenance of brain homeostasis. However, when an exacerbated inflammatory response is triggered, it can promote the development of different brain diseases. Understanding and characterizing their heterogenic responses to the different diseases is important to discover more effective therapeutical targets.

1.1.1.4. Oligodendrocytes

Oligodendrocytes are glial cells responsible for myelinating the neuronal axons in the CNS. These cells go through three different phases: i) migration and proliferation of oligodendrocyte precursor cells, ii) branching of their network of processes, and iii) selection of a neuronal axon, myelin encasing and maintenance of this axon's ensheathment [70]. Oligodendrocytes only produce myelin once their maturation process is complete, which happens in phase three. Once the oligodendrocyte processes connect to an axon, their cellular membrane starts expanding, wrapping around it (Figure 2) [71]. Then, the membrane contracts, ejects cytoplasm and myelin is produced. This lipidic membrane insulates neuronal axons in segments, increasing the speed at which the action potentials travel along the neuronal network [72].

Demyelination is a hallmark of the development of Multiple Sclerosis (MS) [73]. During the onset of this disease, the immune system targets the myelin in the neuronal axons in response to several myelin antigens. Due to age-related changes and limitations inherent to post-mitotic oligodendrocytes, remyelination starts failing as the disease progresses, leaving demyelinated axons vulnerable to neurodegeneration [74,75]. Recently, it has been shown that oligodendrocytes exhibit different functional states in the presence of MS injuries [76].

Oligodendrocytes are essential to the correct functioning of the brain as they are the sole providers of myelin and maintainers of the myelin sheath. When this lipid membrane is damaged or even lost, neurons are left exposed to neurodegeneration. It is therefore important to expand our knowledge about this cell type and the complex mechanisms behind myelin wrapping so we can address the impact of these processes on brain-related diseases.

As we have seen, neurons and glial cells have a very intricate and fine-tunned relationship (Figure 2). While neurons take care of the transmission and processing of information, glial cells have their unique roles to support neuronal function and structure, such as transport and release of neurotransmitters, regulation of inflammation and formation of myelin sheaths. Better understanding these processes and how these brain cell types coordinate their specialized functions would allow us to use this knowledge for developing new therapies and discover new targets for the treatment of neurological diseases.

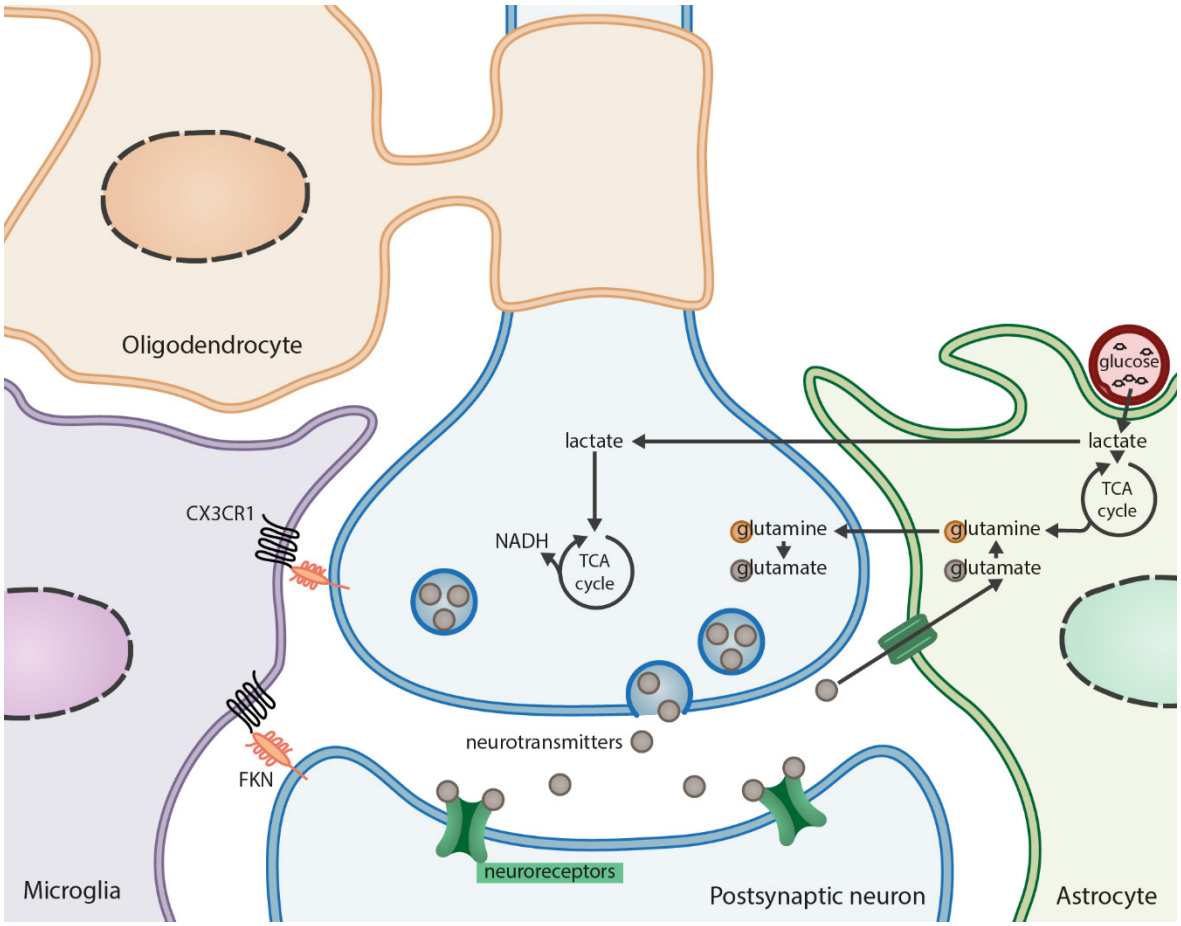


Figure 2: Crosstalk between brain cell types. Neurons are electrically excitable cells responsible for the entire communication network of the nervous system. The electrical signals are converted into chemical ones that are transmitted from the presynaptic to the postsynaptic neuron by the release of neurotransmitters, such as glutamate. Glutamate released by the presynaptic neuron is removed from the synapse by astrocytes, where it is converted into glutamine. Additionally, astrocytes express pyruvate carboxylase which allows for the synthesis of glutamine from glucose. Finally, glutamine is then transported to the presynaptic neuron, where it is converted to glutamate that is released upon new electrical stimulus via synaptic vesicles. When glutamate is cleared from the synaptic cleft, astrocytes uptake glucose from the circulatory system. Glucose is then metabolized via the glycolytic pathway, leading to the production of lactate which is shuttled to neurons. Lactate enters as a substrate in the tricyclic acid cycle (TCA), producing a reduced form of nicotinamide adenine dinucleotide (NADH), which is then used by the electron transport chain to produce adenosine 5′-triphosphate [77]. One of the main functions of microglia is synaptic pruning. In physiological conditions, microglia express CX3C motif chemokine receptor 1 (CX3CR1) which binds to the neuronally expressed fractalkine (FKN), activating these receptors and mediating the release of neuromodulators that promote synaptic plasticity [78]. Finally, oligodendrocytes are essential to the correct functioning of the brain as they are the sole providers of myelin. This lipidic membrane insulates neuronal axons in segments, increasing the speed at which the action potentials travel along the neuronal network.

1.1.2. Deciphering brain complexity at the cell population level

Classifying cell populations in the brain has been a long-term objective in the field of neuroscience. The high level of complexity and relationships inherent to the brain cell types hold the key to understand the correct functioning of this organ. Since the classification of the first brain cell type, multiple other cell types and subtypes have been described. This has been helping us understand how this biological system and their unique properties are able to regulate intricate physical and cognitive activities. Parameters such as morphology, location, electrophysiology, synaptic properties, and expression of marker genes have been used classify the newly observed brain cell types [79] Nevertheless, there are still countless highly specialized cells that remain undescribed. Traditional techniques can only analyze one neuronal phenotype at a time, often in an incomplete manner, which has been slowing down the acquisition of new knowledge. Fortunately, with the advent of next-generation sequencing (NGS), discussed in the next section, we have now considerably advanced the uncovering of brain dynamics, from cellular lineage development to disease onset.

On a larger scale, it is now possible to generate cell type atlases of the human brain, such as the Allen Human Brain Atlas [80]. A comprehensive atlas of the developmental human ventral midbrain and of the human motor cortex have also been generated [81,82]. Revealing not only the cell types, but also characterizing the cell subtypes in the brain would greatly advance our knowledge, especially the understanding of the molecular mechanisms behind disease development. For instance, a recent study has characterized the different cellular subtypes of DANs, where they could clearly see that a particular subtype seems to be more vulnerable to the effects of PD [83]. Two transcriptomically unique reactive astrocytes subtypes have been shown to have different responses in an AD model [84]. Novel high-throughput methods also helped to characterize temporal and spatial differences in disease specific microglia states [85]. Finally, transcriptomic profile of oligodendrocytes allowed the identification of predominant subpopulations of this cell type in MS samples, which may constitute a new target for the development of novel therapeutical approaches for MS [76].

As a result of major advancements in throughput, resolution, and robustness of NGS technologies, we are now able to expand the classification of genetic and phenotypic cell subtypes in the brain. This has revolutionized the current understanding of brain organization and contributed for the characterization of dysregulated mechanisms often associated to disease development.

1.2. Next-generation sequencing techniques for characterizing cell population heterogeneity

In 1953, Dr. James Watson, Dr. Francis Crick, Dr. Maurice Wilkins, and Dr. Rosalind Franklin published three separate scientific articles that are considered to be some of the greatest achievements in human history [86–88]. Together, these articles characterized the structure of the so-called "molecule of life", the Deoxyribonucleic Acid (DNA), and marked a new age in scientific research. Due to the huge potential of this molecule to uncover the mechanisms behind cellular function and dysregulation, multiple efforts have been made to sequence it and extract the important information within. After devoting his scientific career to developing a method for DNA sequencing, Dr. Fred Sanger, in 1977, described a protocol that involved the incorporation of labelled primers with a chemical modification that stopped the extension of the synthesized strand [89]. The fragments obtained from all the chain-terminating reactions were separated in polyacrylamide, visualized by X-rays, and the DNA sequence was assembled. Although this method came into immediate use by the scientific community, it had its drawbacks. The read length obtained from Sanger sequencing is limited to the number of nucleotides that can be amplified for a given primer. For sequencing longer DNA fragments, a new primer is designed in a way that it binds to the end of the known sequence, obtained from the previous step, and Sanger's protocol is repeated. This highly limits the throughput, making it impossible to sequence long DNA sequences, such as the human genome. Several improvements were proposed, namely shotgun sequencing and fluorescence labeling, which led to the development of automated Sanger sequencing devices that could sequence up to 1000 bases daily [90,91]. Consequently, the availability of sequencing data grew exponentially, encouraging the foundation of data repositories and search algorithms, such as GenBank and BLAST, respectively, that are still indispensable tools for current scientific research [92,93].

Together, all these advancements contributed to lowering the time and the cost of sequencing the human genome from 2.7 billion to 10 million dollars, a major milestone in the scientific history [94]. However, there were still some limitations to be addressed. The overlap-based assembly process required a very high processing power and the assembly of repetitive sequences often resulted in errors being introduced, since it was very difficult for the assembly programs to distinguish between repeated copies from flawed base calls and differences of a single base [95]. Despite all efforts to address these challenges, any technical

advancements resulted in minimal benefits, making it soon clear that the efficiency and scalability had reached its peak and new technologies had to be developed.

To address this, NGS platforms were developed. The key difference between Sanger's and NGS, besides its improved accuracy, is multiplexing. With NGS, we scale up from one fragment at a time to a library of millions of DNA templates, bonded to universal adapters and immobilized on a solid surface of microfluidic channels (Figure 3) [91]. This setup completely exposes the complex library to the DNA polymerase and other reaction-catalyzing reagents, allowing for the amplification of multiple DNA fragments in a single reaction. After several steps of amplification, a signal for each of the fragment's sequencing reactions can be digitally detected by the NGS instruments, generating sequencing data for the DNA library of interest. Essentially, NGS platforms perform detection and sequencing at the same time. This massive parallel sequencing approach significantly reduced the time and costs of DNA sequencing and gave us access to large amounts of data, ushering us into the "big data" era. NGS platforms and their continuous improvements led to an exponential growth of both our sequencing scale and throughput, contributing for our increased knowledge of living systems at the genome-wide level. In addition, these technologies helped the characterization and understanding of the vast molecular diversity and heterogeneity of multiple tissues and cellular systems such as cell types, subtypes, and phenotypes.

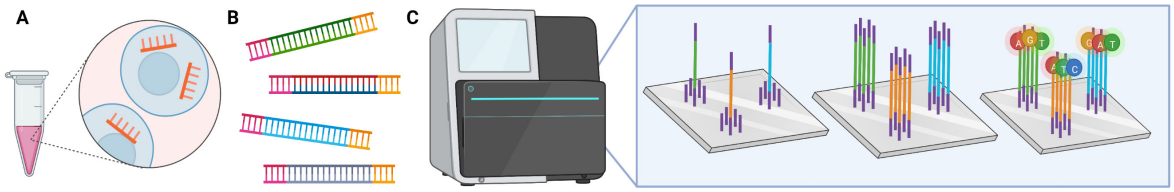


Figure 3: Next-Generation Sequencing (NGS) workflow. NGS protocols can be divided into three steps: **(A)** nucleic acid extraction, **(B)** library preparation, and **(C)** DNA sequencing on microfluidic channels. The immobilized DNA sequences are amplified and a signal for each of the nucleotides can be digitally detected by the NGS instruments, generating sequencing data for the DNA library of interest. Created with BioRender.com.

Bulk sequencing was one of the first applications of NGS technology. These methods can be applied to the study of the genome, transcriptome, and epigenome of millions of cells in distinct biological conditions. Among the several applications of bulk sequencing, RNA-sequencing (RNA-seq), Assay for Transposase Accessible Chromatin using sequencing (ATAC-seq), and Chromatin Immunoprecipitation (ChIP) sequencing (ChIP-seq) are the most commonly used technologies. One of the main applications of bulk RNA-seq is to identify differentially expressed (DE) genes between two distinct conditions, for instance, disease and control, to understand the molecular pathways involved in the development of specific conditions [96]. This technology has been essential in several areas of research, including cancer biology, discovery of new biomarkers, immunotherapy, and clinical applications [97–99]. On the other hand, ATAC-seq and ChIP-seq provide information at the epigenetic level. ATAC-seq determines accessible chromatin regions whereas ChIP-seq identifies loci in the genome that are enriched for DNA-binding proteins, such as transcription factors (TFs), and histone modifications [100,101]. Together, these technologies are widely used for characterization of epigenetic profiles, both by identification of genome wide open chromatin regions and mapping of global protein-DNA binding sites [102,103]. These methods have contributed to the understanding of the epigenetic landscape associated with unique cell types, lineages, and differentiation in both physiological and disease conditions [103–106].

Bulk sequencing has, however, a major limitation. In order to draw conclusions from a bulk experiment, one must assume that the tissue or cellular culture of interest is mostly homogeneous. Indeed, the findings from these studies result from averaging the captured information among all the cells in a given biological sample. It is well established, both at the experimental and theoretical level, that heterogeneity is a characteristic common to most of the tissues and cellular systems [107,108]. In fact, the level of heterogeneity is predicted to increase as we move towards a molecular state, while homogeneity decreases [109]. Therefore, the assumption derived from bulk studies that cells derived from one tissue have homogeneous genetic profiles has been masking the natural cellular heterogeneity in our cultures of interest [110–112]. It soon became clear that developing a platform with a higher resolution and obtain a picture of each individual cell would hold the key to fully characterize unique cellular populations and understand how they affect the homeostasis of a specific tissue.

1.2.1. Single-cell RNA-sequencing

Single-cell NGS technologies provide us with rich datasets that can dissect cellular heterogeneity at a genetic, transcriptomic, and epigenetic level. Understanding this heterogeneity is vital to know how a biological system develops, how it is regulated, both at a homeostatic and disease state, and how it responds to exposure to different stimuli. Indeed, single-cell data shows that cells in a particular tissue are much more heterogeneous than previously thought. This allowed for a deeper classification of previously known cellular populations by uncovering new subpopulations and the characterization of their interactions and phenotypic responses to a given perturbation [113–115].

While DNA can provide information on common variants and mutations for a cell population, much of the cellular heterogeneity is observed at a phenotypic level and reflected on the transcriptome of each individual cell. Due to its unique position in the flow of genetic information, Ribonucleic Acid (RNA) can reveal not only modifications on the functional elements of the DNA code but also capture changes at the level of gene expression, splicing variants, and other post-transcriptional modifications. Therefore, RNA can be considered one of the central sources of cellular heterogeneity and thus cellular transcriptomics became one of the best studied fields at the single-cell level [116]. Single-cell RNA-seq (scRNA-seq) can comprehensively profile the transcriptome of multiple individual cells. The high resolution and statistical descriptive potential of such datasets give us an in-depth picture of gene expression at the single-cell level and allow us to identify rare and unknown cellular subpopulations [117,118]. scRNA-seq has been a staple method for characterizing tissue and tumor cellular heterogeneity, cellular lineages, fates, and trajectories, identifying disease specific cellular populations and respective drug resistances and susceptibilities [114,119–123].

In general, scRNA-seq protocols follow a common methodological backbone that consists of i) isolation of viable single-cells or nuclei, ii) reverse transcription of messenger RNA (mRNA) transcripts, iii) complementary DNA (cDNA) amplification, iv) library preparation, and v) NGS sequencing (Figure 4) [124]. Currently, most of the available protocols can be divided into two main categories: full length or 5’- and 3’-end transcript coverage. Full-length protocols such as SMART-Seq2 and MATQ-seq are preferable to study the transcriptional profile of a specific cellular population, in particular its splicing variants, allele specific expression, isoform quantification, and lowly expressed genes due to their high sensitivity [125–128]. On the other hand, 3’-end protocols such as Chromium, Drop-seq,

and inDrop, can capture a larger amount of cells with a lower sequencing cost [129–131]. While some full-length methods rely on plate-based techniques, most of the 3'-end technologies use microfluidic droplet-based systems [127,128]. Droplet-based platforms require less sample volume and capture more cells per assay, making them more suitable for massive analysis of single-cell expression profiles than plate-based technologies [132–134]. Moreover, droplet-based platforms also allow for sample multiplexing [135]. Multiplexing approaches are based on the hashing technique, which consists of incorporating antibody or lipid-based labels that are unique for each sample before pooling them in one sequencing reaction[136,137]. This approach increases throughput, decreases the amount and consequently the costs of reagents, significantly reducing the cost of library preparation due to the increase of cells processed per reaction [135]. Together with their simple application and optimization, droplet-based 3'-end technologies are currently the most popular cell isolation pipelines for describing the cellular heterogeneity of uncharacterized tissues.

Although scRNA-seq technologies are potent tools to characterize tissue heterogeneity, there are still major challenges at the technical level. When compared to bulk experiments, the amount of captured RNA per cell is lower and the experimental method is more extensive, which causes scRNA-seq protocols to have a higher technical bias and variation. In order to address this issue, several methods have incorporated the use of cellular barcodes with Unique Molecular Identifiers (UMIs). These are used to identify unique mRNA sequences so that each of them is only counted once, preventing PCR amplification bias during the analysis of the generated data [138,139].

Each scRNA-seq experiment produces high dimensional raw data leading to the generation of massive volumes of complex data, driving the development of novel computational methods capable of processing, storing, and analyzing it [140]. Due to the low quantity of processed cells, this sequencing technology has low capture efficiency and high number of dropouts (unsuccessful detection of truly expressed gene transcripts), technical noise, and biological variation [128]. Together, all these factors promoted the development of algorithms, software tools, and packages for the pre-processing and downstream analysis of the generated data [140].

scRNA-seq data analysis follows a standard workflow: quality control (QC), read mapping and alignment, gene expression quantification, normalization and biases correction, dimensionality reduction and clustering (Figure 4). QC can be done to evaluate sequencing

quality, by using for instance FastQC, and eliminate low mapping reads as they probably result from RNA degradation [141]. After gene expression quantification, QC is also performed to filter out low-quality data, such as damaged, dead or aggregated cells. This can be achieved by discarding genes that have minimal expression across cells and cells that have single gene counts [142]. Furthermore, assessing the percentage of mitochondrial genes per cell, which should not be higher than 10%, reveals if there is a leakage of cytoplasmic mRNA, an indication of a ruptured cell [143].

Normalization is performed in order to minimize count and amplification biases. Due to the high heterogeneity and inherent sparsity derived from the numerous dropouts present in scRNA-seq data, new normalization algorithms were developed [144]. scTransform pipeline from the Seurat framework is a state-of-the-art normalization algorithm for UMI-based data and it uses negative binomial regression model [145]. Furthermore, this pipeline also removes potential batch effects, which may introduce biological variability and technical errors that cause the expression of certain genes to be systematically different when compared to others [146].

Since scRNA-seq quantifies the gene expression profile of each captured cell, count matrices have inherently high dimensionality. However, only a few biologically relevant dimensions can define the expression profiles for each cell [147,148]. Therefore, feature selection and dimensionality reduction are performed to identify the most informative features (components) in the data that enable us to analyze and visualize it in a low dimensional space [149]. Seurat is currently the state-of-the-art platform used to perform these steps in scRNA-seq [150]. Based on the assumption that the variation observed at the biological level can be explained by genes that have high cell-to-cell variability, feature selection is performed to identify highly variable genes (HVG). Then, Principal Component Analysis identifies the components that capture most of the variability of the dataset using HVGs as an input. To identify the cellular (sub)populations present in the data, Seurat uses a K-nearest neighbor graph-based clustering method. Finally, this pipeline uses t-distributed Stochastic Neighbor Embedding and Uniform Manifold Approximation and Projection to visualize the identified clusters.

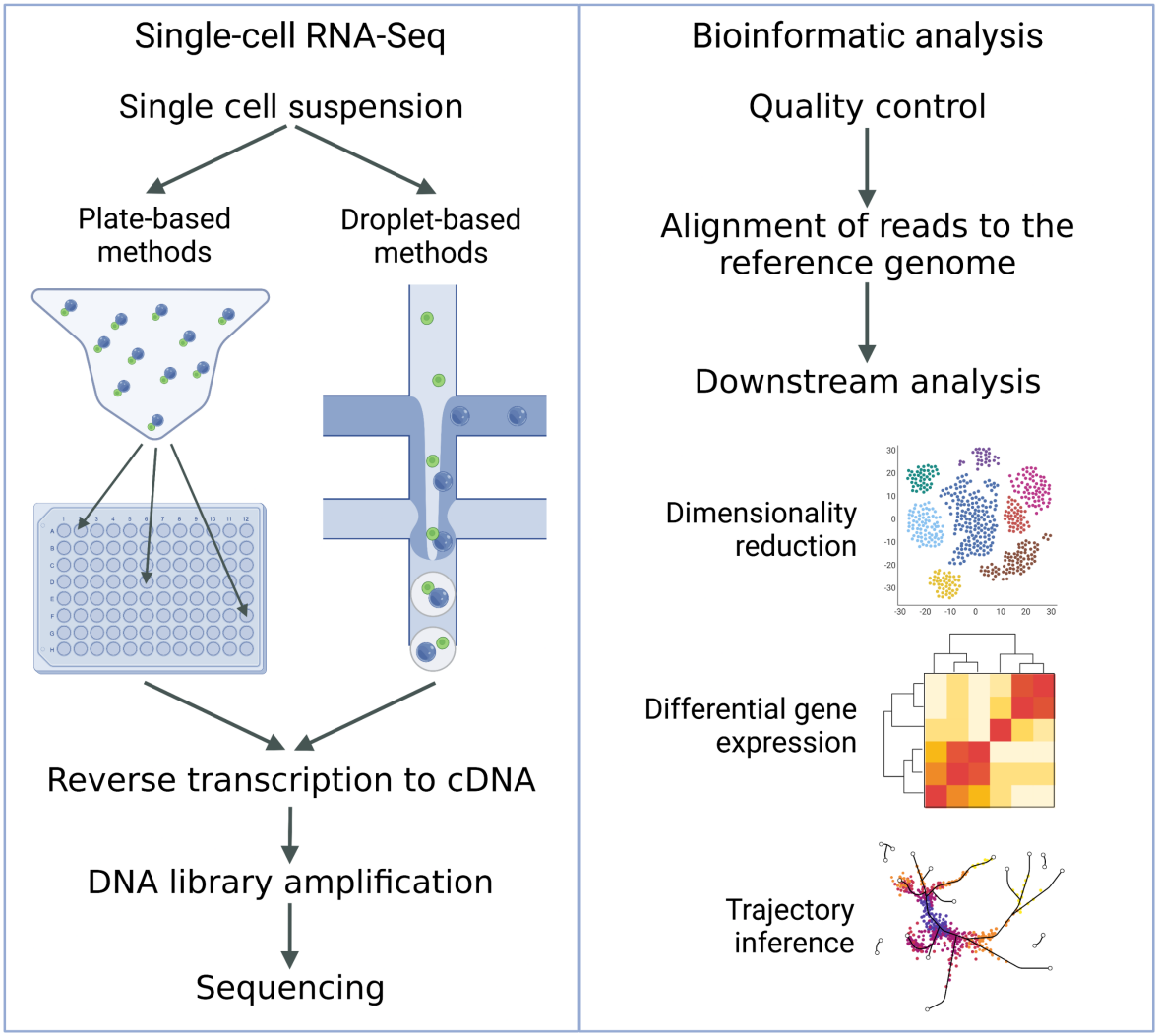


Figure 4: Single-cell RNA-sequencing (scRNA-seq) workflow. The first step on a scRNA-seq protocol is the isolation of viable single-cells or nuclei. Currently, most of the available protocols can be divided into two main categories: full length, such as SMART-Seq2, or 3'-end, such as Chromium. Full-length methods rely mostly on plate-based techniques, while 3'-end technologies use droplet-based systems. Once each cell or nucleus is in a single well or droplet, the protocol follows a general workflow consisting of reverse transcription of mRNA into cDNA sequences, cDNA amplification, DNA library preparation, and NGS sequencing. scRNA-seq data analysis follows a standard workflow: quality control, read mapping and alignment, gene expression quantification, and downstream analysis. Dimensionality reduction is the pillar of this analysis since it is used to identify novel clusters of cell subtypes. These cell subtypes can then be characterized by the identification of marker genes using differential expression analysis or by inferring their trajectory based on pseudotemporal ordering. Created with BioRender.com.

Once the cellular (sub)populations are identified, the downstream analysis is vast and complex (Figure 4). scRNA-seq downstream analysis consists of identifying genes and pathways that are enriched in each cellular (sub)population and the corresponding gene

regulatory networks (GRNs). Seurat is commonly used to identify marker genes in each cluster, Single-Cell Regulatory Network Inference and Clustering (SCENIC) is one of the state-of-the-art algorithms for network reconstruction and CellNet, Mogrify or TransSynW can be used to identify cellular conversion TFs [151–154]. We can also perform cluster annotation, lineage tracing, and trajectory inference. scMap is a well-known tool for literature-based cell type annotation while velocyto and Monocle2 are the most used tools for trajectory inference [155–157].

Although scRNA-seq technologies allowed us to study the uncharted territory of tissue and cellular heterogeneity and its underlying mechanisms, transcriptomics alone does not explain the intricate gene expression profiles characteristic of each cellular population. Indeed, it has been shown that epigenetic profiles, composed of active enhancers, expression of specific DNA binding proteins and open chromatin areas, also play a role in defining individual cell types. Therefore, being able to integrate scRNA-seq with other omics data would allow us to obtain a more representative depiction of the unique biology of each cellular population.

1.2.2. Leveraging the combination of multi-omics data

Single-cell transcriptomics data unravels the cellular heterogeneity of a tissue by capturing cellular subpopulations that express specific gene profiles. However, gene expression is a complex process which is tightly regulated by other mechanisms, such as epigenetic features [158]. Combining transcriptomic information together with available data on epigenetic marks, such as chromatin accessibility, TF-DNA binding interactions and/or histone modifications, can comprehensively describe how the epigenetic landscape modulates gene expression to explain cellular heterogeneity.

In that regard, the combination of scRNA-seq information with ChIP-seq data has been widely used to characterize the regulatory mechanisms behind specific gene expression patterns [159,160]. ChIP-seq is a bulk sequencing method that focuses on identifying DNA-binding proteins, such as TFs, as well as histone modifications [100]. Briefly, ChIP starts with crosslinking proteins bound to DNA regions, followed by DNA fragmentation and immunoprecipitation of protein-DNA complexes using antibodies specific to the protein or histone modification of interest. Finally, the linkage between the precipitated DNA and protein is removed and the released DNA is sequenced. In summary, ChIP-seq can provide

direct evidence of TF binding to DNA sequences, such as promoters and enhancers, an indicator of gene transcription. This data has been extensively used as a gold standard for validating regulatory relationships between TFs and target genes, including the ones predicted by GRN-based methods [161,162].

Another layer of epigenetic information relates to spatial organization of chromatin. It has been shown that the physical proximity between promoters and enhancers increases the transcription levels of tissue specific genes [163,164]. Therefore, Hi-C was developed to capture the conformation of chromatin in a three-dimensional space [165]. Data provided by this technique elucidated regulatory relationships between enhancers and promoters and has been used to develop computational methods that can infer these regulatory interactions [165–168].

Chromatin accessibility is also an important epigenetic feature that is related to activation of gene expression [169]. ATAC-seq is a technique used to profile opened chromatin regions across the genome and provide some insights about the regulatory mechanisms behind unknown gene expression events [101,170]. This method uses hyperactive Tn5 transposase to simultaneously cut and introduce adaptors in accessible chromatin regions which will be used to recognize the DNA fragments for amplification and sequencing. Finally, the sequencing analysis maps these DNA fragments to the genome and identifies "peaks" corresponding to areas of open chromatin. This technique has been extensively used to profile active regulatory regions involved in cell type specific gene expression [171,172].

Important resources have been developed to leverage the combination of these types of epigenomics data. ChIP-Atlas is a comprehensive and most up-to-date database that integrates several layers of epigenetic information, such as ChIP-seq, ATAC-seq, and bisulfite sequencing (BS-seq) for multiple organisms, including human [162,173]. In addition, GeneHancer is an extensive human database that combines bulk multi-omics data, namely ChIP-seq and Hi-C, to identify connections between enhancers and their target genes [174]. Furthermore, computational methods leveraging the combination of single-cell multi-omics, namely scRNA-seq and single-cell ATAC-seq (scATAC-seq), have been extensively used to characterize the regulatory landscape of specific cell (sub)types [159,175,176].

As seen, multi-omics approaches have the potential to extend our knowledge on the epigenetic processes regulating gene expression and to provide a comprehensive image

of the regulatory landscape and respective mechanisms underlying specific cellular functions.

1.3. Computational methods to model cellular systems

Advancements in NGS technologies, namely in scRNA-seq and its integration with other omics data, opened new doors to the development of computational methods that can accurately model the intricate molecular interactions underlying the functioning of complex processes in biological systems. To understand a given molecular process, one has to characterize the interactions between molecules and the effect of these interactions on individual pathways and processes [177]. After determining these features, it is possible to classify the overall impact of each molecule on the biological system in study. However, the dynamics of cellular systems can be convoluted since their molecular processes are extremely intertwined and, thus, difficult to isolate and characterize.

Computational modelling can guide our understanding in biological systems by mimicking their dynamics. Computational methods make use of mathematical algorithms to model molecular interactions and provide insights on how complex biological systems behave under physiological conditions or upon perturbation. These models have been used to describe several types of complex molecular processes, such as regulation of gene expression, signaling pathways, and protein folding [178–180]. Understanding the full complexity underlying each of these processes and how they relate with each other provides crucial insights to better understand cellular systems not only at a physiological level, but also in disease development.

1.3.1. Gene regulatory network inference

Regulation of gene expression is one of the most complex and vital processes in cellular systems. Gene expression levels can be measured by quantification of mRNA, a molecule which production is tightly regulated at many different levels [181]. First, chromatin availability is controlled by epigenetic marks, such as histone acetylation and DNA methylation, which modulate the accessibility of regulatory elements, such as promoters and enhancers [182,183]. Second, the binding of TFs to these regions as well as other cofactors controls the transcription of DNA sequences by regulating the recruitment of the transcriptional machinery towards the transcription starting site [184,185]. In eukaryotic systems,

produced pre-mRNAs undergo processing modifications, such as capping, polyadenylation, and splicing, a process that leads to the production of distinct mature mRNAs from the same initial transcript [184–186]. Once the final mRNAs are produced, they are transported by RNA binding proteins to the cytoplasm where they are either translated by ribosomes or degraded by exoribonucleases [181,187]. Together, these regulatory processes mediate the activation or repression of specific genes, which is reflected in the expression of the corresponding proteins, conferring to the cellular system the ability to perform specific functions.

The interplay between all these regulators of gene expression can be modelled as a GRN, a network describing the regulatory interactions among genes [177]. In that regard, the topology of GRNs depicts the specific interactions between TFs and genes (Figure 5). GRNs are normally represented as directed graphs, with source nodes representing TFs and directed edges describing their regulatory interactions with target nodes (genes) [188]. Furthermore, GRNs can also provide information regarding the type of interaction between nodes, specifically activation or repression. Exploiting the topology of GRNs can guide the identification of hubs or master regulators, that display a high degree of connectivity, and provide additional mechanistic insights about a given cellular system [189,190].

GRNs can be generated using literature-based evidence and/or prior knowledge databases, such as MetaCore from Clarivate Analytics, a manually curated database of molecular interactions [191]. However, the data provided in these resources can be noisy as it derives from the collection of evidence described in distinct experimental conditions, cell types and tissues. Therefore, GRNs solely based on prior knowledge might lack information regarding the specific condition (e.g., healthy or disease) and cell (sub)type, which is important to understand the behavior of a particular cellular system (e.g., disease).

GRNs can be inferred directly from experimental data by identifying significant relationships between genes to accurately capture the dynamics of gene regulation in a specific biological context. To detect statistically relevant regulatory interactions among TFs and genes, measurements such as correlation or mutual information can be used [192,193]. One of the state-of-the-art GRN inference strategies, named SCENIC, uses scRNA-seq data to infer putative TF-gene regulatory relationships [151]. This approach uses GENIE3, a method based on random forest and coexpression, to detect regulatory interactions and then refines these interactions using TF-motif enrichment analysis [151,194]. Most of the currently available GRN inference methods are based on scRNA-seq data, allowing for the generation of highly

contextualized biological networks and providing crucial information for specific conditions or cell (sub)types [148,177].

However, GRN inference methods perform poorly overall when benchmarked against ChIP-seq data, the gold standard strategy to assess the accuracy of the predicted TF-gene regulatory interactions [195]. Indeed, one of the major drawbacks for these methods is that they solely rely on the single-cell data, which is known to be noisy [177,195]. In addition, the measurements and assumptions used to capture these regulatory interactions usually produce a high number of false positive edges. The combination of these limitations lowers the performance of GRN inference methods in predicting accurate regulatory relationships between TFs and genes.

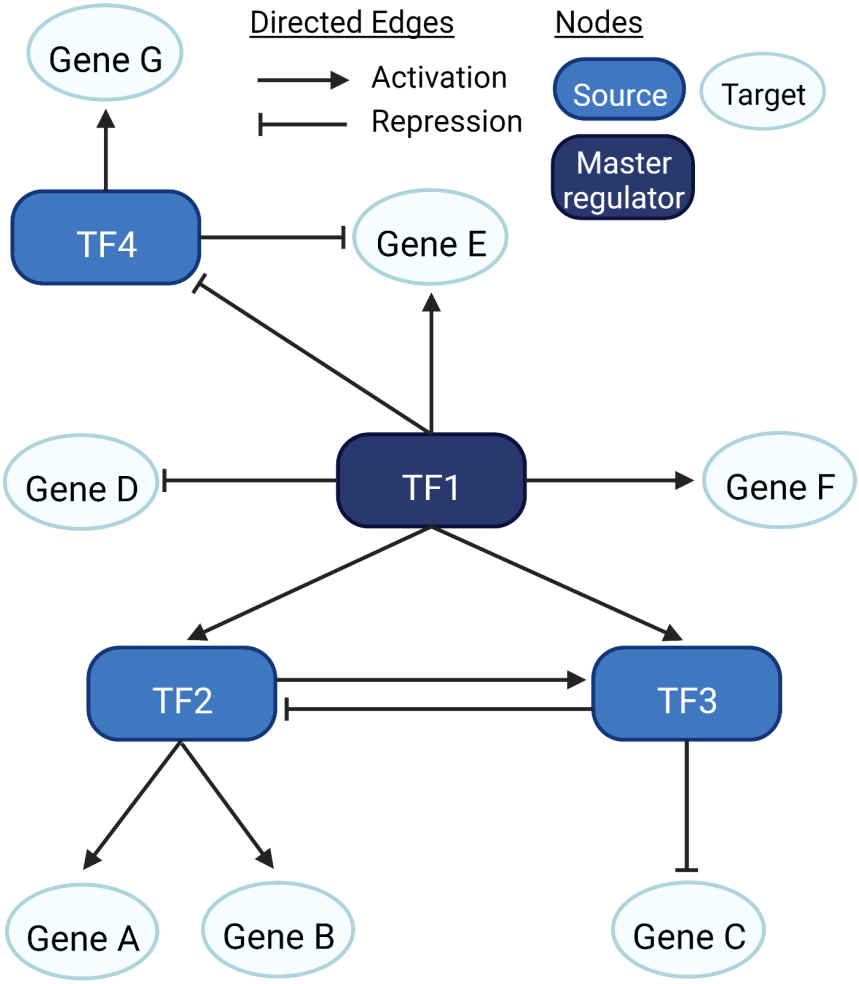


Figure 5: Gene regulatory network (GRN) representation. GRNs shows intricate regulatory interactions between TFs and genes. These networks are represented as directed graphs, with edges connecting source and target nodes. Interactions between nodes can be classified as activation or repression. GRNs can help to identify master regulators based on a high degree of connectivity. Created with BioRender.

The integration of prior information in GRN inference, such as currently available repositories of ChIP-seq data, would improve the accuracy of these methods in modelling transcriptional regulatory landscapes [161,162,195]. Indeed, several GRN-based approaches have been developed to integrate different layers of information [175,196–198]. In that regard, a GRN inference approach integrates several layers of multi-omics data, such as bulk RNA-seq, ATAC-seq, and ChIP-seq, to identify cellular conversion factors that successfully improve cell differentiation protocols [196]. Furthermore, GRNs have also been used to study cis-regulatory interactions in cellular (sub)populations based on the analysis of scRNA-seq and scATAC-seq data obtained from the same cell [175].

Leveraging the information obtained from the combination or integration of multi-omics data has been shown to be a promising strategy to develop computational methods that can accurately model the behavior of different cellular systems and characterize the inherent cellular heterogeneity [193,198].

1.3.2. Transcriptional synergy as an emergent feature of cellular identity

The identity of a given cellular system has been described to be determined by specific sets of TFs that regulate unique gene expression profiles which, in turn, are responsible for conferring distinct features to the cellular system [199]. The precise identification of these groups of TFs, named identity TFs, is still a challenge [79]. Multiple computational methods have been developed to address this limitation by leveraging the high-resolution power of scRNA-seq data [154,200,201]. One of the main strategies used to determine identity TFs is the characterization of master regulators in GRNs [189,190]. However, these methods only capture pair-wise relationships between TFs and genes, and do not account for the potential multifactorial interaction among them [154,201].

TFs have been shown to synergistically regulate gene expression [202–204]. The main regulators of pluripotency and cellular identity in embryonic stem cells (ESCs), *NANOG*, *POU5F1*, and *SOX2*, have been shown to co-occupy the same promoter regions [202]. Also, it was shown that *ISL1* directly interacts with *LHX3* and with *PHOX2A* to control the cellular identity of spinal cord and brain motor neurons, respectively [203]. In hematopoietic stem cells, *GATA2*, *RUNX1*, and *SCL* have been described to act as identity TFs for this lineage and to synergistically interact as protein complexes to bind to adjacent DNA sequences [204].

Literature evidence supports the existence of transcriptional synergy mechanisms that profoundly regulate cellular identity. Synergistic activity is one of the suggested models to explain how just a few identity TFs are able to regulate the expression of such complex and identity-specific gene programs [203,205]. This model proposes that transcriptional regulators, such as TFs, act cooperatively by binding to the same regulatory elements in order to induce cell type or subtype-specific gene expression profiles (Figure 6) [206]. It is also proposed that these transcriptional regulators assemble in a protein complex to perform their regulatory functions and thus it is not possible to predict their effects by considering only their individual activity.

Based on this concept, computational methods have been using Multivariate Mutual Information (MMI) to capture these synergistic interactions among TFs [191,200,201,207].

$$MMI(A;B;C) = I(A;C) + I(B;C) - I(A,B;C)$$

Considering the TFs A, B, and C, MMI quantifies the additional information obtained about C that cannot be described by the sum of the knowledge obtained from the pair-wise interactions between A and C (i.e., *I(A;C)*), and B and C (i.e., *I(B;C)*) [207,208]. The information measured by MMI is based on Shannon's entropy which is calculated, in this case, from the TF's gene expression values obtained from RNA-seq data [209]. Since entropy is an uncertainty measurement, an information gain equals a reduction on the uncertainty and, therefore, a negative MMI value [208,210]. To use this concept as a synergy measurement, one has to assume that a negative MMI value indicates a synergetic relationship between the TFs since the information of A and B together gives more knowledge about C (i.e., *I(A,B;C)*) than the sum of the pair-wise interactions of C with A and with B (Figure 6). Based on this principle, TransSyn, a scRNA-seq based computational algorithm, was able to capture synergistic TFs that control the identity of the human medial floor plate midbrain progenitors, which are precursors of DANs [201].

Focusing on the inherent characteristics of TFs, namely how they cooperate to control cellular identity, is a promising strategy to improve the generation of homogenous and functional cellular (sub)populations and the development of regenerative medicine applications, such as cell replacement therapy (CRT).

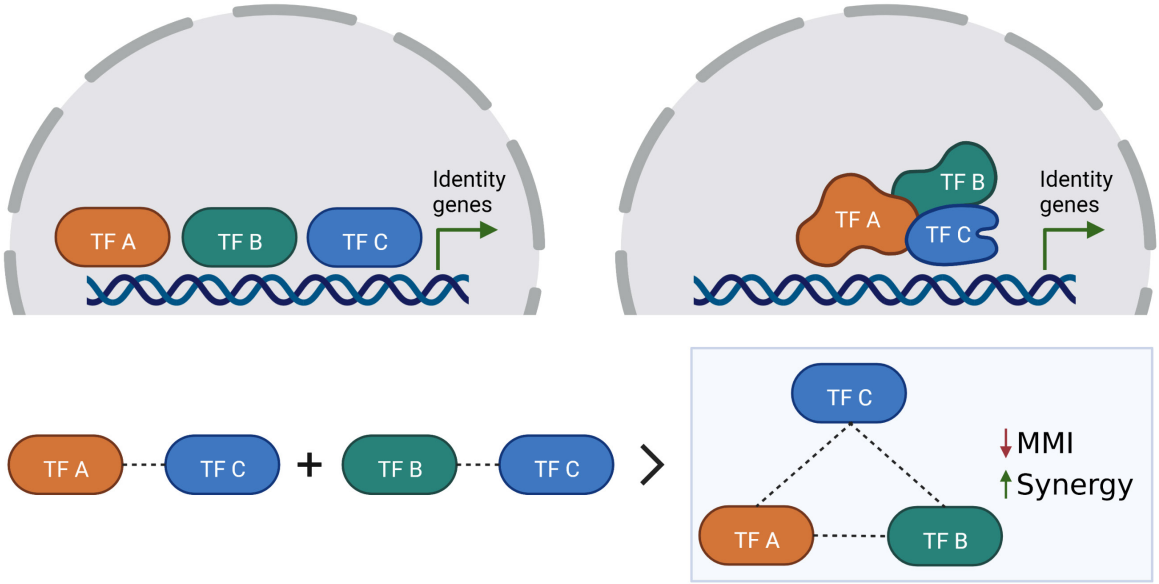


Figure 6: Synergistic interaction between transcription factors (TFs). Mechanisms of cooperative interaction include binding of TFs to the same regulatory region and assembly of TF complexes to regulate the expression of cell (sub)population specific gene programs. Multivariate Mutual Information (MMI) can capture these synergistic interactions. A negative MMI indicates a synergetic interaction between TFs A, B, and C because the information captured by A and B gives more knowledge about the C than the sum of their pair-wise interactions. Created with BioRender.

1.4. Cellular identity and computer guided strategies for cell replacement therapies

During normal development, pluripotent cells start differentiating and acquiring an individual cellular identity by gradually committing to specialized cellular lineages until they become fully differentiated into functional cellular (sub)populations [211]. Cellular identity is defined by the activity of specific genes that maintain the functional and phenotypic features of a given cellular system [199]. Specific sets of TFs, named identity TFs, promote the expression of this unique transcriptomic profile [212–214]. This distinctive profile leads to the production of individual proteins and to the activity of pathways that confers to the cells of a given population their distinct functions. However, gene expression is also determined by another layer of information. Chromatin configuration is key for the expression of certain genes. Even when the right combination of identity TFs is expressed in the cell population of interest, if the chromatin availability is not favorable to the expression of the TF's target genes, the transcriptional program will not be fully activated [215,216]. Thus, combining the

expression of identity TFs together with epigenetic modifiers seems to hold the key for improving current cellular conversion protocols.

1.4.1. Transcriptional and epigenetic regulatory mechanisms

Every cell contains approximately the same DNA sequences [158]. However, each organism has different cell types that carry out unique functions which are determined by specific variations in gene expression [77]. For this reason, fine tuning of gene expression is essential for appropriate cellular development and differentiation. This process is tightly regulated both at the epigenetic and transcriptional level [217,218].

Histone modifications and DNA methylation are two of the main mechanisms behind the control of chromatin conformation [216]. The nucleosome is the central unit of chromatin, and it comprises a core octamer of histones proteins, named H2A, H2B, H3 and H4, wrapped around by DNA [77]. These core histones are vulnerable to covalent modifications in the lysine (K) and arginine residues, namely acetylation and methylation [182]. Acetylation of histone's lysine residues is mediated by histone acetyltransferases. This process removes the positive charge of histone tails, decreasing their binding strength to the negatively charged DNA, inducing chromatin opening and making DNA binding sites available for transcriptional activation [183]. Histone methylation has bimodal effects on gene expression, depending on the methylated residue [219]. For instance, methylation of histone H3 at the K9 and K27 residues is linked to gene repression while methylation of the residues K4, K36 and K79 is associated with gene activation. DNA methylation is mostly linked to repression of gene expression, and it involves methylation of cytosine residues located 5' of a guanosine (CpGs) by DNA methyltransferases [220,221]. The regions of the DNA where CpGs occur at a higher frequency are denominated by CpG islands. These loci are usually not methylated and associated to promoters which implicates them as regulators of gene expression. The combination of these epigenetic modifications confers to each cell type a unique chromatin conformation which strongly conditions the expression of cell type specific genetic programs [222].

Transcriptional regulation is also essential for the maintenance of cellular identity [199]. TFs are the central regulatory factors that control gene expression. These proteins recognize short sequences of DNA in regulatory elements, such as promoters and enhancer regions, where they bind. TFs are composed of several domains, which gives them the ability

to connect to RNA polymerase II, chromatin remolding factors, cofactors, other TFs and transcriptional regulators [223]. Depending on the overall function of the bound factors, the TF complex can behave as an activator or repressor of gene expression [185]. Although about fifty percent of the known TFs are expressed in all cellular populations, it has been shown that a small group of specific TFs is fundamental for the activation of cell type specific gene expression profiles [224–226]. These sets of identity TFs have been identified in multiple cellular systems. For instance, *SOX2*, *KLF4*, *c-MYC*, and *OCT3/4* have been extensively used to reprogram somatic cells into induced pluripotent stem cells (iPSCs) [227]. *MYOD1* has been shown to induce direct reprogramming of fibroblasts to myoblasts while regulating the expression of *FOXA2* and *OTX2* promoted the differentiation of neuroepithelial stem cells to medial floor plate midbrain progenitors [201].

Since these sets of TFs are responsible for controlling cellular identity, one of the most widely used experimental methods to convert between cellular populations is to promote their overexpression. However, it has been shown that, if the chromatin conformation is not favorable to the binding of identity TFs, the conversion process will be hindered [228]. Recently, a specific subset of TFs named pioneer factors (PFs) has been characterized [229]. These proteins have been shown to be main regulators of cellular differentiation and development due to their unique ability to bypass epigenetic constrains and promote conversion between cellular identities [230–232]. Unlike traditional TFs, PFs can bind to their recognition sites in regulatory elements even if these regions are under closed chromatin [230]. PFs induce chromatin opening mechanisms and, consequently, gene expression, by first priming and then promoting the activation of enhancers [233–235]. One of the main features of this group of TFs is their ability to bind to nucleosomal DNA (Figure 7A) [236]. This binding loosens the interaction between core histones and DNA, which is followed by a reduction in DNA methylation and an increase in H3K4 methylation (Figure 7B) [230,237]. These effects occur rapidly, priming the enhancer for its activation [237,238]. Upon suitable stimuli, such as the presence of hormone-responsive factors or the expression of identity TFs, PFs recruit other elements, such chromatin remodeling complexes and non-pioneer TFs, that contribute to the full activation of enhancers (Figure 7C) [239–241]. This activated state is characterized by depletion of nucleosomes, recruitment of the coactivators p300 and CREB-binding protein, and acetylation of H3K27, all markers of an opened chromatin conformation [229,242]. Once the stimulus has ceased, the activated enhancers lose these features, but remained primed by PFs (Figure 7D) [229].

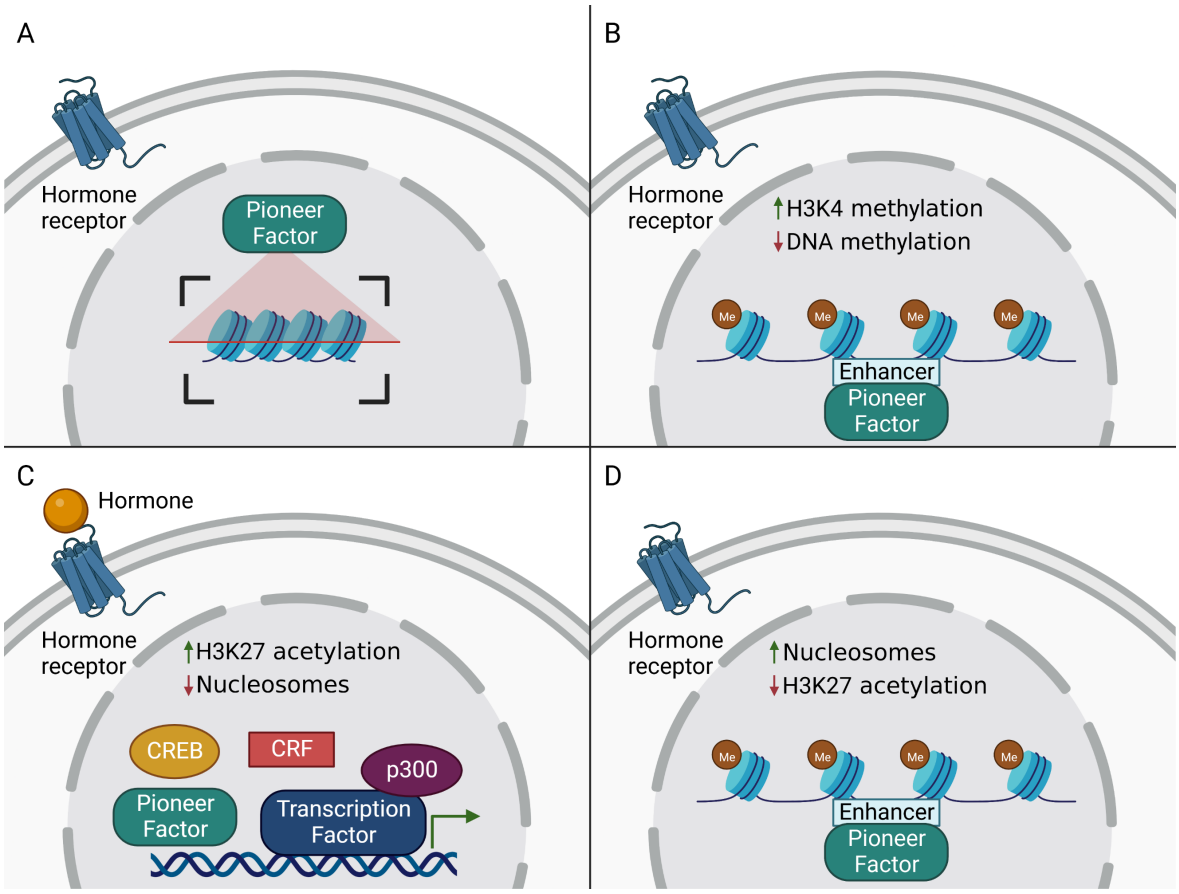


Figure 7: Pioneer factor (PF) mechanism of action. **(A)** PFs start by scanning the genome for their recognition sites in regulatory elements, such as enhancers, even if these regions are under closed chromatin. **(B)** Once bound, PFs prime the enhancer by promoting chromatin opening mechanisms, such as reduction in DNA methylation and an increase in H3K4 methylation. **(C)** Upon suitable stimuli, such as the presence of hormone-responsive factors, PFs recruit other elements, namely chromatin remodeling factors (CRF) and non-pioneer TFs, resulting in the full activation of the enhancer. This activated state is characterized by depletion of nucleosomes, H3K27 acetylation and recruitment of the coactivators p300 and CREB-binding protein. Once the chromatin is fully opened, the transcriptional machinery is recruited to the promoter site and gene expression is initiated. **(D)** When the stimulus is ceased, the activated enhancer loses these features, but remains primed by PFs. Created with BioRender.com.

In summary, PFs are able to bind and open previously inaccessible regions of the chromatin through priming and activation of enhancers. It has been shown that cell type specific enhancers are the primary regulators of cellular identity in cell populations that derive from similar lineages [237,243]. Identity TFs and PFs have been shown to play an important role in defining cellular identity by facilitating the expression of cell type specific genetic programs and, consequently, improving cellular conversion protocols [229].

1.4.2. Identification of cellular conversion transcription factors

In 1957, Dr. Conrad Hal Waddington postulated his view on cellular development [244]. He proposed that, during normal development, ESCs progress down a unidirectional path that leads them to a fully mature and differentiated state. This idea consisted of a marble rolling down from the top of a hill, representing the pluripotent state, to a valley, which depicts the restriction that occurs as cells differentiate. This model, referred to as Waddington's "epigenetic landscape", represents the classical view of lineage specification and cell fate commitment. However, after iPSCs were obtained from fully differentiated cells, it was shown that the somatic state can be reversed to pluripotency [227]. Furthermore, when fibroblasts were directly converted to myoblasts, it was shown that cellular conversion between somatic cell types was possible without going through the pluripotent state [245]. Later, fibroblasts were also shown to be able to be directly reprogrammed into neurons, a somatic cell type that originates from a separate germ layer [246]. Together, these studies showed that, although applicable under normal development, the unidirectionality of Waddington's model could not explain the vast cellular plasticity.

The processes of inducing the conversion in between cellular populations were divided into three classifications, namely differentiation, reprogramming, and direct reprogramming. Reprogramming is defined as the conversion from a somatic cell to iPSCs and direct reprogramming is the conversion between two somatic cell types. These conversion protocols highly rely on the expression of identity TFs which override the gene expression profile of the initial cell type, promoting stable structural and functional changes characteristic of the cell type of interest [227,247]. Identification of these identity TFs, which will be referred to as conversion TFs in this dissertation, is essential to enhance the conversion between cellular populations.

The development of novel computational methods opens doors to an efficient characterization of conversion TFs due to their ability to swiftly screen over the thousands of known TFs [248,249]. The majority of these computational models use gene expression data as an input [152–154]. In particular, GRNs have been extensively exploited to capture conversion TFs [152,153,250]. Most GRN-based methods start by generating a DE profile between source and target cell types based on gene expression data [251]. Then, they built the regulatory network around the overexpressed genes and determine its hubs [152,153]. In graph theory, hubs are identified based on their connectivity degree, meaning the node's outdegree and indegree

[189]. In a directed graph, like GRNs, outdegree corresponds to the number of targets a given node is regulating, while indegree corresponds to the number of nodes regulating it. Nodes with the highest outdegree in the network are usually classified as hubs or master regulator TFs [189,190]. The central role of these main regulators in controlling specific gene expression profiles and consequently cellular identity, makes them the most suitable candidates to induce cellular conversion [193,252].

GRN-based methods have been used successfully to promote the transition between cellular populations [152,153] . For instance, Mogrify identified the TFs necessary to convert human fibroblasts into keratinocytes which were further converted into vascular endothelial cells [153]. CellNet determined *POU2AF1* and *EBF1* as TFs that improve the conversion from human B lymphocytes into macrophages and identified *FOXA1* and *HNF4α* to be inducers of the conversion from mouse fibroblasts to hepatocytes [253].

These strategies are merely based on bulk population studies, where the measurements of gene expression are averaged over a heterogeneous population of cells. Therefore, their individual variability at the transcriptional level is masked within the culture or tissue of interest, making the resolution of the data a limitation for optimization purposes. Ground-breaking developments in gene expression profiling at a single-cell level allowed us to classify cells into distinct subpopulations, and to characterize the TFs that drive cellular transitions [154,200,201]. The generation of single-cell based GRNs led to the discovery of *HOX* and *SOX* as key players in the regulation of the hematopoietic lineages and the identification of *ESRRB*, *NANOG*, and *TBX3* as identity regulators of naïve mouse ESCs [200,254,255].

Additionally, a synergy-based measure of information theory has also been used to determine conversion TFs between cell subtypes and phenotypes using single-cell transcriptomics data [201]. This method identified *FOXA2*, *OTX2*, and *LMX1A* as identity TFs that promote the conversion of neuroepithelial stem cells into medial floor plate midbrain progenitors which further differentiate into DANs [201]. Based on the same concept, another method prioritizes PFs during the identification of conversion TFs [154]. PFs have been extensively demonstrated as being facilitators of epigenetic modifications that lead to the expression of genes inaccessible to TFs without pioneer features [233–235]. Leveraging this subset of TFs has the potential to further improve the outcome of cellular conversion protocols.

Specific cellular subpopulations and phenotypes, such as midbrain DANs, have been shown to play a critical role in disease development due to their selective loss or dysregulation [31,256]. Identifying cell subpopulation-specific conversion TFs holds a great potential for application in the development of CRT, since it would allow us to shift and enrich cell preparations in these subpopulations.

1.4.3. Direct cell reprogramming techniques

Direct reprogramming protocols facilitate the conversion between somatic cellular populations without transitioning through a pluripotent state. This technique has a major potential in clinical application as it allows cellular conversion to occur directly in the damaged tissue [257]. This process avoids proliferation of immature cells and diminishes the risk of tumor formation, has a shorter protocol time, and does not require immunosuppression [258]. Direct reprogramming protocols are an attractive alternative to stem cell-based CRT. Expressing the necessary genetic programs that induce conversion into different cell types requires precise changes in gene expression over a broad set of genes. The most commonly used strategy in these protocols involves the overexpression of conversion TFs using viral vectors, such as lentiviruses and adenoviruses. At first, the selection of these TFs was based solely on developmental biology studies [227,246,259]. Currently, with the development of scRNA-seq technologies and computational algorithms that decipher cellular identity, these TFs can be identified more precisely [152–154].

Ectopic gene expression of the selected TFs has been widely used to directly convert between differentiated cell types. This strategy induced the conversion of mouse fibroblasts into cardiomyocytes by overexpression of *Mef2c*, *Tbx5*, and the PF *Gata4*, and into myoblasts by using the PF *MyoD* [245,259]. Overexpression of *Pou3f2* and *Myt1l* together with the PF *Ascl1* successfully induced the generation of glutamatergic neurons from mouse hepatocytes [260]. However, this approach has some disadvantages. Besides containing the cDNA of the gene of interest, the vectors used in these protocols must also contain the necessary elements for promoting transgene expression. Due to the insert DNA size limit associated to viral vectors, it becomes very difficult to multiplex the expression of three or more genes in one single vector. This limitation forces the delivery of the conversion TFs in different vectors which might decrease the homogeneity of the induced gene expression and, therefore, reduce the efficiency of the direct reprogramming protocols.

Alternatively, endogenous gene regulation mediated by Clustered Regularly Interspaced Short Palindromic Repeats (CRISPR)-based systems has been shown to be an attractive tool for direct cellular reprogramming. Briefly, this approach relies on the recruitment of CRISPR associated protein 9 (Cas9) to genomic sequences defined by short guide RNA (gRNA) molecules [261]. These molecules target the coding region of a specific gene and recruit the CRISPR complex to induce a double strand break in the DNA, permanently silencing the expression of the target gene. On the other hand, a catalytically inactivated version of Cas9 (dCas9) can be fused to a strong repressor or activator complex and used to regulate gene expression instead [261]. In this approach, gRNAs target primarily DNA regions within the promoter of the gene of interest, recruiting the CRISPR-dCas9 complex to control the transcription or accessibility of a target gene.

Since the CRISPR targeting system is based on gRNAs, these approaches have a high multiplexing capacity. Furthermore, these systems are highly flexible, since they can be used to silence, repress and/or activate endogenous gene sets [262–264]. For instance, CRISPR-Cas9 mediated deletion of *MyoD* induced the conversion of mouse myoblasts into brown adipose tissue while upregulation of *Ascl1*, *Pou6f2*, and *Mytl1* genes using CRISPR-dCas9 coupled to the activator complex VP64 facilitated the generation of neurons from mouse fibroblasts [262,263]. Moreover, using a CRISPR-dCas9 dual transactivator system, it was possible to directly reprogram mouse astrocytes into GABAergic neurons *in vivo* by overexpressing *Ascl1*, *Lmx1a*, *Neurod1*, and *Nr4a2* [265].

Direct reprogramming protocols have been successfully used to generate a wide range of functional cellular populations. Although the limited efficiency of these protocols still poses problems, these advancements open doors to improving the efficiency and functionality of the generated cells [266]. Together with the ability of generating cellular populations *in situ*, direct reprogramming holds great potential in the development of CRT and regenerative medicine.

1.4.4. Application in regenerative medicine for brain diseases

The goal of regenerative medicine is to promote tissue regeneration and repair. CRT focuses on creating functional and faithful cell types that can be used to repair damaged tissues. However, reproducing the target cells' phenotype still poses problems. One of the major limitations of this process is the characterization of TFs that induce an effective

conversion into the cellular population of interest. The application of computational modeling to the identification of these conversion TFs greatly benefited the field.

Developing models to identify conversion TFs based on different methods, such as DE gene profiles, GRNs, and synergy-based measure of information theory, has significantly decreased the amount of resources and time spent in selecting and/or screening TFs for their conversion potential [152,153,201]. Furthermore, with the advancements in scRNA-seq based computational methods and the identification of unknown subpopulations, the first steps have been taken to further developing computational models that can predict conversion TFs to generate specific cellular subpopulations (Figure 8). Being able to generate these cell subtypes would have a profound application in different fields of regenerative medicine, such as disease modelling, screening for neuroprotective drugs, and especially in the development of CRTs.

Generating the cell (sub)types of interest has great potential in CRT applications targeting brain-related disorders since a major hallmark of these pathologies is neuronal degeneration (Figure 8) [267]. Advancements in strategies to replenish these cells and the identification of the degenerative mechanisms involved in the disease would increase our ability to regenerate a tissue [268].

Direct reprogramming protocols are one of the main strategies being explored for CRT, since they have the key advantage of being able to be performed *in situ*, avoiding the transplantation process. Directly converting glial cells, such as reactive astrocytes, into neurons would allow the replacement of this cell type [84,269]. Simultaneously, there would be a reduction of the excess of reactive glia, a hallmark in several brain-related pathologies, such as AD and PD, promoting the reestablishment of the balance between astrocytes and neurons in the brain [270]. Notably, human astroglia were successfully reprogrammed into induced DANs (iDANs) by overexpression of *LMX1A*, *micro-RNA 218* and the PFs *ASCL1* and *NeuroD1 in vitro* [271]. The generated iDANs were capable of expressing midbrain-specific TFs and characteristic dopaminergic markers. *In vivo*, this combination of TFs was also successful in generating iDANs from mouse astrocytes and rescuing motor symptoms in a PD mouse model [271]. In the same disease model, direct reprogramming of mouse astrocytes into GABAergic neurons using CRISPR-dCas9 technology partly corrected impaired motor and behavioral functions in mice [265]. Together, these studies demonstrate the potential of *in situ* applications in CRT for the treatment of PD.

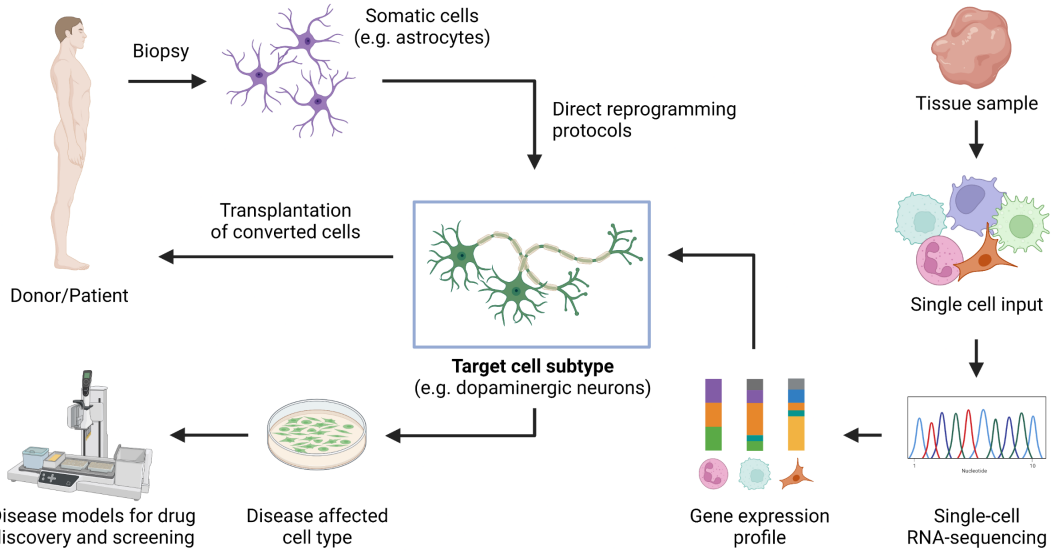


Figure 8: Direct reprogramming strategies based on single-cell RNA-sequencing and their applications in regenerative medicine. With the advent of single-cell technology, it is possible to dissociate a tissue sample into single cells and characterize the gene expression profile of each cell. Based on the similarity between these expression profiles, cells can be clustered into distinct subpopulations. By analyzing the gene expression pattern of the target subpopulation, we can determine the molecular features, such as conversion TFs, that would be required to generate this cell subpopulation. Conversion TFs can be applied to direct reprogramming protocols that induce the conversion between somatic cell types without transitioning through a pluripotent state. Due to their promising *in situ* application, direct reprogramming protocols have become an attractive alternative to stem cell-based cell replacement therapy (CRT). One of the major CRT applications would be to replenish neuronal subtypes lost during the onset of neurodegenerative diseases, such as Parkinson's Disease (PD). Being able to generate, for instance, dopaminergic neurons would not only advance CRT, but also allow to model the effects of PD on these cells and to screen for neuroprotective drugs. Created with BioRender.com.

Besides integrating PFs in cellular conversion protocols, this subset of TFs also has other potential therapeutical applications. Promoting the overexpression of *ASCL1* in glioblastomas has been shown to decrease the tumorigenicity capacity and stem-like features of the tumor as well as to promote neuronal differentiation [272].

Characterizing disease-associated mechanisms behind the destabilization of cellular identity would extend our knowledge of disease onset and progression [273,274]. For instance, stress factors associated with PD were shown to disrupt the expression of the transcriptomic profile of human DANs, which led to the loss of their cellular identity [256]. The development of computational methods that leverage the high resolution of scRNA-seq

to provide mechanistic insights associated with cell (sub)type specific dysregulation would extend our understanding of disease progression and accelerate the development of novel strategies for regenerative medicine.

1.5. Characterization of cell type specific dysregulated mechanisms

Cellular mechanisms, such as transcriptional regulation, signaling pathways, and protein production are complex molecular processes that maintain the correct functioning of the cell. Dysregulation of these mechanisms can trigger disease onset and development [275–277]. For instance, dysregulation of protein folding, oxidative stress, and secretory pathways are some of the molecular processes involved in the development of neurodegenerative diseases, such as AD and PD [275,276]. Another clear example of how cellular dysregulation can promote disease onset is cancer. Dysregulation of cell proliferation, signaling pathways, and cell cycle are some of the hallmarks related with the development of tumor cells and cancer progression [277].

Dysregulated mechanisms in disease do not affect all cellular populations equally [278–282]. For instance, in neurodegenerative diseases, astrocytes become reactive, inducing changes in their secretory pathways and proliferation [49,278]. Neurons experience alterations on their protein folding processes which contributes to their death [281]. On the other hand, DANs in PD are severely affected by protein misfolding, oxidative stress, and impaired cellular transport, prompting the selective degeneration of this cellular subpopulation during the development of this pathology [282].

The dysregulation of these mechanisms may arise due to variants on the genome found both in coding and non-coding regions [283,284]. If these alterations induce the dysregulation of gene expression above a certain threshold, specific cellular mechanisms will be impaired, which ultimately leads to the disruption of cell function and disease onset [284]. These modifications can affect the coding region of the gene, which causes the production of a nonfunctional, mutant protein that will affect the cell's phenotype [285]. On the other hand, these variants can also be present in the regulatory regions of a gene, such as promoters and enhancers, which will impair the regulation of gene expression [283]. Indeed, it has been shown that the majority of single nucleotide polymorphisms (SNPs) occur in regulatory elements, mainly enhancers [286]. This observation can help explain why alterations in the DNA sequence of regulatory regions have been shown to be related with changes in

gene expression detected in disease [283,287]. Based on this evidence, it becomes clear that deciphering cell type specific dysregulated mechanisms would allow us to better understand disease and determine relevant targets for the development of novel therapies.

1.5.1. Dysregulation in gene expression mechanisms

Gene expression regulation is a key biological process that is responsible for controlling the expression of specific gene profiles that confer to the cell its characteristic functions. Modelling the transcriptomic landscape of cellular systems and characterizing its molecular interactions and responses upon perturbation can guide our understanding of these processes in health and disease states.

The main driver of gene expression dysregulation is the occurrence and/or accumulation of genetic variants in the coding and regulatory regions of a gene [288]. These alterations can be classified as i) mutations, including insertions, deletions, and substitutions, ii) or polymorphisms [289]. For a variant to be classified as a polymorphism, it must be present in the genome of at least one percent of a population. The most common polymorphisms are SNPs. To identify these genetic variants, Genome Wide Association Studies (GWAS) have been extensively used to profile the genomes of individuals under different phenotypical conditions [290]. These studies provide information on SNPs identified in genomic loci that are significantly associated with a particular trait, such as a disease [290,291]. Until now, GWAS has identified more than fifteen different risk variants for PD associated to *SNCA* and *LRRK2* genomic loci [292].

Most of the SNPs unveiled by GWAS are located in non-coding regions of the genome, specifically in enhancers [287]. The enrichment of SNPs in regulatory regions suggests that these variants are involved in impairments at the gene regulatory level. This relationship can be analyzed using expression Quantitative Trait Loci (eQTLs) studies [293]. An eQTL is defined as a genomic locus that has a statistically significant association with the variance of expression of a given gene [294]. eQTL analysis can identify eQTLs located closer or further away from the modulated genes. eQTL studies performed at a genome-wide level provided additional insights into the connection between the presence of SNPs and the disruption of cellular functions in autoimmune disorders [295].

Given the effect of SNPs in the regulation of gene expression and their prevalence in enhancer regions, integrating this information in GRNs would give us a more

comprehensive view of the complex mechanisms behind dysregulation in disease. Recent studies have been focusing on studying the impact of disease-related SNPs in tissue specific regulatory landscapes by building networks based on enhancer-gene interactions [296–298]. The effect of SNPs on the binding affinity of TFs to regulatory elements has also been studied [299,300]. Being able to generate GRNs that can convey a complete view of impaired regulatory interactions mediated by TFs and enhancers of regulated genes would provide key mechanistic insights to unravel dysregulation in disease.

1.5.2. Potential in understanding complex disease onset and development

GRNs have been used as a map of molecular interactions to provide crucial insights on dysregulated transcriptomic mechanisms in disease [301]. Comparing GRNs built based on gene expression data obtained from healthy and disease conditions is one of the main strategies used to identify transcriptional regulators involved in disease-related mechanisms [302]. By analyzing the changes that occur between the two GRNs, it is possible to capture regulatory interactions impaired in disease, the master regulators involved in these impairments, and the associated genes. Thus, this analysis can provide us with important therapeutical targets [303]. For instance, *RGS2* was identified as the main regulator of *LRRK2* activity by generating a GRN centered in *LRRK2* using transcriptomics data from PD patients [304]. In that same study, *RGS2* was also identified a promising therapeutical target for individuals with the *LRRK2* mutation, since this signaling gene has been shown to have a protective role against neuronal toxicity.

Several strategies have also been developed to characterize the impact of disease-related SNPs in regulatory networks. Fine mapping identifies genetic variants that have a strong functional association with a given trait, such as a disease [305]. Recently, a study fine mapped SNPs to GRNs in order to identify the master regulators involved in pathways dysregulated in disease [300]. Another study characterized impaired mechanisms associated with genetic variants linked to type 1 diabetes based on cell (sub)population specific GRNs [306]. Finally, a network-based approach leverages single-cell data to characterize the impact of disease-associated genetic variants in cell (sub)types and trajectories [307].

Despite these recent advances, GRN-based approaches lack insights on impaired mechanisms mediated by enhancers and TFs affected by disease-related SNPs. Developing such methods based on single-cell data would allow us to profoundly characterize dysregulated mechanisms associated to disease at the cell (sub)type level and identify potential targets for gene therapy.

2. Scope and Aims

2.1. Scope

As the world population ages, neurodegenerative diseases become more prevalent in society, calling for an immediate action in the development of effective treatments. Since neurodegeneration is a major hallmark of these disorders, regenerative medicine holds great potential in proving the first treatment against neuronal death and provide a cure for brain diseases. However, multiple challenges still need to be addressed before regenerative medicine strategies can be translated to clinical applications. The advent of single-cell technologies provided high resolution data that paved the way towards the development of novel computational platforms that can address these issues. In this PhD dissertation, we leverage single-cell data to achieve two major goals: i) developing computer guided strategies for the generation of specific neuronal subtypes with a phenotype similar to what is found in the adult brain, ii) and unveiling additional molecular insights underlying the development of brain diseases.

2.2. Aims

This PhD dissertation aimed at developing innovative computational and experimental approaches that can address current obstacles in regenerative medicine, at providing potential biomarkers that can improve disease prognosis, and identifying regulatory mechanisms impaired during disease development.

Aim 1: Identification of novel conversion TFs that improve the efficiency and generation of functional cell (sub)types. To achieve this aim, we developed TransSynW, a single-cell based computational approach that prioritizes PFs in the identification of cell conversion TFs. It has been shown that the conformation of the chromatin is one of the major obstacles to the generation of functionally mature cell (sub)types. However, current computational platforms do not consider the impact of epigenetic features in the generation of partially reprogrammed cells. Our method addresses this limitation by prioritizing PFs in the identification of cell conversion TFs. PFs are able to bind to closed areas of the chromatin and induce the expression of their target genes. By adding this novel feature, we are promoting chromatin reorganization, facilitating the binding of TFs to regulatory regions,

inducing the expression of cell (sub)type identity genes. TransSynW is a user-friendly computational platform that we believe will improve the success of cellular conversion protocols and lead to the generation of novel cell (sub)types for regenerative medicine.

Aim 2: Establish a pioneer protocol that leads cell conversion towards specific subtypes. We develop a direct and sequential reprogramming protocol to enrich cell cultures in specific subtypes of DANs. Replacing the DAN population lost during the development of PD has been shown to alleviate the motor symptoms associated to this disease. Some studies suggest that reactive astrocytes, which arise during the development of PD, can play a role in neurogenesis. Therefore, CRT has been an area of focus in the development of potential treatments for neurodegenerative diseases, such as PD. Current direct reprogramming protocols have been able to convert astrocytes into DANs, but with low efficiency and without control over DAN subtype specification. Our protocols are based on ectopic and endogenous expression of TransSynW's conversion TFs. These strategies aim at generating a homogenous cell preparation of a single DAN subtype which can be further used in the development of clinical applications.

Aim 3: Identify specifically expressed TFs to determine important prognosis biomarkers. In this study, we focused on identifying specifically expressed TFs cellular lineages found in glioblastoma (GBM). GBM is one of the most aggressive types of cancer. It has been shown that the cellular lineage from which the tumors derive has an impact on their properties and determines distinct molecular characteristics. The cellular lineage of origin and molecular mechanisms associated to mesenchymal GBM, the most aggressive subtype of GBM, still remain elusive. We identified and validated specifically expressed TFs in the two main cellular lineages of origin found in GBM, providing potential prognosis' biomarkers and therapeutical targets.

Aim 4: Provide additional mechanistic insights on the impact of disease-associated SNPs in the impairment of regulatory interactions. To achieve this, we developed RNetDys, a multi-omics pipeline to identify regulatory mechanisms impaired due to disease-associated SNPs. These have been found mostly at enhancers regions which have been described to control regulatory mechanisms at the cell (sub)type level. RNetDys builds cell (sub)type specific GRNs using a combination of scRNA-seq, scATAC-seq, and prior knowledge data to identify specific dysregulated interactions due to the presence of any SNP. This pipeline identifies impaired mechanistic interactions mediated by TFs and enhancers of regulated

genes, providing important mechanistic insights in disease development with potential applications in gene therapy.

2.3. Originality

The projects presented in this PhD dissertation address two main challenges in the development of clinical applications for brain diseases. They focus on improving cellular conversion protocols by targeting the epigenetic landscape, achieving control over subtype specification, identifying cellular lineage-specific TFs, and deciphering the role of disease-associated SNPs in the impairment of regulatory mechanisms. The findings described on this PhD dissertation leverage single-cell data to develop novel computer guided strategies with potential applications in regenerative medicine. Additionally, we identify novel therapeutical targets and molecular mechanisms underlying the progression of brain diseases, opening doors to the development of clinical applications.

3. Materials and Methods

A detailed description of the materials and methods applied to obtain the outcomes included in this PhD dissertation can be found in each of the sections of the Results chapter (chapter 4). A brief summary of the methodology applied in each of the published articles and manuscripts included in this dissertation is presented below.

In section 4.1, we performed an extensive a literature review concerning transcriptomics-based computational approaches that address current limitations in identifying cellular identity TFs at the subpopulation level, cell fate determinants, and lineage specifiers in the scope of stem cell biology.

In section 4.2, we developed a computational method, TransSynW, to identify cell conversion TFs for cellular populations described in scRNA-seq data. TransSynW starts by identifying specifically expressed TFs in the target cellular population by Jensen-Shannon divergence (JSD). In this context, JSD measures the difference between the observed and the ideal gene expression value of a given TF [191,308]. If a TF is specifically expressed in a cellular population, its expected expression value after normalization should be 1 in this population and 0 in the background populations. The lower the difference between the expression levels of a TF and 1, more specific the TF is for the target cellular population. Then, TransSynW selects the set of specifically expressed TFs in the target cell population with the highest synergy. The same synergy-based calculation is performed to identify the non-specifically expressed PFs. The identified sets of TFs and PFs are then ranked by expression fold change in relation to the starting cellular population. Finally, TransSynW identifies marker genes for the target cellular population by determining specifically expressed genes based on JSD. We validated our method by cross-referencing the identified TFs and markers against literature-based evidence and validated the biological meaning of the newly identified TFs and markers using a manually curated database for molecular interactions.

In section 4.3, we developed novel strategies to directly reprogram human astrocytes into neuronal subtypes by regulating the expression of novel conversion TFs at the endogenous and ectopic level. TransSynW was applied to a previously published scRNA-seq dataset that characterized the human fetal ventral midbrain tissue to obtain the conversion factors for each of the identified DANs subtypes. This method was also used to

determine the conversion factors for a two-step reprogramming protocol. First, we identified the TFs that induce the conversion of astrocytes into DANs, and then the TFs responsible for the specialization into DAN subtypes. For the experimental protocols, we adapted a previously established CRISPR-dCas9 system to induce the overexpression of the predicted conversion factors at the endogenous level. Briefly, we engineered a cell line of human astrocytes to constitutively express dCas9 and infected these cells with lentiviruses containing gRNAs targeting each of TFs together with an activator sequence. We also promoted the expression of our target TFs using an inducible lentiviral expression system containing the corresponding cDNAs. To evaluate the results of our direct reprogramming protocols we performed real-time polymerase chain reaction and immunofluorescence assays for TUBB3 and TH, a neuronal and a dopaminergic marker, respectively.

In section 4.4, we profiled the transcriptome of GBM and low-grade glioma cells at the single-cell level and identified the cellular lineages of origin in each of the patients using a novel neural network approach. Based on these results, we determined the specifically expressed TFs in the two main cellular lineages identified in GBM samples using JSD followed by synergy measurement. The specific expression of the identified TFs was validated by immunofluorescent assays in patient-derived GBM xenografts.

In section 4.5, we developed RNetDys, a multi-omics systematic pipeline to identify impaired regulatory interactions due to disease-associated SNPs. Based on the combination of healthy scRNA-seq and scATAC-seq with prior knowledge data, this pipeline builds cell (sub)type-specific GRNs to profile the regulatory landscape of our target cellular (sub)populations. Based on the constructed GRNs, RNetDys identifies impaired regulatory interactions by mapping disease-associated SNPs and evaluating their impact in the binding affinity of TFs to regulatory regions. As part of the output, we provide a ranked list of TFs mediating the impairment of the regulatory mechanisms based on the topology of the network, the binding affinity score, and the minor allele frequency score calculated for each SNP involved in the dysregulated mechanisms. We examined the accuracy and precision of RNetDys by benchmarking against state-of-the-art methods. We validated our method by collecting SNPs associated with AD, PD, EPI, and diabetes from a publicly available database and infer the relevance of the identified dysregulated interactions against literature-based evidence, GWAS, and eQTL studies.

4. Results

4.1. Computational Methods to Identify Cell-Fate Determinants, Identity Transcription Factors, and Niche-Induced Signaling Pathways for Stem Cell Research

4.1.1. Preface

The advent of NGS technologies increased access to large amounts of data in different omics areas, such as transcriptomics, epigenomics and genomics. Computational biology leveraged this data and developed more accurate, precise and sophisticated algorithms that unravel complex molecular relationships between biological systems. This is particularly relevant in stem cell research, where in the recent years transcriptomics-based approaches allowed us to characterize important cellular phenotypes and understand which factors are responsible for their identity, cellular fate, and how the cellular microenvironment (niche) contributes for maintaining or shifting of cellular phenotypes.

Here, we provide a comprehensive review of several computational methods, based on bulk and scRNA-seq data, that address the mechanisms behind cellular conversion, differentiation, and niche specific pathways. These methods rely on synergistic activity to determine core identity genes, network modelling to detect cell fate determinants, identification of signaling molecules responsible for maintaining cellular phenotypes. We elaborate on how these methods can contribute to address current challenges on the stem cell field and propose how they can be improved to further contribute to important milestones in stem cell research.

In this chapter, I described the methods that determine identify TFs, characterize signaling pathways in cellular niches, and discussed their impact and potential in the stem cell field. The published book chapter is reprinted on the next pages (RightLinks license number 5324211238457).

Chapter 4

Computational Methods to Identify Cell-Fate Determinants, Identity Transcription Factors, and Niche-Induced Signaling Pathways for Stem Cell Research

Muhammad Ali, Mariana Messias Ribeiro, and Antonio del Sol

Abstract

The large-scale development of high-throughput sequencing technologies has not only allowed the generation of reliable omics data related to various regulatory layers but also the development of novel computational models in the field of stem cell research. These computational approaches have enabled the disentangling of a complex interplay between these interrelated layers of regulation by interpreting large quantities of biomedical data in a systematic way. In the context of stem cell research, network modeling of complex gene–gene interactions has been successfully used for understanding the mechanisms underlying stem cell differentiation and cellular conversion. Notably, it has proven helpful for predicting cell-fate determinants and signaling molecules controlling such processes. This chapter will provide an overview of various computational approaches that rely on single-cell and/or bulk RNA sequencing data for elucidating the molecular underpinnings of cell subpopulation identities, lineage specification, and the process of cell-fate decisions. Furthermore, we discuss how these computational methods provide the right framework for computational modeling of biological systems in order to address long-standing challenges in the stem cell field by guiding experimental efforts in stem cell research and regenerative medicine.

Key words Gene regulatory networks, Stem cell research, Cellular reprogramming, Cell-fate determinants, Lineage specifier, Core identity TFs, Systems biology

1 Introduction

The human body comprises different organs, constituted by different cell types, that are supposed to perform dedicated tasks. These cells can be divided into different categories based on, e.g., their ability to differentiate and their function in the body. For example, cells that have the ability to renew themselves and differentiate into any kind of cells are called stem cells—a prominent example is the mammary stem cells (MaSCs) that are known to form the two main

Muhammad Ali and Mariana Messias Ribeiro contributed equally.

Maria dM. Vivanco (ed.), *Mammary Stem Cells: Methods and Protocols*,
Methods in Molecular Biology, vol. 2471, https://doi.org/10.1007/978-1-0716-2193-6_4,

cellular lineages of the human mammary gland [1, 2]. Stem cells further give rise to progenitor cells that have a limited spectrum of differentiation as they can only differentiate into particular specialized cell types. For instance, basal and luminal progenitor cells differentiate into basal/myoepithelial cells that exhibit contractile capacity and luminal cells capable of producing milk, respectively [2]. Although underlying genetic material is the same in each cell of the body, it is the cross-talk between the epigenetic and transcriptional regulatory machinery that controls the identity of different cell types by only allowing specific regions of the genome to be accessible for the transcriptional machinery and being transcribed. Cell identity specification is considered to be determined by cell-specific gene-expression programs that are tightly controlled at the chromatin level. It is widely understood that cell identity is mostly regulated by the action of transcription factors (TFs) that are very particular in their tendency to recognize and bind to specific genomic sequences and regulate transcription [3, 4]. In order for a cell-specific gene-expression program to be expressed, the genomic regions corresponding to the cell type-specific TFs and their distal regulatory regions must be in euchromatin state, harboring active chromatin modifications. Although the number of TFs that are generally expressed in a cell type is considered to be the half of all known TFs [5], a small group of them, commonly known as core TFs, are suggested to be critical for defining the cell identities [6–9].

In some particular cases during cellular differentiation, the expression levels of two inter-dependent genes or TFs determine the fate of the cell through a toggle switch. A toggle switch consists of two genes or TFs that repress each other in a mutual fashion. Usually, this regulatory motif is active during differentiation of a stem/progenitor cell (parent state) into two different lineages (daughter states) and is thought to act as a memory device, being able to choose and maintain cell-fate decisions [10]. In this biomolecular battle, overexpression of each TF corresponds to one of the two mutually exclusive daughter cell fates, where one TF inhibits the other TF and subsequently activates its lineage-determining target TFs and genes. In contrast, a stabilized expression of both TFs maintains the stem/progenitor state [11, 12]. These parent and daughter cellular states are characterized by stable gene-expression programs, determined by underlying gene regulatory networks (GRNs) and the constituent subnetworks (GRN motifs) that are functionally important. A toggle switch is in fact a classical GRN motif that constitutes a molecular mechanism determining the cell-fate decisions and providing stability to transcriptional programs of binary cell-fate choices. A well-known example is the mutual inhibition of an erythroid determinant *GATA1* and a myeloid determinant *SPI1*, two TFs responsible for the development of

erythroid and myeloid blood cells from common myeloid progenitors in the hematopoietic stem cell (HSC) system [13, 14].

The natural ability of stem cells to generate more specialized cell types in a coordinated manner and accurately regulate their activity makes them indispensable for maintaining the tissue homeostasis [15]. This intricate decision of either self-renewing or differentiation into different cell fates is controlled by multiple cell-intrinsic and extrinsic factors. One such important factor is the interactions between stem cells and their in vivo microenvironment, also known as a niche. In their niche, stem cells receive the stimuli that determine their behavior, such as maintaining the dormant state (quiescent state), or either induce self-renewal or differentiate into a particular cell fate (active state) [16]. These stimuli include external cues such as cell–cell and cell–matrix interactions which are translated by the niche into intracellular signaling events that activate and/or repress genes and transcriptional programs. The hypothesis of a specialized stem cell niche was postulated by Schofield in 1978 in his description of the hematopoiesis process [17]. He hypothesized that a cellular niche has a defined anatomical location where a stem cell must be associated with other cells in its environment. This environment determines the stem cell behavior and losing its association with the cellular niche results in the differentiation of stem cells. To this end, molecular signaling pathways in the niche are recognized as important modulators of stem cell maintenance and function. These signaling pathways are redundant in different niches but have different roles according to the specific niches. For example, integrins, heterodimeric transmembrane extracellular matrix (ECM) receptors, consisting of various α and β subunits, have been implicated in the control of stem cell maintenance and progenitor cell differentiation in various embryonic and adult tissues [18, 19]. In particular, α2-, α6-, β1-, and β3-integrin subunits have been shown to serve as surface markers for the isolation of MaSCs populations from basal and luminal mammary epithelial layers, suggesting the important role of ECM in MaSCs microenvironment [20, 21]. Therefore, elucidation of molecular signaling pathways that maintain different stem cell niches, cellular phenotypes, and integrity of the organism is of broad interest and paramount to the design of effective treatments for various human pathologies that are associated with defects in the normal cellular differentiation process.

Challenges exist to identify these properties and single-cell technologies can help with the modeling for addressing these challenges. During the last decade, various experimental techniques have enabled the large-scale generation of high-throughput biological data across different regulatory levels (genomics, epigenomics, and transcriptomics) that led to the development of computational approaches in the field of systems biology and stem cell research. The development of these genome-wide

transcriptomic and epigenetic profiling techniques has enabled more generic endeavors to predict candidate TFs that control cell identity. In this regard, computational methods have been developed to perform genome-wide transcriptomic and/or epigenomic analysis across multiple cell types to identify core TFs critical for cell identity TFs [3, 7, 22–26]. For example, D'Alessio et al. [3] presented a computational approach that relies on transcriptomic data for predicting cell identity TFs and generated an atlas of candidate core TFs for a variety of different human cell types and tissues. Briefly, this computational method searches for the two fundamental characteristics of core TFs: cell-type-specificity and relatively high expression levels. The algorithm they proposed quantifies both these properties by using an entropy-based measure of Jensen–Shannon divergence [27]. They also demonstrated the experimental validation of their core TFs predictions made for reprogramming human fibroblasts into functional retinal pigment epithelial (RPE) cells that possess the morphological and gene-expression features similar to those derived from healthy individuals. Although many computational methods have addressed this challenge of identifying cell identity TFs and several of them have also experimentally verified their predictive power, they all share some important limitations. Foremost, they are unable to systematically integrate the regulatory information from the epigenetic and transcriptomic levels, which has been shown to mediate cell-type-specific gene-expression programs via intricate and interconnected regulatory links as well as controlling the homeostasis of differentiated or pluripotent cells [28, 29]. Most of these methods rely only on transcriptomic data and ignore the fact that cell-type-specificity is also determined by the epigenetic program which is characterized by accessible chromatin regions, active enhancers, and differential binding of regulators [30–32]. Furthermore, these global attempts are broad in their scope as they assess their predictive power using scalable methods and do not systematically evaluate whether predicted factors are sufficient to establish cell identity. In addition, most of the existing methods focus mainly on quantifying the differences between different cell identities and less on the direct identification of TFs and co-factors controlling cell identity.

Similarly, several computational approaches use transcriptomics data for predicting cell-fate determinants during the cellular differentiation process [6, 22, 26, 33, 34]. For example, Okawa et al. presented a computational framework that models the cell differentiation process in the form of GRNs and predicts cell-fate determinants and their GRN motifs [34]. The proposed tool predicted the overexpression of *Esr1* and *Runx2* for the induction of neuronal and astrocytic lineages, respectively, from the mouse neural stem cells (NSCs). Their in silico predictions were also experimentally verified, showing the application of the tool in stem cell research

and regenerative medicine. However, similar to the computational methods aiming for predicting cell identity TFs, these methods are also limited in their usage of bulk transcriptome data while undermining the crucial information from the epigenetic layer of regulation. In addition, none of these methods systematically addresses the problem of cell differentiation efficiency and fidelity, a long-standing challenge in stem cell biology. Unlike most of these approaches that rely on bulk transcriptomic data and large quantities of background or training data sets, only a couple of these methods are able to predict cell-fate determinants at an increased resolution of different cell subpopulations using the single-cell transcriptomic data [35, 36].

The advancements in single-cell sequencing techniques have offered the quantification of expression levels of genes and TFs at single-cell resolution, allowing more reliable measurements of absolute gene-expression levels for a given cell type and its subpopulations. This revolutionary technique has opened new gateways for the development of computational methods that can elucidate complex molecular interaction networks and predict lineage specifiers within a heterogeneous cell population. Similarly, the identification of conserved signaling pathways that provide an accurate description of individual cell behavior while being in a heterogeneous niche requires the accurate characterization of all cellular subpopulations. To this end, sensitive full-length single-cell transcriptome profiling will serve as a basis for the development of single cell-based computational frameworks that can identify niche determinants of different stem cell systems. As the key role of the cell niche in health as well as in several infectious [37] and degenerative diseases [38] is evident, the identification of niche determinants holds the potential to further advance our understanding of underlying disease mechanisms and it can possibly aid in the development of novel therapeutic strategies. Moreover, modeling core GRNs using single-cell data could allow the identification of subpopulations with the highest conversion propensity, thus helping in overcoming the barrier of limited efficiency in directed cellular conversion experiments. Furthermore, single-cell data can help in designing novel experimental strategies to achieve more efficient cellular conversions by predicting cell identity TFs that can initially prime a cell population and then subsequently induce the desired cell type conversion. Thus, guiding experimental attempts for achieving effective in vivo cellular transitions, where limited conversion efficiency is a critical barrier for its application in regenerative medicine.

An overview of currently available computational tools designed to address the above-mentioned three major challenges in stem cell biology is provided in Table 1. Also, a detailed description of state-of-the-art single cell-based computational methods

Table 1
An overview of currently available computational tools designed to use bulk and single-cell transcriptomic and/or epigenetic datasets for addressing challenges in stem cell research, such as predicting cell-fate determinants, lineage specifiers, and identity TFs required for cellular conversions

Method name	Input data	Description	Reference
CellNet	Gene-expression (microarray-based) data from 56 published reports and predefined GRNs	A network biology platform designed to assess the fidelity of cellular engineering and generate hypotheses for improving cell derivations.	[7]
Mogrify	Gene-expression data of 173 human cell types and 134 tissues	A computational framework that combines gene-expression data with regulatory network information to predict the reprogramming factors necessary to induce desired cellular conversion.	[6]
D'Alessio et al.	504 gene-expression profiles, representing 106 cell and tissues types	A computational approach for identifying candidate TFs that control cell identity.	[3]
Crespo et al.	Gene-expression data and predefined GRNs	A computational tool that generalizes the concept of TF cross-repression to predict core TFs for cell conversion.	[39]
Davis et al.	Gene-expression data and chromatin modification ChIP-seq profiles	A computational method for predicting TFs that convert adult cell identity.	[25]
SLICE	Single-cell RNA-seq (scRNA-seq) data	A computational tool that quantitatively measures cellular differentiation states based on single-cell entropy and predicts cell differentiation lineages.	[36]
SeesawPred	Gene-expression data and predefined GRNs	A web application for predicting cell-fate determinants from transcriptomics data.	[33]
Okawa et al.	Gene-expression data and predefined GRNs	A computational algorithm for predicting cell-fate determinants and their GRN motifs.	[34]
Okawa et al.	scRNA-seq data and predefined transcriptional regulatory network (TRNs)	A computational tool for predicting lineage specifiers for different cell subpopulations in binary-fate differentiation events.	[35]
Ravichandran et al.	Gene-expression data, predefined TRNs, and signaling interactome network	A computational approach for identifying niche determinants of cellular phenotypes.	[40]
TransSyn	Single-cell RNA-seq (scRNA-seq) data	A computational platform for the identification of cell population	[41]

(continued)

Table 1
(continued)

Method name	Input data	Description	Reference
		identities by defining their synergistic transcriptional cores.	
SigHotSpotter	Single-cell RNA-seq (scRNA-seq) data and signaling interactome network	A computational web application that predicts key signaling molecules (hotspots) responsible for controlling cell phenotype.	[42]

developed to address these challenges is provided below with their schematic workflow described in Fig. 1.

Cellular differentiation is an intricate process where a less specialized cell type (a stem/progenitor cell, e.g., myoepithelial progenitor cell) evolves into another more specialized cell type (a specific cell, e.g., myoepithelial cell). Despite the recent advances in transcriptomic profiling techniques and increasing research efforts, identification of TFs that determine cell fates during stem cell differentiation (cell-fate determinants) remains a challenge. Identification of cell-fate determinants becomes more challenging for closely related daughter cell types that originate from a common progenitor state such as differentiation of luminal progenitor cells to ductal and alveolar cells [43]. Furthermore, the existence of different cell subpopulations within a particular cell population also hampers our understanding of cellular differentiation processes because of closely related but different gene-expression programs and underlying transcriptional regulatory networks (TRNs) that characterize these states. Therefore, the development of appropriate computational tools is absolutely necessary to utilize the single-cell transcriptomic data for characterizing subpopulation-specific TRNs and identifying TFs that determine specific lineage commitment.

The increasing importance of single-cell gene-expression data in stem cell biology has triggered the development of computational approaches that provide TRN-based modeling of stem cell differentiation at the subpopulation-specific levels and identify lineage specifiers for different cell subpopulations. One prominent example of such a computational framework is the SeesawPred (*see* **Note 1**) [33, 35], a web application that models the cell differentiation process in the form of gene regulatory networks (GRN) and predicts cell-fate determinants. One important feature of this computational method is its flexibility to take into account the single-cell as well as the bulk transcriptomic data for predicting the cell-fate determinants. The application of this computational method to three different stem cell systems predicted already known and experimentally validated lineage specifiers. In addition,

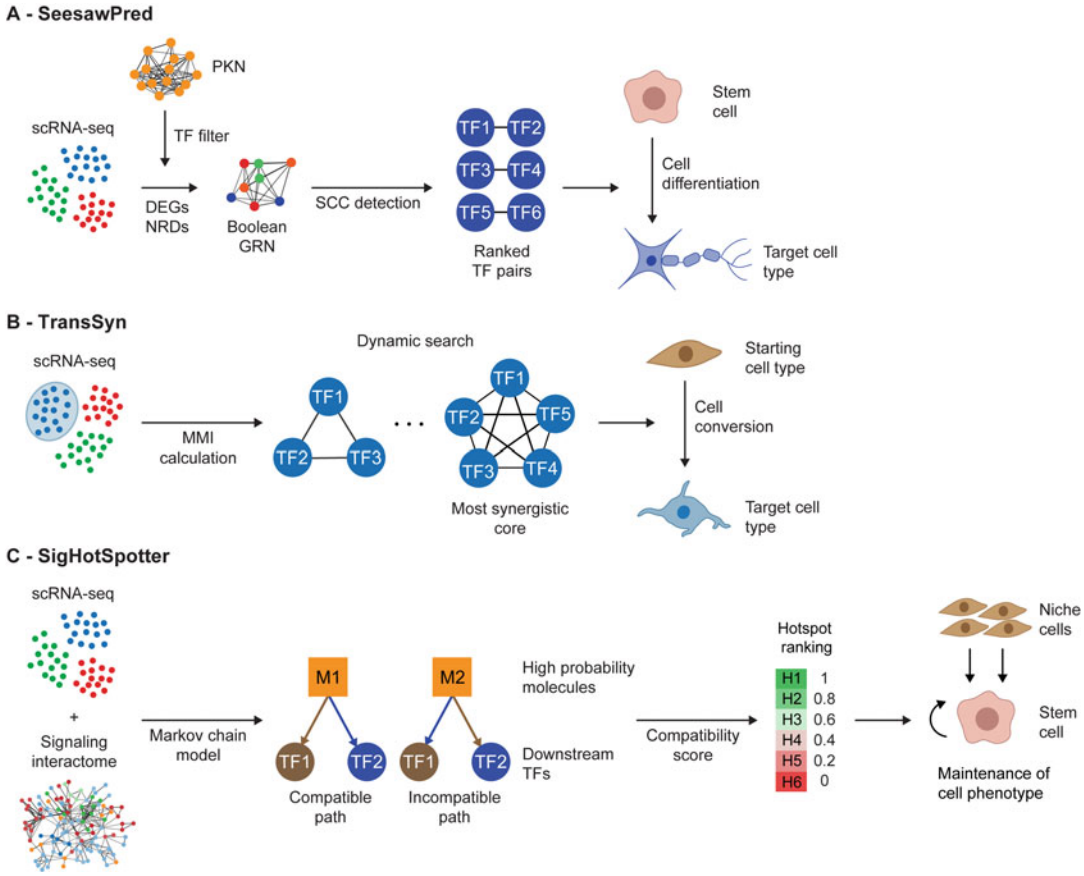


Fig. 1 Workflow of single cell-based computational methods for addressing current challenges in stem cell research. (**a**) SeesawPred workflow overview. This method uses scRNA-seq and a prior knowledge network (PKN) to predict cell-fate determinants, based on a GRN model of cell differentiation. Differentially expressed TFs (DEGs) are identified and the normalized ratio difference (NRD) is calculated for each TF pair. TFs are filtered based on the NRD score and then subdivided based on the provided PKN. The predicted network is pruned using the Boolean GRN formalism and the SCCs identified. Based on the identified SCCs, the final score for each TF pair is calculated and the TF pairs are ranked. The highest-ranking TF pairs are classified as cell-fate determinants and can then be applied in cell differentiation protocols. (**b**) TransSyn workflow overview. This method identifies the most synergistic transcriptional cores of a given population based solely on scRNA-seq data. Based on multivariate mutual information (MMI), TransSyn performs a dynamic search for the most synergistic core by progressively adding TFs to the core, one at a time. The search finishes when the MMI no longer decreases when adding a new TF to the current combination (there is no more increase in the synergy). This last TF combination is identified as the most synergistic core and it can then be applied to cell conversion protocols. (**c**) SigHotSpotter workflow overview. This method predicts key signaling molecules (hotspots) by integrating scRNA-seq data and a pre-compiled signaling interactome in a probabilistic Markov chain model. Based on the compatibility score of the predicted high probability molecules with the respective downstream TFs, SigHotSpotter identifies and ranks signaling hotspots that can then be used to maintain and control a cell phenotype

the experimental validation of two predicted cell-fate determinants confirmed that overexpression of *Runx2* and *Esr1* in mouse NSCs induces astrocyte and neuronal differentiation, respectively [34]. Furthermore, this tool was also employed to interrogate the

differentiation of cardiac progenitor cells to more specialized cell types during normal and abnormal cardiogenesis to study the devastating consequences of disruption in specific cellular subpopulations [44]. The experimental validation of its in silico predictions shows the potential of SeesawPred to guide differentiation experiments in stem cell research and regenerative medicine. The version of SeesawPred that relies on single-cell transcriptomic data is described here [35, 44], whereas the one that uses the bulk transcriptomic data is freely available at https://seesaw.lcsb.uni.lu/.

SeesawPred has been applied to multiple scRNA-seq and bulk transcriptomic datasets, predicting cell-fate determinants and their GRN motifs in multiple stem cell differentiation systems [34, 35, 44]. The application of this approach in stem cell research with therapeutic potential has been demonstrated in cardiogenesis to reveal the basis for organ level developmental defects. The presented computational method revealed the key lineage-specifying TF *Hand2* as a specifier of outflow tract (OFT) cells. Furthermore, it predicted that *Hand2*-null OFT-fated cells have disrupted specification, which was subsequently experimentally verified [44]. In addition, SeesawPred was also used to predict the cell-fate determinants for astrocytic and neuronal lineages from the mouse NSCs. SeesawPred predicted the master regulator *Runx2* and *Esr1* overexpression for astrocytic and neuronal fate, respectively [34]. Moreover, the application of this computational framework to three different stem cell systems predicted already known and experimentally validated lineage specifiers in stem cell subpopulations [35]. For this purpose, three binary-fate stem cell differentiation systems were selected for which high-quality single-cell gene-expression data were available. Precisely, the lineage specifiers were predicted for inner cell mass (ICM), hematopoietic stem cells (HSC), multipotent progenitor (MPP), common myeloid progenitor (CMP), and lung alveolar bipotential progenitor (BP) differentiations. Briefly, SeesawPred predicted the Gata6 and Klf2 to be cell-fate determinants for the primitive endoderm (PE) and epiblast (EPI) differentiation from the inner cell mass (ICM), which is in full agreement with independent experimental observations [45, 46]. Similarly, in line with existing experimental reports, *Hes1* [47] and *Pou6f1* [48] were predicted as lineage specifiers for the alveolar type 1 (AT1) and alveolar type 2 (AT2) subpopulations from the differentiating lung alveolar BP. Taking these facts together, the presented approach addresses crucial challenges in stem cell research as it has the ability to guide experimentalists in the design of new strategies for stem cell therapies and treatment with potential application in regenerative medicine.

On the other hand, recent advances in single-cell RNA sequencing (scRNA-seq) has provided us with a large compilation of high-resolution datasets that reflect the cellular heterogeneity

within a cell population. The high resolution and statistical power of this data has significantly advanced our ability to classify cells into distinct subpopulations, based on their gene-expression profiles. The identity of these cell subpopulations can vary from well-known cell types, subtypes, to uncharacterized subpopulations that, without the scRNA-seq technology, would have remained unknown. Maintaining the identity of different cell populations has been linked to specific sets of TFs [9]. Moreover, the scRNA-seq advent opened doors for the characterization of TFs that act synergistically in order to induce cell transitions between cell populations [41]. Thus, identification of such core TFs is of utmost importance for characterizing and converting any cell population. The quest to advance our knowledge of core TFs identification and characterization of their crucial role in cellular conversions has prompted the development of new computational methods [6–9] that aim at deciphering the underlying regulatory mechanism and interactions responsible for cell identity and transitions.

One such state-of-the-art computational platform is TransSyn (*see* **Note 2**), a tool for the identification of cell population identities by defining their synergistic transcriptional cores [41]. This method is based on the assumption that a cell population identity arises from the synergistic activity of specific TFs that stabilize the expression levels of genes characteristic of that cell population. TransSyn method does not rely on GRNs nor depends on any prior knowledge, and it only requires a scRNA-seq dataset of distinct populations as input. Hence, TransSyn determines cell population identities and promotes the design of experimental strategies by means of predicting core TFs necessary for converting the cell population under consideration to different other cell populations.

As a proof of concept, TransSyn predicted synergistic transcriptional cores recapitulated known identity TFs in 85% of the cases and known synergistic TF interactions related to cell identity [41]. Furthermore, TransSyn cell subpopulation-specific synergistic transcriptional core successfully drove the experimental conversion of human hindbrain neuroepithelial cells (hNES) into medial floor plate midbrain progenitors (hProgFPM), which later differentiated into midbrain dopaminergic (mDA) neurons [41]. TransSyn is freely available at https://sourceforge.net/projects/transsyn/.

TransSyn has been applied to multiple scRNA-seq datasets, predicting synergistic transcriptional cores of 186 cell populations [41]. The predicted synergistic transcriptional cores frequently contained TFs known to maintain the respective cell population identities. For instance, the pluripotency factors *Pou5f1*, *Sox2*, and *Nanog* were predicted as the most synergistic transcriptional core in human embryonic stem cells (ESCs). Indeed, these TFs are known to maintain the ESC phenotype and were described to act

synergistically through enhancer clusters [49]. Also, *Phox2a* and *Isl1*, which have been described to act synergistically to define cranial motor neurons from mouse ESCs [50], were predicted as part of the synergistic core of human fetal oculomotor and trochlear nucleus population. Finally, *Tal1, Runx1, Gata2, and Fli1*, identified as the synergistic transcriptional core of blood cell progenitors, have been shown to act synergistically by interacting at a protein level, stabilizing their cofactor binding to DNA and controlling the cell population identity [51, 52].

TransSyn predictions not only capture the synergy between TFs but also recapitulate several known TFs interactions that control cell population identities. For instance, *Eomes, Otx2, Zic3, Foxa2,* and *Hnf4a* were identified in the synergistic core of embryonic visceral endoderm subpopulation. These TFs have been described to regulate each other and multiple downstream targets specific to that cell population [53, 54]. In addition, *Gata1* and *Ikzf1*, known to functionally regulate each other and maintain the identity of embryonic blood cells [55, 56], were identified in the synergistic TF core of the embryonic erythrocyte population [52]. Differentiation towards vascular endothelial cell fate is regulated by *Id3* and members of the Krüppel-like family of TFs [57, 58], which were included on the predicted synergistic transcriptional core for this population. For mouse enteroendocrine cells, the synergistic TF core consisted of *Foxa1, Foxa2, Insm1, Lmx1a, Neurog1, Neurog3, Nkx2.2*, and *Pax4*, all known to play an important role in the functioning of this cell type [59–62]. Finally, the synergistic transcriptional core obtained for fetal dopaminergic neurons contained *Nurr1* and *Foxa1*, both known for controlling mDA identity and neurogenesis [63, 64].

In order to demonstrate TransSyn efficacy in cell conversion protocols, this computational platform was used to predict the synergistic core of hProgFPM [41]. The results included *Foxa2*, *Otx2*, and *Lmx1a*, TFs which were previously described as important for mouse dopaminergic neuron (mDA) development [64–66]. Since *Otx2* is known to induce *Lmx1a* expression [67], an experiment was designed to shift the identity of hindbrain hNES cell line towards hProgFPM targeting only *Foxa2* and *Otx2* [41]. These two TFs were induced by treatment of hNES with two small molecules: smoothened agonist and Dickoppf1 for *Otx2* and *Foxa2*, respectively. Anteriorization and acquisition of midbrain identity were shown by an increase in the number of *Otx2* positive cells and the increased ratio of *Otx2/Gbx2* expression. In addition, increased levels of *Foxa2* and decreased levels of *Pax6* and *Irx3* were observed, which reveals efficient ventralization. Furthermore, cells obtained by targeting these two TFs generated a significant increase in both Th gene expression and Th positive cells, which were also positive for *Lmx1a*, *Nurr1*, and *Pbx1*, known markers of midbrain identity [63, 66, 68]. Th positive

cells also expressed *Map2*, a mature neuronal marker, and were observed to acquire long processes and varicosities, usually found in mDA neurons. Taken together, such an extensive validation of TransSyn proves its ability to facilitate the design of novel strategies for conversion of cell subpopulation identities with potential applications in regenerative medicine.

Finally, cell rejuvenation strategies are essential to prevent aged and disease cell niches to impair normal cellular function [69–71]. Characterizing and controlling these cell subpopulation phenotypes has great potential for developing novel regenerative medicine strategies. The statistical power and increasing number of scRNA-seq datasets have been helping us to profile distinct cell subpopulations. However, the lack of computational methods that use scRNA-seq data to identify key factors involved in the maintenance of a specific cell phenotype and controlling cell rejuvenation remains a challenge.

In order to bridge this gap in the literature, researchers have introduced SigHotSpotter, a computational web application that integrates scRNA-seq with a probabilistic Markov chain model of signal transduction to predict key signaling molecules (hotspots) that are responsible for the control of a cell phenotype [42]. This method is based on the assumption that in steady-state, single-cell gene expression can be representative of the corresponding protein level [72]. SigHotSpotter integrates signaling and transcriptional networks in order to identify specific signaling hotspots that are involved in sustaining the transmission of external niche signals. These hotspots are predicted based on the highest sustained signal flux instead of their interference in the dysregulation of the entire signaling pathway. In fact, signaling molecules matching these characteristics have been shown to be more likely to transmit stable signals involved in niche maintenance, rather than transient but strong signals, normally related to changes in cell phenotype [73]. Furthermore, SigHotSpotter ranks the predicted hotspots according to their compatibility towards the downstream TFs. This will help experimentalists in prioritizing their targets for further studies and prompt the development of cell rejuvenation strategies for counteracting the loss of activation potential in niche cells due to aging or disease.

SigHotSpotter was applied in silico to four different cell systems where it accurately recapitulated signaling molecules known to be involved in maintaining specific cell phenotypes [42]. Namely, SigHotSpotter predicted that the inhibition of Gsk3β and Map2k1 is responsible for the maintenance of the mouse ESCs in a naive pluripotency state. SigHotSpotter also identified specific signaling pathways involved in fibroblast switching from neonatal to mature state, a process that was shown to prompt the maturation of cardiomyocytes in vivo [74]. Furthermore, the Markov chain applied in this method identified Spr5 as the signaling hotspot mediating a

switch from canonical to non-canonical Wnt activity, inducing quiescence in the aging mice brain [75]. SigHotSpotter is available at https://SigHotSpotter.lcsb.uni.lu and its source code can be accessed at https://gitlab.com/srikanth.ravichandran/sighotspotter.

SigHotSpotter has been applied in four different cell systems based on scRNA-seq datasets from mouse ESCs, hair-follicle stem cells (HFSCs), hematopoietic stem cells (HSCs), and oligodendrocyte progenitor cells [42]. The method correctly predicted signaling hotspots with known experimental validation in those cell systems. For instance, SigHotSpotter accurately predicted the inhibition of Gsk3β and Map2k1 in ESCs under 2i culture conditions. Indeed, different culture conditions are known to maintain ESCs in different phenotypic states. In particular, 2i culture media, which contains selective inhibitors for Gsk3β and Map2k1, is known to maintain ESCs in a naive pluripotency state while leukemia inhibitory factor (LIF) alone is known to maintain ESCs in a primed pluripotency state [76]. In addition, Gsk3β and Bmpr1a, components of, respectively, the Wnt and BMP signaling pathways were identified as inhibited hotspots in HFSCs. Accordingly, Wnt and BMP signaling pathways were reported to play a role in the activation of quiescent HFSCs to generate specialized mesenchymal cells necessary to induce the hair cycle [77]. In young long-term HSCs, Map2k1 and Map3k1, proteins of the ERK signaling pathway, were predicted as activated signaling hotspots together with Gsk3β, a protein regulated by the PI3K signaling pathway. The other two members of the PI3K signaling pathway, Akt1 and Icam1, were identified as inhibited in the same phenotype. Notably, long-term HSCs have been described to be the main cell type involved in the production of blood cells, a process where ERK and PI3K pathways were identified as key regulators of the balance between HSCs dormancy and activation [78–81]. Finally, Notch1 and Gsk3β, that have been reported to inhibit oligodendrocyte differentiation and myelination as inhibited, were also predicted as inhibited in OPCs, while Fyn, a protein reported to be involved in oligodendrocyte migration, was predicted as activated in this cell type [82–84].

The Markov chain applied in this method was used to predict niche signals responsible for age-dependent changes in NSCs [75]. The results predicted Sfrp5, an antagonist of the non-canonical Wnt signaling pathway, as a key signaling intermediate for old quiescent NSCs. Sfrp5 was inhibited in vivo by administering a neutralizing antibody for 14 days. Sfrp5 inhibition significantly decreased the number of proliferating cells in an old mice brain when compared to IgG-treated control old mice. This result indicates a reduced NSC activation, showing that Sfrp5 antagonization leads to increased Wnt canonical activity and, thus, quiescence in the aging mice brain. Furthermore, this shows

that NSCs activation is regulated by a switch from canonical to non-canonical Wnt activity, a new finding that highlights the importance of this method in characterizing different cell phenotypes.

Recently, SigHotSpotter was also applied to further understand the signaling pathways involved in fibroblast identity in mouse heart development [74]. The results identified Foxo, mTOR, and VEGF signaling pathways as specific for neonatal fibroblasts while cell adhesion, estrogen, and TGFβ signaling pathways were identified only in mature fibroblasts. These results suggest that fibroblast switching from neonatal to mature state prompts the in vivo maturation of cardiomyocytes.

In conclusion, in recent years stem cell therapy has become a very promising and important scientific research topic. Advances in this area allowed for a deeper understanding of human disease and prompted the design of novel strategies for tissue regeneration, namely in cell transplantation. The advent of single-cell technology continues to expand our knowledge about molecular mechanisms underlying cell conversion and differentiation, the characterization of cell phenotypes in different conditions, and intrinsic molecular dysregulations in different pathologies. By applying single-cell technology to stem cell research, multiple computational methods were developed to accurately identify key factors in cell conversion, explain differences between cell phenotypes, and predict signaling molecules controlling these processes. To sum up, the discussed computational models provide mechanistic insights into biological processes and generate new predictions that guide experimental research. However, stem cell therapy still has challenges to overcome. For instance, cells transplanted from a different tissue or obtained from in vitro experiments are not always successfully integrated into the patient's tissue [85]. In addition, most in vitro reprogramming and differentiation protocols often have a low conversion efficiency, making it more expensive and time-consuming to collect enough target cells for clinical use or further research. Furthermore, in vitro cell conversion protocols often generate non-functional and immature variants of target cells, failing to obtain the desired cell phenotype and functionalities.

Computational modeling can help address these limitations by developing network-based models that combine GRNs with the hierarchical organization of cell identity TFs, in order to predict specific cell type and subtype identity TFs that lead to in vitro generation of functionally mature target cells [85, 86]. Moreover, these methods can be further used to elucidate mechanisms underlying the dysregulation of cell differentiation associated with, for instance, congenital disorders [44]. Integrating different types of single-cell data in multiscale computer models would allow for a more thorough characterization of a biological system and help experimentalists in designing novel strategies for stem cell therapy.

For instance, combining scRNA-seq and single-cell sequencing assay for transposase-accessible chromatin (scATAC-seq) would provide a better characterization of novel cell types and phenotypes, which would lead to the development of more accurate models for predicting cell-fate determinants.

Modeling biological systems at the tissue level, based on cell–cell or receptor–ligand interactions, could help elucidate the basic processes involved in tissue regeneration and homeostasis and lead to novel predictions of cell–cell interactions that sustain tissue regeneration. For instance, recent scRNA-seq computational methods focused on modeling stem cell niche interactions by integrating scRNA-seq with signaling and transcriptional networks [42, 87]. These approaches acknowledge the fact that predicting key signaling molecules that mediate niche signals responsible for maintaining cell phenotypes is of the utmost importance for stem cell rejuvenation therapies that focus on reverting impaired cell function and improve tissue repair processes in degenerative and age-related diseases. Furthermore, modeling cell-cell communication networks based on scRNA-seq data has helped in predicting key cell–cell interactions involved in the regulation of tissue homeostasis [88–90]. By comparing these reference cell-cell interactomes with the networks of injured or diseased tissues, a computational model could identify which interactions are dysregulated and help develop novel strategies for reverting tissue degeneration. Moreover, multiscale modeling can improve the prediction of relevant cell–cell interactions in tissue homeostasis. To this end, integrating phosphoproteomics data with scRNA-seq would help to identify signaling pathways that maintain stem cell phenotypes while combining scRNA-seq with imaging data could further improve our understanding of cell-cell communication.

In summary, stem cell research can greatly benefit from the increased resolution of scRNA-seq and its integration with new single-cell technologies, such as proteomes, epigenomes, and spatial information in order to have a more comprehensive classification and characterization of cell types, their interactions, and function. Closer collaboration between computational researchers and experimentalists can further stimulate advances in stem cell research. For example, by joining efforts they could come up with a more generic and integrative model that will provide an accurate and deeper insight into the molecular mechanisms underlying the regulation and control of stem cell homeostasis and its dysregulation in disease, therefore, accelerating the translation of our biological knowledge to clinical application.

2 Materials

2.1 SeesawPred: Cell-Fate Determinants

1. SeesawPred requires as an input the gene-expression profiles of the stem/progenitor and two daughter cell types.
2. A prior knowledge network (PKN) that contains interactions between the TFs (*see* **Notes 3** and **4**).

2.2 TransSyn: Cell Identity TFs

1. scRNA-seq gene-expression matrix with column names labeled according to which population each cell belongs, and the rows contain gene-expression values (*see* **Note 5**).
2. List of all known TFs of the species from which the gene-expression matrix was obtained. TFs should be labeled with the same nomenclature used in the gene-expression matrix mentioned in **item 1** (*see* **Note 6**).

2.3 SigHotSpotter: Niche-Induced Signaling Pathways

1. Two scRNA-seq gene-expression matrices from the same cell type in different conditions, such as ESCs cultured in 2i and LIF culture media (condition 1 and condition 2). scRNA-seq gene-expression matrices should have columns representing cell data or replicates and row names labeled according to Gene Symbols nomenclature. Gene-expression matrices data should be normalized or in read counts (TPM/FPKM).
2. Differential expression file containing only the significantly differentially expressed TFs, where "1" denotes upregulated TFs in condition 1 and "−1" denotes upregulated TFs in condition 2. The differential expression file should have Gene Symbols on the first column and differentially expression information on the second column (*see* **Note 7**).
3. Pre-compiled, species-specific signaling interactome network, including receptor/ligand information, a list of TFs and TF interactions (*see* **Notes 8** and **9**).

3 Methods

3.1 SeesawPred: Cell-Fate Determinants

1. The computation of differentially expressed TFs between two daughter cells.
2. The reconstruction of Boolean TRNs from differentially expressed TFs by retrieving interactions from the MetaCore (Clarivate Analytics) database and then prune this network by discarding interactions incompatible with Booleanized gene-expression data of two daughter cells (*see* **Note 10**).
3. The computation of statistically significant normalized ratio difference (NRD) TF pairs to identify pairs of TFs whose

expression ratios showed a significant change in daughter cells in comparison with the stem/progenitor cells (*see* **Note 11**).

4. The identification of strongly connected components (SCC) in the TRN of parental cell subpopulation.
5. TF pairs that are present in the parental SCC and whose TFs are differentially expressed in daughter cell subpopulations are considered candidate opposing lineage specifier pairs (*see* **Note12**).
6. The NRD TF pairs are then tested for significance and filtered. The criteria for keeping a TF pair is that it has to be: (1) differentially expressed between the two daughter cell types, (2) constitute significant NRD TF pair on significance test among all TFs, and (3) show high absolute NRD values in both daughter cell types ($|NRD| > 0.5$) (*see* **Note 13**).
7. Finally, SCCs with significant NRD TF pairs which satisfy the presented criteria are considered the final predictions.

3.2 TransSyn: Cell Identity TFs

1. Discard unclassified populations or populations with less than three cells in the dataset.
2. Filter gene-expression matrix for TFs-only based on the TF list in **item 2** of the Subheading 2.
3. Classify TFs as expressed if their expression values are higher than or equal to 1 in RNA-seq FPKM/RPKM/TPM values, higher than or equal to 10 in normalized read counts, or higher than or equal to 1 in UMI counts. TFs that are below these expression cut-offs are to be considered not expressed (*see* **Note 14**).
4. Select target cell populations for transcriptional synergistic core identification.
5. Calculate the fraction of cells expressing each TF per selected population. Keep the top 10% most frequently expressed TFs that are also expressed in more than 70% of the cells. If the number of TFs is more than 150, calculate the coefficient of variation and select the top 150 TFs (*see* **Note 15**).
6. Convert the zero gene-expression values into one. Log 10-transform the gene-expression values. Convert the gene-expression values with mean equal to zero into one. Discretize obtained gene-expression values using the Freedman–Diaconis rule (*see* **Note 16**). For FPKM/RPKM/TPM values and normalized read counts, set the input for the Freedman–Diaconis rule to the number of cells plus 1. For UMI counts, set it to the number of cells plus 6. Set the range of gene-expression values to be between 0 and the maximum value of each cell population.

7. Compute Shannon's entropy of each TF based on the discretized values and normalize these values by the theoretical maximum entropy (*see* **Note 17**).
8. Calculate multivariate mutual information (MMI) [91] for all combinations of three TFs using the total number of cells in each population. Select the top 1% lowest MMI combinations (*see* **Note 18**). Rank these combinations by total correlation (TC) and select the top 1% highest combinations.
9. Use the selected TC combinations as the initial seeds for the dynamic heuristic search of high-level synergistic TF cores. Add new TFs to each seed combination, one by one, and compute the new MMI. Select the combinations that show lower MMI (less than 0.05) than the seed and compute TC. Select the top ten best TC combinations as seed for the next iteration. The heuristic search is finished when no new combination has a lower MMI (more synergistic) than the seed or when the number of TFs within a combination reaches 15 (*see* **Note 19**).
10. Compute the TC of the final combinations obtained in **step 9**. Calculate MMI for the 20 best TC combinations. If MMI is lower, rank new combinations by TC. Otherwise, rank combinations from **step 9** by TC to obtain the final synergistic transcriptional cores. If there is a tie between combinations, they are ranked according to the highest summed mean gene expression and the top three combinations are kept as the final synergistic transcriptional cores.
11. Identify the cell conversion TFs by calculating the mean gene expression for the TFs in each synergistic transcriptional core and ranking them based on the fold change between the target cell population and the starting cell population (remaining populations in the dataset).

3.3 SigHotSpotter: Niche-Induced Signaling Pathways

1. Model the signal transduction process from the niche to intracellular signaling pathways as a finite discrete time-homogeneous Markov chain, where the signal originates from the niche and propagates successively through a finite set of signaling molecules (*see* **Note 20**).
2. Use the mass action principle to construct the state transition probability matrix, assuming that the probability of interaction between two molecules is proportional to the dot product of their corresponding gene-expression values (*see* **Note 4**). Transition probability values should only be calculated from interactions present in the signaling interactome (*see* **Note 21**).
3. Establish a cut-off parameter that determines the percentage of cells for which both interacting molecules must be expressed for the interaction to be considered in the Markov chain model. For instance, if the cut-off is set to 30%, then interactions

among two molecules are only considered if both genes are simultaneously expressed in at least 30% of the cells in the scRNA-seq.

4. Calculate the stationary distribution of the transition probability matrix by finding the eigenvector of the transposed transition probability matrix with an eigenvalue equal to 1 [92]. This computation gives the steady-state probability distribution of where the signal will be present, at any time point (*see* **Note 22**).
5. Establish a percentile parameter that determines the top fraction of molecules classified as high probability signaling molecules, which will be used for the following calculations.
6. Classify TFs as interface TFs and non-interface TFs. Non-interface TFs are defined as TFs that do not have an incoming edge from a signaling molecule. Interface TFs are defined as TFs that have an incoming edge from a signaling molecule. Filter this list for differentially expressed non-interface TFs (ntDETFs).
7. Evaluate the concordance between the overall effect of the high probability signaling molecules (activation or inhibition) on the expression status of the downstream ntDETFs (up- or downregulated). Trace all directed shortest paths from the high probability signaling molecules to the ntDETFs. Based on these results, calculate the weight for each shortest path, which is obtained from the product of the steady-state probabilities of the nodes contained in the path and its sign. If the number of inhibitor edges is even and the ntDETF is upregulated, the sign of the shortest path is equal to 1. However, if the ntDETF is downregulated, the sign is equal to −1. If the number of inhibitor edges is odd and the ntDETF is downregulated, the sign of the shortest path is equal to 1. On the other hand, if the ntDETF is upregulated, the sign is equal to −1 (*see* **Note 23**).
8. Calculate the compatibility score for a given signaling molecule and a target ntDETF based on the shortest path weight obtained on the previous step. The compatibility score is defined as the fraction of the sum of positive edge weights relative to the sum of absolute edge weights. This score provides a quantitative measurement between a high probability signaling molecule and a target ntDETF. Calculate the overall compatibility score of a signaling molecule for a given phenotype/condition by computing the mean of the compatibility scores over all ntDETFs.
9. Based on the compatibility scores obtained on the previous step, identify the active and inactive signaling hotspots (*see* **Note 24**). Construct the signaling network around these

hotspots by computing all the shortest paths from the niche-node to the predicted signaling hotspots, and then from the hotspots to the ntDETFs (*see* **Notes 25** and **26**).

4 Notes

1. The SeesawPred algorithm was written in R and the web interface has been developed using the Shiny web technology.
2. AnimalTFDB [93] database can be a useful resource to obtain the species-specific TF list.
3. A literature-based manually curated PKN can be retrieved from the MetaCore (Clarivate Analytics) database or any other source of user's choice.
4. TransSyn algorithm was written in C++ and wrapped in R using Rcpp and gtools R packages.
5. It is recommended to use the same population classification and to not reprocess the raw data and gene-expression values defined in the original dataset.
6. AnimalTFDB [93] database can be a useful resource to obtain the species-specific TF list.
7. AnimalTFDB [93] annotation can be useful to define TFs. Differentially expressed TFs can be identified by using a *t*-test. Shortlist the set of obtained differentially expressed TFs by using Benjamini–Hochberg correction and a cut-off for the adjusted *p*-value lower than 0.05.
8. We recommend using Omnipath [94] and ReactomeFI [95] databases to construct the signaling interactome network, since they contain information about the directionality of the interactions and their regulatory nature. For example, given a gene A and a gene B, we can know if A and B interact with each other or if it is a unidirectional interaction, and if these interactions are activations or inhibitions. Gene Ontology classification of the plasma membrane (GO:0005886) and receptor activity (GO:0004872) can be used to compile a list of receptors/ligands. Transcriptional interactions can be obtained from Zaffaroni et al. [96] which contains a high number of manually curated TF interactions from MetaCore (Clarivate Analytics).
9. In order to account for the niche influence on the intracellular signaling and to model sustained signaling, it is recommended to introduce an external niche-node which should be connected to all receptors and ligands as well as all the TFs nodes in the signaling interactome. This step ensures a continuous signal transmission from niche to TFs through signaling intermediates, based on the assumption that, once a signal

reaches a TF, this signal starts once again from the external niche-node. Also, it is recommended to remove non-receptor/ligand nodes that have zero in-degree and non-TF nodes that have a zero out-degree since these nodes do not contribute to sustain the signal transduction.

10. The PKN retrieved from MetaCore contains direct interactions between the TFs obtained from different interaction categories, such as binding, transcriptional regulation, and influence on expression.
11. The required input of SeesawPred web application is a tab-separated value file where columns represent gene-expression (microarray or RNA-Seq) replicates of the stem/progenitor cell type and the two daughter cell types labeled as "Progenitor," "Daughter1," "Daughter2," respectively, and the rows are labeled according to the TF symbols.
12. The PKN retrieved from MetaCore or defined by the user must also be a tab-separated value file containing the list of TF–TF interactions that serve as potential links in the network.
13. Example input files for two cellular differentiation systems (mouse NSCs differentiation into neurons and astrocytes, and Mouse Hematopoietic Stem Cell (HSC) differentiation to erythroid and myeloid) are provided in the online web application of SeesawPred.
14. If for a given dataset setting the cut-offs to the recommended values results in too many expressed TFs, making the following computation steps impracticable, the expression cut-off can be set to 10 instead. If the input dataset is binarized, TFs with mean counts lower than 1 can be discarded.
15. The maximum number of TFs used for the subsequent computations are suggested to be set to 150 since running the algorithm becomes infeasible for a standard desktop computer if this number is higher than that.
16. The Freedman–Diaconis rule was implemented using the R nclass.FD function.
17. Normalizing the TF entropy using the theoretical maximum entropy enables a direct comparison between different TF entropies.
18. MMI measures the information gained by adding a variable (in this case, a TF), which cannot be explained by the sum of the information given by the subsets of variables. Thus, when MMI is negative it means that the TFs are synergistically interacting with each other because the information given by the TFs together is higher than the sum of the information given by the TFs separately. MMI is calculated with all cells in each

population in the dataset, except for the ones in the population for which MMI is being computed.

19. The suggested maximum number of TFs per synergistic core is 15 because continuing the computations with higher numbers is often very demanding and it has been observed that, at that point, most TFs are shared among different combinations.
20. According to this model, the probability of a given signal to propagate from molecule A to molecule B on the next time step only depends on where the signal is currently present and not on where it was on the previous time steps. Thus, the transition probability of a signal to propagate from molecule A to B is defined by a transition probability matrix, calculated based on the scRNA-seq expression matrix.
21. The transition probability value between two molecules is calculated by dividing the dot product of their corresponding single-cell gene-expression vectors (defined as interaction weight) by the sum of the product of the interaction weights between their neighboring molecules.
22. The transition probability matrix will be stochastic since only the interactions present in the signaling interactome will have a non-zero transition probability calculated from the data. The others will have zero probabilities. Since the interaction probability for any two molecules is proportional to the scalar product of their gene-expression values, this probability will be higher only when both the molecules are highly expressed in the same cell and expressed in a large number of cells for that population.
23. The compatibility of the shortest path from a high probability signaling molecule to a ntDETF is determined by the sign. If the sign of the shortest path is negative (-1), it means that the signaling molecule and ntDETF are incompatible and so the weight will be negative. If the sign of the shortest path is positive (1), the signaling molecule and ntDETF are compatible, therefore the weight will be positive.
24. The compatibility score is equal to 1 if all shortest paths between a signaling molecule and a target ntDETF are compatible, making this signaling molecule an active signaling hotspot. If the net effect of the signaling molecule is entirely incompatible with the expression status of the downstream TFs, the compatibility score is equal to 0 and the signaling molecule classified as an inactive signaling hotspot.
25. We recommend using the Igraph implementation of Dijkstra's algorithm [97] to calculate the shortest paths network. This network will show how the signaling hotspots meditate the transmission of the signal from the external niche-node to the ntDETFs.

26. We recommend to output a list of the top ten active and inactive signaling hotspots for each phenotype and a complete list of all ranked predictions, together with the shortest path network of the predicted signaling hotspots in SIF format so it can be easily visualized in, for instance, Cytoscape [98].

References

1. Fu NY, Nolan E, Lindeman GJ, Visvader JE (2020) Stem cells and the differentiation hierarchy in mammary gland development. Physiol Rev 100(2):489–523. https://doi.org/10.1152/physrev.00040.2018
2. Visvader JE, Clevers H (2016) Tissue-specific designs of stem cell hierarchies. Nat Cell Biol 18(4):349–355
3. D'Alessio AC, Fan ZP, Wert KJ et al (2015) A systematic approach to identify candidate transcription factors that control cell identity. Stem Cell Rep 5(5):763–775. https://doi.org/10.1016/j.stemcr.2015.09.016
4. Whyte WA, Orlando DA, Hnisz D et al (2013) Master transcription factors and mediator establish super-enhancers at key cell identity genes. Cell 153(2):307–319. https://doi.org/10.1016/j.cell.2013.03.035
5. Vaquerizas JM, Kummerfeld SK, Teichmann SA, Luscombe NM (2009) A census of human transcription factors: function, expression and evolution. Nat Rev Genet 10(4):252–263
6. Rackham OJL, Firas J, Fang H et al (2016) A predictive computational framework for direct reprogramming between human cell types. Nat Genet 48(3):331–335. https://doi.org/10.1038/ng.3487
7. Cahan P, Li H, Morris SA et al (2014) CellNet: network biology applied to stem cell engineering. Cell 158(4):903–915. https://doi.org/10.1016/j.cell.2014.07.020
8. Buganim Y, Faddah DA, Jaenisch R (2013) Mechanisms and models of somatic cell reprogramming. Nat Rev Genet 14(6):427–439
9. Morris SA, Daley GQ (2013) A blueprint for engineering cell fate: current technologies to reprogram cell identity. Cell Res 23(1):33–48
10. Strasser M, Theis FJ, Marr C (2012) Stability and multiattractor dynamics of a toggle switch based on a two-stage model of stochastic gene expression. Biophys J 102(1):19–29. https://doi.org/10.1016/j.bpj.2011.11.4000
11. Roeder I, Glauche I (2006) Towards an understanding of lineage specification in hematopoietic stem cells: a mathematical model for the interaction of transcription factors GATA-1 and PU.1. J Theor Biol 241(4):852–865. https://doi.org/10.1016/j.jtbi.2006.01.021
12. Huang S, Guo YP, May G, Enver T (2007) Bifurcation dynamics in lineage-commitment in bipotent progenitor cells. Dev Biol 305(2):695–713. https://doi.org/10.1016/j.ydbio.2007.02.036
13. Zhang P, Behre G, Pan J et al (1999) Negative cross-talk between hematopoietic regulators: GATA proteins repress PU.1. Proc Natl Acad Sci U S A 96(15):8705–8710. https://doi.org/10.1073/pnas.96.15.8705
14. Arinobu Y, Mizuno S-i, Chong Y et al (2007) Reciprocal activation of GATA-1 and PU.1 marks initial specification of hematopoietic stem cells into myeloerythroid and myelolymphoid lineages. Cell Stem Cell 1(4):416–427. https://doi.org/10.1016/j.stem.2007.07.004
15. Biteau B, Hochmuth CE, Jasper H (2011) Maintaining tissue homeostasis: dynamic control of somatic stem cell activity. Cell Stem Cell 9(5):402–411
16. Ferraro F, Lo Celso C, Scadden D (2010) Adult stem cells and their niches. Adv Exp Med Biol 695:155–168
17. Schofield R (1978) The relationship between the spleen colony-forming cell and the haemopoietic stem cell. A hypothesis. Blood Cells 4(1–2):7–25
18. Glukhova MA, Streuli CH (2013) How integrins control breast biology. Curr Opin Cell Biol 25(5):633–641
19. Brizzi MF, Tarone G, Defilippi P (2012) Extracellular matrix, integrins, and growth factors as tailors of the stem cell niche. Curr Opin Cell Biol 24(5):645–651
20. Shehata M, Teschendorff A, Sharp G et al (2012) Phenotypic and functional characterisation of the luminal cell hierarchy of the mammary gland. Breast Cancer Res 14(5):R134. https://doi.org/10.1186/bcr3334
21. Stingl J, Eirew P, Ricketson I et al (2006) Purification and unique properties of mammary epithelial stem cells. Nature 439(7079):993–997. https://doi.org/10.1038/nature04496

22. Heinäniemi M, Nykter M, Kramer R et al (2013) Gene-pair expression signatures reveal lineage control. Nat Methods 10(6):577–583. https://doi.org/10.1038/nmeth.2445
23. Lang AH, Li H, Collins JJ, Mehta P (2014) Epigenetic landscapes explain partially reprogrammed cells and identify key reprogramming genes. PLoS Comput Biol 10(8):e1003734. https://doi.org/10.1371/journal.pcbi.1003734
24. Roost MS, Van Iperen L, Ariyurek Y et al (2015) KeyGenes, a tool to probe tissue differentiation using a human fetal transcriptional atlas. Stem Cell Rep 4(6):1112–1124. https://doi.org/10.1016/j.stemcr.2015.05.002
25. Davis FP, Eddy SR (2013) Transcription factors that convert adult cell identity are differentially polycomb repressed. PLoS One 8(5):e63407. https://doi.org/10.1371/journal.pone.0063407
26. Morris SA, Cahan P, Li H et al (2014) Dissecting engineered cell types and enhancing cell fate conversion via cellnet. Cell 158(4):889–902. https://doi.org/10.1016/j.cell.2014.07.021
27. Cabili M, Trapnell C, Goff L et al (2011) Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses. Genes Dev 25(18):1915–1927. https://doi.org/10.1101/gad.17446611
28. Beerman I, Rossi DJ (2015) Epigenetic control of stem cell potential during homeostasis, aging, and disease. Cell Stem Cell 16(6):613–625
29. Avgustinova A, Benitah SA (2016) Epigenetic control of adult stem cell function. Nat Rev Mol Cell Biol 17(10):643–658
30. Neph S, Stergachis AB, Reynolds A et al (2012) Circuitry and dynamics of human transcription factor regulatory networks. Cell 150(6):1274–1286. https://doi.org/10.1016/j.cell.2012.04.040
31. Huang M, Chen Y, Yang M et al (2018) DbCoRC: a database of core transcriptional regulatory circuitries modeled by H3K27ac ChIP-seq signals. Nucleic Acids Res 46(D1):D71–D77. https://doi.org/10.1093/nar/gkx796
32. Saint-André V, Federation AJ, Lin CY et al (2016) Models of human core transcriptional regulatory circuitries. Genome Res 26(3):385–396. https://doi.org/10.1101/gr.197590.115
33. Hartmann A, Okawa S, Zaffaroni G, del Sol A (2018) SeesawPred: a web application for predicting cell-fate determinants in cell differentiation. Sci Rep 8(1):13355. https://doi.org/10.1038/s41598-018-31688-9
34. Okawa S, Nicklas S, Zickenrott S et al (2016) A generalized gene-regulatory network model of stem cell differentiation for predicting lineage specifiers. Stem Cell Rep 7(3):307–315. https://doi.org/10.1016/j.stemcr.2016.07.014
35. Okawa S, del Sol A (2015) A computational strategy for predicting lineage specifiers in stem cell subpopulations. Stem Cell Res 15(2):427–434. https://doi.org/10.1016/j.scr.2015.08.006
36. Guo M, Bao EL, Wagner M et al (2017) SLICE: determining cell differentiation and lineage based on single cell entropy. Nucleic Acids Res 45(7):e54. https://doi.org/10.1093/nar/gkw1278
37. Feingold BJ, Vegosen L, Davis M et al (2010) A niche for infectious disease in environmental health: rethinking the toxicological paradigm. Environ Health Perspect 118(8):1165–1172
38. Decimo I, Bifari F, Krampera M, Fumagalli G (2012) Neural stem cell niches in health and diseases. Curr Pharm Des 18(13):1755–1783. https://doi.org/10.2174/138161212799859611
39. Crespo I, Del Sol A (2013) A general strategy for cellular reprogramming: the importance of transcription factor cross-repression. Stem Cells 31(10):2127–2135. https://doi.org/10.1002/stem.1473
40. Ravichandran S, Okawa S, Martínez Arbas S, del Sol A (2016) A systems biology approach to identify niche determinants of cellular phenotypes. Stem Cell Res 17(2):406–412. https://doi.org/10.1016/j.scr.2016.09.006
41. Okawa S, Saltó C, Ravichandran S et al (2018) Transcriptional synergy as an emergent property defining cell subpopulation identity enables population shift. Nat Commun 9(1):2595. https://doi.org/10.1038/s41467-018-05016-8
42. Ravichandran S, Hartmann A, del Sol A (2019) SigHotSpotter: scRNA-seq-based computational tool to control cell subpopulation phenotypes for cellular rejuvenation strategies. Bioinformatics 36(6):1963–1965. https://doi.org/10.1093/bioinformatics/btz827
43. Lee E, Piranlioglu R, Wicha MS, Korkaya H (2019) Plasticity and potency of mammary stem cell subsets during mammary gland development. Int J Mol Sci 20(9):2357. https://doi.org/10.3390/ijms20092357
44. de Soysa TY, Ranade SS, Okawa S et al (2019) Single-cell analysis of cardiogenesis reveals

basis for organ-level developmental defects. Nature 572(7767):120–124. https://doi.org/10.1038/s41586-019-1414-x

45. Fujikura J, Yamato E, Yonemura S et al (2002) Differentiation of embryonic stem cells is induced by GATA factors. Genes Dev 16(7): 784–789. https://doi.org/10.1101/gad.968802
46. Yeo JC, Jiang J, Tan ZY et al (2014) Klf2 is an essential factor that sustains ground state pluripotency. Cell Stem Cell 14(6):864–872. https://doi.org/10.1016/j.stem.2014.04.015
47. Ito T, Udaka N, Yazawa T et al (2000) Basic helix-loop-helix transcription factors regulate the neuroendocrine differentiation of fetal mouse pulmonary epithelium. Development 127(18):3913–3921
48. Sandbo N, Kregel S, Taurin S et al (2009) Critical role of serum response factor in pulmonary myofibroblast differentiation induced by TGF-β. Am J Respir Cell Mol Biol 41(3): 332–338. https://doi.org/10.1165/rcmb.2008-0288OC
49. Hnisz D, Abraham BJ, Lee TI et al (2013) Super-enhancers in the control of cell identity and disease. Cell 155(4):934–947. https://doi.org/10.1016/j.cell.2013.09.053
50. Mazzoni EO, Mahony S, Closser M et al (2013) Synergistic binding of transcription factors to cell-specific enhancers programs motor neuron identity. Nat Neurosci 16(9): 1219–1227. https://doi.org/10.1038/nn.3467
51. Wilson NK, Foster SD, Wang X et al (2010) Combinatorial transcriptional control in blood stem/progenitor cells: genome-wide analysis of ten major transcriptional regulators. Cell Stem Cell 7(4):532–544. https://doi.org/10.1016/j.stem.2010.07.016
52. Scialdone A, Tanaka Y, Jawaid W et al (2016) Resolving early mesoderm diversification through single-cell expression profiling. Nature 535(7611):289–293. https://doi.org/10.1038/nature18633
53. Chen WS, Manova K, Weinstein DC et al (1994) Disruption of the HNF-4 gene, expressed in visceral endoderm, leads to cell death in embryonic ectoderm and impaired gastrulation of mouse embryos. Genes Dev 8(20):2466–2477. https://doi.org/10.1101/gad.8.20.2466
54. Coffinier C, Thépot D, Babinet C et al (1999) Essential role for the homeoprotein vHNF1/HNF1β in visceral endoderm differentiation. Development 126(21):4785–4794
55. Vassen L, Beauchemin H, Lemsaddek W et al (2014) Growth factor independence 1b (Gfi1b) is important for the maturation of erythroid cells and the regulation of embryonic globin expression. PLoS One 9(5):e96636. https://doi.org/10.1371/journal.pone.0096636
56. Ross J, Mavoungou L, Bresnick EH, Milot E (2012) GATA-1 utilizes ikaros and polycomb repressive complex 2 to suppress Hes1 and to promote erythropoiesis. Mol Cell Biol 32(18): 3624–3638. https://doi.org/10.1128/mcb.00163-12
57. Das JK, Voelkel NF, Felty Q (2015) ID3 contributes to the acquisition of molecular stem cell-like signature in microvascular endothelial cells: its implication for understanding microvascular diseases. Microvasc Res 98:126–138. https://doi.org/10.1016/j.mvr.2015.01.006
58. Suzuki T, Aizawa K, Matsumura T, Nagai R (2005) Vascular implications of the Krüppel-like family of transcription factors. Arterioscler Thromb Vasc Biol 25(6):1135–1141
59. Gierl MS, Karoulias N, Wende H et al (2006) The zinc-finger factor Insm1 (IA-1) is essential for the development of pancreatic β cells and intestinal endocrine cells. Genes Dev 20(17): 2465–2478. https://doi.org/10.1101/gad.381806
60. Ye DZ, Kaestner KH (2009) Foxa1 and Foxa2 control the differentiation of goblet and enteroendocrine L- and D-cells in mice. Gastroenterology 137(6):2052–2062. https://doi.org/10.1053/j.gastro.2009.08.059
61. Larsson LI, St-Onge L, Hougaard DM et al (1998) Pax 4 and 6 regulate gastrointestinal endocrine cell development. Mech Dev 79(1–2):153–159. https://doi.org/10.1016/S0925-4773(98)00182-8
62. Gross S, Garofalo DC, Balderes DA et al (2016) The novel enterochromaffin marker Lmx1a regulates serotonin biosynthesis in enteroendocrine cell lineages downstream of Nkx2.2. Development 143(14):2616–2628. https://doi.org/10.1242/dev.130682
63. Zetterström RH, Solomin L, Jansson L et al (1997) Dopamine neuron agenesis in Nurr1-deficient mice. Science 276(5310):248–250. https://doi.org/10.1126/science.276.5310.248
64. Ferri ALM, Lin W, Mavromatakis YE et al (2007) Foxa1 and Foxa2 regulate multiple phases of midbrain dopaminergic neuron development in a dosage-dependent manner. Development 134(15):2761–2769. https://doi.org/10.1242/dev.000141

65. La Manno G, Gyllborg D, Codeluppi S et al (2016) Molecular diversity of midbrain development in mouse, human, and stem cells. Cell 167(2):566–580.e19. https://doi.org/10.1016/j.cell.2016.09.027
66. Andersson E, Tryggvason U, Deng Q et al (2006) Identification of intrinsic determinants of midbrain dopamine neurons. Cell 124(2):393–405. https://doi.org/10.1016/j.cell.2005.10.037
67. Ono Y, Nakatani T, Sakamoto Y et al (2007) Differences in neurogenic potential in floor plate cells along an anteroposterior location: midbrain dopaminergic neurons originate from mesencephalic floor plate cells. Development 134(17):3213–3225. https://doi.org/10.1242/dev.02879
68. Villaescusa JC, Li B, Toledo EM et al (2016) A PBX1 transcriptional network controls dopaminergic neuron development and is impaired in Parkinson's disease. EMBO J 35(18):1963–1978. https://doi.org/10.15252/embj.201593725
69. del Sol A, Okawa S, Ravichandran S (2019) Computational strategies for niche-dependent cell conversion to assist stem cell therapy. Trends Biotechnol 37(7):687–696
70. Lane SW, Williams DA, Watt FM (2014) Modulating the stem cell niche for tissue regeneration. Nat Biotechnol 32(8):795–803. https://doi.org/10.1038/nbt.2978
71. Neves J, Sousa-Victor P, Jasper H (2017) Rejuvenating strategies for stem cell-based therapies in aging. Cell Stem Cell 20(2):161–175
72. Liu Y, Beyer A, Aebersold R (2016) On the dependency of cellular protein levels on mRNA abundance. Cell 165(3):535–550
73. Wang LD, Wagers AJ (2011) Dynamic niches in the origination and differentiation of haematopoietic stem cells. Nat Rev Mol Cell Biol 12(10):643–655
74. Wang Y, Yao F, Wang L et al (2020) Single-cell analysis of murine fibroblasts identifies neonatal to adult switching that regulates cardiomyocyte maturation. Nat Commun 11(1):2585. https://doi.org/10.1038/s41467-020-16204-w
75. Kalamakis G, Brüne D, Ravichandran S et al (2019) Quiescence modulates stem cell maintenance and regenerative capacity in the aging brain. Cell 176(6):1407–1419.e14. https://doi.org/10.1016/j.cell.2019.01.040
76. Ying QL, Wray J, Nichols J et al (2008) The ground state of embryonic stem cell self-renewal. Nature 453(7194):519–523. https://doi.org/10.1038/nature06968
77. Yang H, Adam RC, Ge Y et al (2017) Epithelial-mesenchymal micro-niches govern stem cell lineage choices. Cell 169(3):483–496.e13. https://doi.org/10.1016/j.cell.2017.03.038
78. Baumgartner C, Toifl S, Farlik M et al (2018) An ERK-dependent feedback mechanism prevents hematopoietic stem cell exhaustion. Cell Stem Cell 22(6):879–892.e6. https://doi.org/10.1016/j.stem.2018.05.003
79. Huang J, Zhang Y, Bersenev A et al (2009) Pivotal role for glycogen synthase kinase-3 in hematopoietic stem cell homeostasis in mice. J Clin Invest 119(12):3519–3529. https://doi.org/10.1172/JCI40572
80. Liu Y-F, Zhang S-Y, Chen Y-Y et al (2018) ICAM-1 deficiency in the bone marrow niche impairs quiescence and repopulation of hematopoietic stem cells. Stem Cell Rep 11(1):258–273. https://doi.org/10.1016/j.stemcr.2018.05.016
81. Kowalczyk MS, Tirosh I, Heckl D et al (2015) Single-cell RNA-seq reveals changes in cell cycle and differentiation programs upon aging of hematopoietic stem cells. Genome Res 25(12):1860–1872. https://doi.org/10.1101/gr.192237.115
82. Liu X, Lu Y, Zhang Y et al (2012) Slit2 regulates the dispersal of oligodendrocyte precursor cells via Fyn/RhoA signaling. J Biol Chem 287(21):17503–17516. https://doi.org/10.1074/jbc.M111.317610
83. Zhang Y, Argaw AT, Gurfein BT et al (2009) Notch1 signaling plays a role in regulating precursor differentiation during CNS remyelination. Proc Natl Acad Sci U S A 106(45):19162–19167. https://doi.org/10.1073/pnas.0902834106
84. Azim K, Butt AM (2011) GSK3β negatively regulates oligodendrocyte differentiation and myelination in vivo. Glia 59(4):540–553. https://doi.org/10.1002/glia.21122
85. Xu J, Du Y, Deng H (2015) Direct lineage reprogramming: strategies, mechanisms, and applications. Cell Stem Cell 16(2):119–134
86. Berneman-Zeitouni D, Molakandov K, Elgart M et al (2014) The temporal and hierarchical control of transcription factors-induced liver to pancreas transdifferentiation. PLoS One 9(2):e87812. https://doi.org/10.1371/journal.pone.0087812
87. Browaeys R, Saelens W, Saeys Y (2020) NicheNet: modeling intercellular communication by linking ligands to target genes. Nat Methods 17(2):159–162. https://doi.org/10.1038/s41592-019-0667-5

88. Camp JG, Sekine K, Gerber T et al (2017) Multilineage communication regulates human liver bud development from pluripotency. Nature 546(7659):533–538. https://doi.org/10.1038/nature22796
89. Skelly DA, Squiers GT, McLellan MA et al (2018) Single-cell transcriptional profiling reveals cellular diversity and intercommunication in the mouse heart. Cell Rep 22(3):600–610. https://doi.org/10.1016/j.celrep.2017.12.072
90. Raredon MSB, Adams TS, Suhail Y et al (2019) Single-cell connectomic analysis of adult mammalian lungs. Sci Adv 5(12):eaaw3851. https://doi.org/10.1126/sciadv.aaw3851
91. Bell AJ (2003) Co-information lattice. In: 4th int symp indep compon anal blind source
92. Stewart JW (1994) An introduction to the numerical solution of Markov chains. Princeton University Press, New Jersey
93. Zhang HM, Chen H, Liu W et al (2012) AnimalTFDB: a comprehensive animal transcription factor database. Nucleic Acids Res 40 (Database issue):D144–D149. https://doi.org/10.1093/nar/gkr965
94. Türei D, Korcsmáros T, Saez-Rodriguez J (2016) OmniPath: guidelines and gateway for literature-curated signaling pathway resources. Nat Methods 13(12):966–967
95. Wu G, Feng X, Stein L (2010) A human functional protein interaction network and its application to cancer data analysis. Genome Biol 11(5):R53. https://doi.org/10.1186/gb-2010-11-5-r53
96. Zaffaroni G, Okawa S, Morales-Ruiz M, Del Sol A (2019) An integrative method to predict signalling perturbations for cellular transitions. Nucleic Acids Res 47(12):e72. https://doi.org/10.1093/nar/gkz232
97. Csardi G, Nepusz T (2006) The igraph software package for complex network research. InterJournal Complex Syst 1695(5):1–9
98. Shannon P, Markiel A, Ozier O et al (2003) Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res 13(11):2498–2504. https://doi.org/10.1101/gr.1239303

4.2. TransSynW: A single-cell RNA-sequencing based web application to guide cell conversion experiments

4.2.1. Preface

Cellular conversion strategies are crucial to generate specific cell types that can potentially open new avenues for the development of regenerative medicine therapies. Recent developments in gene expression profiling at single-cell level provided the means to characterize cellular heterogeneity and identify TFs that promote conversion between cellular (sub)populations. However, conversion TFs might not be able to regulate all their target genes if chromatin conformation makes them inaccessible. PFs are a subset of TFs that have been shown to be able to overcome these limitations and bind to condensed areas of the chromatin to induce the expression of their target genes. Adding PFs to cellular conversion protocols has been shown to improve the success of these approaches.

We developed TransSynW, a computational platform that leverages the high resolution of scRNA-seq data to identify cellular conversion TFs for any cellular (sub)population characterized in this data. The identified sets of conversion factors include specifically expressed TFs as well as non-specific PFs for each target cell (sub)type. Prioritizing PFs among the identified conversion TFs will promote chromatin remodeling, which we foresee to improve the success of cellular conversion protocols. Additionally, TransSynW identifies marker genes for each of the target (sub)populations, allowing researchers to evaluate the performance of cellular conversion protocols. When applying TransSynW to distinct cellular systems, we show that our results well-recapitulated known cellular conversion TFs and marker genes. Moreover, literature search and cross-reference with a database of molecular interactions showed the biological significance of the newly identified conversion TF and markers. TransSynW is a user-friendly platform that has the potential to improve the outcome of cellular conversion protocols for regenerative medicine.

In this study, I implemented and developed the computational method and interface, collected the data, performed the application and literature validation of the precited TFs and markers, and the benchmarking. The published article is reprinted on the next pages (RightLinks license number 5324211471553).

Received: 29 May 2020 | Accepted: 16 August 2020

DOI: 10.1002/sctm.20-0227

ENABLING TECHNOLOGIES FOR CELL-BASED CLINICAL TRANSLATION

STEM CELLS TRANSLATIONAL MEDICINE

TransSynW: A single-cell RNA-sequencing based web application to guide cell conversion experiments

Mariana Messias Ribeiro[1] | Satoshi Okawa[1,2] | Antonio del Sol[1,3,4]

[1]Luxembourg Centre for Systems Biomedicine (LCSB), University of Luxembourg, Belvaux, Luxembourg

[2]Integrated BioBank of Luxembourg, Dudelange, Luxembourg

[3]CIC bioGUNE, Bizkaia Technology Park, Derio, Spain

[4]IKERBASQUE, Basque Foundation for Science, Bilbao, Spain

Correspondence
Antonio del Sol, PhD, Luxembourg Centre for Systems Biomedicine (LCSB), University of Luxembourg, 6, avenue du Swing, L-4367 Belvaux, Luxembourg.
Email: antonio.delsol@uni.lu

Funding information
Fonds National de la Recherche Luxembourg, Grant/Award Number: C17/BM/11662681

Abstract

Generation of desired cell types by cell conversion remains a challenge. In particular, derivation of novel cell subtypes identified by single-cell technologies will open up new strategies for cell therapies. The recent increase in the generation of single-cell RNA-sequencing (scRNA-seq) data and the concomitant increase in the interest expressed by researchers in generating a wide range of functional cells prompted us to develop a computational tool for tackling this challenge. Here we introduce a web application, TransSynW, which uses scRNA-seq data for predicting cell conversion transcription factors (TFs) for user-specified cell populations. TransSynW prioritizes pioneer factors among predicted conversion TFs to facilitate chromatin opening often required for cell conversion. In addition, it predicts marker genes for assessing the performance of cell conversion experiments. Furthermore, TransSynW does not require users' knowledge of computer programming and computational resources. We applied TransSynW to different levels of cell conversion specificity, which recapitulated known conversion TFs at each level. We foresee that TransSynW will be a valuable tool for guiding experimentalists to design novel protocols for cell conversion in stem cell research and regenerative medicine.

KEYWORDS

cellular therapy, clinical translation, differentiation, direct cell conversion, genomics, reprogramming, synergy, transcription factors

1 | INTRODUCTION

Cell conversion is fundamental to many biological processes. Control of cell conversion has significant relevance in stem cell research. For example, generation of functionally specific cells by cell conversion is of clinical interest for cell replacement therapies. However, several roadblocks need to be overcome for achieving optimal cell conversion, such as the accurate characterization of cell populations and the identification of cell conversion factors. Single-cell RNA-sequencing (scRNA-seq) technologies have made it possible to address these challenges. Due to the greater amount of scRNA-seq data generated across the world, experimental researchers are increasingly expressing their interest in deriving novel functional cell types.

Here, we present TransSynW, a scRNA-seq based web application for identifying cell conversion transcription factors (TFs) applicable in stem cell and clinical research (Figure 1A). It prioritizes pioneer factors (PFs) in the prediction of conversion TFs. Evidence suggests that PFs have a key role in chromatin opening, a process often required for cell conversion.[1] Indeed, including PFs on cell conversion protocols has been shown to improve their outcome.[1] Furthermore, it

predicts marker genes for each target cell type, enabling researchers to assess the fidelity of experimentally converted cells. In addition, it is user-friendly, and it does not require users' computer programming or computational resources. We also created a comprehensive video tutorial for guiding users through the web interface.

The application of TransSynW to various cell systems well-recapitulated known cell conversion TFs and made novel predictions, including the phenotypic conversion between cells in organoids and their in vivo counterparts. Moreover, predicted marker genes were consistent with experimentally known ones. These results highlight the applicability of TransSynW to a wide range of cell conversion experiments.

Significance statement

The study proposes a computational web application, TransSynW. To the best of the author's knowledge, it is the only computational tool that can identify cell conversion transcription factors (TFs) for any cell population in single-cell RNA-sequencing data. TransSynW does not require prior biological information, computer programming, and users computational resources. In addition, TransSynW prioritizes pioneer factors among predicted conversion TFs to facilitate chromatin opening often required for cell conversion. Furthermore, TransSynW predicts marker genes for assessing the performance of cell conversion experiments. Thus, TransSynW will be a staple tool for guiding experimentalists to design novel protocols for cell conversion in stem cell research and regenerative medicine.

2 | RESULTS

2.1 | Method overview

The TransSynW algorithm first identifies specifically and nonspecifically expressed TFs, and selects the combination that exhibits the highest synergistic interactions among them (see Methods) (Figure 1B). Notably, here we considered for the nonspecific part only PFs that have previously been reported to be involved in cell conversion protocols (Table S1). Predicted conversion TFs are then ranked by the expression fold change between the target and starting cell populations and users can prioritize the TFs for experimental follow-ups based on this ranking.

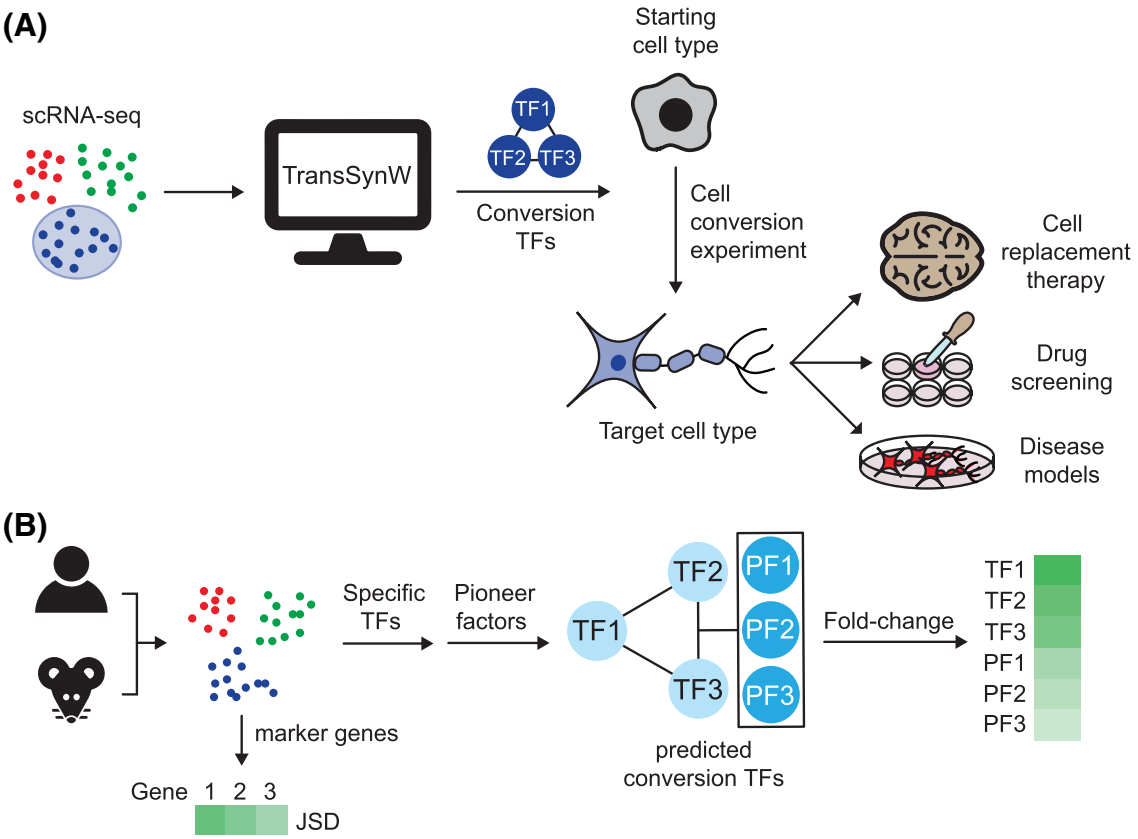


FIGURE 1 A, Application of TransSynW to stem cell research and regenerative medicine. B, Schematic overview of TransSynW algorithm (see also Methods). First, transcription factors (TFs) most specifically expressed in the selected target cell population (specific TFs) and nonspecifically expressed pioneer factors (PFs) are computed. The most synergistic combination of specific TFs and nonspecific PFs is then identified. The predicted set of TFs are ranked by expression fold change between target and starting cell populations. In parallel, top 10 candidate marker genes for target cell population are computed by JSD

We compiled the scRNA-seq data of starting cell types frequently used in cell conversion experiments from various scRNA-seq platforms (Table S2). For optimal results, users are recommended to use starting and target cell type data obtained from the same scRNA-seq platform or, if not available, from the closest sequencing platform. In general, it is recommended to select at least one PF and one specific TF from the predicted conversion TFs. It may be advisable to select more factors if the phenotypic difference between the starting and target cell types is large. Finally, TransSynW also predicts potential marker genes of the target cell populations. This feature enables researchers to select markers for assessing the performance of their cell conversion experiments.

TABLE 1 Predicted specific transcription factors (TFs) and nonspecific PFs

Cell type	Specific TFs	Nonspecific PFs	Annotation in data	Data source (PubMed ID)
(1) Conversion into broad cell type				
Myoblast	**MYF5**, **MYOD1**, **PAX7**, GLIS3, **PAX3**	CEBPB, IRF8, PBX1	1,3,4,5,7	30283141
Keratinocyte	**TRP63**, GATA3, NFIB	**KLF4**, GRHL2, CEBPA	0-16	30283141
Cardiomyocyte	**NKX2-5**, **TBX5**, PROX1, ZFP579, NR0B2	**GATA4**, **MEIS1**, PBX1	9,14	30283141
Hepatocyte	NR1I2, ZFP750, ZFHX4, **HNF1A**, ZBTB48	**HNF4A**, **FOXA3**, **FOXA2**	4,5,10,11,12,15	30283141
HSC	**HLF**, HOXA9, **GATA2**, **TAL1**, **MYCN**	CEBPB, CEBPA, **PBX1**	0,4,8	30283141
Neuron	EOMES, NEUROD6, EGR4, RARB, DLX6	**FOXG1**, **NEUROD1**, PBX1	9,10,12	30283141
Oligodendrocyte/OPC	**NKX6-2**, **OLIG1**, **SOX10**, **OLIG2**, NFE2L3	SOX2	0,6,11	30283141
Macrophage	RUNX3, BATF3, BATF, NFE2, E2F1	**SPI1**, **CEBPA**, ARID3A	Different tissues	30283141
Beta cell	**NKX6-1**, **PDX1**, **MAFA**, OVOL2, MNX1	**NEUROD1**, ISL1, **FOXA2**	0,8,9,11,17	30283141
NSC	ZFP275, **ASCL1**, **TCF3**	**FOXG1**, **SOX2**, PBX1	All young NSCs	30827680
(2) Conversion into subtype				
Dopaminergic neuron	NPAS4, **MYT1L**, EBF3, POU6F1, BNC2	**FOXA2**, **ASCL1**, **GATA3**	hDA	27716510
Medial floorplate progenitor	**LMX1A**, SP2, NR2F6, LMX1B, HMGA2	**FOXA2**, ASCL1, SOX2	hProgFPM	27716510
GABAergic neuroblast	GATA3, SOX14, **MYT1L**, BNC2, ZBTB38	**ASCL1**, **SOX2**, PBX1	hNbGaba	27716510
Oculomotor neuron	**PHOX2B**, **PHOX2A**, **ISL1**, RXRG, NR2F2	FOXA2, ASCL1, "PBX1	hOMTN	27716510
Serotonin neuron	**FEV**, GATA3, SOX1, DPF1, **LMX1B**	**GATA2**, PBX1	hSert	27716510
CD4+ central memory T cell	RBSN, RFX3, NR4A1, KLF9, ID3	GATA3, CEBPB	TCM	29352091
CD8+ memory T cell	**EOMES**, **BACH2**, KLF7, **MYC**, **ID3**	CEBPB, **GATA3**	4,6,11,13	31754020
Memory B cell	KLF13, LMO4, PCBD1, KLF10, ZBTB38	IRF8, SPI1, CEBPB	Memory B cell	31968262
(3) Phenotype conversion				
Primed mESC 1	**LIN28A**, MYC, ID1, FOXP1, ID3	**POU5F1**, **ESRRB**, **KLF4**	FBSLIF	25471879
Naive mESC 1	ZFHX2, MEIS2, ZIC2	**POU5F1**, **ESRRB**, **KLF4**	2iLIF	
Primed mESC 2	**LIN28A**, FOXP1, SOX4	**SOX2**, **POU5F1**, **KLF4**	mES_lif	26431182
Naive mESC 2	SPIC, MITF, MEIS2	**ESRRB**, **KLF4**, **POU5F1**	mES_2i	
Active NSC	CENPS, **EGR1**, INSM1, MXD3, **E2F1**	**ASCL1**, **SOX2**, PBX1	All young aNSCs	30827680
Quiescent NSC	DBP, EPAS1, **ID2**	**FOXG1**, PBX1, **ASCL1**	All young qNSCs	
Fetal hepatocyte	ZGPAT, KLF11, ZBTB20	**GATA4**, **HNF4A**, **CEBPA**	Fetal hepatocyte	30500538
Organoid hepatocyte	HES6, LEF1, THAP8, SOX9, HTT	**FOXA2**, **HNF4A**, MEIS1	Fetal hepatocyte organoid	
Adult hepatocyte 1	KLF9, CEBPD, KLF6	**FOXA2**, **HNF4A**, CEBPB	Hepatocyte	31292543
Adult hepatocyte 2	SCAND1, NR3C1, EDF1	**HNF4A**, **FOXA2**, PBX1	Hepatocyte	30348985
Adult excitatory neuron	MLXIPL, PEG3, HLF, BHLHE40, KLF9	**FOXG1**, CEBPB, PBX1	adult_Ex	31619793
Organoid excitatory neuron	**NEUROG2**, SOX11, SOX4, CSRP2, CARHSP1	**FOXG1**, PBX1	hOrga_EN	
Adult inhibitory neuron	PEG3, MLXIPL, HLF, PPARGC1A, KLF9	**FOXG1**, **SOX2**, PBX1	adult_In	31619793
Organoid inhibitory neuron	SIX3, PAX6, ID4, KLF10, MEIS2	**ASCL1**, **SOX2**, SOX9	hOrga_IN	

Note: Experimentally validated conversion TFs are marked in bold. TFs are ordered from left to right by fold change to MEF/HFF. Cluster IDs annotated to same cell types in PanglaoDB were merged prior to analysis. Macrophage data from different tissues (heart, kidney, lung, muscle, brain, pancreas, skin spleen, trachea) were merged. See Table S3 for literature evidence for predicted conversion TFs.

2.2 | Application to various cell conversions

To demonstrate the applicability of TransSynW, we applied it to different cell systems, which encompassed conversions into broad cell types, subtypes, and phenotypic states (Tables 1 and S3). For example, in the first category, FOXA2, FOXA3, and HNF4A were predicted for the hepatocyte, which, together with HNF1A predicted in the specific part, are known for hepatocyte conversion.[2] The predicted TFs for the beta cells included NKX6-1, MAFA, PDX1, and NEUROD1, which have been shown to induce beta cell conversion.[3-5] Moreover, in both cases the predicted marker genes recapitulated commonly used ones (Tables 2 and S4). Indeed, many predicted conversion TFs are known to regulate each other and the predicted marker genes (Figure 2A,B), supporting the biological relevance of synergistic interactions captured by TransSynW.

Next, we analyzed different subtypes of neurons, as they are one of the most well studied subtypes. Among the predicted TFs for dopaminergic (DA) neurons, MYT1L, ASCL1, FOXA2, and GATA3 have

TABLE 2 Predicted marker genes with documented evidence

Cell type	Predicted marker gene with evidence	Reference (PubMed ID or website)
(1) Conversion into broad cell type		
Myoblast	CALCR, FGFR4, DES, ANKRD1, FITM1	12223412, 26440893, 26492245, 24644428, 8120103
Keratinocyte	KRT5	22028850
Cardiomyocyte	NPPA, MYH6	27123009, https://www.rndsystems.com/cn/research-area/cardiac-stem-cell-markers
Hepatocyte	SRD5A2, FGF21	25974403, 28515909
HSC	ESAM, LHCGR, SLC22A3, TIE1, ANGPT1, RBP1	https://www.rndsystems.com/cn/research-area/hematopoietic-stem-cell-markers 27365425, 27225119
Neuron	HTR2C, NTNG1, HS6ST3	30078709
Oligodendrocyte/OPC	MAG, CLDN11, PLEKHH1, ASPA, TRF	29024657
Macrophage	FOLR2, F13A1, LYZ2, PF4, MGL2, MMP13, CLEC10A	28576768, 29622724, 25477711,
Beta cell	INS1, INS2, G6PC2	22745242, 15133852, 25322827
NSC	NUDC, TUBA1B, TUBA1A	21771589, 29057214, 29281841
(2) Conversion into subtype		
Dopaminergic neuron	ALDH1A1, TH	30096314, http://www.abcam.com/neuroscience/neural-markers-guide
Medial floorplate progenitor	WNT1, MDK	31080111, 24125182, 11750071
GABAergic neuroblast	GAD2	http://www.abcam.com/neuroscience/neural-markers-guide
Oculomotor neuron	PRPH, FGF10, SLIT3, EYA1	24549637, 9221911, 20215354, 31080111
Serotonin neuron	TPH2, SLC6A4	http://www.abcam.com/neuroscience/neural-markers-guide
CD8+ memory T cell	SELL, CXCR5, DRC1	29236683, 18000950, 30243945
Memory B cell	TNFRSF13B, CD27	Company ebioscience, miltenyibiotec
(3) Phenotype conversion		
Primed mESC 1	BMP4	26860365
Active NSC	CENPF	29727663
Quiescent NSC	GJA1	29727663
Fetal hepatocyte	FGB, CYP2E1	28166538, 29622030
Adult hepatocyte 1	CYP3A4	26838674
Adult hepatocyte 2	APOA1	28166538
Adult excitatory neuron	CCK	12815247
Adult inhibitory neuron	CCK, PVALB, CRH	12815247, 2196836, 2843570

Note: See Table S4 for full list of predicted marker genes.

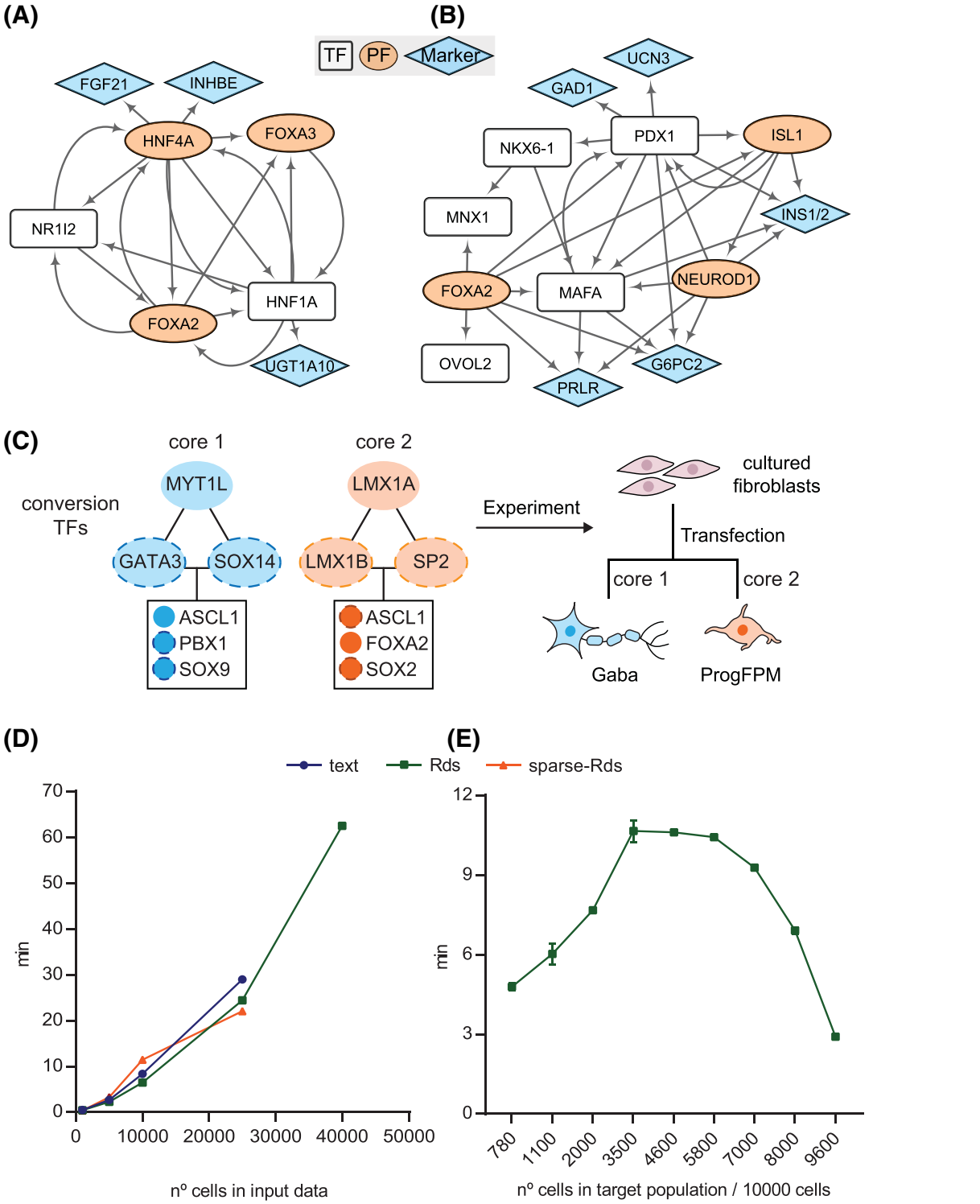


FIGURE 2 Transcriptional regulatory interactions among predicted conversion transcription factors (TFs) and marker genes for, A, hepatocyte and B, beta cell. Interaction data were retrieved from MetaCore from Clarivate Analytic in May/2020. C, Experimental strategy to improve cell conversion protocols for GABAergic neurons (Gaba) and medial floorplate progenitor (ProgFPM) based on TransSynW predicted core TFs. Dashed outlines represent nonvalidated TFs in the literature. D, Processing time vs number of cells in input scRNA-seq file (n = 3). Target population size was fixed to 8% of total size. E, Processing time for Rds files vs number of cells in target population (n = 3). Input population size was fixed to 10 000

been shown to generate DA neurons.[6-8] The predicted TFs for the medial floorplate progenitor, LMX1A and FOXA2, are consistent with the previous attempt to derive this cell subtype.[9] ASCL1 is sufficient to convert fibroblasts into GABAergic neurons.[10] Consistently, the predicted TFs for GABAergic neuroblasts contained ASCL1 and no other TFs known to generate other neuronal subtypes. The predicted

TFs for oculomotor neuron included ISL1, PHOX2A, and PHOX2B which have been reported to generate motor neurons via a synergistic interaction.[11,12] FEV, GATA2, and LMX1B were predicted for serotonergic neurons, which are among the TFs used for deriving this cell subtype.[13] We considered memory T and B cells as subtypes of their naive counterparts. Although a defined set of TFs for generating T cells has not been reported, the nonspecific PFs for both CD4+ and CD8+ T cells contained GATA3 and CEBPB, suggesting that these factors are primary candidates for experimental validation. Indeed, GATA3 is implicated in CD8+ memory T cell conversion.[14] Among the specific TFs, ID3, MYC, BACH2, and EOMES are reported to initiate CD8+ memory T cell conversion.[15-17] The known marker genes, such as SELL and CXCR5, were also identified. Finally, the nonspecific PFs for the memory B cells included IRF8 and SPI1, which together are implicated in the generation of B cell memory.[18]

Another type of cell conversion is phenotypes of a same cell type. The predicted nonspecific PFs for the two mouse embryonic stem cells (mESC) datasets are known to induce pluripotency.[19-21] The specific conversion TFs predicted for both primed mESC populations were LIN28A and FOXP1. LIN28A is known to induce the transition from naive to primed mESCs.[22] FOXP1 is implicated in maintaining pluripotency under non-2i conditions.[23] Whether FOXP1 induces a transition from a naive state to a primed state calls for further investigations. MEIS2 was predicted for both naive mESC populations. Little is known about its role in mESC regulation and hence it constitutes a novel candidate gene. The nonspecific conversion PFs for both active (aNSCs) and quiescent (qNSCs) consisted of known NSC-conversion TFs (eg, ASCL1, SOX2, FOXG1). The specific TFs for aNSCs contained EGR1 known to activate EGFR and accelerate proliferation of NSCs,[24] and E2F1, which is a cell cycle regulator linked to EGFR signaling in NSCs.[25] The conversion TFs for qNSCs included ID2, a BMP effector that has been inferred to regulate qNSCs.[26] Furthermore, CENPF and GJA1 are implicated as markers for late-aNSCs and qNSCs, respectively.[27] Next, the scRNA-seq data of organoid[28] and in vivo hepatocytes[28-30] were analyzed. The nonspecific PFs included general hepatocyte conversion TFs (eg, HNF4A, FOXA2, GATA3). Among the specific TFs for the in vivo hepatocytes were ZBTB20, KLF6, KLF9, CEBPD, and NR3C1. ZBTB20, KLF9 are important for hepatocyte proliferation,[31] whereas KLF6, CEBPD, KLF9, and NR3C1 regulate hepatic glucose and lipid metabolism,[32-34] suggesting that the derivation of in vivo hepatocytes might require sustained cell proliferation and proper metabolization of glucose and lipids. Known hepatocyte marker genes, such as FGB, CYP2E1, CYP3A4, APOA1, were predicted only for the in vivo hepatocytes but none for the in vitro ones. Finally, TransSynW was applied to in vivo and organoid excitatory and inhibitory neurons.[35] TFs predicted only for the in vivo excitatory and inhibitory neurons contained many common TFs (PEG3, KLF9, HLF, and MLXIPL), suggesting a common maturation mechanism. KLF9 is known to be necessary for late-phase maturation of neurons.[36] BHLHE40, which was only predicted for the in vivo excitatory neurons, is implicated in the regulation of neuronal excitability.[37] Moreover, a few known markers (CCK, PVALB, CRH) for excitatory/inhibitory neurons were predicted only for the adult samples. It would be of interest to experimentally test if predicted conversion TFs could indeed convert organoid cells into functional ones.

Taken together, we demonstrated that TransSynW can be effectively applied for identifying conversion TFs for a wide range of cell types. An example experimental strategy for using TransSynW predicted conversion TFs is shown in Figure 2C.

2.3 | Processing speed

The processing speed of TransSynW was assessed using text file, Rds file and a sparse matrix saved as Rds file (sparse-Rds). The time required for the upload of the data was not considered for this analysis. Thus, depending on the users internet connection speed, the overall processing time may vary to a certain degree. Rds files were the most efficient in processing 10 000 cells (6 minutes) (Figure 2D). In addition, up to 40 000 cells were successfully processed with Rds files, whereas only 25 000 cells in the other formats. This is in accordance with the respective file sizes (Table S5). If users wish to use datasets larger than 40 000 cells, we recommend to down-sample them. Next, we benchmarked the execution time against the target cell population size in 10 000 cells. The processing time peaked at 11 minutes for 3500 cells (Figure 2E). Afterwards, it started decreasing due to the reduced size of the background populations. Our general recommendation to users is to use Rds files for datasets with more than 10 000 cells.

3 | DISCUSSION

We have introduced a scRNA-seq based web application, TransSynW, for unbiased identification of cell conversion TFs, following the increasing interest from experimental researchers in generating novel functional cell types identified by scRNA-seq. TransSynW does not require prior biological knowledge, computer programming and computational resources. Moreover, TransSynW identifies potential marker genes for target cell types, which researchers can use for assessing the performance of conversion experiments. Furthermore, prioritization of PFs well recapitulated known conversion TFs in various systems, and predicted novel ones. We foresee that TransSynW will be a valuable tool for the experimental community, particularly for the generation of novel cell populations for stem cell research and regenerative medicine purposes.

4 | MATERIALS AND METHODS

4.1 | Implementation

TransSynW is written in HTML, JavaScript (frontend), PHP and Bash (backend), and runs on a virtual server hosted by Luxembourg Centre of Systems Biomedicine (LCSB, University of Luxembourg).

The frontend allows users to upload all required data, which are then parsed to the backend as different variables. In the backend bash script, the variables are parsed to the TransSynW main R script as different arguments. The output files are compressed into a .zip folder and sent to the user-specified E-mail address.

4.2 | Identification of conversion TFs

The main algorithm is based on the notion that conversion TFs consist of a combination of TFs that are specifically expressed in a target population and TFs that are more broadly expressed in the background population, and that these TFs synergistically interact with each other.[38] The algorithm follows four major steps.

- *Step 1: Identification of candidate TFs*.
  TransSynW first normalizes the data by the total RNA counts. Then TFs whose expression value is 0 across all cells in the target cell population are discarded. Next, it selects top 300 lowest CV (coefficient of variation) TFs as potential candidate TFs, since using more than this number of TFs often resulted in an out-of-memory error during the subsequent computation and conversion TFs usually exhibit low expression variation.
- *Step 2: Identification of most specifically expressed TFs*.
  The set of TFs that are specifically expressed in the target population is determined by Jensen-Shannon Divergence (JSD). JSD is computed for each TF in each cell and the summed JSD value for each TF over all cells is calculated. The top 10 lowest summed-JSD TFs are selected as the most specifically expressed TFs.
- *Step 3: Identification of most synergistic set of specifically expressed TFs*.
  Next, TransSynW identifies the most synergistic subset of TFs among the most specifically expressed TFs by computing MMI.[39]

$$\mathrm{MMI}(S) = -\sum_{T \subseteq S} (-1)^{|T|} H(T),$$

  where $S = \{X_1, X_2, \ldots, X_k\}$, T is a subset of S, $|T|$ denotes the cardinality of T, and H is Shannon's entropies. Negative MMI values imply a synergistic interaction among the TFs.[39] TransSynW first computes MMI of all sets of three TFs among the most specifically expressed TFs. Then a new TF is added to this set and MMI is computed again. If MMI is synergistic, then the next TF is added to the previous set, and so on. This iteration continues until either MMI no longer shows synergy, or when the maximum core size is reached. Here, the maximum core size was set to five.
- *Step 4: Addition of PFs*.
  The specific TF set from step 3 is extended with the nonspecific part, consisting solely of PFs. Every subset of three PFs is added to the specific part. MMI is computed for each set of all TFs and the most synergistic combination is selected as the final conversion TF set.

The final conversion TFs are ranked by the expression fold change calculated between the target cell population and starting cell population.

4.3 | Identification of marker genes

The marker gene set (Table S6) was collected from the following sources; extracellular proteins and membrane receptors,[40] cytoskeletal genes (http://www.informatics.jax.org/), metabolic genes (https://www.vmh.life/#human/all) and CD markers for immune cells (www.abcam.com/CDmarkers). These genes are relatively easily accessible for experimental validation. TransSynW identifies the top 10 candidate marker genes among this compiled set by computing JSD. Literature evidence for predicted markers were collected either manually or from CellMarker (http://biocc.hrbmu.edu.cn/CellMarker/).

4.4 | PF set

Information on PFs that have previously been reported to be involved in cell conversion protocols was manually collected from literature. The list is available in Table S1.

4.5 | scRNA-seq data of starting cell populations

scRNA-seq data of starting cell types were collected from Cell Ranger, GEO and Array Express databases, log 2 transformed and mean gene expression was calculated and compiled in TransSynW (Table S2).

4.6 | scRNA-seq dataset of target cell populations

scRNA-seq data used in this study were obtained from the following sources.[29-31,35,41-48] For References 43, 48, the reprocessed data were retrieved from PangloaDB,[49] as the cell annotation was more accurate than the original one.

ACKNOWLEDGMENTS

We thank Sybille Barvaux for helping gather evidence for pioneer factors. We thank Ernest Arenas, Igor Cervenka, and other anonymous researchers for giving us valuable feedbacks for developing the web application. M.M.R. is supported by Fonds National de la Recherche Luxembourg (C17/BM/11662681).

CONFLICT OF INTEREST

The authors declared no potential conflicts of interest.

AUTHOR CONTRIBUTIONS

M.M.R., S.O.: collection and/or assembly of data, data analysis and interpretation, manuscript writing; A.d.S.: conception and design, manuscript writing, final approval of manuscript.

DATA AVAILABILITY STATEMENT

TransSynW web application is available at https://transsynw.lcsb.uni.lu/. The code repository is available at https://git-r3lab.uni.lu/mariana.ribeiro/transsynw.

ORCID

Antonio del Sol https://orcid.org/0000-0002-9926-617X

REFERENCES

1. Colasante G, Rubio A, Massimino L, Broccoli V. Direct neuronal reprogramming reveals unknown functions for known transcription factors. *Front Neurosci*. 2019;13:283-290.
2. Huang P, He Z, Ji S, et al. Induction of functional hepatocyte-like cells from mouse fibroblasts by defined factors. *Nature*. 2011;475:386-391.
3. Gefen-Halevi S, Rachmut IH, Molakandov K, et al. NKX6.1 promotes PDX-1-induced liver to pancreatic β-cells reprogramming. *Cell Reprogram*. 2010;12:655-664.
4. Guo QS, Zhu MY, Wang L, et al. Combined transfection of the three transcriptional factors, PDX-1, neuroD1, and MafA, causes differentiation of bone marrow mesenchymal stem cells into insulin-producing cells. *Exp Diabetes Res*. 2012;2012:672013-672023.
5. Wang L, Huang Y, Guo Q, et al. Differentiation of iPSCs into insulin-producing cells via adenoviral transfection of PDX-1, NeuroD1 and MafA. *Diabetes Res Clin Pract*. 2014;104:383-392.
6. Hong SJ, Choi HJ, Hong S, Huh Y, Chae H, Kim KS. Transcription factor GATA-3 regulates the transcriptional activity of dopamine β-hydroxylase by interacting with Sp1 and AP4. *Neurochem Res*. 2008;33:1821-1831.
7. Pfisterer U, Kirkeby A, Torper O, et al. Direct conversion of human fibroblasts to dopaminergic neurons. *Proc Natl Acad Sci U S A*. 2011;108:10343-10348.
8. Seok JH, Huh Y, Chae H, Hong S, Lardaro T, Kim KS. GATA-3 regulates the transcriptional activity of tyrosine hydroxylase by interacting with CREB. *J Neurochem*. 2006;98:773-781.
9. Okawa S, Saltó C, Ravichandran S, et al. Transcriptional synergy as an emergent property defining cell subpopulation identity enables population shift. *Nat Commun*. 2018;9:1-10.
10. Chanda S, Ang CE, Davila J, et al. Generation of induced neuronal cells by the single reprogramming factor ASCL1. *Stem Cell Rep*. 2014;3:282-296.
11. Mazzoni EO, Mahony S, Closser M, et al. Synergistic binding of transcription factors to cell-specific enhancers programs motor neuron identity. *Nat Neurosci*. 2013;16:1219-1227.
12. Mong J, Panman L, Alekseenko Z, et al. Transcription factor-induced lineage programming of noradrenaline and motor neurons from embryonic stem cells. *Stem Cells*. 2014;32:609-622.
13. Vadodaria KC, Mertens J, Paquola A, et al. Generation of functional human serotonergic neurons from fibroblasts. *Mol Psychiatry*. 2016;21:49-61.
14. Wang Y, Misumi I, Di Gu A, et al. GATA-3 controls the maintenance and proliferation of T cells downstream of TCR and cytokine signaling. *Nat Immunol*. 2013;14:714-722.
15. Istaces N, Splittgerber M, Lima Silva V, et al. EOMES interacts with RUNX3 and BRG1 to promote innate memory cell formation through epigenetic reprogramming. *Nat Commun*. 2019;10:3306.
16. Ji Y, Pos Z, Rao M, et al. Repression of the DNA-binding inhibitor Id3 by Blimp-1 limits the formation of memory CD8 + T cells. *Nat Immunol*. 2011;12:1230-1237.
17. Roychoudhuri R, Clever D, Li P, et al. BACH2 regulates CD8 + T cell differentiation by controlling access of AP-1 factors to enhancers. *Nat Immunol*. 2016;17:851-860.
18. Carotta S, Willis SN, Hasbold J, et al. The transcription factors IRF8 and PU.1 negatively regulate plasma cell differentiation. *J Exp Med*. 2014;211:2169-2181.
19. Feng B, Jiang J, Kraus P, et al. Reprogramming of fibroblasts into induced pluripotent stem cells with orphan nuclear receptor Esrrb. *Nat Cell Biol*. 2009;11:197-203.
20. Hester ME, Song SW, Miranda CJ, Eagle A, Schwartz PH, Kaspar BK. Two factor reprogramming of human neural stem cells into pluripotency. *PLoS One*. 2009;4:e7044.
21. Wernig M, Meissner A, Cassady JP, Jaenisch R. c-Myc is dispensable for direct reprogramming of mouse fibroblasts. *Cell Stem Cell*. 2008;2:10-12.
22. Zhang J, Ratanasirintrawoot S, Chandrasekaran S, et al. LIN28 regulates stem cell metabolism and conversion to primed pluripotency. *Cell Stem Cell*. 2016;19:66-80.
23. Gabut M, Samavarchi-Tehrani P, Wang X, et al. An alternative splicing switch regulates embryonic stem cell pluripotency and reprogramming. *Cell*. 2011;147:132-146.
24. Alagappan D, Balan M, Jiang Y, Cohen RB, Kotenko SV, Levison SW. Egr-1 is a critical regulator of EGF-receptor-mediated expansion of subventricular zone neural stem cells and progenitors during recovery from hypoxia-hypoglycemia. *ASN Neuro*. 2013;5:183-193.
25. Morizur L, Chicheportiche A, Gauthier LR, Daynac M, Boussin FD, Mouthon MA. Distinct molecular signatures of quiescent and activated adult neural stem cells reveal specific interactions with their microenvironment. *Stem Cell Rep*. 2018;11:565-577.
26. Llorens-Bobadilla E, Zhao S, Baser A, Saiz-Castro G, Zwadlo K, Martin-Villalba A. Single-cell transcriptomics reveals a population of dormant neural stem cells that become activated upon brain injury. *Cell Stem Cell*. 2015;17:329-340.
27. Shah PT, Stratton JA, Stykel MG, et al. Single-cell transcriptomics and fate mapping of ependymal cells reveals an absence of neural stem cell function. *Cell*. 2018;173:1045-1057.e9.
28. Hu H, Gehart H, Artegiani B, et al. Long-term expansion of functional mouse and human hepatocytes as 3D organoids. *Cell*. 2018;175:1591-1606.e19.
29. Aizarani N, Saviano A, Sagar LM, et al. A human liver cell atlas reveals heterogeneity and epithelial progenitors. *Nature*. 2019;572:199-204.
30. MacParland SA, Liu JC, Ma XZ, et al. Single cell RNA sequencing of human liver reveals distinct intrahepatic macrophage populations. *Nat Commun*. 2018;9:1-21.
31. Cvoro A, Devito L, Milton FA, et al. A thyroid hormone receptor/KLF9 axis in human hepatocytes and pluripotent stem cells. *Stem Cells*. 2015;33:416-428.
32. Bechmann LP, Vetter D, Ishida J, et al. Post-transcriptional activation of PPAR alpha by KLF6 in hepatic steatosis. *J Hepatol*. 2013;58:1000-1006.
33. Lai PH, Wang WL, Ko CY, et al. HDAC1/HDAC3 modulates PPARG2 transcription through the sumoylated CEBPD in hepatic lipogenesis. *Biochim Biophys Acta—Mol Cell Res*. 2008;1783:1803-1814.
34. Pei H, Yao Y, Yang Y, Liao K, Wu JR. Krüppel-like factor KLF9 regulates PPARγ transactivation at the middle stage of adipogenesis. *Cell Death Differ*. 2011;18:315-327.
35. Kanton S, Boyle MJ, He Z, et al. Organoid single-cell genomic atlas uncovers human-specific features of brain development. *Nature*. 2019;574:418-422.
36. Scobie KN, Hall BJ, Wilke SA, et al. Krüppel-like factor 9 is necessary for late-phase neuronal maturation in the developing dentate gyrus and during adult hippocampal neurogenesis. *J Neurosci*. 2009;29:9875-9887.
37. Hamilton KA, Wang Y, Raefsky SM, et al. Mice lacking the transcriptional regulator Bhlhe40 have enhanced neuronal excitability and impaired synaptic plasticity in the hippocampus. *PLoS One*. 2018;13:1-22.
38. Okawa S, Del Sol A. A general computational approach to predicting synergistic transcriptional cores that determine cell subpopulation identities. *Nucleic Acids Res*. 2019;47:3333-3343.
39. Bell AJ. The co-information lattice. In: Proceedings of the Fifth International Workshop on Independent Component Analysis and Blind Signal Separation; 2003. p. 921. http://www.kecl.ntt.co.jp/icl/signal/ica2003/cdrom/data/0187.pdf.

40. Ramilowski JA, Goldberg T, Harshbarger J, et al. A draft network of ligand-receptor-mediated multicellular signalling in human. *Nat Commun*. 2015;6:7866.
41. Horns F, Dekker CL, Quake SR. Memory B cell activation, broad anti-influenza antibodies, and bystander activation revealed by single-cell transcriptomics. *Cell Rep*. 2020;30:905-913.e6.
42. Kalamakis G, Brüne D, Ravichandran S, et al. Quiescence modulates stem cell maintenance and regenerative capacity in the aging brain. *Cell*. 2019;176:1407-1419.e14.
43. Kimmel JC, Penland L, Rubinstein ND, Hendrickson DG, Kelley DR, Rosenthal AZ. Murine single-cell RNA-seq reveals cell-identity- and tissue-specific trajectories of aging. *Genome Res*. 2019;29:2088-2103.
44. Kolodziejczyk AA, Kim JK, Tsang JCH, et al. Single cell RNA-sequencing of pluripotent states unlocks modular transcriptional variation. *Cell Stem Cell*. 2015;17:471-485.
45. Kumar RM, Cahan P, Shalek AK, et al. Deconstructing transcriptional heterogeneity in pluripotent stem cells. *Nature*. 2014;516:56-61.
46. La Manno G, Gyllborg D, Codeluppi S, et al. Molecular diversity of midbrain development in mouse, human, and stem cells. *Cell*. 2016;167:566-580.e19.
47. Patil VS, Madrigal A, Schmiedel BJ, et al. Precursors of human CD4+ cytotoxic T lymphocytes identified by single-cell transcriptome analysis. *Sci Immunol*. 2018;3:1-14.
48. Schaum N, Karkanias J, Neff NF, et al. Single-cell transcriptomics of 20 mouse organs creates a Tabula Muris. *Nature*. 2018;562:367-372.
49. Franzén O, Gan LM, Björkegren JLM. PanglaoDB: a web server for exploration of mouse and human single-cell RNA sequencing data. *Database*. 2019;2019:1-9.

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of this article.

How to cite this article: Ribeiro MM, Okawa S, del Sol A. TransSynW: A single-cell RNA-sequencing based web application to guide cell conversion experiments. *STEM CELLS Transl Med*. 2021;10:230–238. https://doi.org/10.1002/sctm.20-0227

4.3. Strategies for *in vitro* direct reprogramming of astrocytes by overexpression of novel identity transcription factors

4.3.1. Preface

Loss of neurons is a hallmark transversal to the several types of neurodegenerative diseases. For instance, PD is characterized by the selective loss of DANs of the SNpc. To replace lost neuronal populations, CRT strategies, such as direct reprogramming, emerged as promising approaches due their potential *in situ* application. Several direct reprogramming protocols were successful in converting astrocytes into distinct types of neurons, including DANs. However, these protocols lack control over subtype specification, which results on the generation of a heterogeneous cellular preparation, with different phenotypes and levels of functionality.

We developed a direct reprogramming strategy that leverages the multiplexing capability of a previously established CRISPR-dCas9 system to induce the endogenous expression of multiple conversion TFs. To identify these conversion factors, we applied TransSynW to a previously published scRNA-seq dataset of the human midbrain tissue. By applying these protocols, we were able to convert human astrocytes into TUBB3-positive cells. We also established a sequential reprogramming protocol which is divided into two steps. First, we induced the expression of conversion TFs to obtain DANs, and then we overexpress the TFs identified to be involved in subtype specialization. Based on the ectopic expression of the conversion TFs identified by TransSynW for this two-step protocol, we were able to convert human astrocytes into cells with a neuronal-like morphology. After performing further optimization studies, these cellular conversion approaches have the potential to overcome current limitations and advance the field of regenerative medicine.

In this study, I applied TransSynW to obtain the conversion TFs, validated the gRNAs that promote overexpression of the target TFs, engineered the DNA vectors and lentiviruses, implemented the CRISPR-dCas9 and cDNA-based approaches, performed the direct and sequential reprogramming, and validated the overexpression of the target TFs and marker genes.

4.3.2. Manuscript

Strategies for direct reprogramming of astrocytes into neurons by overexpression of novel identity transcription factors

Mariana Messias Ribeiro[1,2], Anqi Xiong[2], Satoshi Okawa[1], Antonio del Sol[1,3,4,*] and Ernest Arenas[2,*]

[1] Luxembourg Centre for Systems Biomedicine (LCSB), University of Luxembourg, L-4362 Esch-sur-Alzette, Luxembourg

[2] Laboratory of Molecular Neurobiology, Department of Medical Biochemistry and Biophysics, Karolinska Institutet, 171 77 Stockholm, Sweden

[3] CIC bioGUNE-BRTA (Basque Research and Technology Alliance), Bizkaia Technology Park, 48160 Derio, Spain

[4] IKERBASQUE, Basque Foundation for Science, 48012 Bilbao, Spain

[*] To whom correspondence should be addressed: Antonio.delsol@uni.lu (AdS) and Ernest.arenas@ki.se (EA)

Abstract

One of the main goals of regenerative medicine is to create functional and mature cell types that can replace damaged tissues. In that regard, direct reprogramming is a promising approach to replace the loss of neuronal cells in neurodegenerative diseases, such as Parkinson's disease. Here, we applied a computation tool, TransSynW, to a single-cell RNA-sequencing dataset from the developing human ventral midbrain tissue to predict the most suitable transcription factors (TFs) to directly reprogram human astrocytes into three subtypes of midbrain dopaminergic neurons (DANs). We also predicted conversion TFs for a novel sequential reprogramming protocol to make first generic DANs and then DAN subtypes. To induce the endogenous expression of candidate TFs *in vitro*, we used a programmable CRISPR-dCas9 system. Although most of the TFs could be efficiently expressed, one required treatment with epigenetic modifiers, and others could not be

expressed using this system, even after testing multiple guide RNAs. These results indicate that the success of the CRISPR-dCas9 system in cellular conversion protocols is partially limited by epigenetic mechanisms. Based on the ectopic expression of TFs, we developed a direct and a sequential approach to convert human astrocytes into neuronal-like cells. Optimization approaches to generate midbrain DAN subtypes are suggested. We also discuss current limitations as well as potential therapeutic applications.

Introduction

Neurodegenerative diseases are characterized by a continuous loss of neuronal function and structure in the brain (Hung et al., 2010). The selective loss of nigrostriatal dopaminergic neurons (DANs) in the substantia nigra (SN) *pars compacta* is one of the hallmarks of Parkinson's Disease (PD) (McGregor & Nelson, 2019). Due to the intricate mechanisms behind neuronal loss and the lack of regenerative ability of the midbrain, it has not been possible to develop a treatment to efficiently prevent or reverse the damage caused by disease (Sivandzade & Cucullo, 2021). Cell replacement therapy (CRT) has emerged as a promising approach to fight neurodegenerative diseases. Indeed, clinical trials have shown that the motor symptoms of PD patients can be alleviated by grafting embryonic midbrain tissue containing DANs (Lindvall et al., 1988). In a few cases, patients have been able to stop taking their levodopa medication and remained asymptomatic for 15 years (Kefalopoulou et al., 2014), achieving near-normal levels of dopaminergic innervation after 24 years (Li et al., 2016). More recently, pluripotent stem cells have been used to generate DANs and successfully rescue motor deficits in animal models of this disease (Kikuchi et al., 2017; Kim et al., 2021; Kirkeby et al., 2017). This work has led to the initiation of clinical trials in PD (Barker et al., 2013; Doi et al., 2020; Kim et al., 2021; Piao et al., 2021; Schweitzer et al., 2020; Tao et al., 2021). However, current differentiation protocols generate multiple cell (sub)types and selective generation of subtype specific midbrain DANs has not yet been achieved.

Direct reprogramming of glial cells, such as astrocytes, into induced DANs (iDANs) emerged as a promising therapeutic strategy to replace DANs in animal models of PD (Rivetti Di Val Cervo et al., 2017). The main advantage of this approach is its potential *in situ* application, which bypasses the need for an external source of cells, extensive *ex vivo* cultivation, and the transplantation procedure (H. Wang et al., 2021). Several studies

successfully reprogrammed astrocytes into distinct types of functional neurons by overexpressing different combinations of transcription factors (TFs) (Berninger et al., 2007; Heinrich et al., 2011; Rivetti Di Val Cervo et al., 2017; Torper et al., 2015). More recently, CRISPR-dCas9 based systems have become an increasingly attractive tool for direct reprogramming (Black et al., 2016; Giehrl-Schwab et al., 2022; C. Wang et al., 2017). These approaches rely on the delivery of dCas9, an activator and short guide RNA (gRNA) molecules to induce the expression of genes from their endogenous locus. The small size of the gRNAs facilitates the expression of multiple TFs simultaneously (Dominguez et al., 2016). In this paper, we adapted a programmable CRISPR-dCas9 system (Zalatan et al., 2015) with RNA scaffolds that encompass not only the gRNAs, but also the activator sequences that act as recruiters of the transcriptional machinery towards the promoter region of the target TFs.

The advent single-cell RNA-sequencing (scRNA-seq) has enabled the discovery of previously unsuspected cell heterogeneity in multiple tissues. In a pioneering study, La Manno *et. al.* identified three different subtypes of human DANs in the fetal ventral midbrain, named hDA0, hDA1, and hDA2 (la Manno et al., 2016). This study defined TF combinations expressed in each cell (sub)type, opening the door for the design of cell conversion strategies. However, the question of which of the multiple factors expressed in each cell (sub)type are sufficient for conversion of one cell type into another still remains to be answered.

In this study we applied TransSynW, a recently described computational tool, to identify the best combination of TFs to convert between cell subtypes (Ribeiro et al., 2021). Using this tool and scRNA-seq of the human ventral midbrain (la Manno et al., 2016), we identify the most suitable TFs to convert astrocytes into three subtypes of midbrain DANs. We hereby report variable success in direct *in vitro* reprogramming of human astrocytes into neurons and we propose strategies to improve the conversion of astrocytes into specific subtypes of DANs.

Results

Identification of the candidate conversion transcription factors to generate dopaminergic neuron subtypes

The application of TransSynW to scRNA-seq data of the developing human midbrain (La Manno et al., 2016) revealed that the most favorable TFs for the conversion of astrocytes into human DAN subtypes were: *EBF2*, *EN1*, *FOXA2* and *MYT1L* for hDA0; *ASCL1*, *EBF3*, *FOXA2* and *NPAS4* for hDA1; and *BNC2*, *PBX1* and *POU6F1* for hDA2 subtypes (Figure 1A and Table 1). Interestingly, it was previously shown that the distinct expression of *SOX6* and *OTX2* regulates the specification of DAN subpopulations in the midbrain (Panman et al., 2014). Specifically, *SOX6* is selectively expressed in DANs located in the SN while *OTX2* expression is confined to the ventral tegmental area. When evaluating the expression of these TFs in the DAN subtypes identified in the human midbrain scRNA-seq dataset, hDA2 was found to share the expression of *SOX6*, amongst other features, with SN DANs (la Manno et al., 2016). Since PD is characterized by the loss of DANs in the SN, the hDA2 subtype is an interesting target for CRT strategies for PD.

Sequential cell conversion strategies have been shown to generate cells with higher functionality and improved efficiency (Bonora-Centelles et al., 2009; Gaeta et al., 2013; X. Liu et al., 2013; Morris et al., 2014). To test this approach, we developed a two-step reprogramming protocol and examine whether it is possible to generate DANs of the hDA2 subtype with enhanced functionalities and/or efficiency. In the first step of this protocol, TransSynW was applied to the scRNA-seq dataset to identify the conversion factors for generating generic DANs, without subtype identity. Then, TransSynW was used to determine the specific conversion factors leading to their specialization into the hDA2 subtype (Figure 1B). For this sequential strategy, we used a combination of three TFs to make generic DANs (*ASCL1*, *NR4A2,* and *PBX1*) followed by the induction of three other TFs to make the hDA2 subtype (*FOXA2*, *POU6F1*, and *SOX6*) (Table 1).

As positive control for the newly predicted conversion TFs, we overexpressed a combination of factors that has been previously shown to reprogram astrocytes to DANs both *in vitro* and *in vivo* (Rivetti Di Val Cervo et al., 2017), namely the TFs *ASCL1*, *LMX1A* and *NEUROD1* (NeAL), together with miR218 (referred to as NeAL218).

Establishment of a CRISPR-dCas9 activation system to directly reprogram human astrocytes *in vitro*

To achieve a reliable CRISPR-dCas9 activation (CRISPRa) in astrocytes, we engineered an immortalized human fetal astrocyte cell line to stably express dCas9 (hIA-dCas9). The expression of dCas9 was confirmed by immunocytochemistry (Supplemental Figure 1A). Next, RNA scaffold lentiviruses (LVs) containing gRNAs and an activator were employed to activate the endogenous expression of the selected TFs as described (Konermann et al., 2015; Zalatan et al., 2015). Specific gRNAs were found to induce the expression of most of the target genes in hIA-dCas9 by 48 hours after LV infection, including *ASCL1*, *EBF2*, *EN1*, *FOXA2*, *LMX1A*, *NEUROD1*, *NR4A2*, *NPAS4*, and *SOX6* (Figure 1C). However, the expression of *BNC2*, *MYT1L*, *PBX1*, and *POU6F1* could not be increased despite testing multiple gRNAs.

Previous studies have shown that chromatin regulators can act as barrier to reprogramming, and that by adding small molecules capable of modifying the epigenetic state, it is possible to improve cell conversion (Basu & Tiwari, 2021; Becker et al., 2017; Jin et al., 2021; Rivetti Di Val Cervo et al., 2017). We therefore included an overnight treatment with the histone deacetylase inhibitor valproic acid (VPA), and the DNA methyltransferase inhibitor (DNMTi) 5-aza-2'-deoxycytidine (Dec), prior to lentiviral infection (Huangfu et al., 2008; Pennarossa et al., 2013). Notably, by adding these chromatin remodeling agents, the expression of *MYT1L* increased and we thus had the complete set of gRNAs necessary to reprogram astrocytes into hDA0 and hDA1 subtypes (Table 1, Figure 1C).

Direct reprogramming of human astrocytes stably expressing dCas9 into hDA0 and hDA1 subtypes

To directly reprogram human astrocytes into hDA0 and hDA1 subtypes, we infected hIA-dCas9 astrocytes with RNA scaffold LVs encoding the validated gRNAs for each of our target genes. Each DNA construct encompassed the gRNA sequence to target a single TF as well as the activator (PCP-p65-HSF1). After LV infection, cells were cultivated and treated as previously described (Rivetti Di Val Cervo et al., 2017) with some minor modifications (Figure 2A and Methods). The expression of endogenous TFs being targeted by the programmable CRISPR-dCas9 system was examined four days after LV infection (Figure 2B). We found that the expression of *EBF2*, *EN1*, and *MYT1L* (required for hDA0 conversion) and that of *EBF3* (for hDA1) were lower than expected. To evaluate the

efficiency of each of the different TF combinations in directly reprogramming astrocytes, the presence of TUBB3 and TH was determined. All the protocols tested gave rise to TUBB3-positive cells (Figure 2C). TUBB3-positive cells in the control condition (NeAL) exhibited long processes and neuronal-like morphology. However, when we examined the expression of the dopaminergic marker tyrosine hydroxylase (TH), no positive cells were observed in either the control or any of the experimental conditions. Since expression levels of the TFs in the control experiments were acceptable (Figure 2B), the main difference with the experiments by Rivetti Di Val Cervo *et. al.* and the ones in the next section (Figure 3) was the use of the cell line hIA-dCas9 instead of the unmodified hIA cell line. We suspect that changes associated to the generation of the hIA-dCas9 cell line (i.e insertional mutagenesis, growth at clonal dilution, higher passage number, etc.) may have negatively affected the capacity of this cell to reprogram. The unexpected absence of a positive reprogramming control means that we were not able to ascertain whether the predicted TFs can promote the conversion of astrocytes into iDAN subtypes.

Reprogramming of human astrocytes into dopaminergic neurons or into hDA2 neurons by TF overexpression

To overcome the limitations associated to the inability of the CRISPR-dCas9 system to activate *BNC2*, *PBX1*, and *POU6F1*, and the problems with the hIA-dCas9 astrocytes, we decided to ectopically express the TFs predicted for reprograming hIA astrocytes into hDA2 neurons. For this purpose, we engineered inducible lentiviral expression vectors containing the coding sequences (cDNA) of the TFs determined for the direct and sequential hDA2 protocols (Figure 3A, Supplemental Table 1). As control, we used the NeAL218 combination, as described (Rivetti Di Val Cervo et al., 2017).

In the direct reprogramming protocol, the hIA cell line was infected with a LV encoding the reverse tetracycline transactivator together with distinct LVs encoding the conversion factors determined for the hDA2 subtype (Table 1). Ectopic gene expression was induced four days after LV infection by treating the cells with doxycycline and the rest of the protocol was performed as previously described (Figure 3B and Methods).

In the sequential protocol, we first used LVs containing the cDNA sequences for *ASCL1*, *NR4A2*, and *PBX1* to reprogram astrocytes into DANs (Figure 3B, Supplemental Table 1). Two days after the beginning of the doxycycline treatment, we infected cells with the second set of conversion TFs (*FOXA2*, *POU6F1*, and *SOX6*) to induce the hDA2 subtype. Transgene

expression was confirmed four days after doxycycline treatment by immunofluorescence (Supplemental Figure 1B) and quantified by RT-qPCR (Figure 3C). The gene expression levels obtained were comparable to those previously found to be sufficient to induce hIA astrocyte conversion into DANs (Rivetti Di Val Cervo et al., 2017).

Next, we evaluated the efficiency of the direct and the sequential hDA2 conversion protocols. Preliminary immunocytochemical analysis of the NeAL218 control cells at the end of the protocol revealed the presence of TH-positive cells exhibiting neuronal morphology (Figure 3D). In contrast, none of the two hDA2 protocols (direct or sequential) gave rise to any TH-positive cells. However, while the direct hDA2 protocol did not significantly change the morphology of the cells, the sequential hDA2 approach gave rise to cells exhibiting neuronal-like morphology, similar to the one obtained with the NeAL218 combination. These results, albeit very preliminary, confirm previous results indicating that it is possible to generate iDANs by overexpression of NeAL218 (Rivetti Di Val Cervo et al., 2017). On the other hand, the preliminary nature of the experiments does not allow determining whether the TFs predicted by TransSynW have contributed to change the phenotype of the cells towards the acquisition of hDA2 neurons. Further experiments would be necessary to confirm whether this is the case.

Discussion

Neurodegenerative diseases are characterized by the progressive loss of specific neuronal subpopulations. For instance, cholinergic neurons are the most affected during the progression of Alzheimer's Disease while DANs in the SN are the selectively lost during the development of PD (McGregor & Nelson, 2019; Niikura et al., 2006). So far, the treatment for these diseases mostly focuses on symptom management since the underlying cause of neurodegeneration is largely unknow and we lack therapeutic tools capable of addressing pathogenic mechanisms. An alternative approach is the development of treatments that can modify the course of disease, such as CRTs. Despite advances in the field, it is still not possible to replace the cell subtype lost during the development of PD. Generating the specific subtype of DANs lost in the ventral tier of the SN *pars compacta* still remains a challenge and a goal of CRT for PD. In this study, we focused on the development of two novel direct and one sequential reprogramming strategies to generate the DAN subtype lost in PD. We applied TransSynW to determine the most suitable combination of TFs to convert

astrocytes into human embryonic DANs of the hDA2 subtype which is thought to give rise to SN DANs.

Direct reprogramming protocols published so far have aimed at converting astrocytes into DANs by either delivering TFs through ectopic overexpression of cDNAs (Rivetti Di Val Cervo et al., 2017; Torper et al., 2015) or by inducing endogenous expression with CRISPRa-based systems (Giehrl-Schwab et al., 2022). In both cases, the success of the strategy largely depends on the levels of expression of TFs, either *in vitro* or *in vivo*. *In vitro* delivery of TFs generally offers better control of gene expression and results in high levels of expression and conversion to DANs (Giehrl-Schwab et al., 2022; Rivetti Di Val Cervo et al., 2017; Torper et al., 2015). In contrast, *in vivo* studies generally achieve lower levels of expression and shown variable results. To date, *in vivo* delivery of TFs by LV has been the only successful method at reprogramming adult endogenous astrocytes into iDANs (Rivetti Di Val Cervo et al., 2017), while both adeno-associated virus and CRISPRa delivery systems converted astrocytes into GABAergic neurons, despite using the same TFs as *in vitro* (Giehrl-Schwab et al., 2022; Torper et al., 2015). Similarly, in our experiments, we observe that the LV system gave rise to iDANs, while the CRISPRa system did not, which is consistent with previous findings.

Other factors may have also contributed to the discrepancies in reprogramming when employing the cDNA overexpression versus the CRISPRa systems. For instance, we suspect that differences in the strength of heterologous and endogenous promoters as well as in copy number of the target genes might have an influence in our results (Chavez et al., 2016; Fontana et al., 2020; Pang et al., 1999). To address this issue, alternative activator domains, such as SAM, SunTag, VPR or SPH systems (Clow et al., 2022; Konermann et al., 2015; Tanenbaum et al., 2014; H. Zhou et al., 2018) and alternative module strategy, such as the Casilio platform (Cheng et al., 2016), should be considered in future CRISPRa-based strategies. Additionally, engineering a polycistronic lentiviral vector encoding multiple gRNAs would allow for a more consistent expression of all the TFs in each infected cell. Moreover, the same gene could be targeted with multiple gRNAs (Chavez et al., 2016; Perez-Pinera et al., 2013; Shakirova et al., 2020) and both transcriptional and epigenetic control mechanisms could be regulated simultaneously in order to achieve higher levels of endogenous gene expression.

Additional factors may also limit the efficiency of the application of the CRISPRa system in reprogramming. It has been shown that epigenetic mechanisms associated to closed chromatin conformation can impair the binding of Cas9 to its targets and therefore compromise its function (Baumann et al., 2019; Jensen et al., 2017). This may explain why we were unable to regulate the endogenous expression of four TFs (*BNC2*, *MYT1L*, *PBX1*, and *POU6F1*). In support of this possibility, treatment with the epigenetic modifiers VPA and Dec allowed the upregulation of *MYT1L* in the hIA-dCas9 cells. However, treatment with zebularine, a strong DNMTi (L. Zhou et al., 2002), did not improve the expression of the remaining genes (data not shown). As multiple epigenetic marks can regulate gene expression, a better strategy for the future would be to characterize the epigenetic landscape of hIA-dCas9 cells by ATAC-seq and ChIP-seq. Upon identification of the chromatin barriers limiting the expression of these genes, epigenetic regulators could be recruited to specific locations in the chromatin using CRISPR-dCas9 systems. Factors such as Tet1, an enzyme controlling DNA demethylation, or p300, a histone acetyltransferase (Baumann et al., 2019; Delvecchio et al., 2013; X. Liu et al., 2013; Rada-Iglesias et al., 2011; Shrimp et al., 2018) could then be recruited to specific locations to enable transcription of the gene of interest.

With regards to the TFs predicted by TransSynW, experiments based on ectopic expression allowed us to evaluate the possible usefulness of the hDA2 reprogramming strategy, independently of the issues associated to the CRISPRa system described above and in the results section. In the direct hDA2 protocol *ASCL1*, *BNC2,* and *FOXA2* were overexpressed. While *ASCL1* and *FOXA2* have been previously expressed in the context of reprogramming, *BNC2* was tested for the first time. Interestingly, *BNC2* has been found to be constitutively expressed in endogenous midbrain DANs (la Manno et al., 2016), and mutations in this gene have been associated with the development of PD (Hook et al., 2018). However, we observed that upregulation of *BNC2* induced cell death in our cultures (Supplemental Figure 1B and C). Notably, in the context of cancer, *BNC2* downregulation has been connected to the development of hepatocellular, ovarian, and squamous cell carcinoma (Akagi et al., 2009; Cesaratto et al., 2016; Chahal et al., 2016; Wu et al., 2016). Accordingly, *BNC2* has been described as a tumor suppressor gene and its upregulation is known to prevent cell proliferation (Akagi et al., 2009; Cesaratto et al., 2016). The mechanisms behind the regulation of cell proliferation by *BNC2* remain elusive, however it has been shown that BNC2 physically interacts with both p53 and retinoblastom (pRb) proteins (Benevolenskaya

et al., 2005; J. Liu et al., 2020; NCBI Entry: BNC2, 2022). This is of relevance for our study because the hIA cell line was immortalized with the Simian virus 40 large T antigen, a protein that forms complexes with p53 and pRb (DeCaprio et al., 1988; Lane & Crawford, 1979; Lin & Simmons, 1991). We speculate that by overexpressing *BNC2* and promoting its interaction with p53 and pRB we may have inadvertently reversed the immortalization of the cell line, resulting in senescence and cell death. Future experiments should thus test this hypothesis by using non-immortalized astrocytes to examine whether *BNC2* can indeed contribute to hDA2 conversion.

In the two-step reprogramming protocol, we tested two sets of TFs (first *ASCL1*, *NR4A2*, and *PBX1*, and then *FOXA2*, *POU6F1*, and *SOX6*). Cells undergoing this protocol acquired neuronal-like morphology, suggesting a possible shift towards a neuronal fate, but did not acquire a TH-positive phenotype. Since cells reprogrammed with control factors (NeAL218) became both neuron-like and TH-positive, our results suggest that some of the factors currently used in the control protocol may be required. Future experiments should thus examine whether a combination of NeAL218 and subtype-specific TFs may allow conversion of astrocytes into DAN subtypes. For these experiments, longer differentiation times should be examined to allow for additional maturation and detection of functionality, such as dopamine release or electrophysiological properties defining DANs.

It would also be of importance to perform a more detailed characterization of gene expression at the single cell level during conversion. These experiments may contribute the resolve several key questions such as: Do astrocytes lose their original astrocyte phenotype during reprogramming? Do the reprogrammed cells acquire complete or partial neuronal phenotypes? Do converted cells resemble endogenous human midbrain neurons? What is the conversion trajectory? What are the combinations of TFs and their expression levels giving rise to either GABAergic neurons (Giehrl-Schwab et al., 2022; Torper et al., 2015) or DANs (Rivetti Di Val Cervo et al., 2017)? This information will be very valuable both to understand the reprogramming process and to further improve future protocols and strategies for the conversion of astrocytes into neurons.

Finally, applying these conversion protocols *in vivo* and developing novel systems that would ensure the safety and efficacy of the application of these strategies in patients will be essential to achieve clinical translation.

Material and Methods

Cell culture

hIA were a gift from Dr. Eugene O. Major, National Institute of Health (10Bl, a SVGmm single-cell clone). These cells were cultured and expanded in Eagle's Minimum Essential Medium (EMEM, ATCC #30-2003) with 20% Fetal Bovine Serum (FBS, Gibco #10270106) previously heat-inactivated for 30 mins at 56°C, and 50µg/mL of gentamicin (Gibco #15750060). hIA were maintained in a 5% CO2 incubator at 37 °C and used for experiments for a maximum of four passages.

Transformation and plasmid purification

Plasmids were transformed into Stbl3 chemically competent E. coli (Invitrogen #C737303) by heat shock and spread on agar plates containing 100µg/mL of ampicillin (Sigma #A5354). Plates were incubated at 37°C overnight and individual colonies were picked for further validation. Plasmid purification was performed using NucleoSpin Plasmid (Macherey-Nagel #740588.250) and NucleoBond Xtra Maxi EF (Macherey-Nagel #740424.50) when isolating lentiviral vectors. Purified plasmids were sent for Sanger sequencing using Eurofins Genomics' TubeSeq Service to confirm the correct assembly.

Lentiviral production

LVs were produced in HEK293FT cells (Invitrogen #R70007), cultured in Dulbecco's Modified Eagle Medium (DMEM with GlutaMAX Supplement, Gibco #31966021) with 10% FBS and 500µg/mL of geneticin (Gibco #10131027). HEK293FT cells were used for lentiviral production for a maximum of ten passages. On the day before the transfection, $5.0x10^6$ HEK293FT cells were plated in a T-75 flask ($6.6 x10^4$ cells/cm^2). On the following day the cells were transfected using 2.5mg/mL of polyethylenimine (PEI, Polysciences #23966-2). The lentiviral vectors of interest were co-transfected with 2^{nd} generation packaging plasmids psPAX2 (Addgene #12259) and pMD2.G (Addgene #12259) in a ratio of 5µg:5µg:2.5µg of DNA, respectively. After an overnight incubation, the culture medium containing the transfection mix was removed and fresh medium without geneticin was supplied to the cells. Two days after transfection, the culture medium was collected and stored at 4°C and fresh medium without geneticin was supplied to the cells. Three days after transfection, the culture medium was pooled together with the one collected on the previous day and centrifuged at 3000rpm to pellet cell debris. Supernatants were further cleared

through a 0.45µm filter and concentrated by ultrafiltration. Briefly, the filtered supernatants were added to Amicon filter tubes (Amicon Ultra-15, PLHK, membrane Ultracel-PL, 100 kD, Millipore # UFC910024) and centrifuged at 4750rpm in two rounds of 45 minutes each, at 4°C. The remaining volumes were aliquoted and kept at −80°C. LVs were subsequently titrated in hIA (10Bl, a SVGmm single-cell clone) using the Lenti-X Provirus quantitation kit (Takara Bio, #631239) according to the manufacturer's instructions.

Generation of the hIA-dCas9 cell line

The lentiviral vector pGH125_dCas9-Blast was purchased from Addgene (#85417). This lentiviral plasmid was packaged as previously described. hIA cells (10Bl, a SVGmm single-cell clone) were transduced overnight with pGH125_dCas9-Blast LV in their culture medium supplemented with 4µg/ml of polybrene (Sigma #H9268). Selection of dCas9-expressing cells was done by treating the transduced cells with 5 ug/mL of blasticidin (Gibco #A1113903) for ten days. hIA-dCas9 were cultured and expanded in the same culture medium as hIA and cryopreserved in FBS with 10% DMSO. hIA-dCas9 were maintained in a 5% CO2 incubator at 37°C and used for experiments for a maximum of four passages.

Identification of conversion TFs

For determining the most suitable conversion factors to generate subtypes of DANs, namely hDA0, hDA1, hDA2, the scRNA-seq expression matrix and cluster annotation files were obtained from NCBI's Gene Expression Omnibus (GEO) data repository using the accession #GSE76381 (la Manno et al., 2016). As a starting cell population, we used the bulk RNA-sequencing data of hIA from the GEO data repository using the accession #GSE93528 (specifically, GSM2454244, GSM2454245, GSM2454246 and GSM2454247) (Rivetti Di Val Cervo et al., 2017). For each cell subpopulation, TransSynW (Ribeiro et al., 2021) identified eight conversion TFs, namely five TFs and three PFs, ordered by increased expression fold change to hIA. The criteria used to select which combination of conversion factors to apply in our protocols was done based on fold change ranking and literature research on the molecular mechanisms each of the TFs is involved.

For the direct reprogramming protocol, the above-mentioned files were uploaded on TransSynW web interface and the target subpopulations hDA0, hDA1, and hDA2 were selected. As a background, we used all the identified cell (sub)populations in this scRNA-seq dataset. For the sequential protocol, we started by analyzing the hierarchical clustering of the scRNA-seq data, also provided by TransSynW (Supplemental Figure 2). To obtain the

conversion factors to generate DANs, we used as background the cell populations located on two hierarchical levels above, namely hGaba, hNbGaba, hNbML5, and hSert. For the target cell population (DANs), we merged the data from the hDA0, hDA1, and hDA2 subpopulations using TransSynW. To obtain the specific TFs to specialize DANs into hDA2, we used as background the hDA0 and hDA1 cell subtypes and run TransSynW selecting the hDA2 subpopulation.

gRNA design

gRNAs were designed using the computational tools CRISPick and CHOP-CHOP (Doench et al., 2014; Labun et al., 2019). gRNAs were designed for CRISPRa, using the human GRCh38 reference genome, considering the protospacer adjacent motif sequence of SpyoCas9 (NGG), the tracrRNA GTTTV (V = not T), and the default target regions of each tool (Chen et al., 2013). The selected gRNA sequences are listed in Supplemental Table 2.

Cloning of RNA scaffold lentiviral expression vectors

Oligo annealing

Complementary oligo sequences containing overhangs for the Esp3I restriction site were obtained for each of the tested gRNAs and annealed using T4 ligase buffer (New England BioLabs #B0202S). The reaction was incubated in a thermocycler for 30 minutes at 37°C, followed by 5 minutes at 95°C and a final ramp rate of 6°C per minute to 25°C.

Golden Gate assembly

The RNA scaffold backbone vector contains the PCP-p65-HSF1 activator complex under the regulation of the elongation factor 1α promoter, followed by the gRNA region. This region comprises the bivalent PP7 RNA hairpin recruiter and a unique Esp3I cloning site for gRNA insertion, driven by the human U6 promoter. Briefly, the RNA scaffold constructs contain a single target-specific gRNA together with a bivalent PP7 RNA hairpin that recruits the RNA-binding protein PCP fused to the transcriptional activators p65 and HSF1. This vector was synthesized by the CRISPR Functional Genomics facility (SciLifeLab), and its DNA sequence can be obtained upon request.

Golden Gate assembly was performed using 25ng/μL of the RNA scaffold backbone vector, 50nM of the annealed oligo and catalyzed by the restriction enzyme Esp3I (New England BioLabs #R0734S) and the T7 ligase (Enzymatics #L6020L). The thermocycling protocol consisted of 15 cycles of 5 minutes at 37°C and 5 minutes at 20°C.

Transfection of gRNA plasmids

hIA-dCas9 were transfected using Lipofectamine 2000 (Invitrogen #11668019). Prior to transfection, lipofectamine was diluted in OptiMEM (Gibco #31985070) and incubated for 20 minutes at room temperature. In the meantime, DNA was also diluted in an equal volume of OptiMEM. After the lipofectamine incubation, this mix was added to each of the DNA solutions and incubated for another 20 minutes at room temperature. Finally, the total volume of lipofectamine-DNA mix was added in small drops to each well. For each 12-well, 600ng of DNA were transfected with 1µL of lipofectamine in a total of 500µL of OptiMEM. After four hours, the transfection medium was removed, and fresh culture medium was added to the hIA-dCas9. Two days after transfection, the cells were harvested for RNA isolation.

PCR amplification

gRNA regions that were able to induce the highest expression of the target TFs were PCR amplified using the Q5 Hot Start High-Fidelity 2x Master Mix (New England BioLabs #M0494S) with the primers below.

gRNA-lenti_F: 5’ GGATTAATTAAGAGGGCCTATTTCCC 3’

gRNA-lenti_R: 5’ CCAAGAATTCAAAAAAAGCACCGA 3’

The activator sequence PCP-p65-HSF1 was amplified from the RNA scaffold backbone vector using the same method with the primers below. These primers were obtained from NEBuilder Assembly Tool.

PCP-Puro_F: 5′ AGATCCTAGAGTCGACCCGGACCATGTCGCGGAGG 3’

PCP-Puro_R: 5’ CGTCGCTTGGTCGGTCATTTCGTTCAGGCACCGG 3’

All PCR products were purified from agarose gel using QIAquick Gel Extraction Purification kit (Qiagen #28706) according to manufacturer’s instructions.

Restriction enzyme assays

All restriction enzymes and purification kits were used according to manufacturer’s instructions. lentiCRISPR v2-dCas9 was purchased from Addgene (#112233) and used as the lentiviral backbone for cloning the PCR amplified regions described above. First, the lentiviral backbone was digested using NheI-HF (New England BioLabs #R3131S) and

BamHI-HF (New England BioLabs #R3136S). The digested product was purified from an agarose gel using QIAquick Gel Extraction Purification kit. The same enzymes were used to digest the purified PCR product containing the amplified activator sequence PCP-p65-HSF1. The resulting product of the enzymatic digestion was purified using QIAquick PCR Purification kit (Qiagen #28104). The purified PCR product containing the activator sequence and the digested lentiviral backbone were ligated using Gibson assembly to form the modified vector (lentiCRISPR-PCP-p65-HSF1).

Purified PCR products containing the selected gRNA regions were digested using PacI (New England BioLabs #R0547S) and EcoRI-HF (New England BioLabs #R3101S) and the resulting products were purified using QIAquick PCR Purification kit. lentiCRISPR-PCP-p65-HSF1 vector was digested using the same enzymes and purified from an agarose gel using QIAquick Gel Extraction Purification kit. Each of the selected gRNA regions was ligated to lentiCRISPR-PCP-p65-HSF1 plasmid, forming the final lentiviral vectors used for LV production.

DNA ligation

DNA fragments were ligated using Quick Ligation kit (New England BioLabs #M2200S) using 0.020pmol of vector DNA and a molar ratio of 1:3 vector/insert. The reaction was incubated at room temperature for 10 minutes and chilled on ice prior to transformation.

Gibson assembly

Gibson assembly was performed to ligate the purified PCR amplified activator sequence (PCP-p65-HSF1) with the NheI and BamHI digested lentiviral backbone using the NEBuilder HiFi DNA Assembly Master Mix (New England BioLabs, #E2621S). NEBuilder Assembly Tool was used to design the overlapping primers required for this method. A molar ratio of 1:2 vector/insert and 0.02pmol of vector DNA were used for this reaction, which was incubated in a thermocycler for 20 minutes at 50°C and chilled on ice prior to transformation.

Design of the inducible cDNA lentiviral expression vectors

FUW-tetO-MCS was purchased from Addgene (#84008) and was used as the lentiviral backbone for constructing the inducible cDNA expression vectors. cDNA sequences were obtained from NCBI's Consensus CDS platform. Each inducible vector contains two target TF coding regions with an incorporated Kozak sequence, followed by a fluorescent reporter

(Figure 3A, Supplemental Table 1). The coding sequences are separated by P2A and a T2A sequences preceding the fluorescent reporter. For cDNAs larger than 3.2 kb, such as BNC2, the inducible plasmids only contained a single cDNA sequence separated of a fluorescent reporter by a P2A sequence. The expression of the cDNAs is under the regulation of a tetracycline-inducible promoter. The assembly of these inducible cDNA lentiviral expression vectors was outsourced to BioCat GmbH.

Direct reprogramming protocol

CRISPRa-based protocol

hIA-dCas9 were transduced overnight with the generated RNA scaffold LVs (MOI 25) in EMEM supplemented with 20% FBS and 4 µg/ml of polybrene (Sigma #H9268). After transduction, hIA-dCas9 were treated with 3ng/mL TGFβ1 (PeproTech #100-21) for 36 hours, followed by a culture medium change to KON2 medium and an overnight treatment with 0.5mM VPA (Sigma #P4543) and 0.25mM Dec (Sigma #189826). KON2 medium is composed of KnockOut DMEM (Gibco #10829018) with 15% KnockOut Serum Replacement (Gibco #10828028), and supplemented with N2 (Gibco, #17502048), 150 uM of ascorbic acid (AA) (Sigma #A4544), and with the following growth factors (GFs): 1ng/mL TGFβ3 (R&D Systems #243-B3), 2mM LM-22A4 (Tocris #4607), 2ng/mL GDNF (R&D Systems #212-GD), 10ng/mL NT3 (R&D Systems #267-N3), and 0.5mM db-cAMP (Sigma # D0627) as previously described (Chung et al., 2010; Nolbrant et al., 2020). Next, hIA-dCas9 cells were treated with dual SMAD inhibitors, namely 2uM of SB431542 (SB) and 0.25uM of LDN193189 (LDN), for 36 hours. Finally, the culture medium was changed and supplemented with midbrain patterning signals, specifically, 0.6uM of CT99021 (CT) and 0.5uM of purmorphamine (Pur) and replaced every 2 days until the end of the protocol.

cDNA-based protocol

hIA were transduced overnight with the generated cDNAs LVs (MOI 5 for BNC2 LV and MOI 15 for remaining LVs) in EMEM supplemented with 20% FBS and 4µg/ml of polybrene (Sigma #H9268). After four days, 2µg/mL of doxycycline (Sigma #D9891) were added to the culture medium to activate the expression of the transgenes. The remaining protocol was followed as described above. In the sequential protocol, the hIA were infected with second set of cDNA LVs two days after transgene activation with doxycycline.

RT–qPCR

Total RNA was extracted using TRIzol reagent (Invitrogen #15596026) according to manufacturer's instructions. RNA concentration and quality were measured by spectrophotometry at UV light and 260:280 ratio, respectively, using Nanodrop 2000 (Thermo Scientific). RNA samples were treated with DNaseI RNase-free (Thermo Scientific #EN0521) according to manufacturer's instructions. Reverse transcription of 500µg of DNAse-treated RNA was performed with High-Capacity RNA-to cDNA kit (Applied Biosystems #4387406) according to manufacturer's instructions. The reverse transcribed cDNA was amplified using Fast SYBR Green Master Mix (Applied Biosystems #4385616) in a StepOnePlus real-time PCR system (Applied Biosystems). Primers used are listed in Supplemental Table 2. Data analysis is based on the ΔΔCt method with normalization of the mRNA expression levels to the housekeeping gene GAPDH. All qPCR reactions were performed in technical triplicates.

Immunocytochemistry

Cells were briefly washed with Dulbecco′s Phosphate Buffered Saline (DPBS, Gibco #14190094) and fixed with 4% paraformaldehyde for 15 minutes at room temperature. Subsequently, the cells were washed with Phosphate Buffered Saline (PBS, Gibco #70011069) two times for 10 minutes, permeabilized and blocked in a solution with 0.1% Triton X-100 and 10% normal donkey serum (Jackson ImmunoResearch #017-000-121) in PBS for 30 minutes at room temperature. Then, cells were incubated overnight with the primary antibodies at 4ºC in a humidified chamber. The primary antibodies were diluted in 10% normal donkey serum in PBS as follows: rabbit anti-TH 1:800 (Pel-Freez #P40101) and mouse anti-TUBB3 1:2000 (Promega #G7121). After incubation with the primary antibody, cells were washed two times for 10 minutes and incubated for one hour at room temperature with Alexa Fluor secondary antibodies, diluted 1:1000 in 10% normal donkey serum in PBS. Cells were again washed as described and the nuclei were stained with 300nM of DAPI for 5 minutes. After a new washing step, cells were kept in a 50/50 solution of PBS-Glycerol. Fluorescence images were acquired under 5x, 10x and 20x magnifications using a AxioScope.A1 fluorescence microscope with integrated camera AxioCam HR (Carl Zeiss) and a Zeiss LSM980-Airy confocal microscope. The attained images were analyzed with the ImageJ software, version 1.53q for Windows.

Author contributions

M.M.R collected the single-cell data, applied the computational tool, performed the experiments, processed the data, and wrote the manuscript. AX performed the experiments. S.O processed the single-cell data. A.d.S conceived the project, obtained funds, and wrote the manuscript. E.A conceived the project, obtained funds, supervised the experimental work, and wrote the manuscript.

Conflicts of interest

None declared.

Acknowledgments

We thank Emilia Síf Ásgrimsdottir, Ka Wai Lee, and Dr. Shanzheng Yang for the insightful discussions about the project.

Funding

M.M.R. is supported by Fonds National de la Recherche Luxembourg (C17/BM/11662681). Work in the EA Lab was supported by Vetenskapsrådet (2020-01426), EU grant Neurostemcell-reconstruct (H2020, 874758), ERC advanced grant (884608), Knut and Alice Wallenberg Foundation (KAW scholar 2018.0232), Karolinska Institutet StratRegen (SFO2018), Cancerfonden (CAN 2016/572), Parkinsonfonden (900/16), and Hjärnfonden (FO2019-0068). M.M.R. was supported by Fonds National de la Recherche Luxembourg (C17/BM/11662681).

References

Akagi, T., Ito, T., Kato, M., Jin, Z., Cheng, Y., Kan, T., Yamamoto, G., Olaru, A., Kawamata, N., Boult, J., Soukiasian, H. J., Miller, C. W., Ogawa, S., Meltzer, S. J., & Koeffler, H. P. (2009). Chromosomal abnormalities and novel disease-related regions in progression from Barrett's esophagus to esophageal adenocarcinoma. International Journal of Cancer, 125(10), 2349–2359. https://doi.org/10.1002/ijc.24620

Barker, R. A., Barrett, J., Mason, S. L., & Björklund, A. (2013). Fetal dopaminergic transplantation trials and the future of neural grafting in Parkinson's disease. In The Lancet Neurology (Vol. 12, Issue 1, pp. 84–91). Lancet Publishing Group. https://doi.org/10.1016/S1474-4422(12)70295-8

Basu, A., & Tiwari, V. K. (2021). Epigenetic reprogramming of cell identity: lessons from development for regenerative medicine. In Clinical Epigenetics (Vol. 13, Issue 1). BioMed Central Ltd. https://doi.org/10.1186/s13148-021-01131-4

Baumann, V., Wiesbeck, M., Breunig, C. T., Braun, J. M., Köferle, A., Ninkovic, J., Götz, M., & Stricker, S. H. (2019). Targeted removal of epigenetic barriers during transcriptional reprogramming. Nature Communications, 10(1), 1–12. https://doi.org/10.1038/s41467-019-10146-8

Becker, J. S., McCarthy, R. L., Sidoli, S., Donahue, G., Kaeding, K. E., He, Z., Lin, S., Garcia, B. A., & Zaret, K. S. (2017). Genomic and Proteomic Resolution of Heterochromatin and Its Restriction of Alternate Fate Genes. Molecular Cell, 68(6), 1023-1037.e15. https://doi.org/10.1016/j.molcel.2017.11.030

Benevolenskaya, E. v., Murray, H. L., Branton, P., Young, R. A., & Kaelin, W. G. (2005). Binding of pRB to the PHD protein RBP2 promotes cellular differentiation. Molecular Cell, 18(6), 623–635. https://doi.org/10.1016/j.molcel.2005.05.012

Berninger, B., Costa, M. R., Koch, U., Schroeder, T., Sutor, B., Grothe, B., & Götz, M. (2007). Functional properties of neurons derived from in vitro reprogrammed postnatal astroglia. Journal of Neuroscience, 27(32), 8654–8664. https://doi.org/10.1523/JNEUROSCI.1615-07.2007

Black, J. B., Adler, A. F., Wang, H. G., D'Ippolito, A. M., Hutchinson, H. A., Reddy, T. E., Pitt, G. S., Leong, K. W., & Gersbach, C. A. (2016). Targeted Epigenetic Remodeling of Endogenous Loci by CRISPR/Cas9-Based Transcriptional Activators Directly Converts Fibroblasts to Neuronal Cells. Cell Stem Cell, 19(3), 406–414. https://doi.org/10.1016/j.stem.2016.07.001

Bonora-Centelles, A., Jover, R., Mirabet, V., Lahoz, A., Carbonell, F., Castell, J. v., & Gómez-Lechón, M. J. (2009). Sequential hepatogenic transdifferentiation of adipose tissue-derived stem cells: Relevance of different extracellular signaling molecules, transcription factors involved, and expression of new key marker genes. Cell Transplantation, 18(12), 1319–1340. https://doi.org/10.3727/096368909X12483162197321

Cesaratto, L., Grisard, E., Coan, M., Zandonà, L., de Mattia, E., Poletto, E., Cecchin, E., Puglisi, F., Canzonieri, V., Mucignat, M. T., Zucchetto, A., Stocco, G., Colombatti, A., Nicoloso, M. S., & Spizzo, R. (2016). BNC2 is a putative tumor suppressor gene in high-grade serous ovarian carcinoma and impacts cell survival after oxidative stress. Cell Death and Disease, 7(9), e2374–e2374. https://doi.org/10.1038/cddis.2016.278

Chahal, H. S., Lin, Y., Ransohoff, K. J., Hinds, D. A., Wu, W., Dai, H. J., Qureshi, A. A., Li, W. Q., Kraft, P., Tang, J. Y., Han, J., & Sarin, K. Y. (2016). Genome-wide association study identifies novel susceptibility loci for cutaneous squamous cell carcinoma. Nature Communications, 7. https://doi.org/10.1038/ncomms12048

Chavez, A., Tuttle, M., Pruitt, B. W., Ewen-Campen, B., Chari, R., Ter-Ovanesyan, D., Haque, S. J., Cecchi, R. J., Kowal, E. J. K., Buchthal, J., Housden, B. E., Perrimon, N., Collins, J. J., & Church,

G. (2016). Comparison of Cas9 activators in multiple species. Nature Methods, 13(7), 563–567. https://doi.org/10.1038/nmeth.3871

Chen, B., Gilbert, L. A., Cimini, B. A., Schnitzbauer, J., Zhang, W., Li, G. W., Park, J., Blackburn, E. H., Weissman, J. S., Qi, L. S., & Huang, B. (2013). Dynamic imaging of genomic loci in living human cells by an optimized CRISPR/Cas system. Cell, 155(7), 1479–1491. https://doi.org/10.1016/j.cell.2013.12.001

Cheng, A. W., Jillette, N., Lee, P., Plaskon, D., Fujiwara, Y., Wang, W., Taghbalout, A., & Wang, H. (2016). Casilio: A versatile CRISPR-Cas9-Pumilio hybrid for gene regulation and genomic labeling. In Cell Research (Vol. 26, Issue 2, pp. 254–257). Nature Publishing Group. https://doi.org/10.1038/cr.2016.3

Chung, T. L., Brena, R. M., Kolle, G., Grimmond, S. M., Berman, B. P., Laird, P. W., Pera, M. F., & Wolvetang, E. J. (2010). Vitamin C promotes widespread yet specific DNA demethylation of the epigenome in human embryonic stem cells. Stem Cells, 28(10), 1848–1855. https://doi.org/10.1002/stem.493

Clow, P. A., Du, M., Jillette, N., Taghbalout, A., Zhu, J. J., & Cheng, A. W. (2022). CRISPR-mediated multiplexed live cell imaging of nonrepetitive genomic loci with one guide RNA per locus. Nature Communications 2022 13:1, 13(1), 1–10. https://doi.org/10.1038/s41467-022-29343-z

DeCaprio, J. A., Ludlow, J. W., Figge, J., Shew, J. Y., Huang, C. M., Lee, W. H., Marsilio, E., Paucha, E., & Livingston, D. M. (1988). SV40 large tumor antigen forms a specific complex with the product of the retinoblastoma susceptibility gene. Cell, 54(2), 275–283. https://doi.org/10.1016/0092-8674(88)90559-4

Delvecchio, M., Gaucher, J., Aguilar-Gurrieri, C., Ortega, E., & Panne, D. (2013). Structure of the p300 catalytic core and implications for chromatin targeting and HAT regulation. Nature Structural and Molecular Biology, 20(9), 1040–1046. https://doi.org/10.1038/nsmb.2642

Doench, J. G., Hartenian, E., Graham, D. B., Tothova, Z., Hegde, M., Smith, I., Sullender, M., Ebert, B. L., Xavier, R. J., & Root, D. E. (2014). Rational design of highly active sgRNAs for CRISPR-Cas9-mediated gene inactivation. Nature Biotechnology, 32(12), 1262–1267. https://doi.org/10.1038/nbt.3026

Doi, D., Magotani, H., Kikuchi, T., Ikeda, M., Hiramatsu, S., Yoshida, K., Amano, N., Nomura, M., Umekage, M., Morizane, A., & Takahashi, J. (2020). Pre-clinical study of induced pluripotent stem cell-derived dopaminergic progenitor cells for Parkinson's disease. Nature Communications 2020 11:1, 11(1), 1–14. https://doi.org/10.1038/s41467-020-17165-w

Dominguez, A. A., Lim, W. A., & Qi, L. S. (2016). Beyond editing: Repurposing CRISPR-Cas9 for precision genome regulation and interrogation. In Nature Reviews Molecular Cell Biology. https://doi.org/10.1038/nrm.2015.2

Fontana, J., Dong, C., Kiattisewee, C., Chavali, V. P., Tickman, B. I., Carothers, J. M., & Zalatan, J. G. (2020). Effective CRISPRa-mediated control of gene expression in bacteria must overcome strict target site requirements. Nature Communications, 11(1), 1–11. https://doi.org/10.1038/s41467-020-15454-y

Gaeta, X., Xie, Y., & Lowry, W. E. (2013). Sequential addition of reprogramming factors improves efficiency. Nature Cell Biology, 15(7), 725–727. https://doi.org/10.1038/ncb2800

Giehrl-Schwab, J., Giesert, F., Rauser, B., Lao, C. L., Hembach, S., Lefort, S., Ibarra, I. L., Koupourtidou, C., Luecken, M. D., Truong, D. J., Fischer-Sternjak, J., Masserdotti, G., Prakash, N., Ninkovic, J., Hölter, S. M., Vogt Weisenhorn, D. M., Theis, F. J., Götz, M., & Wurst, W. (2022). Parkinson's disease motor symptoms rescue by CRISPRa-reprogramming astrocytes into GABAergic neurons. EMBO Molecular Medicine, 14(5). https://doi.org/10.15252/emmm.202114797

Heinrich, C., Gascón, S., Masserdotti, G., Lepier, A., Sanchez, R., Simon-Ebert, T., Schroeder, T., Ötz, M., & Berninger, B. (2011). Generation of subtype-specific neurons from postnatal astroglia of the mouse cerebral cortex. Nature Protocols, 6(2), 214–228. https://doi.org/10.1038/nprot.2010.188

Hook, P. W., McClymont, S. A., Cannon, G. H., Law, W. D., Morton, A. J., Goff, L. A., & McCallion, A. S. (2018). Single-Cell RNA-Seq of Mouse Dopaminergic Neurons Informs Candidate Gene Selection for Sporadic Parkinson Disease. American Journal of Human Genetics, 102(3), 427–446. https://doi.org/10.1016/j.ajhg.2018.02.001

Huangfu, D., Maehr, R., Guo, W., Eijkelenboom, A., Snitow, M., Chen, A. E., & Melton, D. A. (2008). Induction of pluripotent stem cells by defined factors is greatly improved by small-molecule compounds. Nature Biotechnology, 26(7), 795–797. https://doi.org/10.1038/nbt1418

Hung, C. W., Chen, Y. C., Hsieh, W. L., Chiou, S. H., & Kao, C. L. (2010). Ageing and neurodegenerative diseases. In Ageing Research Reviews (Vol. 9, Issue SUPPL., pp. S36–S46). Elsevier. https://doi.org/10.1016/j.arr.2010.08.006

Jensen, K. T., Fløe, L., Petersen, T. S., Huang, J., Xu, F., Bolund, L., Luo, Y., & Lin, L. (2017). Chromatin accessibility and guide sequence secondary structure affect CRISPR-Cas9 gene editing efficiency. FEBS Letters, 591(13), 1892–1901. https://doi.org/10.1002/1873-3468.12707

Jin, T., Rehani, P., Ying, M., Huang, J., Liu, S., Roussos, P., & Wang, D. (2021). scGRNom: a computational pipeline of integrative multi-omics analyses for predicting cell-type disease genes and regulatory networks. Genome Medicine, 13(1). https://doi.org/10.1186/s13073-021-00908-9

Kefalopoulou, Z., Politis, M., Piccini, P., Mencacci, N., Bhatia, K., Jahanshahi, M., Widner, H., Rehncrona, S., Brundin, P., Björklund, A., Lindvall, O., Limousin, P., Quinn, N., & Foltynie, T. (2014). Long-term clinical outcome of fetal cell transplantation for Parkinson disease: two case reports. JAMA Neurology, 71(1), 83–87. https://doi.org/10.1001/JAMANEUROL.2013.4749

Kikuchi, T., Morizane, A., Doi, D., Magotani, H., Onoe, H., Hayashi, T., Mizuma, H., Takara, S., Takahashi, R., Inoue, H., Morita, S., Yamamoto, M., Okita, K., Nakagawa, M., Parmar, M., & Takahashi, J. (2017). Human iPS cell-derived dopaminergic neurons function in a primate Parkinson's disease model. Nature, 548(7669), 592–596. https://doi.org/10.1038/NATURE23664

Kim, T. W., Piao, J., Koo, S. Y., Kriks, S., Chung, S. Y., Betel, D., Socci, N. D., Choi, S. J., Zabierowski, S., Dubose, B. N., Hill, E. J., Mosharov, E. v., Irion, S., Tomishima, M. J., Tabar, V., & Studer, L. (2021). Biphasic Activation of WNT Signaling Facilitates the Derivation of Midbrain Dopamine Neurons from hESCs for Translational Use. Cell Stem Cell, 28(2), 343-355.e5. https://doi.org/10.1016/J.STEM.2021.01.005

Kirkeby, A., Nolbrant, S., Tiklova, K., Heuer, A., Kee, N., Cardoso, T., Ottosson, D. R., Lelos, M. J., Rifes, P., Dunnett, S. B., Grealish, S., Perlmann, T., & Parmar, M. (2017). Predictive Markers Guide Differentiation to Improve Graft Outcome in Clinical Translation of hESC-Based Therapy for Parkinson's Disease. Cell Stem Cell, 20(1), 135–148. https://doi.org/10.1016/J.STEM.2016.09.004

Konermann, S., Brigham, M. D., Trevino, A. E., Joung, J., Abudayyeh, O. O., Barcena, C., Hsu, P. D., Habib, N., Gootenberg, J. S., Nishimasu, H., Nureki, O., & Zhang, F. (2015). Genome-scale transcriptional activation by an engineered CRISPR-Cas9 complex. Nature, 517(7536), 583–588. https://doi.org/10.1038/nature14136

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., Lönnerberg, P., Ryge, J., Barker, R. A., Arenas, E., & Linnarsson, S. (2016). Molecular Diversity of Midbrain Development in Mouse, Human, and Stem Cells. Cell, 167(2), 566-580.e19. https://doi.org/10.1016/j.cell.2016.09.027

Labun, K., Montague, T. G., Krause, M., Torres Cleuren, Y. N., Tjeldnes, H., & Valen, E. (2019). CHOPCHOP v3: Expanding the CRISPR web toolbox beyond genome editing. Nucleic Acids Research, 47(W1), W171–W174. https://doi.org/10.1093/nar/gkz365

Lane, D. P., & Crawford, L. v. (1979). T antigen is bound to a host protein in SY40-transformed cells. In Nature (Vol. 278, Issue 5701, pp. 261–263). Nature Publishing Group. https://doi.org/10.1038/278261a0

Li, W., Englund, E., Widner, H., Mattsson, B., van Westen, D., Lätt, J., Rehncrona, S., Brundin, P., Björklund, A., Lindvall, O., & Li, J. Y. (2016). Extensive graft-derived dopaminergic innervation is maintained 24 years after transplantation in the degenerating parkinsonian brain. Proceedings of the National Academy of Sciences of the United States of America, 113(23), 6544–6549. https://doi.org/10.1073/PNAS.1605245113

Lin, J. Y., & Simmons, D. T. (1991). The ability of large T antigen to complex with p53 is necessary for the increased life span and partial transformation of human cells by simian virus 40. Journal of Virology, 65(12), 6447–6453. https://doi.org/10.1128/jvi.65.12.6447-6453.1991

Lindvall, O., Gustavii, B., Åstedt, B., Lindholm, T., Rehncrona, S., Brundin, P., Widner, H., Björklund, A., Leenders, K. L., Frackowiak, R., Rothwell, J. C., Marsden, C. D., Johnels, B., Steg, G., Freedman, R., Hopper, B. J., Seiger, Å., Strömberg, I., & Olson, M. B. L. (1988). Fetal dopamine-rich mesencephalic grafts in Parkinson's disease. Lancet (London, England), 2(8626–8627), 1483–1484. https://doi.org/10.1016/S0140-6736(88)90950-6

Liu, J., Guan, D., Dong, M., Yang, J., Wei, H., Liang, Q., Song, L., Xu, L., Bai, J., Liu, C., Mao, J., Zhang, Q., Zhou, J., Wu, X., Wang, M., & Cong, Y. S. (2020). UFMylation maintains tumour suppressor p53 stability by antagonizing its ubiquitination. Nature Cell Biology, 22(9), 1056–1063. https://doi.org/10.1038/s41556-020-0559-z

Liu, X., Sun, H., Qi, J., Wang, L., He, S., Liu, J., Feng, C., Chen, C., Li, W., Guo, Y., Qin, D., Pan, G., Chen, J., Pei, D., & Zheng, H. (2013). Sequential introduction of reprogramming factors reveals a time-sensitive requirement for individual factors and a sequential EMT-MET mechanism for optimal reprogramming. Nature Cell Biology, 15(7), 829–838. https://doi.org/10.1038/ncb2765

McGregor, M. M., & Nelson, A. B. (2019). Circuit Mechanisms of Parkinson's Disease. In Neuron (Vol. 101, Issue 6, pp. 1042–1056). Cell Press. https://doi.org/10.1016/j.neuron.2019.03.004

Morris, S. A., Cahan, P., Li, H., Zhao, A. M., San Roman, A. K., Shivdasani, R. A., Collins, J. J., & Daley, G. Q. (2014). Dissecting engineered cell types and enhancing cell fate conversion via Cellnet. Cell, 158(4), 889–902. https://doi.org/10.1016/j.cell.2014.07.021

NCBI entry: BNC2. (2022). NCBI. https://www.ncbi.nlm.nih.gov/gene/54796

Niikura, T., Tajima, H., & Kita, Y. (2006). Neuronal Cell Death in Alzheimers Disease and a Neuroprotective Factor, Humanin. Current Neuropharmacology, 4(2), 139–147. https://doi.org/10.2174/157015906776359577

Nolbrant, S., Giacomoni, J., Hoban, D. B., Bruzelius, A., Birtele, M., Chandler-Militello, D., Pereira, M., Ottosson, D. R., Goldman, S. A., & Parmar, M. (2020). Direct Reprogramming of Human Fetal- and Stem Cell-Derived Glial Progenitor Cells into Midbrain Dopaminergic Neurons. Stem Cell Reports, 15(4), 869–882. https://doi.org/10.1016/j.stemcr.2020.08.013

Pang, K. M., Lynes, M. A., & Knecht, D. A. (1999). Variables controlling the expression level of exogenous genes in Dictyostelium. Plasmid, 41(3), 187–197. https://doi.org/10.1006/plas.1999.1391

Panman, L., Papathanou, M., Laguna, A., Oosterveen, T., Volakakis, N., Acampora, D., Kurtsdotter, I., Yoshitake, T., Kehr, J., Joodmardi, E., Muhr, J., Simeone, A., Ericson, J., & Perlmann, T. (2014). Sox6 and Otx2 Control the Specification of Substantia Nigra and Ventral Tegmental Area Dopamine Neurons. Cell Reports, 8(4), 1018–1025. https://doi.org/10.1016/J.CELREP.2014.07.016

Pennarossa, G., Maffei, S., Campagnol, M., Tarantini, L., Gandolfi, F., & Brevini, T. A. L. (2013). Brief demethylation step allows the conversion of adult human skin fibroblasts into insulin-secreting cells. Proceedings of the National Academy of Sciences of the United States of America, 110(22), 8948–8953. https://doi.org/10.1073/pnas.1220637110

Perez-Pinera, P., Kocak, D. D., Vockley, C. M., Adler, A. F., Kabadi, A. M., Polstein, L. R., Thakore, P. I., Glass, K. A., Ousterout, D. G., Leong, K. W., Guilak, F., Crawford, G. E., Reddy, T. E., & Gersbach, C. A. (2013). RNA-guided gene activation by CRISPR-Cas9-based transcription factors. Nature Methods, 10(10), 973–976. https://doi.org/10.1038/nmeth.2600

Piao, J., Zabierowski, S., Dubose, B. N., Hill, E. J., Navare, M., Claros, N., Rosen, S., Ramnarine, K., Horn, C., Fredrickson, C., Wong, K., Safford, B., Kriks, S., el Maarouf, A., Rutishauser, U., Henchcliffe, C., Wang, Y., Riviere, I., Mann, S., … Tabar, V. (2021). Preclinical Efficacy and Safety of a Human Embryonic Stem Cell-Derived Midbrain Dopamine Progenitor Product, MSK-DA01. Cell Stem Cell, 28(2), 217-229.e7. https://doi.org/10.1016/J.STEM.2021.01.004

Rada-Iglesias, A., Bajpai, R., Swigut, T., Brugmann, S. A., Flynn, R. A., & Wysocka, J. (2011). A unique chromatin signature uncovers early developmental enhancers in humans. Nature, 470(7333), 279–285. https://doi.org/10.1038/nature09692

Ribeiro, M. M., Okawa, S., & del Sol, A. (2021). TransSynW: A single-cell RNA-sequencing based web application to guide cell conversion experiments. Stem Cells Translational Medicine, 10(2), 230–238. https://doi.org/10.1002/sctm.20-0227

Rivetti Di Val Cervo, P., Romanov, R. A., Spigolon, G., Masini, D., Martín-Montañez, E., Toledo, E. M., la Manno, G., Feyder, M., Pifl, C., Ng, Y. H., Sánchez, S. P., Linnarsson, S., Wernig, M., Harkany, T., Fisone, G., & Arenas, E. (2017). Induction of functional dopamine neurons from human astrocytes in vitro and mouse astrocytes in a Parkinson's disease model. Nature Biotechnology, 35(5), 444–452. https://doi.org/10.1038/nbt.3835

Schweitzer, J. S., Song, B., Herrington, T. M., Park, T.-Y., Lee, N., Ko, S., Jeon, J., Cha, Y., Kim, K., Li, Q., Henchcliffe, C., Kaplitt, M., Neff, C., Rapalino, O., Seo, H., Lee, I.-H., Kim, J., Kim, T., Petsko, G. A., … Kim, K.-S. (2020). Personalized iPSC-Derived Dopamine Progenitor Cells for Parkinson's Disease. The New England Journal of Medicine, 382(20), 1926–1932. https://doi.org/10.1056/NEJMOA1915872

Shakirova, K. M., Ovchinnikova, V. Y., & Dashinimaev, E. B. (2020). Cell Reprogramming With CRISPR/Cas9 Based Transcriptional Regulation Systems. In Frontiers in Bioengineering and Biotechnology (Vol. 8, p. 882). Frontiers Media S.A. https://doi.org/10.3389/fbioe.2020.00882

Shrimp, J. H., Grose, C., Widmeyer, S. R. T., Thorpe, A. L., Jadhav, A., & Meier, J. L. (2018). Chemical Control of a CRISPR-Cas9 Acetyltransferase. ACS Chemical Biology, 13(2), 455–460. https://doi.org/10.1021/acschembio.7b00883

Sivandzade, F., & Cucullo, L. (2021). Regenerative stem cell therapy for neurodegenerative diseases: An overview. In International Journal of Molecular Sciences (Vol. 22, Issue 4, pp. 1–21). MDPI AG. https://doi.org/10.3390/ijms22042153

Tanenbaum, M. E., Gilbert, L. A., Qi, L. S., Weissman, J. S., & Vale, R. D. (2014). A protein-tagging system for signal amplification in gene expression and fluorescence imaging. Cell, 159(3), 635–646. https://doi.org/10.1016/j.cell.2014.09.039

Tao, Y., Vermilyea, S. C., Zammit, M., Lu, J., Olsen, M., Metzger, J. M., Yao, L., Chen, Y., Phillips, S., Holden, J. E., Bondarenko, V., Block, W. F., Barnhart, T. E., Schultz-Darken, N., Brunner, K., Simmons, H., Christian, B. T., Emborg, M. E., & Zhang, S. C. (2021). Autologous transplant therapy alleviates motor and depressive behaviors in parkinsonian monkeys. Nature Medicine, 27(4), 632–639. https://doi.org/10.1038/S41591-021-01257-1

Torper, O., Ottosson, D. R., Pereira, M., Lau, S., Cardoso, T., Grealish, S., & Parmar, M. (2015). In Vivo Reprogramming of Striatal NG2 Glia into Functional Neurons that Integrate into Local Host Circuitry. Cell Reports, 12(3), 474–481. https://doi.org/10.1016/j.celrep.2015.06.040

Wang, C., Liu, W., Nie, Y., Qaher, M., Horton, H. E., Yue, F., Asakura, A., & Kuang, S. (2017). Loss of MyoD Promotes Fate Transdifferentiation of Myoblasts Into Brown Adipocytes. EBioMedicine, 16, 212–223. https://doi.org/10.1016/j.ebiom.2017.01.015

Wang, H., Yang, Y., Liu, J., & Qian, L. (2021). Direct cell reprogramming: approaches, mechanisms and progress. Nature Reviews Molecular Cell Biology 2021 22:6, 22(6), 410–424. https://doi.org/10.1038/s41580-021-00335-z

Wu, Y., Zhang, X., Liu, Y., Lu, F., & Chen, X. (2016). Decreased Expression of BNC1 and BNC2 Is Associated with Genetic or Epigenetic Regulation in Hepatocellular Carcinoma. International Journal of Molecular Sciences, 17(2), 153. https://doi.org/10.3390/ijms17020153

Zalatan, J. G., Lee, M. E., Almeida, R., Gilbert, L. A., Whitehead, E. H., la Russa, M., Tsai, J. C., Weissman, J. S., Dueber, J. E., Qi, L. S., & Lim, W. A. (2015). Engineering complex synthetic transcriptional programs with CRISPR RNA scaffolds. Cell, 160(1–2), 339–350. https://doi.org/10.1016/j.cell.2014.11.052

Zhou, H., Liu, J., Zhou, C., Gao, N., Rao, Z., Li, H., Hu, X., Li, C., Yao, X., Shen, X., Sun, Y., Wei, Y., Liu, F., Ying, W., Zhang, J., Tang, C., Zhang, X., Xu, H., Shi, L., … Yang, H. (2018). In vivo simultaneous transcriptional activation of multiple genes in the brain using CRISPR–dCas9-activator transgenic mice. Nature Neuroscience 2018 21:3, 21(3), 440–446. https://doi.org/10.1038/s41593-017-0060-6

Zhou, L., Cheng, X., Connolly, B. A., Dickman, M. J., Hurd, P. J., & Hornby, D. P. (2002). Zebularine: A novel DNA methylation inhibitor that forms a covalent complex with DNA methyltransferases. Journal of Molecular Biology, 321(4), 591–599. https://doi.org/10.1016/S0022-2836(02)00676-9

Tables

Table 1. Selected conversion TFs for direct and sequential reprogramming protocols from human astrocytes into different cell subtypes. TF combinations selected to obtain DANs subtypes were determined using TransSynW computational method (Ribeiro et al., 2021). As control TF combination, we used the validated conversion TFs in Rivetti Di Val Cervo et al., 2017.

Protocol	**Target TFs/PFs**		**Target cell subtype**
Control (Rivetti Di Val Cervo et al., 2017)	*ASCL1, LMX1A, NEUROD1*		DANs
Direct reprogramming (TransSynW)	*EBF2, EN1, FOXA2, MYT1L*		hDA0
	ASCL1, EBF3, FOXA2, NPAS4		hDA1
	ASCL1, BNC2, FOXA2		hDA2
Sequential reprogramming (TransSynW)	*ASCL1, NR4A2, PBX1*	*FOXA2, POU6F1, SOX6*	

Figures

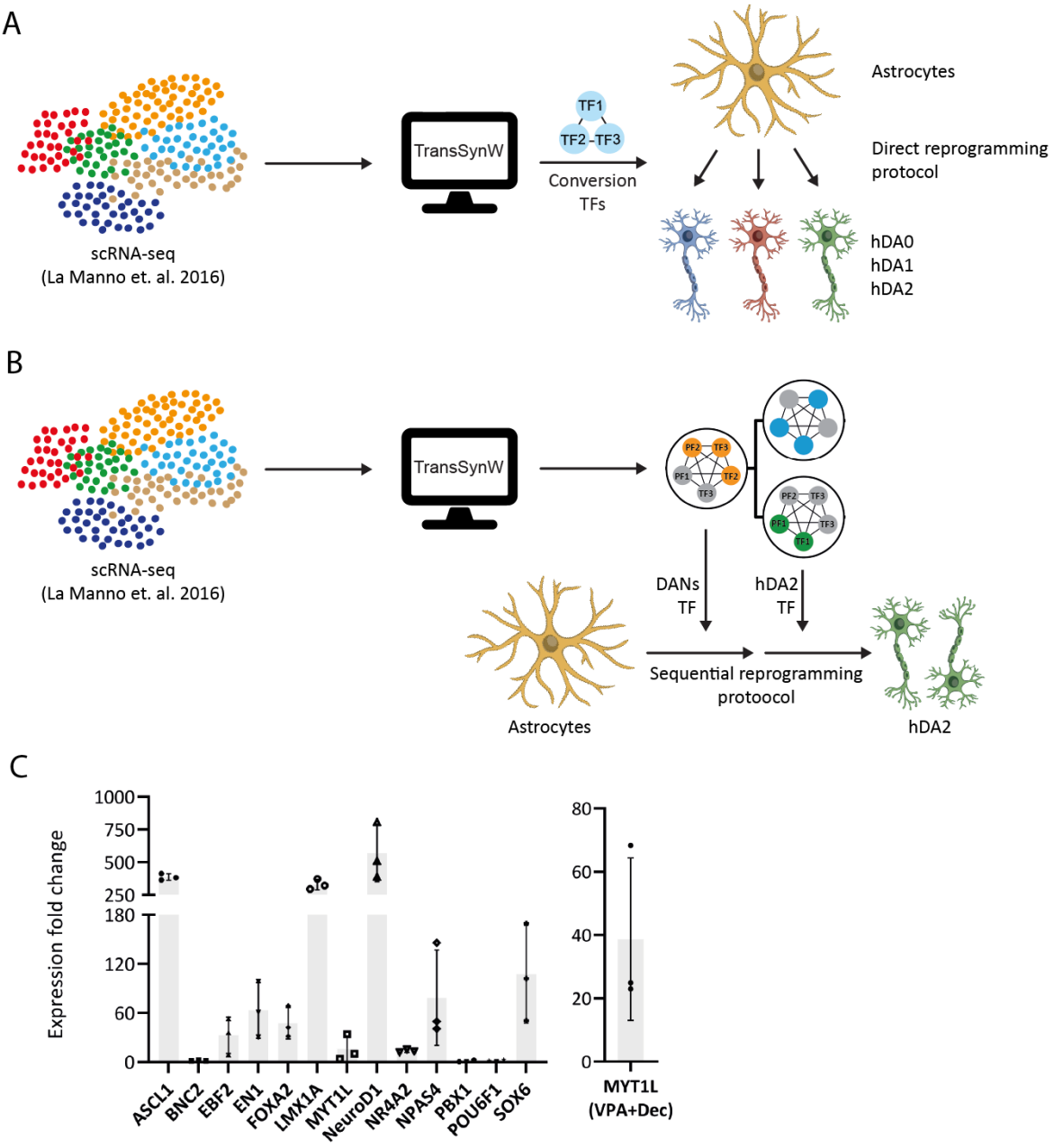


Figure 1. Identification of conversion TFs to directly reprogram hIAs into DAN subtypes and validation of their activation by gRNA expression. (A) TransSynW identification of TFs to generate hIAs into hDA0, hDA1, and hDA2 using a direct reprogramming protocol, using on a previously published scRNA-seq dataset (la Manno et al., 2016). **(B)** TransSynW application to a two-step sequential reprogramming protocol to obtain first TFs that induce the cell conversion of hIAs into DANs, and then the TFs to promote their specialization into hDA2, based on the same scRNA-seq data. **(C)** Activation of the endogenous gene expression of the conversion TFs in hIA-dCas9 by RT-qPCR two days after lentiviral transduction (n=3, biological replicates). On the right, *MYT1L* gene

expression is activated in the presence of 0.5 mM of VPA and 0.25 mM of Dec. Activation of CRISPR-dCas9 mediated gene expression is represented as fold change between hIA-dCas9 transduced with and without gRNAs. Gene expression values are normalized to GAPDH, and error bars depict mean ± standard deviation between biological replicates.

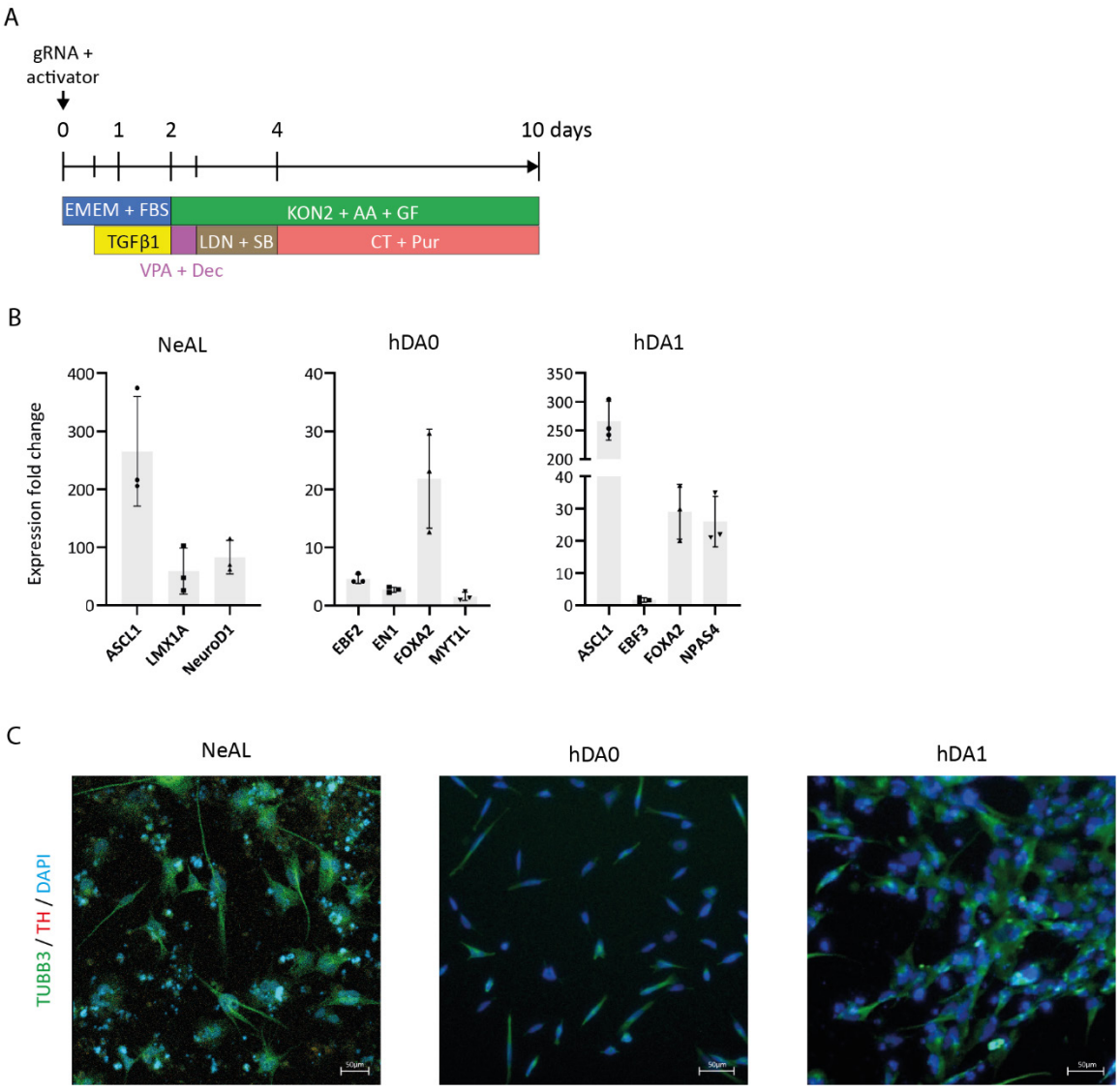


Figure 2. CRISPRa system induces direct reprogramming of hIA-dCas9 into TUBB3-positive cells. (A) Direct reprogramming protocol mediated by CRISPR-dCas9 system adapted from Rivetti Di Val Cervo et al., 2017. **(B)** Multiplexed endogenous gene expression activation of the TF combinations NeAL, hDA0, and hDA1 in hIA-dCas9 by RT-qPCR four days after lentiviral transduction (n=3, biological replicates). CRISPR-dCas9 gene expression activation is represented as fold change between hIA-dCas9 transduced with and without gRNAs. Gene expression values are normalized to GAPDH, and error bars depict

mean ± standard deviation between biological replicates. **(C)** Fluorescence microscope pictures of directly reprogrammed hIA-dCas9 show immunoreactivity for TUBB3, a neuronal marker, ten days after lentiviral transduction. Scale bar, 50µm.

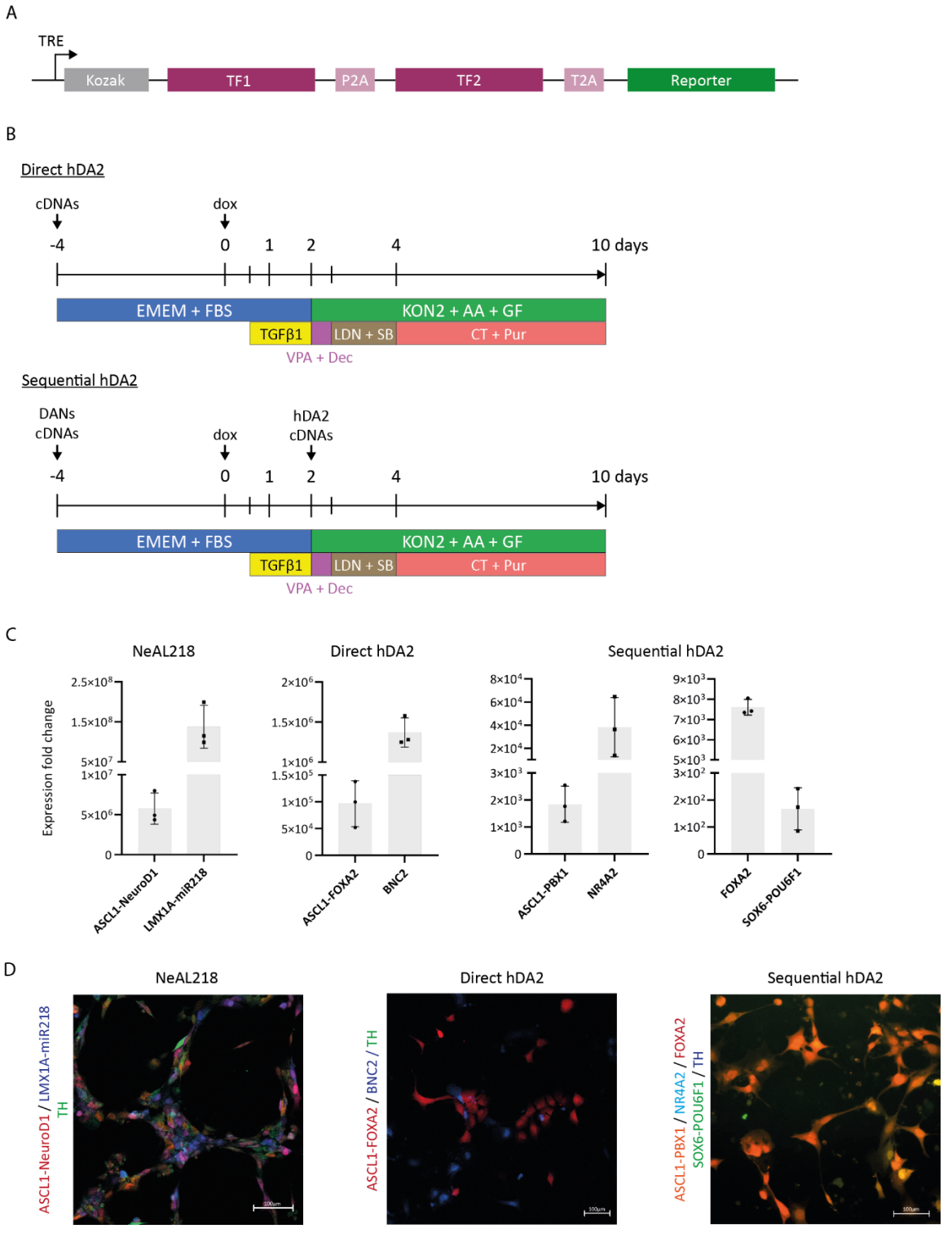


Figure 3. Sequential reprogramming protocol induces direct reprogramming of hIA into neuronal-like cells. **(A)** Schematic representation of the tetracycline-regulated cDNA expression system. The induction of our target TFs cDNAs and reporter genes is regulated by a tetracycline response element (TRE). A Kozak sequence was introduced before the coding regions and each coding sequence will be separated by a self-cleaving 2A peptide. Ectopic gene expression of up to three coding sequences can be promoted using this vector design. **(B)** Direct and two-step sequential reprogramming protocol to convert hIA into hDA2, adapted from Rivetti Di Val Cervo et al., 2017. **(C)** Ectopic gene expression activation of the TF combinations NeAL218, Direct hDA2, and Sequential hDA2 in hIA by RT-qPCR four days after doxycycline (dox) induction (n=3, biological replicates). Gene expression is represented as fold change between hIA transduced with and without cDNAs. Gene expression values are normalized to GAPDH, and error bars depict mean ± standard deviation between biological replicates. **(D)** Confocal pictures of directly reprogrammed hIA show immunoreactivity for TH ten days after dox induction. NeAL218 condition shown expression of TH and the hDA2 sequential protocol generated cells with neuronal morphology. Scale bar, 100µm.

Supplementary Information

Supplemental Tables

Supplemental Table 1. List of cDNA combinations used in the direct and sequential reprogramming protocols to obtain hDA2 subtype. cDNA sequences were cloned in the order presented below. EGFP, mCardinal, tdTomato and TagBFP are fluorescent reporters.

TF combination	**cDNA constructs**			
NeAL218	ASCL1, NeuroD1, TagBFP		LMX1A, miR218, mCardinal	
Direct hDA2	ASCL1, FOXA2, mCardinal		BNC2, TagBFP	
Sequential hDA2	ASCL1, PBX1, tdTomato	NR4A2, mCerulean3	FOXA2, mCardinal	SOX6, POU6F1, EGFP

Supplemental Table 2. List of gRNA oligonucleotide sequences used in the CRISPR-dCas9 mediated direct reprogramming protocols, and respective computational design tools.

Target Gene	Target Sequence	Oligo 1	Oligo 2	gRNA designer
ASCL1	AGCCGCTCGC TGCAGCAGCG	CACCGAGCCGCT CGCTGCAGCAG CG	AAACCGCTGCTG CAGCGAGCGGC TC	CRISPick
BNC2	GCCCGCGCGC GCCCTCGCCG	CACCGGCCCGC GCGCGCCCTCGC CG	AAACCGGCGAG GGCGCGCGCGG GCC	CRISPick
EBF2	AGTCCCGGCA ACGCAGAGCG	CACCGAGTCCCG GCAACGCAGAG CG	AAACCGCTCTGC GTTGCCGGGACT C	CRISPick
EBF3	CAGTCAGTCG GCGAGCGCGG CGG	CACCGCAGTCA GTCGGCGAGCG CGGCGG	AAACCCGCCGC GCTCGCCGACTG ACTGC	CHOP-CHOP
EN1	TGAAGGCAAA AAGTGTGCCT	CACCGTGAAGG CAAAAAGTGTG CCT	AAACAGGCACA CTTTTTGCCTTC AC	CRISPick
FOXA2	GAGGGAGCGC GAGAGAGGGA	CACCGGAGGGA GCGCGAGAGAG GGA	AAACTCCCTCTC TCGCGCTCCCTC C	CRISPick
LMX1A	GAGCACCACG GCCCCGCCAC	CACCGGAGCAC CACGGCCCCGCC AC	AAACGTGGCGG GGCCGTGGTGCT CC	CRISPick
MYT1L	AAAGCCTAGG AGAGGATGAG	CACCGAAAGCC TAGGAGAGGAT GAG	AAACCTCATCCT CTCCTAGGCTTT C	CRISPick
NeuroD1	AGAACGGGGA GCGCACAGCC	CACCGAGAACG GGGAGCGCACA GCC	AAACGGCTGTG CGCTCCCCGTTC TC	CRISPick
NPAS4	GAGCCCCCCT CCCCAGTCAG	CACCGGAGCCC CCCTCCCCAGTC AG	AAACCTGACTG GGGAGGGGGGC TCC	CRISPick
NR4A2	GTCCAGGGAG CGCGGCAGCG	CACCGGTCCAG GGAGCGCGGCA GCG	AAACCGCTGCC GCGCTCCCTGGA CC	CRISPick
PBX1	GGGGCAAAGG GAAGGGGAGG	CACCGGGGGCA AAGGGAAGGGG AGG	AAACCCTCCCCT TCCCTTTGCCCC C	CRISPick
POU6F1	GGAGCCAGGA GCGAGGGGTG	CACCGGGAGCC AGGAGCGAGGG GTG	AAACCACCCCTC GCTCCTGGCTCC C	CRISPick

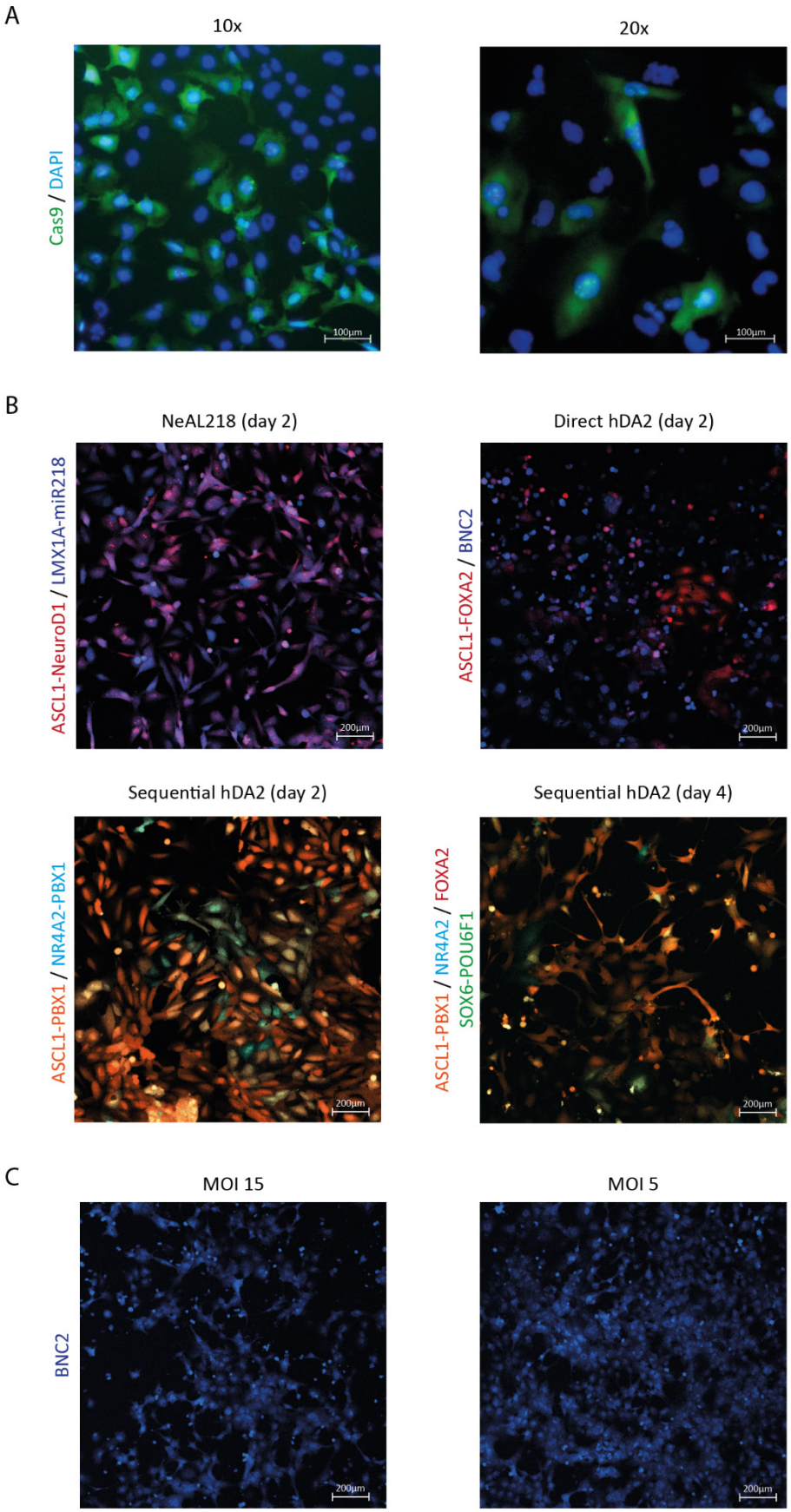
SOX6	TCATCTTGCCTGTGTTGTCT	CACCGTCATCTTGCCTGTGTTGTCT	AAACAGACAACACAGGCAAGATGAC	CRISPick

Supplemental Table 3. List of human primers used for RT-qPCR.

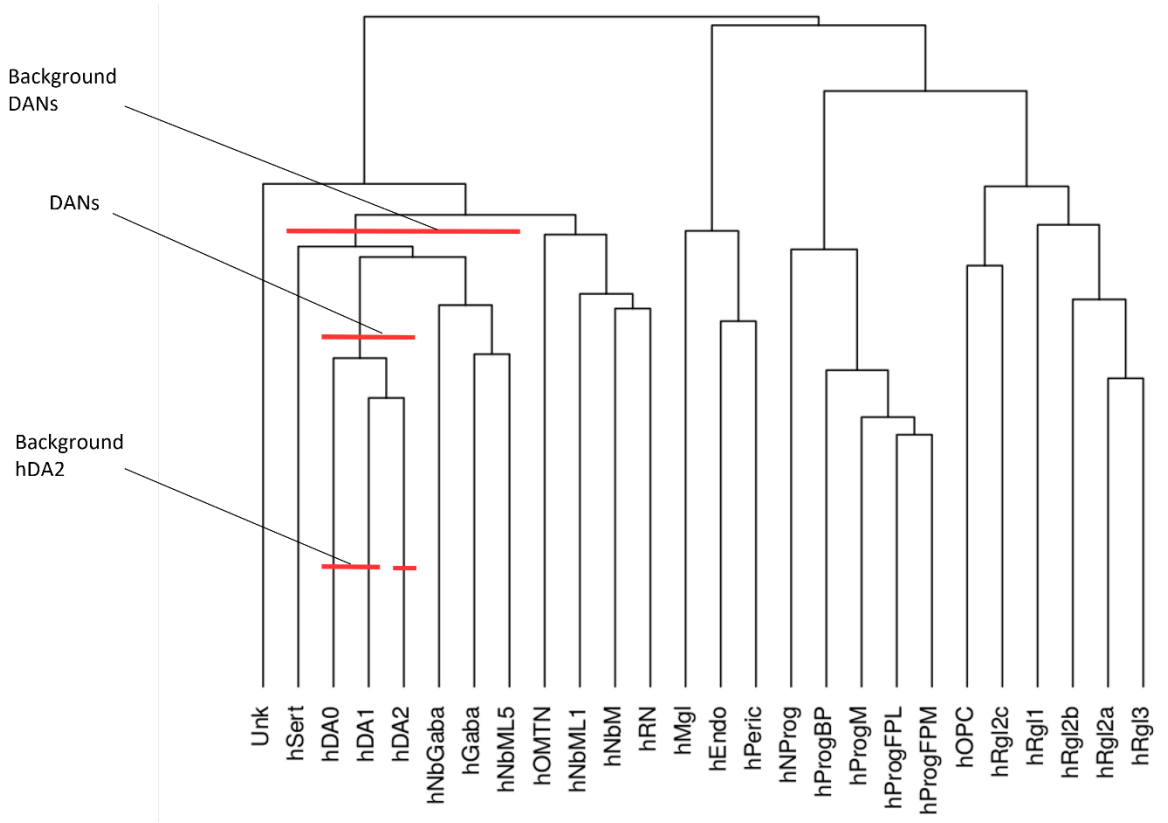
Target Gene	Forward Primer	Reverse Primer
ASCL1	GCAGGAGCTTCTCGACTTCACC	AAAGATGCAGGTTGTGCGATCA
BNC2	TCAGCAGCCTGATGCTCTATGG	GCCGTGCAGTATGTGTGCAGTACC
EBF2	TAGGAAGAGGACCAACTCTGAAA	CGACATTAGCGTCCACCACTC
EBF3	CCGCTAACTCTCCCTACGGC	AAAATGCCGTGTGTGCTGCT
EN1	TCGCAGCAGCCTCTCGTATG	CCTGGAACTCCGCCTTGAGT
FOXA2	CCGTTCTCCATCAACAACCT	GGGGTAGTGCATCACCTGTT
GAPDH	GAAGGTGAAGGTCGGAGTCA	GACAAGCTTCCCGTTCTCAG
LMX1A	GATCCCTTCCGACAGGGTCTC	GGTTTCCCACTCTGGACTGC
MYT1L	CTCGGCAAAATCGCTGAGGAT	TCCAGACTATTGGAGGTATTGCT
NEUROD1	AGCCCCAAGGTCCTCCAA	CGTGCTCCTCGTCCTGAGA
NPAS4	CAGATCAACGCCGAGATCCG	GACGCCCTTGCGAGTGTAGAT
NR4A2	ACCACTCTTCGGGAGAATACA	GGCATTTGGTACAAGCAAGGT
PBX1	TAAAAAGCCTTGGTGCTTCCCA	GCTCGTCCATCTCCAAAGGCTA
POU6F1	GCCTACAGCCAGTCAGCCATCT	GTTCCGCAGTTCAGCTTCGTTT
SOX6	TACCTCTACCTCACCACATAAGC	ACATCGGCAAGACTCCCTTTG
ASCL1-NEUROD1-BFP	GTCTCACGGCAGCATCTTCT	CGGACATCACTCTCTCGTGG
BNC2-BFP	TTCGGCGAGACAAAGAGCAT	TCGCTGTCGGTCTTAGAGGA
LMX1A-miR218-mCardinal	CCACCTGCACTGTAACCTGA	TCTGTCCACGAAGTACACGC

FOXA2-mCardinal	GAAGCTGAGAGGCGTGAACT	GTACAGTGTCTCGGTGGTGG
EN1-tdTomato	GGAACAAAGTACCCCGAGCA	TGCTGGGAATCGGTCTTCAC
MYT1L-EGFP	AACTACGATGAGCTGGTGGC	GGAGTTCATCTCGCTCTCGG
FOXA2-EBF2-mCerulean	ATGTGTAGGGTGCTGCTGAC	TCGGAGGGGGTCTCATTTCT
ASCL1-EBF3-mCardinal	GTACTCCAACGGCGTGAGAA	CGGGATTCTTATCCTGGCCC
FOXA2-NPAS4-EGFP	TCTTCTTCAAGGACGACGGC	CAGCTCGATCCTGTTCACCA
ASCL1-PBX1-tdTomato	ATGTGAGGCCGTGATGATCC	CTCGTTCAGGATCTCGGTGG
ASCL1-FOXA2-mCardinal	GACGGCTGCCTGATCTACAA	TCTTCTTCTGCATCACGGGG
NR4A2-mCerulean	GCTGATCTTCTGCAATGGCG	GGAGGAGAACTCCACGATGC
SOX6-POU6F1-EGFP	GACGCCATGACACAGGATCT	GCTTGTTGTGCATGATGGGG

Supplemental Figures



Supplemental Figure 1. Efficiency of lentiviral transduction. (A) Fluorescence microscope pictures show immunoreactivity for Cas9 in hIA-dCas9 cell line. Scale bar, 100µm. **(B)** Confocal pictures show lentiviral transduction efficiency of ectopic expression of conversion TFs two and four days after doxycycline induction and **(C)** difference in surviving cells two days after lentiviral transduction of BNC2 expression vector with different MOIs. Scale bar, 200µm.



Supplemental Figure 2. Hierarchical clustering of scRNA-seq data used for identifying conversion TFs for the two-step sequential reprogramming protocol. Hierarchical clustering analysis was performed based on Pearson correlation coefficients for each of the cell subtypes in the scRNA-seq dataset (la Manno et al., 2016). Red lines show the different branches used to perform the identification of conversion TFs using TransSynW.

4.4. Neural network learning defines glioblastoma features to be of neural crest perivascular or radial glia lineages

4.4.1. Preface

GBM is one of the most aggressive types of brain tumor and patients are often given a poor prognosis due to the lack of effective treatments. High tumor heterogeneity is one of the major obstacles to the identification of potential therapeutical targets. Characterizing the molecular profile of the cellular populations present in GBM samples at the single-cell level can elucidate important mechanisms underlying the development of this type of cancer. The Cancer Genome Atlas (TCGA) was able to classify GBM tumors into three main subtypes based on their molecular profiles. The mesenchymal subtype is highly resistant to currently available therapies and patients with this subtype have lower life expectancy. It has been shown that the cellular lineage of origin of GBM tumors can determine important molecular characteristics. However, the molecular mechanisms behind the development of the mesenchymal subtype remain elusive. Therefore, identifying the cellular lineage from which this GBM subtype derives can help stratify GBM in different subtypes and significantly improve disease prognosis.

To address this issue, we performed scRNA-seq of patient-derived GBMs to characterize their cellular lineages of origin. We found the perivascular (PeriV) lineage to be specifically present in GBM and the molecular features of the mesenchymal subtype to be very similar to the ones of this cell lineage. We validated *PROX1* and *FOXC1* as a specifically expressed TFs in tumors derived from radial glia (Rgl) and PeriV lineage, respectively, in patient-derived xenografts. These findings provide potential biomarkers to more accurately predicted GBM patients' prognosis.

In this study, I performed the data analysis for identifying the specifically expressed TFs in the PeriV and Rgl cellular lineages, which results can be found on Figure 4 of the published article. This article is reprinted on the next pages (AAAS Author License to Publish (standard), section III.A.2).

CELL BIOLOGY

Neural network learning defines glioblastoma features to be of neural crest perivascular or radial glia lineages

Yizhou Hu[1]†, Yiwen Jiang[1]†, Jinan Behnan[1], Mariana Messias Ribeiro[2], Chrysoula Kalantzi[1], Ming-Dong Zhang[1], Daohua Lou[1], Martin Häring[1], Nilesh Sharma[1], Satoshi Okawa[2], Antonio Del Sol[2,3,4], Igor Adameyko[5,6], Mikael Svensson[7,8], Oscar Persson[7,8], Patrik Ernfors[1]*

Glioblastoma is believed to originate from nervous system cells; however, a putative origin from vessel-associated progenitor cells has not been considered. We deeply single-cell RNA–sequenced glioblastoma progenitor cells of 18 patients and integrated 710 bulk tumors and 73,495 glioma single cells of 100 patients to determine the relation of glioblastoma cells to normal brain cell types. A novel neural network–based projection of the developmental trajectory of normal brain cells uncovered two principal cell-lineage features of glioblastoma, neural crest perivascular and radial glia, carrying defining methylation patterns and survival differences. Consistently, introducing tumorigenic alterations in naïve human brain perivascular cells resulted in brain tumors. Thus, our results suggest that glioblastoma can arise from the brains' vasculature, and patients with such glioblastoma have a significantly poorer outcome.

INTRODUCTION

Glioblastoma is the most common brain tumor (*1*), and it has an invariably poor prognosis despite aggressive therapy. A combination of high-throughput genomic and epigenetic data with bioinformatic analyses has provided a comprehensive view of genetic mechanisms underlying glioblastoma oncogenesis and progression (*2*, *3*). Analyzing transcriptional intertumor heterogeneity within The Cancer Genome Atlas (TCGA) project identified three main subtypes, which are tightly associated with genomic alterations: TCGA-classical, TCGA-proneural, and TCGA-mesenchymal (TCGA-mes) (*4*). However, there is also notable intratumoral heterogeneity where different cells from the same tumor can be classified into different TCGA subtypes (*5*).

Gliomas are believed to arise from one of the two major types of neural cells of the brain: neuronal or glial by a reactivation of stem-like developmental gene programs. This cancer stem cell (CSC) hypothesis implicates a hierarchical continuum of differentiating cells within the tumor, with the CSC at the apex, having tumor-initiating and -propagating properties with resistance to therapy (*6*). Single-cell RNA sequencing (scRNA-seq) studies support this conjecture, and transcriptional profiles of various types of gliomas are consistent with neural progenitor–like, oligodendrocyte precursor (OPC)–like, or astrocytic-like cells (*5*, *7*–*10*). Introducing identical glioblastoma driver mutations into human glial or neuronal progenitor cells results in molecular distinct subtypes, highlighting the importance of the originating cell lineage for tumor phenotype and stratification (*11*, *12*). However, less is known of the cellular origin of the highly malignant glioblastoma with mesenchymal features (*5*, *13*).

Thus, previous computational cell-of-origin classifications mapped most glioblastoma to neuronal and glial cell types (*5*, *9*, *10*) and additional studies have identified possible mechanisms for these to transition into mesenchymal-like glioblastoma. However, the relation of mesenchymal glioblastoma to alternative nonneural progenitor cells residing in the brain has not been explored. Perivascular mural cells of the brains' blood vessels are of neural crest origin (*14*, *15*). As blood vessels descend into the brain parenchyma during development, vessel-attached neural crest–derived cells differentiate into the different perivascular cell types, with those remaining behind differentiating into leptomeningeal cells (*14*, *16*). Recently, a previously unknown perivascular fibroblast (vFB)–like cell type was identified (*17*), which appears to function as a restricted stem-like cell type that generates pericytes and mesenchymal smooth muscle cells (SMCs) in both the developing and adult brain (*18*, *19*).

Here, we deeply sequenced 4073 glioblastoma progenitor cells from 18 patients and integrated data from an additional 8443 tumor cells from 16 patients with low-grade glioma and 60,979 tumor cells from 66 patients with glioblastoma in the analysis. A novel neural network–based projection was used to learn the transcriptional features from normal brain cell types and thereafter used to assign individual tumor cells as well as deconvoluted bulk tumors at the level of both the cellular steady state and the developmental trajectory dynamics. Our analysis revealed two principal cell lineage patterns in glioblastoma—neural and perivascular. The most undifferentiated adult naïve cell type correlate in the neural cell lineage pattern was radial glia (Rgl), and in the vascular, it is the vFB cell type. Patients with perivascular glioblastoma exhibited significantly poorer survival. Animals with xenografts of naïve human perivascular cells harboring targeted genetic changes observed in glioblastoma present with tumors, indicating that the brain perivascular cells are competent to initiate brain tumors.

RESULTS

Neural network classifier maps glioblastoma tumor progenitor cells to two principally different endogenous cell lineages of the brain

We enriched tumor progenitor cells from 18 patients of high-grade glioblastoma for scRNA-seq (data file S1) and validated the

[1]Division of Molecular Neurobiology, Department of Medical Biochemistry and Biophysics, Karolinska Institutet, Stockholm, Sweden. [2]Computational Biology Group, Luxembourg Centre for Systems Biomedicine (LCSB), University of Luxembourg, 4362 Esch-sur-Alzette, Luxembourg. [3]CIC bioGUNE, Bizkaia Technology Park, 48160 Derio, Spain. [4]IKERBASQUE, Basque Foundation for Science, 48013 Bilbao, Spain. [5]Department of Molecular Neurosciences, Center for Brain Research, Medical University Vienna, Vienna, Austria. [6]Department of Physiology and Pharmacology, Karolinska Institutet, Stockholm, Sweden. [7]Department of Clinical Neuroscience, Karolinska Institutet, Stockholm, Sweden. [8]Department of Neurosurgery, Karolinska University Hospital, Stockholm, Sweden.
*Corresponding author. Email: patrik.ernfors@ki.se
†These authors contributed equally to this work.

tumorigenicity of these cells by intracranial orthotopic xenografts with follow-up histological analyses (fig. S1A). Fourteen of the 18 patient samples reduced overall survival in the xenograft experiment (fig. S1B). A total of 4073 high-quality single cells (median 2.87 million total reads per cell; fig. S1C) were included in a copy number variation (CNV) analysis, confirming alterations associated with brain tumors (data file S1) and subsequently clustered. Excluding a cluster of $CD45^+$ immune cells, the remaining 19 clusters were assigned into TCGA subtypes by a neural network classifier trained by the original TCGA data and subsequently named after TCGA subtype names (MS1-8, CL1-8, PN1-2, and NL1) (fig. S1D). Most clusters dominantly differed among individual patients, except for two cell clusters of TCGA-mes subtypes (MS3 and MS5) that spanned across different patients (fig. S1E, left). The cell clusters were organized into two clouds of coclustered cells when using Uniform Manifold Approximation and Projection (UMAP). Cells of the TCGA-mes subtype were in one cloud, while cells of all other TCGA subtypes were located in another cloud (Fig. 1A and fig. S1E, right).

To identify the endogenous brain cell-type correlates of the patients' glioma cells, we applied the machine learning classifiers with learned transcriptional features from normal brain reference cell types derived from the neurogenic niche of the developing mouse brain (*20*). After comparing four classifiers driven by logistic regression, support vector machine, vanilla neural network, and node-level graph neural networks, we decided to use a vanilla neural network classifier for further studies according to the prediction accuracy, time consumption, and overfitting control, as described in Materials and Methods. The classifier accuracy was further validated by an independent integrated dataset of normal cells from human embryonic midbrain (*21*) and cortex (fig. S1F) (*22*), and a randomized expression matrix (fig. S1G). Throughout the study, we refer to previously annotated cell types as "reference" cell types, and such closely related reference cell types were annotated in this study into cell lineages on the basis of the known differentiation trajectories. Using this neural network classifier, most tumor cells of the TCGA-mes subtype were assigned to the reference pericytes and vascular leptomeningeal cells (VLMCs), both of the perivascular lineage, while tumor cells of other TCGA subtypes were similar to reference neuronal or glial cells (i.e., reference Rgl, neuroblasts, astrocytes, oligodendrocyte cells, and immature granule neurons) (Fig. 1B and fig. S1H). Cells that failed to assign into one single cell type were located in the center of the radar plot, indicating cells of unknown cell type or a transcriptional plasticity of multiple cell types.

The neural crest–derived perivascular cells (reference pericytes and VLMCs) of the brain and the reference radial glia–derived neural cells (all neuronal and glial cell types of the brain) represent entirely different developmental cell lineages. When stratifying patients into either an Rgl-lineage type or a perivascular (PeriV)–lineage type based on the dominant cell percentage of one type and nonsignificant cell percentage of the other type in each patient, we did not observe significant differences of overall survival in the xenograft experiment (fig. S1, A and B). To further increase the resolution of reference brain cell types, we applied the machine learning classifier with learned transcriptional features from human developing brain cell types (*23*) and validated the observation of the existence of both Rgl-lineage–type and PeriV-lineage–type glioblastoma cells (fig. S1I). Thus, these results suggest that glioblastoma cells share molecular features with either the Rgl-lineage [including Rgl-like tumor progenitor cells; a neuronal sublineage including neuroblasts and neurons; an oligodendrocyte-sublineage (Olig-sublineage) including oligodendrocytes and its precursors, the OPCs and newly formed oligodendrocytes (NFOL); and an astrocyte-sublineage including differentiating and adult astrocytes] or the PeriV-lineage including perivascular cells and VLMCs.

Analysis of the differentially expressed genes between Rgl-lineage– and PeriV-lineage–type glioblastoma cells that were also expressed in their respective naïve cell types (i.e., normal reference brain Rgl and PeriV cells) revealed the existence of mutually exclusive expression between lineages but highly shared features with their corresponding endogenous reference cell types of each lineage in glioblastoma cells (Fig. 1C) and in the naïve cell types of the developing mouse brain (fig. S1J).

Perivascular lineage–type tumors exclusive to high-grade glioma

The previously analyzed cells were from high-grade glioma. We therefore made use of scRNA-sequenced cells obtained from resected and dissociated high- and low-grade gliomas (*5*, *7*, *9*, *10*, *24*–*26*) to validate our results and to compare the cell-type composition of PeriV- and Rgl-lineage tumor cells between high- and low-grade gliomas. A total of 8443 cells from low-grade glioma and 65,052 cells from glioblastoma originally defined as tumor cells were applied for the neural network classifier described in Fig. 1B. We found that low-grade glioma contains tumor cells with higher cell-type similarity to native reference cell types (high cell-type probability) than high-grade glioblastoma (Fig. 1, D and E, left, and fig. S1K, left). To exclude the fact that this result is caused by variability of sequencing quality between platforms of scRNA-seq, and to exclude a bias due to required threshold in the similarity scoring, we also validated this observation using only data generated from the same technical platform and applied different threshold requirements (fig. S1K, right). Low-grade glioma cells were most similar to reference Rgl, OPCs/NFOLs, and astrocytes, which together accounted for 99.48% of all tumor cells (Fig. 1D, right). In contrast, almost all glioblastomas were composed of multiple cell types, including high similarity to reference pericytes/VLMCs, to Rgl (i.e., Rgl-like tumor cells), as well as substantial numbers to the more differentiated progenies (astrocytes of the Astro-sublineage; OPC, NFOLs, and oligodendrocytes of the Olig-sublineage; neuroblasts and immature granule cells "Granule" of the Neuronal-sublineage) (Fig. 1E, right). Among the glioblastoma cells, 11.1% were assigned to the reference PeriV-lineage, while none of the low-grade glioma cells were assigned to these (Fig. 1F and fig. S1L). Thus, the existence of glioma assigning to the PeriV-lineage reference cells is specific to high-grade glioma among all 100 patients.

Rgl-lineage glioblastoma cells acquire higher cellular plasticity after mesenchymal transition but rarely transition into PeriV-lineage cell types

The acquisition of a mesenchymal transcriptional profile in glioblastoma cells can be forced by the microenvironment or by an intrinsic transition under certain selective pressure (*13*). To examine whether the PeriV-lineage tumor cells can transition from Rgl-lineage glioblastoma cells, we applied the neural network classifier on a recent published scRNA-seq dataset containing spontaneous mouse glioblastoma that was initiated from glial fibrillary acidic protein (*GFAP*)–expressing cells (*27*). In this model, a mesenchymal cell

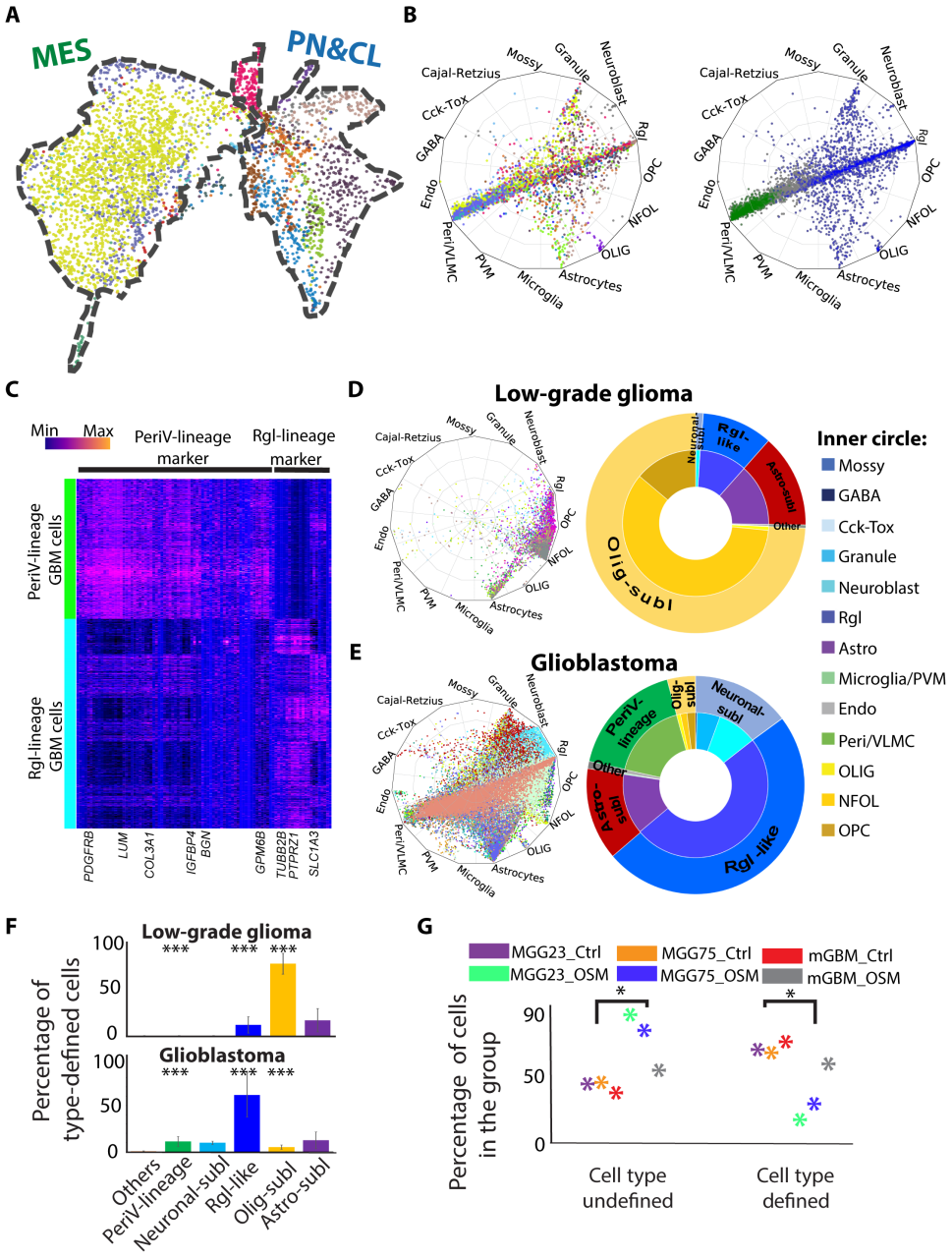


Fig. 1. Cell-type assignment of high- and low-grade glioma revealed that perivascular lineage tumor cells are present only in high-grade glioma. (**A**) UMAP visualization of patient-derived glioblastoma cells. Color coding based on cell clusters. The contours of two main clouds of cells outlined with a dashed line and labeled with TCGA subtypes on the top. CL, classical; MES, mesenchymal; PN, proneural. (**B**) Radar plot visualization of the cell-type scores of glioblastoma cells in relation to the trained reference brain cell types. Color coding based on cell clusters (left) or cell-type lineages (right, blue: Rgl-lineage; green, PeriV-lineage). The position of each dot indicates the cell-type score between that cell and the trained reference cell types, which are indicated outside each wheel bend. Abbreviations are as in fig. S1F. (**C**) Heatmap of differential gene expression between PeriV-lineage and Rgl-lineage glioblastoma cells. Selected gene symbols are at the bottom. Color bar indicates the expression intensity at the top left. (**D** and **E**) Left: Radar plots show the cell-type scores of low-grade glioma and glioblastoma cells in relation to the trained reference brain cell types. Right: Donut charts show the quantitative distribution of cell type–defined glioblastoma cells. The inner donut layer represents the reference cell types that tumor cells are assigned to, and the outer layer represents the normal cell-type lineages. (**F**) The distribution of low-grade glioma and glioblastoma cells to defined reference cell-type lineages. ***$P < 0.001$. (**G**) Scatter chart represents the significant cell-type score of control (Ctrl) and oncostatin M (OSM)–treated glioblastoma multiforme (GBM) cells against each defined reference brain cell type. "Cell type defined" represents glioblastoma cells with high cell-type scores above the cutoff, and "cell type undefined" represents cells with low scores. Dot colors are indicated at the top. *$P < 0.05$.

transition from the *GFAP*$^+$ Rgl-lineage could be induced by oncostatin M (OSM) (*27*). Thus, if the identified PeriV-lineage glioblastoma represents a transition from the Rgl-lineage through this mechanism, we expected to identify PeriV-lineage glioblastoma cells in this dataset. Nearly all *GFAP*-derived glioblastoma cells were assigned to reference Rgl-lineage cells (Rgl, neuroblasts, and granule cells), but none to pericytes/VLMCs (fig. S1M). Because OSM induced a mesenchymal transition of these glioblastoma cells (*27*, *28*), we compared three glioblastoma cell lines with or without OSM treatment in our classifier of TCGA subtypes and observed that OSM significantly increased mesenchymal features and inhibited proneural features (fig. S1N), in line with previous findings. Nevertheless, our classifier of endogenous reference brain cells did not recognize the OSM-transformed mesenchymal cells as PeriV cells, and instead assigned these mesenchymal cells into an undefined state (Fig. 1G and fig. S1O). These results corroborate that OSM initiates plasticity of glioblastoma cells including initiation of mesenchymal features and that this mechanism could account for some glioblastoma classified as mesenchymal. However, our results suggest that glioblastoma with perivascular features as defined using our classifier cannot be explained by an OSM-driven cell state transition.

Clinical relevance and CpG methylation of PeriV-lineage and Rgl-lineage glioblastoma

To explore the clinical relevance of tumors with PeriV-lineage and Rgl-lineage signatures, we scored the data of 161 bulk RNA-sequenced glioblastoma from the TCGA using the classifier. However, the bulk data reflect transcriptional features of multiple cell types (fig. S2A) that are highly heterogeneous, consistent with previous results (*5*). To identify the dominant cell types, the bulk data were transformed (*29*) and deconvoluted into single-cell resolution (fig. S2B) (*30*), and the deconvoluted data were then scored and visualized in a radar plot (Fig. 2A). The majority of the TCGA classified glioblastoma subtypes (TCGA-mes, TCGA-proneural, TCGA-classical, and TCGA-neural) were robustly assigned into four endogenous reference brain cell types: 19 tumors were assigned to reference cells of the PeriV-lineage (perivascular cells and VLMCs) and the remaining tumors were assigned to Rgl-lineage reference cells, including 53 to astrocytes, 32 to Rgl, and 9 to OPCs/NFOLs, accounting for 70.19% of all tumors. The lack of assignment of tumors to reference granule and neuroblast cells in bulk sequenced data likely reflects that these differentiated cells are rare in the tumors and might therefore become dwarfed when bulk-sequenced. In line with previous results obtained from scRNA-seq data, 9 of 10 top scRNA-seq enriched marker genes of PeriV-lineage–type and Rgl-lineage–type reference cells (data file S2) were found to be differentially expressed between PeriV-lineage–type and Rgl-lineage–type glioblastoma tumors sequenced in the TCGA framework (Fig. 2B). We next examined the relation between PeriV- and Rgl-lineage tumor types to TCGA subtypes by cross annotation. PeriV-lineage glioblastoma was overwhelmingly composed from the TCGA-mes subtype (Fig. 2C, top). In contrast, only 44.4% of TCGA-mes subtypes were of the PeriV-lineage, while the rest were most similar to the reference Rgl-lineage (including Rgl-like cells and cells in sublineages of Rgl) (Fig. 2C, bottom), indicating that the TCGA-mes subtype might consist of two different transcriptional states, one but not the other showing high similarity to the reference PeriV cells. The TCGA-classified proneural and glioma cytosine-phosphate-guanine (CpG) island methylator phenotype (G-CIMP) subtype mostly shared features with reference Rgl, while TCGA-classical and TCGA-neural subtypes mostly shared features with reference astrocyte cells (fig. S2C). To exclude that this finding was a result of a distortion due to analysis of bulk RNA-sequenced data, we classified the merged set of all scRNA-seq high-grade glioblastoma cells into TCGA subtypes and thereafter cross-annotated the cells of the TCGA-mes subtype to native reference brain cell types (Fig. 2D and fig. S2D). This analysis confirmed that glioblastoma cells of the TCGA-mes subtype are mainly assigned to PeriV cells, with most of the remaining cells showing the greatest similarity to reference Rgl and astrocytes of the brain. Furthermore, we re-examined glioblastoma cells from a public dataset (*7*) in our classifier of endogenous brain cells. In this study, tumor cells were assigned as "glial progenitor cancer cell," "oligo-lineage cancer cell," "astrocytic cancer cell," "mesenchymal cancer cell," and "neuronal cancer cell" on the basis of the similarity to developing brain cell types (*7*). Our classifier confirmed these previous results (Fig. 2E) and, in addition, corroborated that their annotated mesenchymal cancer cells are assigned to either PeriV-lineage or Rgl-lineage reference cells (Fig. 2F).

In the bulk RNA-sequenced glioblastoma of the TCGA, 106 of 113 cell type–defined *IDH1* wild-type (wt) glioblastoma patients with survival information were used for survival analysis. Glioblastoma with a dominant PeriV-lineage–type phenotype predicted markedly shorter survival than the Rgl-lineage type, and 18 of 19 patients' life spans were <24 months (Fig. 2G). This observation was further validated when stratifying the Rgl-lineage into sublineages on the basis of assignment to the dominating reference cell types (Rgl-like, Astro-sublineage, and Olig-sublineage) (fig. S2E).

We next explored the mutational burden among the glioblastoma defined by PeriV-lineage– and Rgl-lineage–type signatures. Thirty-two genes with high frequency of mutation were significantly enriched (fig. S2F and data file S3). PeriV-lineage–type and Rgl-lineage–type glioblastoma carried a shared enrichment in mutations of *TTN*, *PKHD1*, *TP53*, *PTEN*, and *FLG* genes, and a differential mutational burden with *NF1* gene strongly associated to the PeriV-lineage type and *EGFR* gene to the Rgl-lineage type, especially the astrocyte subtype.

In addition to the transcriptional level, we tested if the methylation status can be used to predict the lineage-based classification of glioblastoma. We first enriched the differential methylation sites with PeriV-lineage–type and Rgl-lineage–type signatures. Hierarchical clustering using these signature methylation sites confirmed a classification congruent to transcription for nearly all patients (Fig. 2H, fig. S2G, and data file S4). Examining the signature methylation sites revealed that tumors of the PeriV-lineage type displayed, for example, increased methylation of *GFAP* gene and *S100B* gene, while *MGMT* gene and *STAT6* (signal transducer and activator of transcription 6) gene were more unmethylated, indicating a suppression of glial genes and an enhanced malignant expression pattern. In agreement, STAT6 has been shown as a unique marker and driver of meningeal hemangiopericytoma, a type of brain tumor that originates from pericytes (*31*). Thus, the methylation signatures reflected the innate cell-type features of PeriV-lineage– and Rgl-lineage–type glioblastoma.

We examined if the methylation status can predict tumor type using machine learning. A neural network classifier was generated by training transcriptionally defined PeriV-lineage– or Rgl-lineage–type glioblastoma with the methylation signatures. Similar to the hierarchical clustering (Fig. 2H), the methylation-based classifier assigned the majority of tumors to the corresponding transcriptionally defined PeriV-lineage–type and Rgl-lineage–type glioblastoma with high

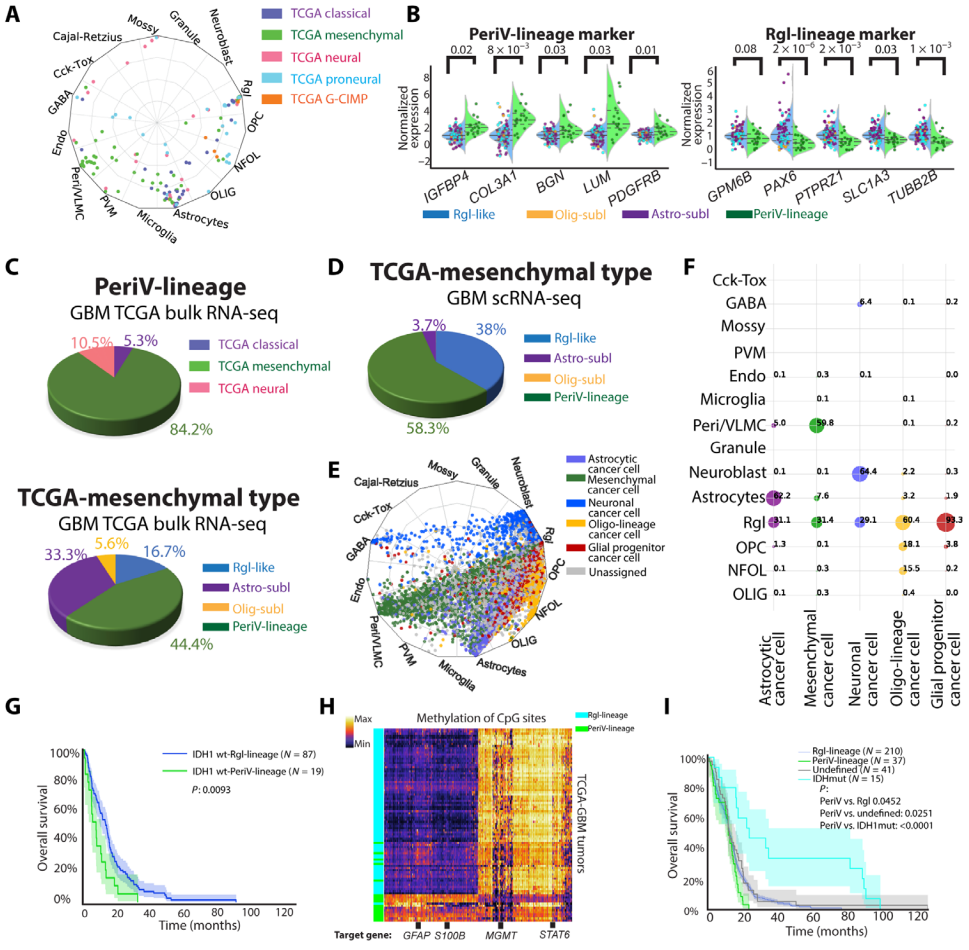


Fig. 2. Tumor subtype assignment, methylation status, and survival of deconvoluted bulk tumor data from TCGA/DFKN. (**A**) Radar plot visualizes the cell-type scores for deconvoluted bulk glioblastoma in relation to trained reference brain cell types. Colors represent the TCGA-defined subtype of each tumor. (**B**) Violin swarm plot of the original gene expression of selected marker genes in the PeriV-lineage and Rgl-lineage of TCGA glioblastoma; blue background represents Rgl-lineage tumors and green background represents PeriV-lineage tumors. Dot colors represent the defined reference brain cell types of each tumor in (A). The dashed line in each violin plot represents the distribution quartiles. *P* value of Student's *t* test on top. Abbreviations are as in fig. S2C. (**C** and **D**) Pie plots representing the composition of TCGA-classified subtypes in the PeriV-lineage (C, top), cell-type sublineages identified in the TCGA-mes subtype (C, bottom) of bulk glioblastoma, or cell-type sublineages identified in the TCGA-mes subtype of scRNA-seq glioblastoma cells (D). (**E**) Radar plot visualizes cell-type scores of state-defined glioblastoma cells in relation to trained reference brain cell types. (**F**) Dot plot represents the percentage of the defined cell states of glioblastoma cells in each originally defined cell-type state. Dot sizes from small to big represent the percentage from low to high. (**G**) Patient survival of isocitrate dehydrogenase 1 (*IDH1*) wild-type glioblastoma from the TCGA assigned as belonging to the Rgl-lineage and PeriV-lineage. (**H**) Heatmap representing the differential methylated site–based hierarchical clustering of the TCGA glioblastomas assigned to the PeriV-lineage and Rgl-lineage type. Selected target genes of the methylated sites are listed at the bottom. Color bar indicates the expression intensity at the top left. *STAT6*, signal transducer and activator of transcription 6. (**I**) Patient survival of glioblastoma from TCGA assigned to Rgl-lineage, PeriV-lineage, *IDH1*-mutant types, and nonclassified based on methylation.

accuracy (fig. S2H). Next, we used this trained classifier for scoring 559 glioblastomas from a merged TCGA/DKFZ dataset (data file S4) and evaluated patient survival. Consistent with previous studies, isocitrate dehydrogenase 1 (*IDH1*)–mutant glioblastoma predicted a better outcome. In the remaining 288 *IDH1* wt patients that include life span information, the PeriV-lineage type predicted the poorest patient survival with 0% 2-year survival (Fig. 2I). We also applied the same classifier for an independent dataset of 151 patients from the CGGA (Chinese Glioma Genome Atlas) (*32*) and further evaluated the *IDH1* wt patient survival. A comparable survival to that of the TCGA/DKFZ studies was observed. Although the difference was not significant, none of the glioblastoma patients with PeriV-lineage–type signatures were alive after 2 years (fig. S2I).

Perivascular lineage–type glioblastoma consists of cells similar to vFBs, pericytes, and vascular SMCs

To examine whether cells of PeriV-lineage glioblastoma cells can be assigned to a specific perivascular cell type, we used a high-quality dataset of reference brain vascular cells, generated by Smart-seq2 scRNA-seq (*17*). Thus, we trained a neural network classifier with learned features from this dataset (fig. S3, A and B), and then assigned the merged dataset of low- and high-grade glioma cells to the reference vascular cell types. Consistent with our previous finding (Fig. 1), glioblastoma cells that were previously assigned to pericytes/VLMCs (fig. S3C, left) were robustly assigned to one of the three perivascular cell types: the immature stem-like vFBs, SMCs, and pericytes. Bulk sequenced data from TCGA were robustly assigned

to vFBs (fig. S3C, middle). In contrast, low-grade glioma cells were rarely assigned to any vascular cell types (fig. S3C, right).

Reconstruction of glioblastoma cells along the developmental trajectory of the radial glia and neural crest cell lineages

Meningeal cells as well as the brain perivascular cell types arise from mesenchymal neural crest cells (*15*, *19*) attaching to blood vessels descending into the brain parenchyma during development (*19*). We therefore next examined the similarity of glioblastoma cells to cranial neural crest and neural tube cells captured from the developing mouse embryo at the time when neural crest cells delaminate from the neural tube (*33*) to meningeal cells (*34*) and to perivascular cells (*17*), as well as cells of the Rgl-lineage including adult Rgl, neuroblasts (*35*), oligodendrocytes, and astrocytes. All these data were generated using the Smart-Seq2 platform. On the basis of our previous analyses, these cell types together represent the endogenous cell types that glioblastoma displays similarities to. To track the developmental location of each glioblastoma cell along the lineage trajectory of brain cells, we developed a neural network–based projection model, SWAPLINE (Single-cell Weighted Assignment and Projection on developmental LINEages) (fig. S3D). We first visualized the normal reference brain cell types in a UMAP (Fig. 3A). Each cell-type cluster's position in the UMAP reflects its transcriptional status in the relatively flattened topology in partition-based graph abstraction (PAGA) and the predicted cells must be assigned according to the limited PAGA nodes supervised by machine learning (fig. S3E). Nevertheless, the result is consistent with previous experimental lineage tracing studies, confirming the validity of the model. Consistently, all assigned tumor cells via SWAPLINE exhibited marker expression consistent with their position and naïve reference cell types (see below). This UMAP was later used as reference map for the projection of glioblastoma cells onto the brain's normal differentiation trajectories.

The accuracy of the SWAPLINE model was tested and confirmed using the independent sets of human brain cells (fig. S3, F and G) (*21*, *22*). SWAPLINE assigned cells correctly in the lineage trajectories, while unrelated control cells (endothelial cells and microglia) were filtered out automatically in the model because of low scores. Next, we applied the model to project each glioblastoma cell into the differentiation trajectories of brain cell types (fig. S3H). The relative tumor cell position in relation to the background map plot of reference developmental/endogenous cell types was visualized (Fig. 3B). To disentangle the transcriptional roadmap of glioblastoma cells, we generated a statistical ensemble of principal branching tree trajectories (*36*) from the high-dimensional transcriptional space (Fig. 3C). The main tree structure summarized glioblastoma cell distribution and comprehensively showed the progression of glioblastoma cells along each developmental lineage trajectory. Two main glioblastoma lineage structures were observed with differentiated cells at termini, after which each branch was named. One lineage was organized around a shared center of Rgl reference cells with branches of cancer cells toward reference astrocytes (Astro-sublineage glioblastoma cells), neuroblasts (Neuronal-sublineage glioblastoma cells), and oligodendrocyte cells (Olig-sublineage of glioblastoma cells). Here, reference Rgl from two developmental stages was included (adult Rgl and developmental Rgl). The other lineage structure was the PeriV-lineage represented as a single line structure, with PeriV-lineage glioblastoma cells positioned from the most undifferentiated early reference migratory neural crest cells to differentiated reference perivascular mural cells.

Cross-annotation of patients and lineage branches revealed that all patients dominantly contained glioblastoma cells assigned either to the reference Rgl-lineage (Astro-sublineage, Neuronal-sublineage, or Olig-sublineage) or to the reference PeriV-lineage cells (fig. S3I). For patients with an Rgl-lineage–type glioblastoma, all subbranches coexisted in all patients, although at different proportions, revealing the intratumor lineage heterogeneity among patients with an Rgl-lineage signature.

To further explore the most similar cell type of PeriV-lineage glioblastoma cells along the differentiation trajectory from undifferentiated reference migratory neural crest cells to differentiated reference perivascular mural cells, we constructed a new cranial neural crest cell reference dataset via integrating the migrating cranial neural crest cells, neural crest mesenchymal progenitor cells (*33*), meningeal cells (*34*), and brain perivascular cells (*17*), which should represent all known neural crest derivatives in the brain region. After training with this reference dataset in the neural network model, we found that the PeriV-lineage tumor cells are most similar to vFBs and migrating neural crest cells (fig. S3J).

The existence of two lineages in glioblastoma cells was further confirmed by SWAPLINE lineage reconstruction for two independently published glioblastoma datasets, including (*5*) (fig. S3, K to N) and (*7*) (fig. S3, O to R). Moreover, we applied SWAPLINE assignment for glioblastoma cells with or without OSM treatment and found that almost all cells were assigned to Rgl-lineage cells (fig. S3, S and T), indicating that the cell-type state of glioblastoma cells remains conserved even after the OSM-induced transition to a more mesenchymal-like state. However, OSM-treated cells exhibited an increased feature of delaminating neural crest cell (fig. S3U) and reduced feature of radial glia, suggesting that the mesenchymal signature induced by OSM reflects features of the epithelial-mesenchymal transition of premigratory neural crest cells (*37*).

Next, we enriched pseudo-time marker genes that associated with each branch trajectory (data file S5), and the normalized expression of the selected marker genes along the Rgl-lineage branches was visualized in the branching tree (Fig. 3D). For example, *STMN2* and *SOX10* were specifically expressed in glioblastoma cells at the distal part of the neuronal- and Olig-sublineages, respectively, suggesting the existence of stable transcriptional status along these two branches. In contrast, Rgl-like tumor cells and glioblastoma cells at the distal part of the Astro-sublineage and Rgl-enriched *SOX9* and *GFAP* were, albeit at lower levels, also expressed across all branches, indicating lack of unique markers for these glioblastoma cells. Consistently, *RGS4*, which is transiently expressed during neural crest differentiation (*38*), was also expressed in PeriV-lineage glioblastoma, specifically enriched in the progenitor-like cells of such tumors (Fig. 3E), while expression of lumican (*LUM*) and actin alpha 2, smooth muscle (*ACTA2*) was consistently enriched in glioblastoma cells corresponding to the more differentiated brain vFBs and SMCs, respectively.

Cell cycle and differentiation potential along differentiation branches of glioblastoma cells

Tumor initiation and propagation requires cell division. In our dataset and two independent glioblastoma datasets (*5*, *7*), cycling tumor cells were mainly observed at the region of reference Rgl and between the reference migrating neural crest and vFB cells, while tumor cells in all branch termini were relatively quiescent (fig. S4, A to C). These observations suggest that the mitotic hyperactivity

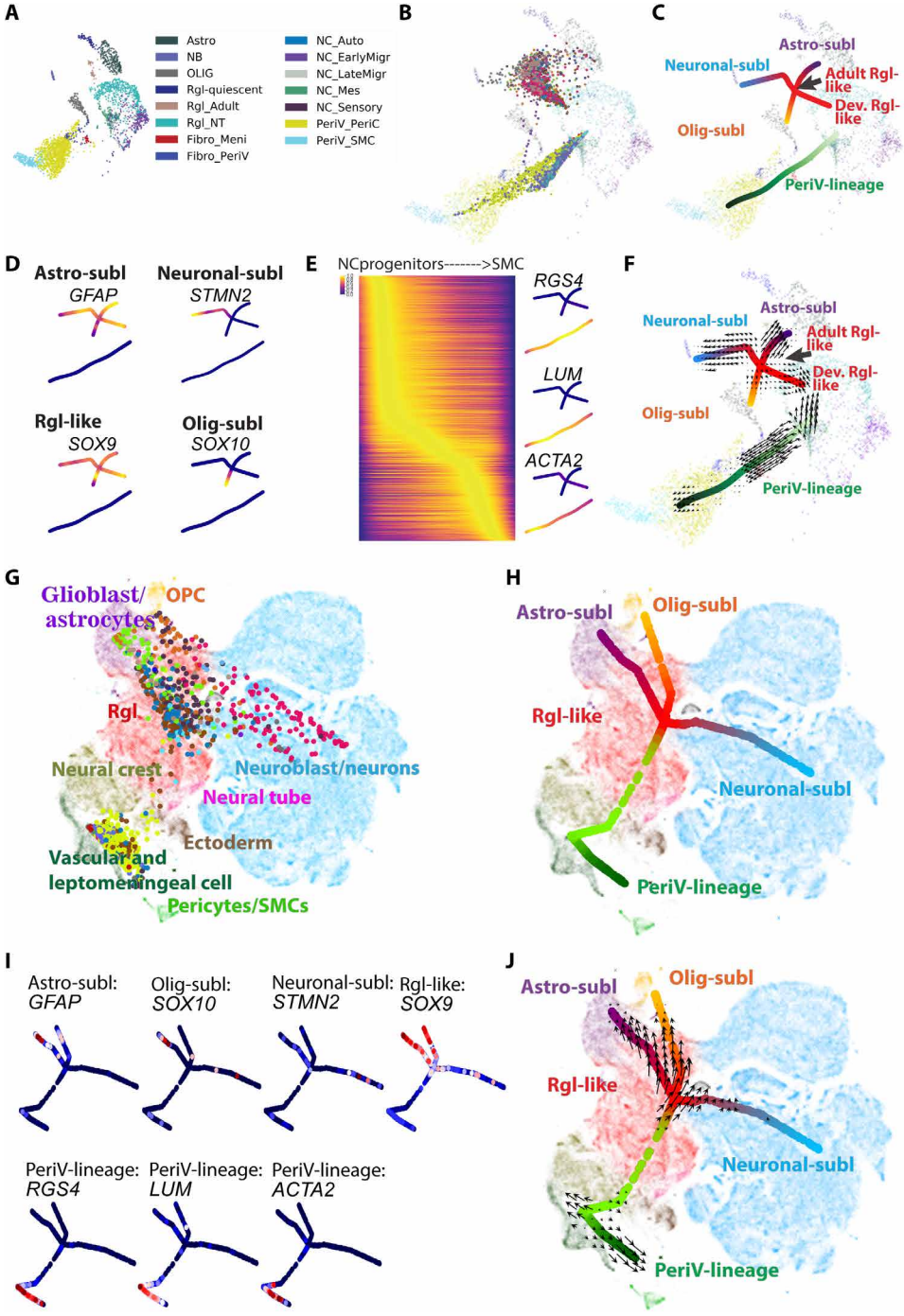


Fig. 3. Relation of glioblastoma cells to the developing central nervous system and neural crest. (**A**) Plot of reference cells. UMAP visualization of cell clusters from the developing central nervous system and neural crest lineages (*17*, *33*–*35*). Abbreviations are as in fig. S3E. (**B**) Projection of all glioblastoma cells to the reference plot. Reference cells are indicated by "×" and glioblastoma cells are indicated by "dot," which represent the projected developmental position of the individual glioblastoma cells to native reference cell types. (**C**) Principal tree plot summarizing the developmental status trajectory of the glioblastoma cells. Lineages are indicated by colors and text. Abbreviations are as in fig. S3M. (**D**) Visualization of normalized expression in tumor cells of pseudo-time marker genes for branches in the Rgl-lineage. (**E**) Left: Heatmap shows the normalized expression of pseudo-time genes according to the voltage peak along the neural crest trajectory. Right: Projection of the normalized expression in tumor cells of selected marker genes on the branching tree plot. Dark purple to yellow represents the minimal to maximal expression. (**F**) Quiver visualization of RNA velocity of glioblastoma cells on the branching tree plot. The arrow of each glioblastoma cell points to the direction of future status, extrapolated from RNA velocity estimates. (**G** and **H**) SWAPLINE projection and branching tree visualization of glioblastoma cells onto developmental mouse brain and neural crest reference plot from the mouse developmental brain atlas (*16*). Abbreviations are as in fig. S4J. (**I**) Marker gene expression in glioblastoma cells and visualized in the branching tree projected on the reference developmental mouse brain plot. Dark blue to red represents the minimal to maximal gene expression. Abbreviations are as in fig. S4J. (**J**) Quiver visualization of RNA velocity of glioblastoma cells onto developmental mouse brain and neural crest reference plot.

of progenitor-like tumor cells is a general rule for tumors with an Rgl-lineage–type and PeriV-lineage–type transcriptional signature. Mitotic events developmentally couple with cell differentiation and fate decision (*39*). RNA velocity analysis (*40*) revealed that the main trend of differentiational status change along each sublineage branch was from the progenitor region to differentiated termini (Fig. 3F and fig. S4D). Both the neuroblast and the oligodendrocyte branch of glioblastoma cells showed reduced differentiation at the developmental terminus, consistent with pseudo-gene results in Fig. 3D. Tumor cells at the terminus of the astrocyte branch exhibited lineage reversal, indicating bidirectional glioblastoma cell differentiation along the reference Rgl to astrocyte differentiation trajectory. In the PeriV-lineage, the main differentiation trend of glioblastoma cells was from reference migrating neural crest cells to perivascular cells. We also found that some of the most undifferentiated glioblastoma cells assigned to the PeriV-lineage displayed differentiation vectors toward reference spinal cord Rgl cells.

The most undifferentiated glioblastoma cells are expected to be enriched at the regions of the reference Rgl and neural crest cells (fig. S4E). To enhance the resolution of the reference map for a subsequent annotation of the most undifferentiated stem-like glioblastoma cells, we extracted these cells according to the density estimation and performed a zoom-in projection on the recently released mouse developmental brain atlas (*16*) again using the SWAPLINE projection (fig. 3G and fig. S4, F to I). The summarized tree structures and RNA velocity estimation further disentangled the progression of glioblastoma progenitor-like cells along each embryonic developmental brain lineage (Fig. 3, H to J, and fig. S4J). Confirming the above results, some tumor cells clustered with reference Rgl cells as well as along branches of reference cell differentiation into astrocytes, neuroblasts, and oligodendrocytes. Other glioblastoma cells were mainly located at the reference embryonic neural crest/VLMC region of the map with a branch toward reference perivascular cells. Reference cell lineage markers further confirmed that the tumor cells assigned to a developmental position also expressed the expected markers of naïve cells in that differentiation branch of the embryonic brain (Fig. 3I and data file S5). Furthermore, the relation of glioblastoma cells to these reference embryonic developmental lineages was further validated by SWAPLINE lineage reconstruction for two independent published glioblastoma datasets from (*5*) (fig. S4, K to M) and (*7*) (fig. S4, N to P), with similar results. To enhance the resolution of the reference brain cell types, we applied the machine learning classifiers with learned transcriptional features from early human developing brain cell types (fig. S4Q) (*41*), further validating our observation (fig. S4R). Combined, these results indicate that heterogeneity in glioblastoma can be explained by two main cell-type lineages of the brain, the radial glia and the PeriV-lineage, with tumor cell transcriptional programs at large recapitulating normal transcriptional routes of differentiation.

The direct lineage relationship of glioblastoma cells to developmental and adult brain cells indicates that transcription factors (TFs) that define cell types and thereby drive differentiation in the developing brain also contribute to the diversity of glioblastoma cells along the lineage trajectories. Thus, we divided our tumor cells into six lineage clusters according to their lineage branches and progenitor feature relationship to reference cells. Subsequently, we enriched the differentially expressed TFs from each glioblastoma lineage cluster as described in fig. S4E. Next, we applied the same enrichment for the published glioblastoma dataset (*5*), as well as for the annotated reference dataset of normal brain cell types (*20*). By comparing these three datasets, we identified unique TFs defining Rgl-lineage (30 TFs) and PeriV-lineage tumor cells (6 TF genes: *FLI1*, *FOXC1*, *STAT6*, *KLF2*, *TFAP2C*, and *MSC*) shared with normal development and a few glioblastoma-specific factors within each of the lineages (Fig. 4, A and B, and data file S6). Next, we applied SCENIC for identifying gene networks regulated by master TFs (regulon activity) in both Rgl-lineage and PeriV-lineage cells. After comparing the enriched TFs, 20 master TFs were identified with significant regulon activity (fig. S4S). The Rgl-lineage consisted of 14 TF regulons, including some known Rgl-specific TF genes, such as *HES5*, *RFX4*, and *SOX10*. We identified six TF regulons specific for PeriV-lineage, including *STAT4*, *STAT6*, *TFAP2C*, *FOXC1*, *FLI1*, and *MSC*. Furthermore, analysis showed that shared features between the two lineages (PeriV and Rgl) all relate to the cell cycle, including five cell cycle–regulating TFs (*FOXM1*, *MIS18BP1*, *MYBL1*, *MYBL2*, and *WDHD1*) (Fig. 4C and data file S6). Two lineage-specific TFs, *PROX1* for Rgl-lineage and *FOXC1* for PeriV-lineage, were validated in the tumor tissue of patient-derived xenografts (Fig. 4D). *SOX2* and *POU3F2* are driver genes in glioblastoma-propagating cells (*42*) that are induced during oncogenesis since they are not expressed in normal perivascular cells but present in migrating neural crest (*43*). Therefore, we also validated these two genes as lineage-shared TFs (Fig. 4E).

Initiation of PeriV brain tumors from perivascular cells

Mouse models have indicated that glioblastoma can efficiently be initiated from the glial and stem cell compartments of the brain (*11*). The notable similarity of PeriV-lineage–type tumor cells to endogenous reference perivascular cells suggests that perivascular cells can also be susceptible for malignant transformation. To test whether perivascular cells might initiate brain tumors when carrying genetic alterations mimicking glioblastoma, we first investigated the expression profiles of the spontaneous glioblastoma tumors from both *Nes*-CreERt2 *Pten*/*Trp53*/*Nf1* KO mice and *NG2*-CreERt2 *Pten*/*Trp53*/*Nf1* KO mice (*11*, *12*). Nestin is predominantly expressed in neural stem cells (i.e., radial glia cells), but *NG2* is typically expressed in oligodendrocytes as well as in perivascular cells in the mouse brain (*44*). Thus, we hypothesized that tumors from *NG2*-CreERt2 *Pten*/*Trp53*/*Nf1* KO mice can arise from either naïve oligodendrocytes or perivascular cells of the brain, while tumors from *Nes*-CreERt2 *Pten*/*Trp53*/*Nf1* KO mice should arise only from radial glia cells. Hierarchical clustering revealed that two of the seven sequenced tumors derived from $NG2^{+}$ cells were PeriV-lineage and the other five were Rgl-lineage. Furthermore, none of the seven glioblastomas induced from Nes^{+} cells carried any perivascular signature pattern (fig. S5A).

Platelet-derived growth factor (PDGF) acting through PDGF receptors induces proliferation and migration of perivascular cells (*45*). We therefore estimated the tumorigenesis potential of human brain perivascular cells by introducing *PDGFB* and depleting *CDKN2A* (*p16INK4A* and *p14ARF*) in primary human brain pericytes ($Peri^{PDGFB/CDKN2A}$) and introducing *PDGFB* and co-depleting *NF1*/*TP53* in human primary brain vFBs ($fibroblast^{PDGFB/NF1/TP53}$) with green fluorescent protein (GFP) introduced into both cell types (fig. S5, B to D). These alterations led to marked increases in in vitro growth compared to naïve cells and significantly promoted the colony formation in vFBs (Fig. 5, A and B, and fig. S5E). To explore the consequences of these genetic alterations on cell identity, we scRNA-sequenced vFBs with and without the alterations. We observed

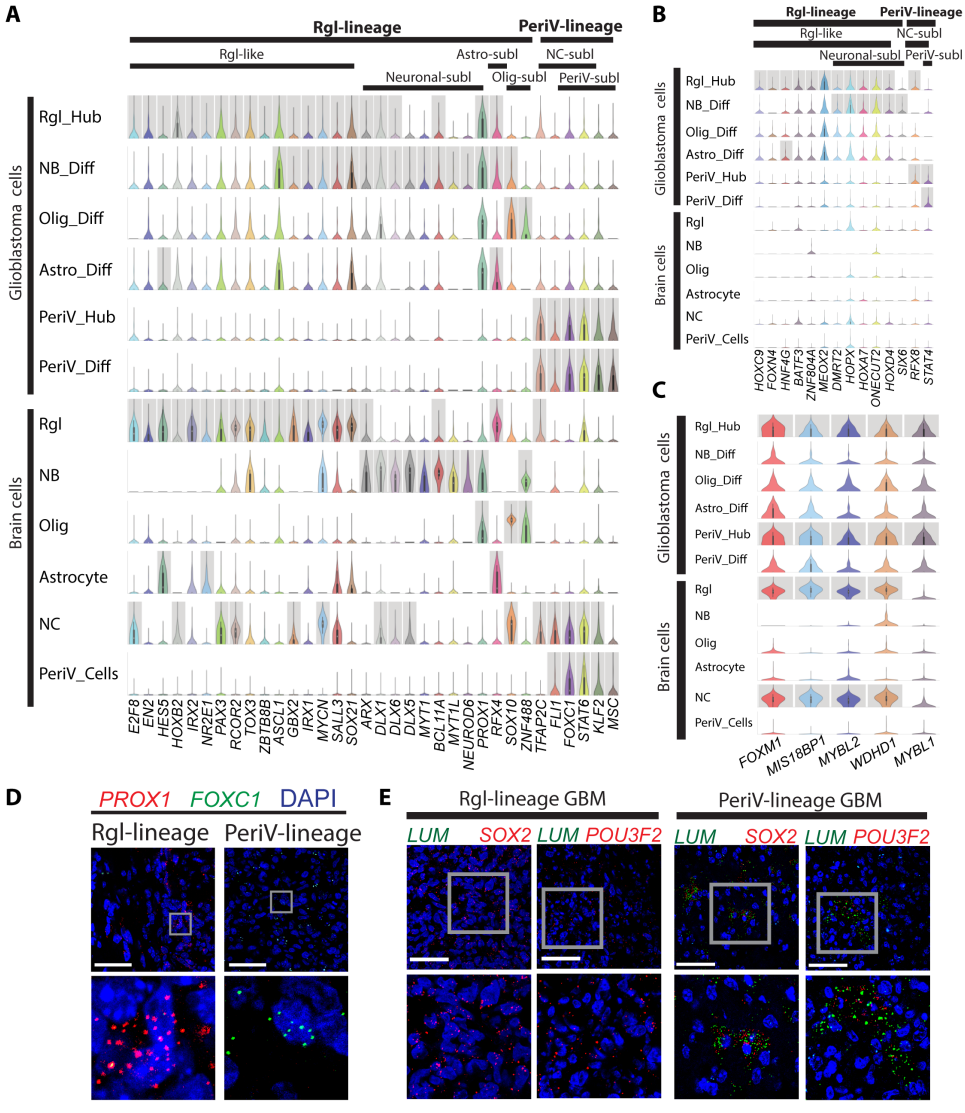


Fig. 4. Conserved TF signatures between naïve brain and neural crest cells with Rgl- and PeriV-lineage glioblastoma. (**A** to **C**) Violin plot of TF expression shared between tumor cells and normal reference cell types (A), of TFs unique to glioblastoma cells (B), and of TFs shared between Rgl- and PeriV-lineage glioblastoma cells (C). *y* axis, the relative expression level; *x* axis, TF gene names. Cell types and lineages are indicated at the top of the chart. Gray columns represent the significantly differential expression. "Diff" indicates tumor cells at the distal differentiation of the sublineage trajectories and "Hub" indicates stem-like cells of the Rgl and perivascular lineages corresponding to native radial glia and neural crest cells, respectively. (**D** and **E**) Validation of *PROX1* and *FOXC1* mRNA expression in Rgl-lineage– and PeriV-lineage–type patient-derived glioblastoma xenografts, respectively (D). Validation of *SOX2* and *POU3F2* mRNA expression in both PeriV-lineage–type and Rgl-lineage–type patient-derived glioblastoma xenografts, *LUM* was used as a marker of PeriV-lineage tumor (E). Tumor lineage type and gene names are at the top. Each bottom figure is a higher magnification from the gray frame of the top figure. Scale bars, 50 μm.

comprehensive CNV changes in genetically modified vFBs (Fig. 5C and fig. S5F), with the significant deletion of Chr.4q, 1q, 9q, and 18q, and amplification of Chr.12q and 5q, indicating that a few founding mutations can lead to large genetic alterations. In particular, the alterations of Chr.18q and 5q have been identified in mesenchymal glioblastoma (*5*) and meningioma (*46*)—another type of brain tumor derived from the neural crest lineage. SWAPLINE projection of the control and genetically modified vFBs in the developmental adult reference plot revealed a marked dedifferentiation of the modified vFBs toward reference neural crest progenitors (Fig. 5, D and E). Consistently, more G_2-M cycling cells were observed in modified vFBs (Fig. 5F and fig. S5G). By comparing the transcriptional profile between control and modified vFBs, we identified 773 up-regulated and 638 down-regulated genes (data file S7). Pathway enrichment revealed that "cell cycle and chromatin reorganization" and "neural crest differentiation" were significantly increased, while "HOX gene–related tissue patterning" was suppressed, indicating a dedifferentiation toward a neural crest stem cell state and a loss of anterior-posterior positioning information (Fig. 5G and data file S7). The cells were introduced into the brain in the orthotopic mouse model to test for tumor initiation. Both the modified pericytes and vFBs generated tumors, and the mice exhibited poorer tumor-associated survival than the control group receiving naïve cells (fig. S5H). Consistently, none of the control groups transplanted with

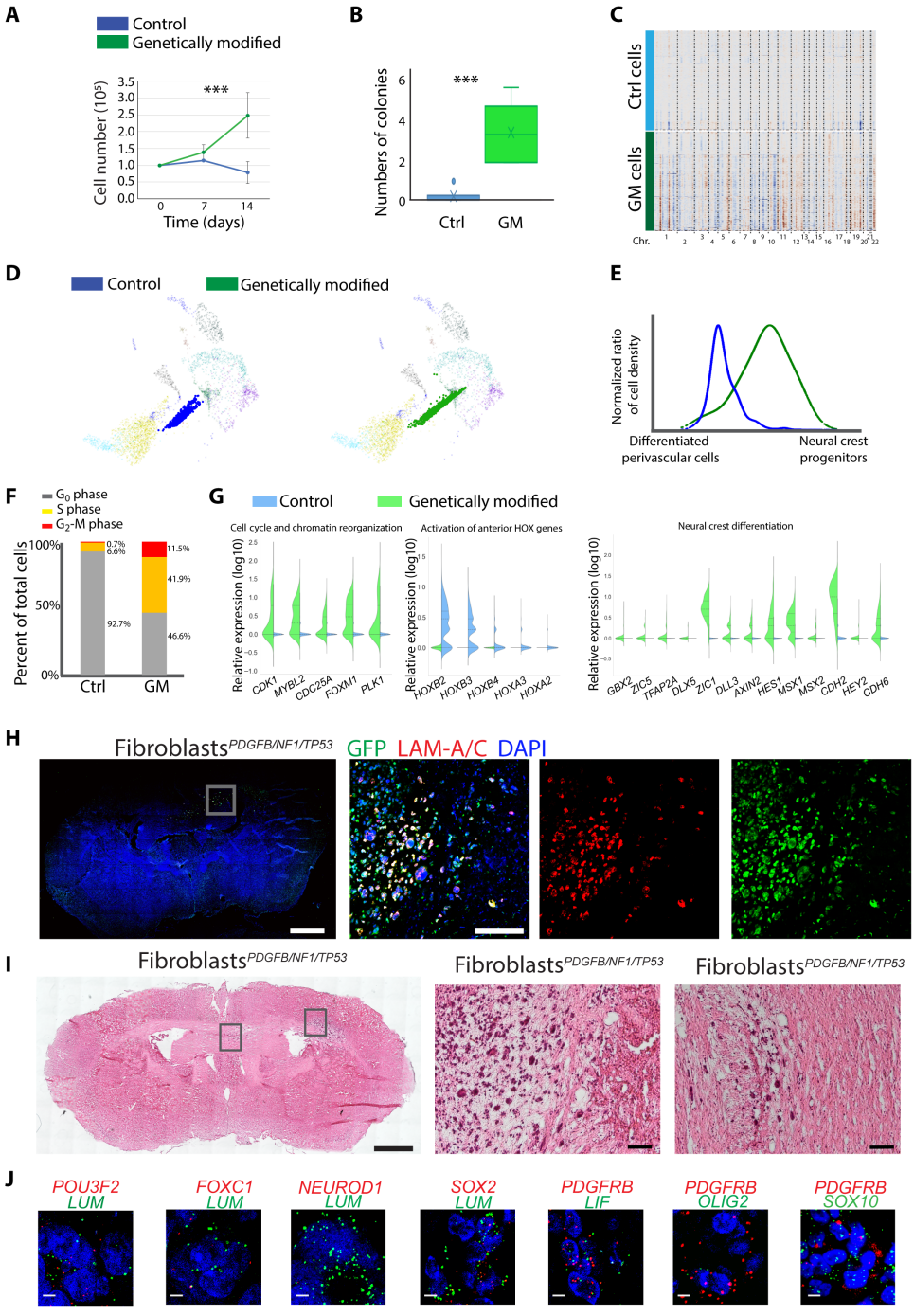


Fig. 5. In vivo initiation of tumors from perivascular cells. (**A** and **B**) In vitro proliferation (A) and colony formation (B) of brain vFB with/without carrying genetic alterations of patient-derived glioblastoma [genetically modified (GM), green]. Means ± SD, three independent measurements. Student's *t* test, ***$P < 0.001$. (**C**) CNV analysis of control (blue) and GM fibroblasts (green). (**D**) Projection of control and genetic modified fibroblasts to the reference plot of normal reference cell types from Fig. 3A. (**E**) Quantification of the differentiation status of control (blue) and GM fibroblasts (green) along the developmental trajectory of in vivo differentiation of reference perivascular cells. The *y* axis represents the normalized cell density of projected fibroblasts in (D). The *x* axis represents the linearized developmental position between differentiated brain perivascular cells and neural crest progenitors. (**F**) Quantification of cycling phases of control (Ctrl) and GM fibroblasts. (**G**) Gene expression of top significant pathways enriched by up- and down-regulated genes in GM fibroblasts as compared to the naïve fibroblasts. (**H** and **I**) Representative fluorescence (H) or hematoxylin and eosin (I) staining of the coronal section from mouse xenograft of GM fibroblasts. Magnified tumor regions boxed. Green, GFP; red, anti-human lamin (LAM) A/C; blue, 4′,6-diamidino-2-phenylindole (DAPI). Scale bars (H and I): 1000 μm, whole section; 100 μm, magnified figures. (**J**) In vivo mRNA expression of indicated marker genes in xenograft tumor tissues of genetic modified fibroblasts. Human *LUM* and *PDGFRB* were used to label tumor cells. Gene names and color are indicated in each panel. Scale bars, 10 μm.

the corresponding naïve cell types had a confirmed brain tumor by histological analysis, while all genetically altered perivascular cells did. Fluorescence staining confirmed that the brain tumors were of human cell origin (anti-human lamin A/C and GFP; Fig. 5, H and I, and fig. S5, I to K). Both Peri$^{PDGFB/CDKN2A}$ mice and fibroblast$^{PDGFB/NF1/P53}$ mice exhibited extensive neoplastic growth and most animals displayed a diffuse and infiltrative phenotype. The xenograft tumor tissue exhibited cellular mitotic activity (Ki67), altered microvascular patterns (CD31), and abnormal remodeling of extracellular matrix proteins (fibronectin and collagen VI) (fig. S5L). Furthermore, the expression of PeriV-lineage tumor marker genes (*POU3F2*, *FOXC1*, *SOX2*, and *LIF*) in the tumor tissue of the grafted mice was observed, while Rgl-lineage genes *NEUROD1* and *OLIG2* were rarely observed (Fig. 5J). We observed some tumor cells coexpressing the neural crest progenitor marker *SOX10*, in line with our in silico observation of a cellular dedifferentiation in transformed tumor cells (Fig. 5, D and E).

DISCUSSION

scRNA-seq has provided unparalleled insights into the molecular nature of glioblastoma cells and has offered new means to explain the cell of origin, tumor phenotype, cell heterogeneity, and patient outcome (*47*). In this study, we combined the application of a neural network classifier and the trajectory analysis of native brain cells to identify the relation of glioblastoma cells to normal brain cells. Our results identified that some glioblastomas display high similarities to radial glia and its progenies (Rgl-lineage), consistent with previous studies assigning tumor cells to neural cell types using a list of defined marker genes, hierarchical clustering, or reference cells in principal components analysis (PCA) (*5*, *7*, *8*, *10*, *26*). Unexpectedly, we identified the remaining glioblastoma to be similar to perivascular cells (PeriV-lineage), and consistently, tumor cells were robustly allocated along one of the two cell lineages. Furthermore, we validated the tumor-propagating ability of naïve brain perivascular cells. According to our neural network classification of scRNA-seq data as well as deconvolution of bulk data, glioblastoma of a PeriV-lineage type represents a proportion of the TCGA-mes subtype. Furthermore, consistent results were obtained on patient survival using gene expression– or methylation-based patient stratification into Rgl-lineage or PeriV-lineage. Patients with a PeriV-lineage–type signature show significantly poorer survival than those with an Rgl-lineage type. Combined, our results suggest the existence of a subgroup of glioblastoma with similarities to perivascular cells of the brain, which is distinct from the Rgl-lineage.

Although transcription can be affected by both mutations driving transformation as well as the microenvironment (*5*), the originating cell lineage can represent an important determinant of glioblastoma molecular characteristics (*12*). Among the conserved markers expressed in most cell types of each of the lineage (Fig. 1C), there is a high expression in Rgl-lineage cells of *PTPRZ1* and *SLC1A3*, which previously have been shown to contribute to glioblastoma initiation and progression (*10*). Furthermore, the expressions of PeriV-lineage markers, *LUM* and platelet-derived growth factor receptor beta (*PDGFRB*), have also been previously evidenced in glioblastoma (*48*, *49*). Because glioblastoma tumors exhibit cells with features consistent with precursor populations, shared developmental determinants of the progenitor cell fates could contribute to oncogenesis. Cell cycle analysis along the lineage trajectories revealed both Rgl-lineage and PeriV-lineage tumor cells to be rapidly dividing with markedly reduced proliferation of the more differentiated cells within each lineage. When we identified shared features between the two progenitor cell populations, nearly all shared genes were cell cycle–regulating transcriptional activators. This suggests that a major shared feature in the progenitor cells of the two lineages (PeriV- and Rgl-lineage) involves cell cycle control. Thus, transcriptional determinants contributing to oncogenesis in the two different lineages unrelated to cell cycle control are for the most part unique to each lineage and coincides with those in normal brain lineage trajectories.

RNA-velocity analyses show that the main flow in glioblastoma is from progenitor cells to differentiated cell types, and hence, glioblastoma develops along conserved neurodevelopmental gene programs, in agreement with a recent similar analysis (*7*). However, unlike that study, we find lineage reversal of tumor cells in the astrocyte branch of differentiation as well as of PeriV-lineage tumor cells carrying similarity to reference vFB cells. This difference may be a consequence of the fact that we performed a comprehensive RNA velocity with all assigned glioblastoma cells on the lineage branching tree plot, instead of on selected individual patients or selected reference brain cell types, thus overall increasing resolution. Furthermore, the standard dimensional reduction (such as PCA and *t*-distributed stochastic neighbor embedding) in a previous analysis could be too strict for estimating RNA velocity across tumor patients, due to the individual variance (*5*, *10*). Instead, a score-based branch plot may better reflect the roadmap of developmental programs for cancer studies (*50*). The finding of lineage reversal of some more differentiated cells is consistent with a high degree of plasticity observed in glioblastoma cells (*5*, *8*, *10*) and suggests that, within glioblastoma, tumor cells with astrocyte and vFB features along with the glioblastoma resident progenitor populations can be originators of the cancer cell hierarchy and, thus, driving cancer growth. This is also consistent for the PeriV-lineage–type glioblastoma in experimental data, since recapitulating in perivascular cells genetic changes of glioblastoma is sufficient to initiate tumors with perivascular cell expression features in orthotopic grafted mice, including a derepression of the stemness maintenance factor SOX2 (*51*).The profound impact of a limited set of TFs on the fate of perivascular cells is illustrated by the direct reprogramming of pericytes to neurons through a neural stem cell intermediate by forced expression of SOX2 and the proneural ASCL1 TF (*52*), suggesting that re-expression of SOX2 alone is sufficient for a dedifferentiation of pericytes to a stem-like cell state from which ASCL1 induces neurogenesis. Thus, our results are consistent with the notion that some glioblastoma can originate from neural crest–derived leptomeningeal and perivascular cells. It appears that, within these, a few acquired mutations can start a process involving genetic instability and re-expression of developmental TFs shifting differentiated perivascular cells into more progenitor-like cells within the differentiation trajectory of the neural crest.

MATERIALS AND METHODS

The reagents, software, and public datasets are listed in data file S8. The machine learning models, training datasets, testing datasets, main lineages, sublineages, and assigned cell types are listed in data file S9.

Human GC cultures

Surgical tissue samples and clinical information for glioma patients were obtained from Karolinska Hospital in accordance with the

protocol approved by the regional ethical review board. An informed written consent was obtained from all patients. We have used 18 human glioblastoma cell lines between passages 1 and 5. Tumors were classified by a neuropathologist on the basis of the World Health Organization classification. Human glioblastoma tissues were cultured as previously described (*53*) with some modification. The tissue was minced with a scalpel, digested in Accutase/TrypLE (1:1) at 37°C for 15 min, and triturated through 18G and 21G needles. The dissociated cells were resuspended in NeuroCult NS-a basal medium (STEMCELL Technologies) with the addition of 1% B27 (Invitrogen), 0.5% N2 (Invitrogen), and 10 ng/ml each of EGF and fibroblast growth factor 2 (PeproTech), plated on laminin-coated Primaria dishes (Corning), and cultured as adherent cells.

Lentiviral-based genetic modifications of human pericytes and fibroblasts

Human brain vascular pericytes (HBVPs) and human brain vascular adventitial fibroblasts (HBVAFs) were purchased from ScienCell and cultured following the instructions provided by the company. The lentiviral construct, shCDKN2A pGFP-c-shLenti vector, was purchased from OriGene Technologies, and shNF1/P53 dual shRNA (CS-LvRU6GP) expressing GFP and pEZ-Lv151 vector expressing PDGFB were purchased from GeneCopoeia. The viral particles were produced in 293T cells through cotransfection of pMD2.G and psPAX2 at a ratio of 4:2:3. Supernatants were harvested 48 and 72 hours after transfection and concentrated using Lenti-X Concentrator solution (ClonTech). Viral pellets were resuspended in phosphate-buffered saline (PBS) and stored at −70°C until further use. HBVPs or HBVAFs were infected for 48 hours and then selected.

Colony formation assay

A total of 1×10^4 cells were mixed in 1.5 ml of 0.4% agarose as the top layer with a bottom base of 1.5 ml of 0.6% agarose, cultured in a six-well plate. The 0.4% and 0.6% agarose are the mixtures of low–melting point agarose and NeuroCult NS-a basal medium above. Every culture well is photographed for at least two views randomly; then, the pictures were counted for colony numbers after 20 days. The average counts were taken as counts of one sample. Triplicate wells were included in each analysis and at least three independent experiments were conducted.

Intracranial transplantation

Animal experiments were performed in accordance with the rules and regulations of Karolinska Institute and approved by the local animal ethics committee. Intracranial transplantation of human germinal center (GC) cultures was performed in neonatal nonobese diabetic–severe combined immunodeficient (NOD-SCID) mice as previously described (*54*). Human GCs were dissociated in TrypLE, and the number of cells was determined using a Coulter Counter (Coulter Electronics). Stereotaxic injections of 2×10^5 genetic-modified HBVP or HBVAF cells in 4 μl of Dulbecco's PBS were performed on 8- to 10-week-old female NOD-SCID mice. The coordinates were 0.5 mm anterior of bregma, 1.1 mm lateral, and 2.5 mm ventral. Injected mice were monitored every second day and euthanized upon symptoms of disease. After euthanizing the mice, their brain was collected and fixed with 4% paraformaldehyde in PBS for overnight. The tissue was then washed with PBS and incubated with 15% sucrose for 24 hours, and 30% sucrose for another 24 hours. After that, the tissue was embedded into optimal cutting temperature compound (Sakura Biotech) in a Cryomold (Sakura Biotech) and frozen using liquid nitrogen. The frozen tissue blocks were stored in −80°C. Ten- to 12-μm-thin cryo-sections of xenograft tumor tissue were prepared on Superfrost Plus slides and slides were either stored in −80°C or processed immediately for immunofluorescence, fluorescence in situ hybridization, or hematoxylin and eosin staining.

Immunofluorescence analysis of mouse brains

Frozen sections were blocked in PBS containing 0.2% Triton X-100 (PBS-T), 3% bovine serum albumin, and 5% normal goat serum and incubated with primary antibodies for 1 hour at room temperature or at +4° for 4 hours in a humidified chamber. The sections were then washed with PBS-T three times and incubated with secondary antibodies (1:500) at +4° for 4 hours. After finally washing three times in PBS-T, sections were mounted in Immu-Mount (Thermo Fisher Scientific) containing 4′,6-diamidino-2-phenylindole. The pictures were taken using an LSM 700 confocal microscope (Carl Zeiss).

Fluorescence in situ hybridization (RNAscope)

Transcripts were detected using the RNAscope assay for fresh-frozen tissue (Advanced Cell Diagnostics). The probes were designed and provided commercially by Advanced Cell Diagnostics Inc. For the complete list of probes and genes, see Resource and Reagent List. The staining was performed using the RNAscope Fluorescent Multiplex Reagent Kit (catalog no. 320850), reagents, and probes according to the manufacturer's instructions. Imaging was performed using LSM 700 confocal microscopes (Carl Zeiss).

Single-cell isolation and cDNA synthesis

A Fluidigm C1 Autoprep System microfluidic chip was used to capture the cells. Immediately after the image acquisition, cell lysis, reverse transcription, and polymerase chain reaction (PCR) amplification were performed as previously described (*55*). The amplified cDNA was harvested with 13 μl of Harvest Reagent and cDNA library quality was measured on an Agilent Bioanalyzer.

Preparation of sequencing library and Illumina sequencing

For patient-derived glioblastoma cells, we used 5′ single-cell–tagged reverse transcription sequencing (STRT-seq). Cell barcoding and fragmentation were performed in a single step using Tn5 DNA transposase ("tagmentation") as described previously. One microliter of Dynabeads MyOne Streptavidin C1 beads (Invitrogen) was resuspended in binding and blocking buffer (10 mM tris, 250 mM NaCl, 5 mM EDTA, and 0.5% SDS) at the ratio of 1:20 and then added to each well. After incubation at room temperature for 15 min, all wells were pooled, and the beads were washed once with 100 μl of washing buffer (10 mM tris–150 mM NaCl and 0.02% Tween 20), once with 100 μl of QIAGEN Qiaquick PB, and then twice with 100 μl of washing buffer. Restriction was performed to cleave 3′ fragments: The beads were incubated in 100 μl of restriction mix [1× NEB CutSmart and PvuI-HF enzyme (0.4 U/μl)] for 1 hour at 37°C. Last, the beads were washed three times with the washing buffer, and then resuspended in 30 μl of ddH_2O and incubated for 10 min at 70°C to elute the DNA. AMPure beads XP (Beckman Coulter) were used at 1.8× volume and eluted in 30 μl to remove short fragments. The molar concentrations of the libraries were determined with KAPA Library Quant qPCR (Kapa Biosystems) and the size distribution was evaluated after PCR (12 cycles) using an Agilent Bioanalyzer. Sequencing was performed on an Illumina

HiSeq 2000 with C1-P1-PCR2 as read 1 primer and C1-TN5-U as index read primer. Reads of 50 base pairs (bp) as well as 8-bp index reads corresponding to the cell-specific barcodes were generated. For genetic-modified perivascular cells, the scRNA-seq was performed by using Chromium Single Cell 3′ Reagent Kits (10x Genomic, version 3) according to the manufacturer's instruction.

Bioinformatics preprocessing, copy number analysis, and clustering

For STRT-seq, the reads were aligned by STAR using GRCh38.p12 genome assembly and processed as described previously (*55*). The cells harboring less than 1000 detected transcripts or less than 450 detected genes were filtered out. After these quality control procedures, 4073 cells were left with the median detected protein coding genes of 3531 counts. For 10x scRNA-seq, data preprocessing was performed via Cell Ranger. The copy number analysis was performed with CONICS following the instruction (*56*). Briefly, genes expressed in <5 cells were excluded. After centering the gene expression in each cell around the mean, the *z*-score of the centered gene expression was calculated across all cells. Next, the bimodal distribution of gene expression in any regions across cells was determined by a Gaussian mixture model mode, and the regions containing more than 100 expressed genes were identified for the next step. Then, the reported mixture models were chosen following the criteria of the Bayesian information criterion >5 and the *P* value of likelihood ratio test <0.05. To detect the existence of CNVs, the threshold of posterior probabilities was set as 0.55, and the gain or loss was determined by comparing the average expression in the normal cells. The heatmap visualizations of chromosomal alterations were generated in every single cell across the genome for all calculated patients.

Before clustering, we removed the cell cycle–related genes and then computed the coefficient variation (CV) (SD divided by the mean) versus the predicted CV (estimated by a nonlinear noise model) and applied the fit of noise distribution to select the most variable features that are greater than the expected CV. Support vector regression (SVR) from scikit-learn package was used for this analysis. The most variable features were used for calculating the top 20 PCs, and the top 10 nearest neighbors, 0.5 minimum distance, and Euclidean distance were used for UMAP.

The most variable genes were then used for cell clustering via different algorithms including the DBSCAN algorithm (Seurat V1.2) and the Louvain method for community detection with a resolution value of 1 (Seurat V3.0+) (*55*, *57*). Furthermore, we applied several rounds of clustering, zoom-in clustering, and cluster recombining to make sure that all clusters are biologically meaningful and exhibited significant markers. Eventually, cells were grouped into 20 clusters, and the marker genes of every cluster were determined via enrichment score as described in (*44*). The enrichment score $E_{i,j}$ for gene i and cluster j was defined as

$$E_{i,j} = \left(\frac{\alpha_{i,j} + \varepsilon 1}{\alpha_{i,\bar{j}} + \varepsilon 1}\right)\left(\frac{\beta_{i,j} + \varepsilon 2}{\beta_{i,\bar{j}} + \varepsilon 2}\right)$$

Here, $\alpha_{i,j}$ represents the score of nonzero expression for the cells in this cluster, and $\alpha_{i,\bar{j}}$ represents the score of nonzero expression for the cells that are not in this cluster. $\beta_{i,j}$ represents the mean expression for the cells in this cluster, and $\beta_{i,\bar{j}}$ represents the mean expression for cells that are not in the cluster. A small value of the constants ε_1 and ε_2 is added to prevent the divisor from having a value of zero.

Scoring analysis of cell-type identity

For this analysis, our goal was to score the probabilistic cell identity of each cell relative to the defined cell types at the transcriptional level (*21*). We built an L2-regularized logistic regression model, a C-support vector classification model, and a vanilla neural network model (PyTorch framework with Skorch package) for classification tasks and trained the model to learn the general prototypes of defined cell types. To train the model, we removed the cell cycle–related genes, and then computed the CV (SD divided by the mean) versus the predicted CV (estimated by a nonlinear noise model), and applied the fit of noise distribution to select the most variable features that are greater than the expected CV. SVR from the scikit-learn package was for this analysis. The overdispersed genes were further ranked by two heuristics for the cell-type specificity of both fold change and enrichment score change (*44*). For TCGA subtype classification, the originally defined TCGA subtypes were used as reference cell types, and the originally identified marker genes of the four subtypes were manually added as feature genes for training the neural network classifier. For the lineage classification based on the differential methylation sites, the defined lineages at the transcriptional level were used as the reference cell types, and the identified differential methylation sites were used as the features for training the neural network classifier. The cross-species alignment was performed as described in (*21*). To compare the data from UMI-based platforms and the Smart-seq2 platform, data were scaled by SD owing to the potentially larger gene variation in Smart-seq2 (*58*). Subsequently, the ranked marker genes of the defined cell types were log-transformed and scaled by Minmax normalization, and then used for the different learning models:

1) The L2-regularized logistic regression model was as described in (*59*).

2) To test the adequate strength of the regularization in the C-support vector classification model, the C regularization parameter and three kernel types, "linear," "sigmoid," and "rbf," were inspected via GridSearchCV. The classifier accuracy was estimated by a *k*-fold cross-validation, of which the dataset was randomly split (25% test_size). The value of the C regularization parameter and the kernel type were chosen corresponding to the maximum point of the learning curve reaching the accuracy plateaus.

3) The neural network model contains an input layer with the number of neuron nodes being the same as the number of marker genes, a hidden layer with the number of neuron nodes being the same as 20% of marker gene numbers, and an output layer with the number of neuron nodes being the same as the number of defined cell types. Linear regression was performed between each layer, and 30% of dropouts were set to reduce the overfitting. Rectified linear unit (ReLu) was used as the activation function of the hidden layer, and Softmax was used for the output layer to evaluate the probabilities. Nesterov momentum was used as a stochastic gradient descent (SGD) optimizer. To choose the adequate regularization strength, the classifier accuracy and the loss value were inspected against epoch numbers. The classifier accuracy was estimated by a *k*-fold cross-validation, of which the dataset was randomly split ($k = 3$). The learning rate, epoch number, and momentum were chosen corresponding to the maximum point of the learning curve reaching the accuracy plateaus.

4) The node-level graph neural network (GNN) model contains an input layer with the number of node features being the same as the number of marker genes, two hidden layers with the number of neuron nodes being the same as 25% of marker gene numbers, and an output layer with the number of neuron nodes being the same as the number of defined cell types. The edge indexes were selected as the top 10 nodes upon K-nearest neighboring (KNN) calculation of the top 30 principal components.

GCNConv (message passing) was performed between each layer, and 20% of dropouts were set. ReLu was used as the activation function of the hidden layer, and Softmax was used for the output layer to evaluate the probabilities. Momentum γ was set to 0.9 in the SGD optimizer. To choose the adequate regularization strength, the classifier accuracy and the loss value (CrossEntropyLoss) were inspected against epoch numbers. The learning rate, epoch number, and momentum were chosen corresponding to the maximum point of the learning curve reaching the accuracy plateaus.

We set the same learning steps for all four models and found that the learning accuracy and running period were 97.62% and 1390.83 s for the L2-regularized logistic regression model; 97.59% and 3084.81 s for the C-support vector classification model; 99.6% and 131.14 s for the vanilla neural network model; and 99.13% and 349.81 s for the node-level GNN. Thus, the ready vanilla neural network model was further used to predict the probabilities of each cell belonging to each trained reference cell type. The permutation test of dataset was applied to qualify the significance of the prediction, and the *P* value was calculated by false discovery rate. The prototype threshold of a defined cell type was determined as the larger value of significant probability ($P < 0.05$) and dominant probability (>60). If the probability of a predicted cell to one cell type is over this cell type's prototype threshold, this predicted cell was considered as "cell type defined" and was assigned to this cell type. Data were visualized in the radar plot. The radar plot consists of a sequence of equiangular polygon spokes with the distal vertex representing each trained reference cell type. The distance between the polygon center and each vertex of the polygon represents the relative probabilities of each trained reference cell assigned to the defined reference cell types. Thus, the position of each predicting cell was calculated as a linear combination of the probabilities against all reference cell types and then visualized as the relative position to all vertices of the polygon.

Deconvolution of bulk tumor RNA sequencing

A bulk tumor tissue contains both the malignant cells and various microenvironment cells that disturb the transcriptional profile of the endogenous tumor cells. In addition, the intratumor heterogeneity of glioblastoma tissue further blurs the expression matrix. To enrich/denoise the gene expression of the dominant tumor cells from glioblastoma bulk tissue, we applied the deconvolution method via the power-law transformations and the autoencoder of convolutional neural network (CNN) (*60*). The RNA-seq data of TCGA were obtained from the UCSC Cancer Browser, and our scRNA-seq data were used as the reference dataset for deconvolution. Genes in the reference dataset were prefiltered by the count frequency as described in BACKSPIN (*55*), and then used for the deconvolution of bulk tissue. Each gene was scaled by Minmax normalization and visualized by a curved line plot; the *x* axis represents the cell/sample that was sorted by the expression value of the gene. Thus, we obtained the distribution of gene expression of these datasets and visualized them in a curve line plot. The mean values of all curves were calculated for the least squares polynomial approximation via Numpy, and the square root was used as weights to find the γ value of the curve. By comparing the γ values of both bulk tissue data and reference glioblastoma single-cell data, the expression matrix of bulk sequencing was fit to the same distribution of single-cell sequencing via power-law transformations (fig. S2B, step 1).

Next, the CNN autoencoder was applied for denoising the transformed datasets. The autoencoder contains two layers of convolution and four layers of transposed convolution in the PyTorch framework. The hyperbolic tangent activation function (Tanh) was used as the activation function between each layer, and sigmoid was used for the output layer. The mean squared error between each element in the input (MSELoss) was evaluated against the epoch. The learning rate and epoch number were chosen corresponding to the minimum point of loss_value curve after reaching the loss_value plateaus (fig. S2B, step 2). After the training of the reference glioblastoma scRNA-seq data, the model was performed for the deconvolution of the transformed dataset of glioblastoma bulk tissue. The deconvoluted dataset was scaled and visualized in a curve line plot as described above for evaluation and subsequently used for further analysis.

Single-cell Weighted Assignment and Projection on developmental LINEages

The aim of SWAPLINE is to place each test cell into a trajectory position of normal developmental lineage(s), via combining both KNN and the scoring of probabilistic cell identity. The workflow is described in fig. S3D.

To construct the reference lineage trajectory, the endogenous mouse brain cell types were from developmental brain atlas (*16*) or collected from different datasets generated via the Smart-seq2 scRNA-seq platform, including adult Rgl/neural stem cells, neuroblasts (*35*), meningeal cells derived from neural crest (*34*), neural crest and neural tube cells captured from the developing embryo (*33*), and oligodendrocytes, astrocytes, and perivascular mural cells (*17*). These cell types should together represent possible endogenous brain cell types to which glioblastoma cells display similarities. Meningeal cells, embryonic neural crest cells, and perivascular mural cells theoretically belong to neural crest lineage in brain, while other cell types follow the CNS neural development. UMAP was used to build the reference plot that reflects the transcriptional relations among all reference cell types. PAGA analysis (*61*) further confirmed the lineage relations among the reference cell types. Subsequently, two steps of quantification were applied in parallel: First, we used all these reference cell types to perform the cell scoring of probabilistic similarity. Next, we divided the prototype probabilities into two groups according to the developmental lineages of the reference cell types: a neural crest lineage and a CNS neural lineage as described above. For each predicted cell type, the mean value of the prototype probabilities of the two lineage groups was used to estimate the lineage similarity of this predicted cell type, the higher lineage probability assigned, and the predicted cell type into this lineage for further lineage-specific SWAPLINE analysis. Since there are two major lineages in the reference cells during neural/neural crest development, we assigned each predicted cell type into its normal developmental lineage by referring to the top *N* ($N = 3$ or 4 here) closest reference cell type in PAGA. For each lineage, the top connected reference cell types and predicted cells were used for

probabilistic scoring. The permutation test was applied as the negative control and background noise. Second, we used KNN to evaluate the putative position of each predicted cell corresponding to every reference cell types in the UMAP. Briefly, we first calculated the top principal components of all cells following the Elbow method, and then used these principal components to access the pairwise distances of Euclidean metric among all cells. For each predicted cell, we selected the top 25 nearest cells in each reference cell type and calculated the median UMAP coordinates of these top nearest cells. Thus, we obtained the KNN putative positions of the predicted cells in each top N connected reference cell type. Furthermore, the prototype probabilistic score of each cell was normalized to the median value of randomized probabilities that were generated from the permutation test and further rescaled by Minmax. The cells with global prototype similarity (putatively low-quality cells or extremely high-plasticity cells) were excluded if one predicted cell's SD of probability among prototypes was lower than the permutation test. Subsequently, a linear combination of both KNN putative positions and cell probabilities of top N related and connected reference cell types represents the developmental trajectory position of each predicted cell: Let N be the total number of prototypes, let p_m be the probability of a cell belonging to prototype m, let c_{mj} be the coordinate of nearest neighboring cell j of the predicted cell from prototype m, and let k be the top closest constant; the predicted coordinates of test cell $\vec{a}$ upon the origin of coordinates then was defined

$$\vec{a} = \sum_{m=1}^{N} p_m \left(\frac{1}{k} \sum_{j=1}^{k} c_{\mathrm{mj}} \right)$$

Disentangling trajectory analysis of the branching tree

The principal branching tree was constructed to elucidate the fundamental lineages of glioblastoma cells via a simplified elastic principal graphs. Elastic principal graphs are a generalization of the elastic map algorithm for approximating principal manifolds from the data with a given topology (*36*). A principal manifold is an undirected graph (B) composed of nodes (N) and edges (E). The nodes are embedded into the data space by minimizing both the approximation error (mean squared distance) to the data points and the elastic energy [U^{Φ}(B)], defined as

$$U^{\Phi}(\mathrm{D},\mathrm{B}) = \frac{1}{\mathrm{Num}} \sum_{j=1}^{|N|} \sum_{Pn(i)=j} \min \{ \| D_i - \Phi(N_j) \|^2, T_r^2 \} + U^{\Phi}(\mathrm{B})$$

k-star in graph G defines a subgraph that contains $k + 1$ nodes, $n_{0,1,\ldots,k} \in \mathrm{N}$, and k edges $\{(n_0, n_i) | i = 1, \ldots, k\}$. D represents the structured data points, and Num is the number of data points. ϕ (N_j) is the map Φ: $N \rightarrow \mathrm{R}^{\mathrm{m}}$, which represents an embedding of each j node in the data space. The data point partitioning Pn was defined as $Pn(i) = \arg\min_{j = 1\ldots|\mathrm{V}|} (D_i - \phi(\mathrm{V}j))^2$, and it provides an index of a node that is the closest to the ith data point in the graph. Each iteration provides the initial guess of ϕ, the partitioning $Pn(i)$ is computed, and U ϕ (D, B) is minimized via exploring new node positions in the data space. T_r represents the trimming radius, a distance dropout parameter in the limit, of which the data points were used for graph optimization. For the comprehensive evaluation, we set T_r as infinite here. The edges among the nodes define the elastic energy, which serves as a penalty for the graph embedding. The elastic energy is manifested by two main factors: the stretching and non-equal distance of node-to-node positions [U_E^{Φ}(B), weighted by the λ] and the deviation from harmonic embedding [U_R^{Φ}(B), weighted by μ], defined as

$$U^{\Phi}(\mathrm{B}) \coloneqq U_E^{\Phi}(\mathrm{B}) + U_R^{\Phi}(\mathrm{B})$$

$$U_E^{\Phi}(\mathrm{B}) \coloneqq \sum_{E_i} \{ \lambda + \alpha(\max(2, k_{E_i(0)}, k_{E_i(1)}) - 2) \} \| \Phi(E_i(0)) - \Phi(E_i(0)) \|^2$$

$$U_R^{\Phi}(\mathrm{B}) \coloneqq \mu \sum_{S_i} \left(\Phi(S_i(0)) - \frac{1}{k_i} \sum_{j=1}^{k_i} \Phi(\mathrm{S}_i(j)) \right)^2$$

An elastic principal tree contains selected families of k-stars S_k. Each graph edge $E^{(i)}$ has two nodes $E^{(i)}(0)$ and $E^{(i)}$ (1). $S_k^{(j)}(0)$ to $S_k^{(j)}(k)$ denote the nodes of a star $S_k^{(j)}$ in the graph, and $S_k^{(j)}(0)$ represents the center node that links to all other nodes. According to the equation, the elastic energy is regulated by two weighted factors: λ, regularizing the overall length of the edges, and μ, the deviation of the star nodes from harmonic embedding. Thus, we evaluated the construction of a principal tree upon different combinations of λ and μ. Besides these two, the parameter α independently regulates the appearance of branches via perturbing the edges of higher-order star nodes. To avoid excessive branching, we use a small value (0.01) here according to the formal description. As the SWAPLINE coordinates of each glioblastoma cell represent its status within the developmental trajectory of normal brain cells, we use the SWAPLINE coordinates to perform the low-dimensional construction of the principal branching tree. To test the robustness of the principal graph, we inspected different combinations of the elastic stretching (λ; range, 0.001 to 0.02) and the deviation from harmonicity penalty (μ; range, 0.05 to 0.5). A total of 2565 rounds of the principal graph were tested and visualized. To obtain the minimum branching and the maximum elastic stretching, we chose the principal tree produced with λ = 0.01 and μ =0.2 for subsequent analyses; alternatively, the principal tree can be obtained from the PCA of the parameter tests described above. Each edge of the principal tree was smoothened by one-dimensional interpolation via the interp1d package from SciPy. In addition, the small branch with only one single link between two nodes was merged into the neighboring larger branch. Next, we used the Shapely package to project all cell dots onto the principal edge by evaluating the shortest distance at the two dimensions and adjusted the cell positions to keep the same intercellular distance along each branch. To identify the branching related genes, we separated the principal tree to five branches according to the branch point and the branch lineage. For each branch, the smoothed expression for each gene along the branch was determined by using a Gaussian filter or a generalized linear model (SciPy package). Significant branching genes were determined by three heuristics: (i) significant distribution based on the cumulative distribution function comparing the branching position and the smoothed expression, (ii) significant correlation (Spearman's) between the branching position and the smoothed expression before and after peak value, and (iii) the gene expression should fit the criteria that at least 5% of the cells express two molecular counts and at least 20% of the cells express one molecular count. All smoothed expression was normalized to the central branching point for further comparison.

Analysis of cell cycle

A list of genes has been assigned to two major phases (S and G_2-M) of the cell cycle (*9*). The significant phase activation was evaluated

by comparing the expression of phase-related genes and the expression of random genes as described in Seurat, with small modification. Briefly, the overdispersed genes of a dataset were evaluated by estimating the mean and coefficients of variation. The overdispersed cell cycle genes were selected for phase scoring, and the rest of the genes were ranked by the expression and separated into 25 intervals according to the rank. In each interval, we selected the first 50 genes for randomization and thus generate the random gene matrix. The phase scores were generated by estimating the differential mean expression of the phase genes and the randomized genes. Phase G_0-G_1 was decided if the expression of phase genes was lower than randomized values. The activation of other phases was decided by the larger value of the phase score. Thus, each cell was assigned to different phases of the cell cycle and subsequently projected to the plot.

Comprehensive RNA velocity of all glioblastoma cells on STRT-seq/STRT-seq-2i

Spliced and unspliced counts of glioblastoma cells were quantified as described by La Manno *et al.* (*40*) using the RNA velocity package, with modification for 5′ STRT-seq. We extracted the barcode and UMI with the fault tolerance of 1 base mismatch from the FASTQ file. Meanwhile, we added the first 4 bases of the transcript sequence to the original 6 bases of UMI to generate 10 new bases of UMI for each read. The barcode tag and UMI tag were defined via SAMtools (pysam). The reads were aligned by STAR using GRCh38.p12 genome assembly and processed as described previously (*55*). We calculated spliced and unspliced counts using the built-in package of Velocyto (session of "any technique-advanced use") with masking expressed repetitive elements. A total of 2451 cells were selected with the criteria of 200 unspliced molecules and 200 spliced molecules, and most variable genes were filtered with the criteria of four minimum unspliced molecules detected in a minimum of three cells. PCAs were selected according to 0.55% ratio of variance explained by each of the selected components. Data were smoothened via balanced KNN imputation with K = 500, b_sight = 4*K, b_maxl = 3*K. The variance normalizing transform was performed in log value. The time step for extrapolation is 5, and kernel scaling was set as 0.05 in calculating the transition probability to project the velocity direction on the embedding. The embedding scatter plot was forked from the branching tree plot as described above, and the branching tree plot widened along each axis for better visualization.

Extraction of core/hub glioblastoma cells via density estimation

To estimate the density of glioblastoma cells in the lineage plot, the coordinate of each cell in both scatter plot and branching tree plot was stacked vertically and applied for kernel-density estimation using Gaussian kernels. Bandwidth vector was generated via the rule of thumb of Scott. Relative density was calculated by comparing the overall density in the plot. Cells with the top 50% density were defined as hub/core cells, and the rest of the cells were defined as branch cells.

SCENIC analysis

To infer the TFs and their target gene networks, SCENIC analysis was performed according to the authors' vignette. Briefly, the TF-targeted gene sets were identified via the following criteria: first, coexpression with TFs and, second, enriched in the direct motif of the TF. Then, the regulon activities were scored and binarized to determine whether the gene sets of each regulon were significantly enriched in cells.

Quantification and statistical analysis

Statistical analysis between groups was performed using two-tailed Student's *t* test. Kaplan-Meier survival was calculated via log-rank test. Experiments were representative of at least three independent and biological replicates. Error bars in figures represent means ± SEM. *P* values were indicated in figures or marked as $*P < 0.05$ and $**P < 0.01$.

SUPPLEMENTARY MATERIALS

Supplementary material for this article is available at https://science.org/doi/10.1126/sciadv.abm6340

View/request a protocol for this paper from *Bio-protocol*.

REFERENCES AND NOTES

1. G. P. Dunn, M. L. Rinne, J. Wykosky, G. Genovese, S. N. Quayle, I. F. Dunn, P. K. Agarwalla, M. G. Chheda, B. Campos, A. Wang, C. Brennan, K. L. Ligon, F. Furnari, W. K. Cavenee, R. A. Depinho, L. Chin, W. C. Hahn, Emerging insights into the molecular and cellular basis of glioblastoma. *Genes Dev.* **26**, 756–784 (2012).
2. C. W. Brennan, R. G. W. Verhaak, A. McKenna, B. Campos, H. Noushmehr, S. R. Salama, S. Zheng, D. Chakravarty, J. Z. Sanborn, S. H. Berman, R. Beroukhim, B. Bernard, C. J. Wu, G. Genovese, I. Shmulevich, J. Barnholtz-Sloan, L. Zou, R. Vegesna, S. A. Shukla, G. Ciriello, W. K. Yung, W. Zhang, C. Sougnez, T. Mikkelsen, K. Aldape, D. D. Bigner, E. G. van Meir, M. Prados, A. Sloan, K. L. Black, J. Eschbacher, G. Finocchiaro, W. Friedman, D. W. Andrews, A. Guha, M. Iacocca, B. P. O'Neill, G. Foltz, J. Myers, D. J. Weisenberger, R. Penny, R. Kucherlapati, C. M. Perou, D. N. Hayes, R. Gibbs, M. Marra, G. B. Mills, E. Lander, P. Spellman, R. Wilson, C. Sander, J. Weinstein, M. Meyerson, S. Gabriel, P. W. Laird, D. Haussler, G. Getz, L. Chin, C. Benz, J. Barnholtz-Sloan, W. Barrett, Q. Ostrom, Y. Wolinsky, K. L. Black, B. Bose, P. T. Boulos, M. Boulos, J. Brown, C. Czerinski, M. Eppley, M. Iacocca, T. Kempista, T. Kitko, Y. Koyfman, B. Rabeno, P. Rastogi, M. Sugarman, P. Swanson, K. Yalamanchii, I. P. Otey, Y. S. Liu, Y. Xiao, J. T. Auman, P. C. Chen, A. Hadjipanayis, E. Lee, S. Lee, P. J. Park, J. Seidman, L. Yang, R. Kucherlapati, S. Kalkanis, T. Mikkelsen, L. M. Poisson, A. Raghunathan, L. Scarpace, B. Bernard, R. Bressler, A. Eakin, L. Iype, R. B. Kreisberg, K. Leinonen, S. Reynolds, H. Rovira, V. Thorsson, I. Shmulevich, M. J. Annala, R. Penny, J. Paulauskis, E. Curley, M. Hatfield, D. Mallery, S. Morris, T. Shelton, C. Shelton, M. Sherman, P. Yena, L. Cuppini, F. DiMeco, M. Eoli, G. Finocchiaro, E. Maderna, B. Pollo, M. Saini, S. Balu, K. A. Hoadley, L. Li, C. R. Miller, Y. Shi, M. D. Topal, J. Wu, G. Dunn, C. Giannini, B. P. O'Neill, B. A. Aksoy, Y. Antipin, L. Borsu, S. H. Berman, C. W. Brennan, E. Cerami, D. Chakravarty, G. Ciriello, J. Gao, B. Gross, A. Jacobsen, M. Ladanyi, A. Lash, Y. Liang, B. Reva, C. Sander, N. Schultz, R. Shen, N. D. Socci, A. Viale, M. L. Ferguson, Q. R. Chen, J. A. Demchok, L. A. L. Dillon, K. R. M. Shaw, M. Sheth, R. Tarnuzzer, Z. Wang, L. Yang, T. Davidsen, M. S. Guyer, B. A. Ozenberger, H. J. Sofia, J. Bergsten, J. Eckman, J. Harr, J. Myers, C. Smith, K. Tucker, C. Winemiller, L. A. Zach, J. Y. Ljubimova, G. Eley, B. Ayala, M. A. Jensen, A. Kahn, T. D. Pihl, D. A. Pot, Y. Wan, J. Eschbacher, G. Foltz, N. Hansen, P. Hothi, B. Lin, N. Shah, J. G. Yoon, C. Lau, M. Berens, K. Ardlie, R. Beroukhim, S. L. Carter, A. D. Cherniack, M. Noble, J. Cho, K. Cibulskis, D. DiCara, S. Frazer, S. B. Gabriel, N. Gehlenborg, J. Gentry, D. Heiman, J. Kim, R. Jing, E. S. Lander, M. Lawrence, P. Lin, W. Mallard, M. Meyerson, R. C. Onofrio, G. Saksena, S. Schumacher, C. Sougnez, P. Stojanov, B. Tabak, D. Voet, H. Zhang, L. Zou, G. Getz, N. N. Dees, L. Ding, L. L. Fulton, R. S. Fulton, K. L. Kanchi, E. R. Mardis, R. K. Wilson, S. B. Baylin, D. W. Andrews, L. Harshyne, M. L. Cohen, K. Devine, A. E. Sloan, S. R. VandenBerg, M. S. Berger, M. Prados, D. Carlin, B. Craft, K. Ellrott, M. Goldman, T. Goldstein, M. Grifford, D. Haussler, S. Ma, S. Ng, S. R. Salama, J. Z. Sanborn, J. Stuart, T. Swatloski, P. Waltman, J. Zhu, R. Foss, B. Frentzen, W. Friedman, R. McTiernan, A. Yachnis, D. N. Hayes, C. M. Perou, S. Zheng, R. Vegesna, Y. Mao, R. Akbani, K. Aldape, O. Bogler, G. N. Fuller, W. Liu, Y. Liu, Y. Lu, G. Mills, A. Protopopov, X. Ren, Y. Sun, C. J. Wu, W. K. A. Yung, W. Zhang, J. Zhang, K. Chen, J. N. Weinstein, L. Chin, R. G. W. Verhaak, H. Noushmehr, D. J. Weisenberger, M. S. Bootwalla, P. H. Lai, T. J. Triche Jr., D. J. van den Berg, P. W. Laird, D. H. Gutmann, N. L. Lehman, E. G. VanMeir, D. Brat, J. J. Olson, G. M. Mastrogianakis, N. S. Devi, Z. Zhang, D. Bigner, E. Lipp, R. McLendon, The somatic genomic landscape of glioblastoma. *Cell* **155**, 462–477 (2013).
3. D. Sturm, H. Witt, V. Hovestadt, D. A. Khuong-Quang, D. T. W. Jones, C. Konermann, E. Pfaff, M. Tönjes, M. Sill, S. Bender, M. Kool, M. Zapatka, N. Becker, M. Zucknick, T. Hielscher, X. Y. Liu, A. M. Fontebasso, M. Ryzhova, S. Albrecht, K. Jacob, M. Wolter, M. Ebinger, M. U. Schuhmann, T. van Meter, M. C. Frühwald, H. Hauch, A. Pekrun, B. Radlwimmer, T. Niehues, G. von Komorowski, M. Dürken, A. E. Kulozik, J. Madden,

A. Donson, N. K. Foreman, R. Drissi, M. Fouladi, W. Scheurlen, A. von Deimling, C. Monoranu, W. Roggendorf, C. Herold-Mende, A. Unterberg, C. M. Kramm, J. Felsberg, C. Hartmann, B. Wiestler, W. Wick, T. Milde, O. Witt, A. M. Lindroth, J. Schwartzentruber, D. Faury, A. Fleming, M. Zakrzewska, P. P. Liberski, K. Zakrzewski, P. Hauser, M. Garami, A. Klekner, L. Bognar, S. Morrissy, F. Cavalli, M. D. Taylor, P. van Sluis, J. Koster, R. Versteeg, R. Volckmann, T. Mikkelsen, K. Aldape, G. Reifenberger, V. P. Collins, J. Majewski, A. Korshunov, P. Lichter, C. Plass, N. Jabado, S. M. Pfister, Hotspot mutations in H3F3A and IDH1 define distinct epigenetic and biological subgroups of glioblastoma. *Cancer Cell* **22**, 425–437 (2012).

4. Q. Wang, B. Hu, X. Hu, H. Kim, M. Squatrito, L. Scarpace, A. C. deCarvalho, S. Lyu, P. Li, Y. Li, F. Barthel, H. J. Cho, Y.-H. Lin, N. Satani, E. Martinez-Ledesma, S. Zheng, E. Chang, C.-E. G. Sauvé, A. Olar, Z. D. Lan, G. Finocchiaro, J. J. Phillips, M. S. Berger, K. R. Gabrusiewicz, G. Wang, E. Eskilsson, J. Hu, T. Mikkelsen, R. A. De Pinho, F. Muller, A. B. Heimberger, E. P. Sulman, D.-H. Nam, R. G. W. Verhaak, Tumor evolution of glioma-intrinsic gene expression subtypes associates with immunological changes in the microenvironment. *Cancer Cell* **32**, 42–56.e6 (2017).
5. C. Neftel, J. Laffy, M. G. Filbin, T. Hara, M. E. Shore, G. J. Rahme, A. R. Richman, D. Silverbush, M. L. Shaw, C. M. Hebert, J. Dewitt, S. Gritsch, E. M. Perez, L. N. G. Castro, X. Lan, N. Druck, C. Rodman, D. Dionne, A. Kaplan, M. S. Bertalan, J. Small, K. Pelton, S. Becker, D. Bonal, Q.-D. Nguyen, R. L. Servis, J. M. Fung, R. Mylvaganam, L. Mayr, J. Gojo, C. Haberler, R. Geyeregger, T. Czech, I. Slavc, B. V. Nahed, W. T. Curry, B. S. Carter, H. Wakimoto, P. K. Brastianos, T. T. Batchelor, A. Stemmer-Rachamimov, M. Martinez-Lage, M. P. Frosch, I. Stamenkovic, N. Riggi, E. Rheinbay, M. Monje, O. Rozenblatt-Rosen, D. P. Cahill, A. P. Patel, T. Hunter, I. M. Verma, K. L. Ligon, D. N. Louis, A. Regev, B. E. Bernstein, I. Tirosh, M. L. Suvà, An integrative model of cellular states, plasticity, and genetics for glioblastoma. *Cell* **178**, 835–849.e821 (2019).
6. L. F. Parada, P. B. Dirks, R. J. Wechsler-Reya, Brain tumor stem cells remain in play. *J. Clin. Oncol.* **35**, 2428–2431 (2017).
7. C. P. Couturier, S. Ayyadhury, P. U. le, J. Nadaf, J. Monlong, G. Riva, R. Allache, S. Baig, X. Yan, M. Bourgey, C. Lee, Y. C. D. Wang, V. Wee Yong, M. C. Guiot, H. Najafabadi, B. Misic, J. Antel, G. Bourque, J. Ragoussis, K. Petrecca, Single-cell RNA-seq reveals that glioblastoma recapitulates a normal neurodevelopmental hierarchy. *Nat. Commun.* **11**, 3406 (2020).
8. Q. Weng, J. Wang, J. Wang, D. He, Z. Cheng, F. Zhang, R. Verma, L. Xu, X. Dong, Y. Liao, X. He, A. Potter, L. Zhang, C. Zhao, M. Xin, Q. Zhou, B. J. Aronow, P. J. Blackshear, J. N. Rich, Q. He, W. Zhou, M. L. Suvà, R. R. Waclaw, S. S. Potter, G. Yu, Q. R. Lu, Single-cell transcriptomics uncovers glial progenitor diversity and cell fate determinants during development and gliomagenesis. *Cell Stem Cell* **24**, 707–723.e8 (2019).
9. I. Tirosh, A. S. Venteicher, C. Hebert, L. E. Escalante, A. P. Patel, K. Yizhak, J. M. Fisher, C. Rodman, C. Mount, M. G. Filbin, C. Neftel, N. Desai, J. Nyman, B. Izar, C. C. Luo, J. M. Francis, A. A. Patel, M. L. Onozato, N. Riggi, K. J. Livak, D. Gennert, R. Satija, B. V. Nahed, W. T. Curry, R. L. Martuza, R. Mylvaganam, A. J. Iafrate, M. P. Frosch, T. R. Golub, M. N. Rivera, G. Getz, O. Rozenblatt-Rosen, D. P. Cahill, M. Monje, B. E. Bernstein, D. N. Louis, A. Regev, M. L. Suvà, Single-cell RNA-seq supports a developmental hierarchy in human oligodendroglioma. *Nature* **539**, 309–313 (2016).
10. A. Bhaduri, E. D. Lullo, D. Jung, S. Müller, E. E. Crouch, C. S. Espinosa, T. Ozawa, B. Alvarado, J. Spatazza, C. R. Cadwell, G. Wilkins, D. Velmeshev, S. J. Liu, M. Malatesta, M. G. Andrews, M. A. Mostajo-Radji, E. J. Huang, T. J. Nowakowski, D. A. Lim, A. Diaz, D. R. Raleigh, A. R. Kriegstein, Outer radial glia-like cancer stem cells contribute to heterogeneity of glioblastoma. *Cell Stem Cell* **26**, 48–63.e6 (2020).
11. S. Alcantara Llaguno, D. Sun, A. M. Pedraza, E. Vera, Z. Wang, D. K. Burns, L. F. Parada, Cell-of-origin susceptibility to glioblastoma formation declines with neural lineage restriction. *Nat. Neurosci.* **22**, 545–555 (2019).
12. Z. Wang, D. Sun, Y.-J. Chen, X. Xie, Y. Shi, V. Tabar, C. W. Brennan, T. A. Bale, C. D. Jayewickreme, D. R. Laks, S. A. Llaguno, L. F. Parada, Cell lineage-based stratification for Glioblastoma. *Cancer Cell* **38**, 366–379.e8 (2020).
13. Y. Kim, F. S. Varn, S. H. Park, B. W. Yoon, H. R. Park, C. Lee, R. G. W. Verhaak, S. H. Paek, Perspective of mesenchymal transformation in glioblastoma. *Acta Neuropathol. Commun.* **9**, 50 (2021).
14. H. C. Etchevers, C. Vincent, N. M. Le Douarin, G. F. Couly, The cephalic neural crest provides pericytes and smooth muscle cells to all blood vessels of the face and forebrain. *Development* **128**, 1059–1068 (2001).
15. K. Ando, S. Fukuhara, N. Izumi, H. Nakajima, H. Fukui, R. N. Kelsh, N. Mochizuki, Clarification of mural cell coverage of vascular endothelial cells by live imaging of zebrafish. *Development* **143**, 1328–1339 (2016).
16. G. La Manno, K. Siletti, A. Furlan, D. Gyllborg, E. Vinsland, A. M. Albiach, C. M. Langseth, I. Khven, A. R. Lederer, L. M. Dratva, A. Johnsson, M. Nilsson, P. Lönnerberg, S. Linnarsson, Molecular architecture of the developing mouse brain. *Nature* **596**, 92–96 (2021).
17. M. Vanlandewijck, L. He, M. A. Mäe, J. Andrae, K. Ando, F. del Gaudio, K. Nahar, T. Lebouvier, B. Laviña, L. Gouveia, Y. Sun, E. Raschperger, M. Räsänen, Y. Zarb, N. Mochizuki, A. Keller, U. Lendahl, C. Betsholtz, A molecular atlas of cell types and zonation in the brain vasculature. *Nature* **554**, 475–480 (2018).
18. M. Mravic, G. Asatrian, C. Soo, C. Lugassy, R. L. Barnhill, S. M. Dry, B. Peault, A. W. James, From pericytes to perivascular tumours: Correlation between pathology, stem cell biology, and tissue engineering. *Int. Orthop.* **38**, 1819–1824 (2014).
19. K. Ando, W. Wang, D. Peng, A. Chiba, A. K. Lagendijk, L. Barske, J. G. Crump, D. Y. R. Stainier, U. Lendahl, K. Koltowska, B. M. Hogan, S. Fukuhara, N. Mochizuki, C. Betsholtz, Peri-arterial specification of vascular mural cells from naïve mesenchyme requires Notch signaling. *Development* **146**, dev165589 (2019).
20. H. Hochgerner, A. Zeisel, P. Lönnerberg, S. Linnarsson, Conserved properties of dentate gyrus neurogenesis across postnatal development revealed by single-cell RNA sequencing. *Nat. Neurosci.* **21**, 290–299 (2018).
21. G. La Manno, D. Gyllborg, S. Codeluppi, K. Nishimura, C. Salto, A. Zeisel, L. E. Borm, S. R. W. Stott, E. M. Toledo, J. C. Villaescusa, P. Lönnerberg, J. Ryge, R. A. Barker, E. Arenas, S. Linnarsson, Molecular diversity of midbrain development in mouse, human, and stem cells. *Cell* **167**, 566–580.e19 (2016).
22. R. D. Hodge, T. E. Bakken, J. A. Miller, K. A. Smith, E. R. Barkan, L. T. Graybuck, J. L. Close, B. Long, N. Johansen, O. Penn, Z. Yao, J. Eggermont, T. Höllt, B. P. Levi, S. I. Shehata, B. Aevermann, A. Beller, D. Bertagnolli, K. Brouner, T. Casper, C. Cobbs, R. Dalley, N. Dee, S. L. Ding, R. G. Ellenbogen, O. Fong, E. Garren, J. Goldy, R. P. Gwinn, D. Hirschstein, C. D. Keene, M. Keshk, A. L. Ko, K. Lathia, A. Mahfouz, Z. Maltzer, M. McGraw, T. N. Nguyen, J. Nyhus, J. G. Ojemann, A. Oldre, S. Parry, S. Reynolds, C. Rimorin, N. V. Shapovalova, S. Somasundaram, A. Szafer, E. R. Thomsen, M. Tieu, G. Quon, R. H. Scheuermann, R. Yuste, S. M. Sunkin, B. Lelieveldt, D. Feng, L. Ng, A. Bernard, M. Hawrylycz, J. W. Phillips, B. Tasic, H. Zeng, A. R. Jones, C. Koch, E. S. Lein, Conserved cell types with divergent features in human versus mouse cortex. *Nature* **573**, 61–68 (2019).
23. X. Fan, Y. Fu, X. Zhou, L. Sun, M. Yang, M. Wang, R. Chen, Q. Wu, J. Yong, J. Dong, L. Wen, J. Qiao, X. Wang, F. Tang, Single-cell transcriptome analysis reveals cell lineage specification in temporal-spatial patterns in human cortical development. *Sci. Adv.* **6**, eaaz2978 (2020).
24. S. Darmanis, S. A. Sloan, D. Croote, M. Mignardi, S. Chernikova, P. Samghababi, Y. Zhang, N. Neff, M. Kowarsky, C. Caneda, G. Li, S. D. Chang, I. D. Connolly, Y. Li, B. A. Barres, M. H. Gephart, S. R. Quake, Single-cell RNA-seq analysis of infiltrating neoplastic cells at the migrating front of human glioblastoma. *Cell Rep.* **21**, 1399–1410 (2017).
25. A. S. Venteicher, I. Tirosh, C. Hebert, K. Yizhak, C. Neftel, M. G. Filbin, V. Hovestadt, L. E. Escalante, M. K. L. Shaw, C. Rodman, S. M. Gillespie, D. Dionne, C. C. Luo, H. Ravichandran, R. Mylvaganam, C. Mount, M. L. Onozato, B. V. Nahed, H. Wakimoto, W. T. Curry, A. J. Iafrate, M. N. Rivera, M. P. Frosch, T. R. Golub, P. K. Brastianos, G. Getz, A. P. Patel, M. Monje, D. P. Cahill, O. Rozenblatt-Rosen, D. N. Louis, B. E. Bernstein, A. Regev, M. L. Suvà, Decoupling genetics, lineages, and microenvironment in IDH-mutant gliomas by single-cell RNA-seq. *Science* **355**, eaai8478 (2017).
26. J. Yuan, H. M. Levitin, V. Frattini, E. C. Bush, D. M. Boyett, J. Samanamud, M. Ceccarelli, A. Dovas, G. Zanazzi, P. Canoll, J. N. Bruce, A. Lasorella, A. Iavarone, P. A. Sims, Single-cell transcriptome analysis of lineage diversity in high-grade glioma. *Genome Med.* **10**, 57 (2018).
27. T. Hara, R. Chanoch-Myers, N. D. Mathewson, C. Myskiw, L. Atta, L. Bussema, S. W. Eichhorn, A. C. Greenwald, G. S. Kinker, C. Rodman, L. N. G. Castro, H. Wakimoto, O. Rozenblatt-Rosen, X. Zhuang, J. Fan, T. Hunter, I. M. Verma, K. W. Wucherpfennig, A. Regev, M. L. Suvà, I. Tirosh, Interactions between cancer cells and immune cells drive transitions to mesenchymal-like states in glioblastoma. *Cancer Cell* **39**, 779–792.e11 (2021).
28. K. Natesh, D. Bhosale, A. Desai, G. Chandrika, R. Pujari, J. Jagtap, A. Chugh, D. Ranade, P. Shastry, Oncostatin-M differentially regulates mesenchymal and proneural signature genes in gliomas via STAT3 signaling. *Neoplasia* **17**, 225–237 (2015).
29. T. Lu, C. M. Costello, P. J. P. Croucher, R. Häsler, G. Deuschl, S. Schreiber, Can Zipf's law be adapted to normalize microarrays? *BMC Bioinformatics* **6**, 37 (2005).
30. P. Vincent, H. Larochelle, Y. Bengio, P.-A. Manzagol, Extracting and composing robust features with denoising autoencoders, in *Proceedings of the 25th International Conference on Machine Learning* (ACM, 2008), pp. 1096–1103.
31. L. A. Doyle, M. Vivero, C. D. Fletcher, F. Mertens, J. L. Hornick, Nuclear expression of STAT6 distinguishes solitary fibrous tumor from histologic mimics. *Mod. Pathol.* **27**, 390–395 (2014).
32. Z. Zhao, K.-N. Zhang, Q. Wang, G. Li, F. Zeng, Y. Zhang, F. Wu, R. Chai, Z. Wang, C. Zhang, W. Zhang, Z. Bao, T. Jiang, Chinese Glioma Genome Atlas (CGGA): A comprehensive resource with functional genomic data for Chinese Glioma patients. *Genomics Proteomic Bioinformatics* **19**, 1–12 (2021).
33. R. Soldatov, M. Kaucka, M. E. Kastriti, J. Petersen, T. Chontorotzea, L. Englmaier, N. Akkuratova, Y. Yang, M. Häring, V. Dyachuk, C. Bock, M. Farlik, M. L. Piacentino, F. Boismoreau, M. M. Hilscher, C. Yokota, X. Qian, M. Nilsson, M. E. Bronner, L. Croci, W. Y. Hsiao, D. A. Guertin, J. F. Brunet, G. G. Consalez, P. Ernfors, K. Fried, P. V. Kharchenko, I. Adameyko, Spatiotemporal structure of cell fate decisions in murine neural crest. *Science* **364**, (2019).
34. F. Bifari, I. Decimo, A. Pino, E. Llorens-Bobadilla, S. Zhao, C. Lange, G. Panuccio, B. Boeckx, B. Thienpont, S. Vinckier, S. Wyns, A. Bouché, D. Lambrechts, M. Giugliano, M. Dewerchin,

A. Martin-Villalba, P. Carmeliet, Neurogenic radial glia-like cells in meninges migrate and differentiate into functionally integrated neurons in the neonatal cortex. *Cell Stem Cell* **20**, 360–373.e7 (2017).
35. E. Llorens-Bobadilla, S. Zhao, A. Baser, G. Saiz-Castro, K. Zwadlo, A. Martin-Villalba, Single-cell transcriptomics reveals a population of dormant neural stem cells that become activated upon brain injury. *Cell Stem Cell* **17**, 329–340 (2015).
36. L. Albergante, E. Mirkes, J. Bac, H. Chen, A. Martin, L. Faure, E. Barillot, L. Pinello, A. Gorban, A. Zinovyev, Robust and scalable learning of complex intrinsic dataset geometry via ElPiGraph. *Entropy Switz* **22**, 296 (2020).
37. R. Kalluri, R. A. Weinberg, The basics of epithelial-mesenchymal transition. *J. Clin. Invest.* **119**, 1420–1428 (2009).
38. N. Grillet, V. Dubreuil, H. D. Dufour, J.-F. Brunet, Dynamic expression of RGS4 in the developing nervous system and regulation by the neural type-specific transcription factor Phox2b. *J. Neurosci.* **23**, 10613–10621 (2003).
39. M. Jakoby, A. Schnittger, Cell cycle and differentiation. *Curr. Opin. Plant Biol.* **7**, 661–669 (2004).
40. G. La Manno, R. Soldatov, A. Zeisel, E. Braun, H. Hochgerner, V. Petukhov, K. Lidschreiber, M. E. Kastriti, P. Lönnerberg, A. Furlan, J. Fan, L. E. Borm, Z. Liu, D. van Bruggen, J. Guo, X. He, R. Barker, E. Sundström, G. Castelo-Branco, P. Cramer, I. Adameyko, S. Linnarsson, P. V. Kharchenko, RNA velocity of single cells. *Nature* **560**, 494–498 (2018).
41. U. C. Eze, A. Bhaduri, M. Haeussler, T. J. Nowakowski, A. R. Kriegstein, Single-cell atlas of early human brain development highlights heterogeneity of human neuroepithelial cells and early radial glia. *Nat. Neurosci.* **24**, 584–594 (2021).
42. M. L. Suva, E. Rheinbay, S. M. Gillespie, A. P. Patel, H. Wakimoto, S. D. Rabkin, N. Riggi, A. S. Chi, D. P. Cahill, B. V. Nahed, W. T. Curry, R. L. Martuza, M. N. Rivera, N. Rossetti, S. Kasif, S. Beik, S. Kadri, I. Tirosh, I. Wortman, A. K. Shalek, O. Rozenblatt-Rosen, A. Regev, D. N. Louis, B. E. Bernstein, Reconstructing and reprogramming the tumor-propagating potential of glioblastoma stem-like cells. *Cell* **157**, 580–594 (2014).
43. E. N. Schock, C. LaBonne, Sorting Sox: Diverse roles for sox transcription factors during neural crest and craniofacial development. *Front. Physiol.* **11**, 606889 (2020).
44. A. Zeisel, H. Hochgerner, P. Lönnerberg, A. Johnsson, F. Memic, J. van der Zwan, M. Häring, E. Braun, L. E. Borm, G. L. Manno, S. Codeluppi, A. Furlan, K. Lee, N. Skene, K. D. Harris, J. Hjerling-Leffler, E. Arenas, P. Ernfors, U. Marklund, S. Linnarsson, Molecular architecture of the mouse nervous system. *Cell* **174**, 999–1014.e22 (2018).
45. M. Hellström, M. Kalen, P. Lindahl, A. Abramsson, C. Betsholtz, Role of PDGF-B and PDGFR-beta in recruitment of vascular smooth muscle cells and pericytes during embryonic blood vessel formation in the mouse. *Development* **126**, 3047–3055 (1999).
46. R. G. Weber, J. Boström, M. Wolter, M. Baudis, V. P. Collins, G. Reifenberger, P. Lichter, Analysis of genomic alterations in benign, atypical, and anaplastic meningiomas: Toward a genetic model of meningioma progression. *Proc. Natl. Acad. Sci. U.S.A.* **94**, 14719–14724 (1997).
47. M. L. Suva, I. Tirosh, Single-cell RNA sequencing in cancer: Lessons learned and emerging challenges. *Mol. Cell* **75**, 7–12 (2019).
48. C. Farace, J. A. Oliver, C. Melguizo, P. Alvarez, P. Bandiera, A. R. Rama, G. Malaguarnera, R. Ortiz, R. Madeddu, J. Prados, Microenvironmental modulation of decorin and lumican in temozolomide-resistant glioblastoma and neuroblastoma cancer stem-like cells. *PLOS ONE* **10**, e0134111 (2015).
49. A. Balbous, U. Cortes, K. Guilloteau, C. Villalva, S. Flamant, A. Gaillard, S. Milin, M. Wager, N. Sorel, J. Guilhot, A. Bennaceur-Griscelli, A. Turhan, J. C. Chomel, L. Karayan-Tapon, A mesenchymal glioma stem cell profile is related to clinical outcome. *Oncogenesis* **3**, e91 (2014).
50. M. L. Suva, I. Tirosh, The glioma stem cell model in the era of single-cell genomics. *Cancer Cell* **37**, 630–636 (2020).
51. L. H. Pevny, S. K. Nicolis, Sox2 roles in neural stem cells. *Int. J. Biochem. Cell Biol.* **42**, 421–424 (2010).
52. M. Karow, J. G. Camp, S. Falk, T. Gerber, A. Pataskar, M. Gac-Santel, J. Kageyama, A. Brazovskaja, A. Garding, W. Fan, T. Riedemann, A. Casamassa, A. Smiyakin, C. Schichor, M. Götz, V. K. Tiwari, B. Treutlein, B. Berninger, Direct pericyte-to-neuron reprogramming via unfolding of a neural stem cell-like program. *Nat. Neurosci.* **21**, 932–940 (2018).
53. Y. Xie, T. Bergström, Y. Jiang, P. Johansson, V. D. Marinescu, N. Lindberg, A. Segerman, G. Wicher, M. Niklasson, S. Baskaran, S. Sreedharan, I. Everlien, M. Kastemar, A. Hermansson, L. Elfineh, S. Libard, E. C. Holland, G. Hesselager, I. Alafuzoff, B. Westermark, S. Nelander, K. Forsberg-Nilsson, L. Uhrbom, The human glioblastoma cell culture resource: Validated cell models representing all molecular subtypes. *EBioMedicine* **2**, 1351–1363 (2015).
54. Y. Jiang, V. D. Marinescu, Y. Xie, M. Jarvius, N. P. Maturi, C. Haglund, S. Olofsson, N. Lindberg, T. Olofsson, C. Leijonmarck, G. Hesselager, I. Alafuzoff, M. Fryknäs, R. Larsson, S. Nelander, L. Uhrbom, Glioblastoma cell malignancy and drug sensitivity are affected by the cell of origin. *Cell Rep.* **18**, 977–990 (2017).
55. A. Zeisel, A. B. Muñoz-Manchado, S. Codeluppi, P. Lönnerberg, G. la Manno, A. Juréus, S. Marques, H. Munguba, L. He, C. Betsholtz, C. Rolny, G. Castelo-Branco, J. Hjerling-Leffler, S. Linnarsson, Cell types in the mouse cortex and hippocampus revealed by single-cell RNA-seq. *Science* **347**, 1138–1142 (2015).
56. S. Muller, A. Cho, S. J. Liu, D. A. Lim, A. Diaz, CONICS integrates scRNA-seq with DNA sequencing to map gene expression to tumor sub-clones. *Bioinformatics* **34**, 3217–3219 (2018).
57. T. Stuart, A. Butler, P. Hoffman, C. Hafemeister, E. Papalexi, W. M. Mauck III, Y. Hao, M. Stoeckius, P. Smibert, R. Satija, Comprehensive integration of single-cell data. *Cell* **177**, 1888–1902.e21 (2019).
58. C. Ziegenhain, B. Vieth, S. Parekh, B. Reinius, A. Guillaumet-Adkins, M. Smets, H. Leonhardt, H. Heyn, I. Hellmann, W. Enard, Comparative analysis of single-cell RNA sequencing methods. *Mol. Cell* **65**, 631–643.e4 (2017).
59. B. He, P. Chen, S. Zambrano, D. Dabaghie, Y. Hu, K. Möller-Hackbarth, D. Unnersjö-Jess, G. G. Korkut, E. Charrin, M. Jeansson, M. Bintanel-Morcillo, A. Witasp, L. Wennberg, A. Wernerson, B. Schermer, T. Benzing, P. Ernfors, C. Betsholtz, M. Lal, R. Sandberg, J. Patrakka, Single-cell RNA sequencing reveals the mesangial identity and species diversity of glomerular cell transcriptomes. *Nat. Commun.* **12**, 2141 (2021).
60. Y. Yuan, Z. Bar-Joseph, Deep learning for inferring gene relationships from single-cell expression data. *Proc. Natl. Acad. Sci. U.S.A.* **116**, 27151–27158 (2019).
61. F. A. Wolf, F. K. Hamey, M. Plass, J. Solana, J. S. Dahlin, B. Göttgens, N. Rajewsky, L. Simon, F. J. Theis, PAGA: Graph abstraction reconciles clustering with trajectory inference through a topology preserving map of single cells. *Genome Biol.* **20**, 59 (2019).

Acknowledgments: We thank S. Linnarsson (Karolinska Institute, Sweden) and G. La Manno (École Polytechnique Fédérale de Lausanne, Switzerland) for the technical support and the discussion; P. Lönnerberg for technical assistance; and D. Usoskin for help on animal work. We thank Science for Life Laboratory, the National Genomics Infrastructure funded by the Swedish Research Council, and Uppsala Multidisciplinary Center for Advanced Computational Science for providing assistance in massively parallel sequencing and access to the UPPMAX computational infrastructure. **Funding:** This research was funded by the Swedish Medical Research Council, the Knut and Alice Wallenberg Foundation (Wallenberg Scholar), the Swedish Cancer Society 18 0635 to P.E., and Swedish Society for Medical Research (SSMF) fellowship to Y.H. **Author contributions:** P.E. supervised the study. P.E., Y.H., and Y.J. designed the overall study. O.P. and M.S. provided samples and clinical annotation and reviewed the clinical data. I.A. coordinated the data acquisition. Y.H., M.M.R., S.O., and A.D.S. performed and interpreted the computational analyses. Y.J., Y.H., C.K., M.H., and N.S. performed and analyzed the in vitro experiments. Y.J., Y.H., J.B., M.-D.Z., and D.L. performed the in vivo experiments. P.E., Y.H., and Y.J. interpreted the data and wrote the manuscript. All authors reviewed and approved the final manuscript. **Competing interests:** The authors declare that they have no competing interests. **Data and materials availability:** The accession number for the sequencing data reported in this paper is GEO (https://www.ncbi.nlm.nih.gov/geo): GSE159416 and GSE171287. Codes and Jupyter notebooks showing key steps of the analysis are publicly available in GitHub (https://github.com/ernforslab/Hu-et-al._GBMlineage2022) and Zenodo (https://doi.org/10.5281/zenodo.6321370). All data needed to evaluate the conclusions in the paper are present in the paper and/or the Supplementary Materials.

Submitted 30 September 2021
Accepted 20 April 2022
Published 8 June 2022
10.1126/sciadv.abm6340

4.5. RNetDys: identification of disease-related impaired regulatory interactions due to single nucleotide polymorphisms

4.5.1. Preface

Gene expression is regulated by complex mechanisms at the transcriptomic and epigenomic level. The accessibility of promoter and enhancer regions is controlled by chromatin conformation which may restrict the binding of TFs to these regulatory regions and the recruitment of the transcriptional machinery. The dysregulation of these mechanisms has been linked to disease onset and development.

SNPs have been characterized as genetic risk factors due to their association to disease-related genes. However, how SNPs affect the regulation of gene expression and promote disease development is still unclear. The majority of the characterized SNPs has been found in enhancers which are described to have a cell type-specific function. Providing additional mechanistic insights regarding the impact of SNPs in the impairment of regulatory interactions would allow for a deeper understanding of disease development and advance gene therapy approaches.

We developed RNetDys, a systematic pipeline based on multi-omics data that characterizes impaired regulatory mechanisms due to disease-associated SNPs. This pipeline combines scRNA-seq, scATAC-seq, and prior knowledge data to infer cell (sub)type specific GRNs. We determine impaired regulatory interactions by mapping disease-associated SNPs to the regulatory regions in the inferred GRN. We collected SNPs associated to five diseases and validated the relevance of our results by cross-referencing with GWAS, eQTL and literature-based evidence. RNetDys is a user-friendly systematic pipeline that leverages multi-omics data to identify regulatory mechanisms impaired due to disease-associated SNPs, providing promising targets for the development of gene therapies.

In this study, I collect the datasets used for benchmarking, performed data analysis for the generation of healthy cell (sub)type specific GRNs, curated the table reporting the impaired regulatory interactions in each of the case studies, and performed the validation of impaired interactions (Figure 4, Figure S2-7, and Table S2-4 of the manuscript).

4.5.2. Preprint

RNetDys: identification of disease-related impaired regulatory interactions due to SNPs

Céline Barlier[1], Mariana Messias Ribeiro[1], Sascha Jung[2], Antonio del Sol[1,2,3,*]

[1] Luxembourg Centre for Systems Biomedicine (LCSB), University of Luxembourg, L-4362 Esch-sur-Alzette, Luxembourg

[2] CIC bioGUNE-BRTA (Basque Research and Technology Alliance), Bizkaia Technology Park, 48160 Derio, Spain

[3] IKERBASQUE, Basque Foundation for Science, Bilbao, 48012 Bilbao, Spain

[*] To whom correspondence should be addressed: Antonio.delsol@uni.lu

Abstract

Motivation: The dysregulation of regulatory mechanisms due to Single Nucleotide Polymorphisms (SNPs) can lead to diseases and does not affect all cell (sub)types equally. Current approaches to study the impact of SNPs in diseases lack mechanistic insights. Indeed, they do not account for the regulatory landscape to decipher cell (sub)type specific regulatory interactions impaired due to disease-related SNPs. Therefore, characterizing the impact of disease-related SNPs in cell (sub)type specific regulatory mechanisms would provide novel therapeutical targets, such as promoter and enhancer regions, for the development of gene-based therapies directed at preventing or treating diseases.

Results: We present RNetDys, a pipeline to decipher cell (sub)type specific regulatory interactions impaired by disease-related SNPs based on multi-OMICS data. RNetDys leverages the information obtained from the generated cell (sub)type specific GRNs to provide detailed information on impaired regulatory elements and their regulated genes due to the presence of SNPs. We applied RNetDys in five disease cases to study the cell (sub)type differential impairment due to SNPs and leveraged the GRN information to guide the

characterization of dysregulated mechanisms. We were able to validate the relevance of the identified impaired regulatory interactions by verifying their connection to disease-related genes. In addition, we showed that RNetDys identifies more precisely dysregulated interactions linked to disease-related genes than expression Quantitative Trait Loci (eQTL) and provides additional mechanistic insights.
Availability: RNetDys is a pipeline available at https://github.com/BarlierC/RNetDys.git
Contact: Antonio.delsol@uni.lu

Introduction

Gene regulation is largely controlled by the binding of transcription factors (TFs) to regulatory elements, such as promoters and enhancers, to control cell (sub)type specific functions. Notably, it has been shown that most of these functions are strongly regulated by enhancer activity (Latchman, 2011; Andersson et al., 2014). Therefore, the impairment of the regulatory interactions between TFs and enhancers of regulated genes can lead to dysregulations that trigger pathological gene expression changes that contribute to disease development (Lee and Young, 2013). In that regard, Single Nucleotide Polymorphisms (SNPs) have been shown to be associated with regulatory dysregulations driving complex diseases, such as diabetes and Alzheimer's disease (AD) (Hiramoto et al., 2015; Akhlaghipour et al., 2022). Standard approaches such as Genome-Wide Association Studies (GWAS) and expression Quantitative Trait Loci (eQTLs) have been used to study the association between SNPs and genes (Visscher et al., 2017; Bryois et al., 2022; Gazal et al., 2022). In particular, GWAS successfully deciphered thousands of disease-related SNPs (Claringbould and Zaugg, 2021). GWAS showed that the majority of these SNPs were found in non-coding regions, particularly in enhancer regions, and thus were most likely involved in gene regulation (Nica and Dermitzakis, 2013). Moreover, eQTLs have been useful to provide further insights in understanding the influence of SNPs in diseases by associating them to their target genes, based on the statistical association of gene expression variation to these genetic polymorphisms (Jeng et al., 2020). However, these approaches only provide information on SNP-gene relationships. Leveraging multi-OMICS data to construct and exploit the regulatory landscape in order to gather additional mechanistic insights would significantly contribute to a better understanding of the impact of disease-related SNPs on gene regulation and disease development. Notably, GRNs have been widely used to gain

insights into diseases (Emmert-Streib et al., 2014; Ament et al., 2018; Bakker et al., 2021) but the characterization of underlying regulatory mechanisms dysregulated due to SNPs and the cell (sub)types specifically impaired remains elusive. The resolution of cell (sub)type specific regulatory mechanisms impaired due to SNPs in disease would provide additional mechanistic insights and pave the way towards the development of gene-based therapies for disease prevention and treatment (Uddin et al., 2020).

We present RNetDys, a multi-OMICS pipeline that identifies impaired regulatory mechanisms due to the presence of disease-related SNPs at the cell (sub)type level. In particular, RNetDys combines scRNA-seq, scATAC-seq, ChIP-seq and prior-knowledge to build comprehensive cell (sub)types or state specific GRNs that are leveraged to capture impaired interactions due to disease-related SNPs. Compared to existing strategies to study SNPs (Farh et al., 2014, Yu et al., 2022; Nathan et al., 2022), this pipeline provides a comprehensive view of the impaired regulatory landscape, including interactions mediated by TFs and enhancers of regulated genes and activation or repression mechanisms to provide additional mechanistic insights. In particular, RNetDys provides the binding affinity score of impaired TFs, the type of mechanism dysregulated, and a list of ranked TFs based on their importance in the impaired network topology, the strength of the binding impairment and the frequency of SNPs occurring in the global population.

We applied RNetDys in five disease case studies and showed that it was able to accurately capture impaired regulatory interactions and provide additional mechanistic insights by leveraging the information obtained from the GRN inference.

Material and methods

General workflow of RNetDys

We implemented a systematic pipeline integrating different type of OMICS data to decipher impaired regulatory mechanisms due to SNPs in disease by leveraging the GRN information. The pipeline was divided in two main parts composed of the cell (sub)type specific GRN inference and the capture of impaired regulatory interactions due to disease-related SNPs to gain regulatory mechanistic insights for the disease.

Cell (sub)type specific regulatory interactions inference

The cell (sub)type specific regulatory network inference was based on a multi-OMICS approach that used single cell transcriptomics and single cell chromatin accessibility, not necessarily matched, as well as prior-knowledge, including ChIP-seq data and reported enhancers interactions. First, using the scRNA-seq we selected genes that were conserved at least in 50% of the cells for further analyses. Then, we ensured the accessibility of the corresponding promoter regions using scATAC-seq data and predicted TF-promoter interactions by intersecting the ChIP-seq TF-binding evidence with the open promoter regions using BEDTools (Quinlan and Hall, 2010). Then, we performed a peak correlation using the scATAC-seq data and carried out a statistical test, as well as a BH multiple correction, to select the significant interactions such as p-adjusted value < 0.05. The identified enhancer-promoter interactions were then intersected with GeneHancer (Fishilevich et al., 2017), used as a backbone, and interactions involving active promoters were kept. Then, TF-enhancers interactions were inferred by intersecting the ChIP-seq and scATAC-seq data. Finally, the regulatory interactions were signed to distinguish activations from repressions by computing the Pearson correlation between TFs and genes using the scRNA-seq dataset (Fig. S1). Correlation scores for enhancer-promoter interactions were computed such as:

$$corV_{E_a \to G_b} = \sum_x corV_{TF_x \to G_b}$$

With corV corresponding to the correlation value, E denoting the enhancer and G corresponding to the gene. And, correlation scores for TF-enhancer were computed such as:

$$corV_{TF_a \to E_b} = \sum_x corV_{TF_a \to G_x}$$

With corV denoting the correlation value, E corresponding to the enhancer and G to the gene. Finally, positive correlation scores were considered to be activations whereas negative ones were considered to be repressions. Further details are provided in Supplementary Information.

Identify candidate impaired regulatory interactions

Using the cell (sub)type specific GRN inferred in healthy condition, we then contextualized the GRN towards the disease condition. The contextualization required a list of SNPs for the disease studied and the cell (sub)type GRN of interest. The SNPs were mapped to the GRN

by using their coordinates and interactions for which a SNP was falling into a TF binding region of an enhancer or promoter were considered as candidates to be impaired in the disease. We then performed a TF binding analysis using PERFECTOS-APE (E. Vorontsov et al., 2015) to refine the candidate interactions by selecting the ones having at least one binding site significantly impaired by the SNP (Supplementary Information). Finally, we ranked TFs by their involvement in the regulatory impairments based on the network topology and the MAF score of SNPs such as:

$$Rank_{TF} = RE \times \frac{NG}{RE} \times \left(\sum |AI|_i^r \times \left(MAF_i^r \times \sum MAF^r \right) \right)$$

Where RE denote the number of regulatory elements regulated by the TF, NG corresponds to the number of downstream genes across RE, AI denotes the binding affinity impairment log2FC and i corresponds to the SNPs and r the regulatory element.

Prior-knowledge collection and processing

RNetDys relied on prior-knowledge data that were collected and processed to be integrated in the pipeline. The ChIP-seq bed files were downloaded from ChIP Atlas (Oki et al., 2018) for human hg19 and hg38 assemblies. Bed files were annotated using HOMER (Heinz et al., 2010) with the latest GTF file for each assembly. Enhancer regions and their connected genes were obtained from the GeneHancer database (Fishilevich et al., 2017). Of note, GeneHancer database provided information for hg38 coordinates and hence, we used LiftOver (https://genome.ucsc.edu/cgi-bin/hgLiftOver) to convert these coordinates for hg19 to provide more flexibility to our pipeline.

Data collection and analysis

First, to perform the benchmarking analysis, we collected 20 publicly available scRNA-seq and 11 scATAC-seq datasets from six human cell lines including BJ, GM12878, H1-ESC, A549, Jurkat and K562 (Table S1). Then, we collected scRNA-seq and scATAC-seq healthy data from pancreas and brain tissues to extract cell (sub)types using Seurat (Hao et al., 2021) and Signac (Stuart et al., 2020), and then generated the GRNs (Supplementary information). Finally, we collected SNPs from ClinVar (Landrum et al., 2018) for five diseases including Alzheimer's disease (AD), Parkinson's disease (PD), Epilepsy (EPI), Diabetes type I (T1D) and type II (T2D) to perform the network contextualization towards the disease condition. Notably, SNPs were defined as being single nucleotide variants found at least in 1% of the

global population such as MAF >= 0.01 (Supplementary Information). In addition, we performed an outdegree analysis for three main TFs involved in the regulatory impairments. The outdegree ratios were computed by scaling each TF outdegree by the maximum outdegree in each cell (sub)type.

Validation and comparison to state-of-the-art

We first assessed the performances of RNetDys in identifying cell (sub)type specific regulatory interactions and compared them to state-of-the-art GRN inference methods (Aibar et al., 2017; Chan et al., 2017; Kim, 2015; Huynh-Thu et al., 2010) (Supplementary Information). First, we benchmarked the performances of each method to infer cell (sub)type specific TF-promoter interactions. The gold standards (GS) were compiled using cell line specific ChIP-seq from Cistrome (Mei et al., 2017) by selecting only the highest quality data. Then, we assessed the performances of RNetDys for capturing cell (sub)type specific enhancer-promoter regulatory interactions compared to Cicero, a widely used method to identify cis-interactions based on scATAC-seq data (Pliner et al., 2018). The GS networks were built using promoter capture Hi-C data from 3DIV (Yang et al., 2018) for three of the human cell lines. For both benchmarking analyses, we computed the precision (PPV) and F1-score (F1) to assess the performances such as:

$$PPV = \frac{TP}{(TP+FP)} \text{ and } F_1 = \frac{2\times TP}{2\times TP+FP+FN}$$

With TP = True Positive (predicted and found in the GS), FP = False Positives (predicted but not in the GS) and FN = False Negatives: (not predicted but in the GS).

We then compared the ability of RNetDys to precisely capture gene-disease relationships in cell (sub)types, compared to eQTL (Bryois et al., 2022). First, we downloaded Online Mendelian Inheritance in Man (OMIM) Morbid Map (Amberger et al., 2019), filtered for gene-disease interactions reported in the five diseases in study (AD, EPI, PD, T1D, and T2D), and removed the interactions reported as provisional. Then, we matched these gene interactions to the SNP-associated genes identified by eQTL and RNetDys. The ratio of matched genes in eQTL and RNetDys was calculated by dividing the number of matched genes by the total of genes identified in each of the methods across all cell (sub)types.

Results

RNetDys, a multi-OMICS pipeline to decipher impaired regulatory mechanisms

We implemented RNetDys, a systematic pipeline based on multi-OMICS data to decipher impaired regulatory interactions due to SNPs in diseases by leveraging the information of cell (sub)type specific GRNs (Fig.1).

RNetDys is an integrative approach relying on single cell transcriptomics and single cell chromatin accessibility from a specific cell (sub)type or state, as well as prior-knowledge information including extensive ChIP-seq data (Oki et al., 2018) and reported enhancer-promoter relationships (Fishilevich et al., 2017). The pipeline is composed of two main parts: (i) the cell (sub)type specific GRN inference and (ii) the identification of impaired regulatory mechanisms due to SNPs in diseases (Fig. 1, Fig. S1). The first part consists of the GRN inference for a healthy cell (sub)type or state based on scRNA-seq and scATAC-seq data as an input. Notably, the two single cell datasets do not need to be matched but they need to contain the same cell (sub)type. The second part takes as an input a cell (sub)type or state specific GRN and a list of SNPs of particular interest for the disease studied (Visscher et al., 2017; Landrum et al., 2018). In particular, the SNPs provided could have been described as related to the disease of interest in prior-knowledge databases (Landrum et al., 2018) or identified by genotyping analyses (Nielsen et al., 2011). As a result, RNetDys provides the impaired regulatory mechanisms, the corresponding SNPs, the affinity scores of TF having their binding site impaired, and a list of ranked TF regulators based on their involvement in the observed impairments (Fig. 1).

RNetDys is more accurate to infer cell (sub)type specific GRNs

RNetDys mainly relies on the cell (sub)type specific regulatory landscape to identify impaired regulatory interactions due to disease-related SNPs. Therefore, we assessed the performance of RNetDys in predicting cell (sub)type specific GRNs (Fig. 2). In this regard, we performed the benchmarking of both TF-gene and enhancer-promoter interactions, compared to current methods. We showed that our approach overcame the state-of-the-art GRN inference methods for predicting cell (sub)type specific TF-gene interactions with an average precision of 0.20 and average accuracy of 0.28 (Fig. 2A, B).

This assessment highlighted the strength of combining different regulatory layers with prior-knowledge to provide predictions with a higher confidence. Moreover, we showed that

RNetDys outperformed Cicero in capturing cell (sub)type specific enhancer-promoter interactions with a median precision of 0.76 and median accuracy of 0.72, supporting the confidence provided by the prior-knowledge leveraged by our approach (Fig. 2C, D). This analysis demonstrated the accuracy of the cell (sub)type specific GRN information leveraged by our pipeline to capture impaired transcriptional regulatory mechanisms due to SNPs in diseases.

RNetDys provides additional insights into the mechanistic dysregulation enhanced by SNPs

Validation of the SNPs impairment and comparison to state-of-the-art approaches

We applied RNetDys to five diseases, including AD, PD, EPI, T1D and T2D, by collecting disease-related SNPs from ClinVar (Landrum et al., 2018) and cell (sub)type specific GRNs generated from human pancreas and brain tissues. First, we supported the relevance of predicted SNP-gene interactions identified by RNetDys using available GWAS data from ClinVar database and recently published cell-type specific eQTL information (Bryois et al., 2022). Across the five diseases, we were able to find support for 90% of the SNP-target gene relationships identified by our pipeline (Table S4). Furthermore, by using cell type specific eQTL data, we were able to validate the occurrence of certain SNPs and their impact on the predicted target genes in specific cell types. For instance, our results show that the PD-associated SNPs rs11538371, rs2072814 and rs8137714 are found to be linked to TIMP3 in astrocytes (Table S4). In fact, TIMP3 is an inhibitor of metalloproteinases, enzymes secreted by astrocytes (Yin et al., 2006), that are implicated in several PD-associated processes such as dopaminergic neuron degeneration, neuroinflammation, and proteolysis of α-synuclein (Sung et al., 2005; Choi et al., 2011; Annese et al., 2015). Second, we evaluated the precision of RNetDys in capturing gene-disease relationships at the cell(sub)type level using the OMIM database (Amberger et al., 2019) (Fig. S2). When compared to the eQTL data, we observed that the genes captured by our approach as being impaired due to the presence of SNPs are more often linked to disease than the genes captured by eQTL. Although eQTL captures a larger number of SNP-gene interactions, few of them are actually described to be involved in the disease, thus explaining the low ratio. On the other hand, RNetDys identifies more genes linked to each SNP that have been described as related to the disease, demonstrating the higher precision of RNetDys compared to eQTL.

Cell (sub)type differential dysregulation in diseases

Then, we studied the differential impairment across cell (sub)types in the five diseases as it has been reported that some cell (sub)types were more involved in disease mechanisms (Muratore et al., 2017; Kamath et al., 2022). We observed that cell (sub)types shared few impaired interactions in the studied diseases, especially in EPI and PD (Fig. 3). Interestingly, in EPI, astrocytes, OPCs and inhibitory neurons seem to be the most impaired cell types. This is consistent with literature evidence that shows that modifications in GABA receptors, which are expressed in inhibitory neurons, are closely linked to epilepsy (Tanaka et al., 2012). Furthermore, impairment of antiquin expression, encoded by the gene ALDH7A1, in astrocytes has been described to be linked with dysregulation of neurotransmitter shuttling and recycling, one of the major causes of neurological deficits (David et al., 2009; Jansen et al., 2014). Finally, studies showed that myelinated neuronal axons are damaged in epileptic patients and the ability of OPCs to proliferate is reduced in samples obtained from patients with dysplasia (Luo et al., 2015; Donkels et al., 2020).

Insights into the cell (sub)type specific regulatory impairments

We finally aimed at exploiting the GRN information provided by RNetDys to further analyse the regulatory impairments of cell (sub)types (Fig. 4, Fig. S3-S6). We observed that in AD (Fig. 4), the same enhancers were involved in all cell (sub)types specific networks with an impact on the expression of APP and presenilin 1 (PSEN1). Alterations in the expression of these genes are primarily linked to the development of AD (Dewachter et al., 2002; Matsui et al., 2007). Furthermore, recent studies have shown that not only neurons, but also astrocytes and microglia to be involved in the accumulation of β-amyloid plaques (Palop and Mucke, 2010; Frost and Li, 2017). However, the impairment of the TFs and enhancers regulating these two genes seems to be different across cell (sub)types (Fig. 4). Indeed, most of the SNPs in astrocytes and microglia would induce a repression of APP whereas this gene seems to be activated in other cell (sub)types (Fig. 4). It has been described that these two cell types provide protective effects, with microglia facilitating the clearance of β-amyloid overproduced by neurons in AD (Fakhoury, 2018).

To provide better insights on the main regulatory TFs behind disease dysregulation, we ranked the impaired TFs based on network topology and impact of each involved SNP (see Methods). Notably, we could observe that certain TFs, such as CREB1, MXI1, and STAT3, are ranked as top regulators across different brain diseases and diabetes (Table 1). To

investigate this, we evaluated the outdegree distribution of these TFs and we observed different outdegrees values across cell (sub)types, which demonstrates that our ranking has no bias towards highly connected TFs (Fig. S7). CREB1, MXI1, and STAT3 participate in common cell mechanisms involved disease development, such as cell death and inflammation. However, each of these TFs has been described to play a different function in these mechanisms in different diseases.

For instance, MXI1 has been shown to be involved in the aging of the neurovascular unit, which contributes for the progression of AD (Zhao et al., 2022). On the other hand, the same TFs seems to be part of the unique transcriptomic signature of T2D, which we can also observe in our results as this TF does not show as a key regulator in T1D (Table 1) (Cubillos-Angulo et al., 2020). Finally, MXI1 was found to be one of the main regulators involved in impaired regulatory interactions for PD, apart from dopaminergic neurons (Fig. S3). MXI1 has been described to be involved in the mitochondrial homeostasis, dysregulated in PD and known to be involved with neurodegeneration (Lestón Pinilla et al., 2021; Malpartida et al., 2021).

CREB1 has been extensively shown to regulate gluconeogenesis through the coactivator PGC-1, playing a vital role in the regulation of efficient glucose sensing and insulin exocytosis and in the development of diabetes (Herzig et al., 2001). Our results show this TF to be the main regulator involved in AD and EPI in all cell (sub)types, apart from astrocytes (Table 1, Fig. 4, Fig. S4). CREB1 is a TF responsible for regulating the major pathways that mediate neurotrophin-associated gene expression, a group of proteins that promotes survival and neuronal development (Shaywitz and Greenberg, 1999). Indeed, increased CREB activity promotes hyperexcitability, one of the main triggers of seizures, while reduced levels seem to prevent epilepsy (Zhu et al., 2012; Wang et al., 2020) (Fig. S4). PSEN1 has been shown to be a downstream target of CREB1 (Cui et al., 2022), which further supports the results obtained by our pipeline as CREB1 was predicted to regulate PSEN1. PSEN1 upregulation leads to myelin dysfunction in OPCs in cases of familial AD (Desai et al., 2011). Notably, our pipeline predicts a decrease in CREB1 binding affinity to the promoter and enhancer regions of PSEN1 in the presence of rs1800839, potentially elucidating one of the possible mechanisms behind PSEN1 upregulation previously observed in AD (Fig. 4).

Finally, STAT3 was overall found to be the main regulator involved in impaired interactions of T1D and T2D (Table 1, Fig.s S5 and S6). In the pancreas, STAT3 has been shown to regulate insulin secretion and islet development (Saarimäki-Vire et al., 2017). In addition, in T2D, exacerbated STAT3 signalling has been shown to lead to insulin resistance in skeletal muscle of diabetic patients (Mashili et al., 2013), supporting its importance as a regulator of the dysregulations involved in the disease. In neurodegenerative diseases, STAT3 activation has been shown to promote astrogliosis, which is reflected in our results by an increase of the binding affinity of this TF to distinct regulatory regions (Fig. 4A and Fig. S4A) (Toral-Rios et al., 2020).

Discussion

The study of cell (sub)type or state specific regulatory interactions impaired due to disease-related SNPs is required to pave the way towards the development of gene-based therapies to prevent or treat diseases (Rao et al., 2021). In addition, the comprehensive view of the regulatory landscape, including interactions mediated by TFs and enhancers of regulated genes, is critical to study dysregulated mechanisms in diseases (Emmert-Streib et al., 2014; Chiou et al., 2021). In that regard, existing strategies to study the impact of SNPs do not exploit the GRN information to provide additional mechanistic insights into the disease-related dysregulations (Rao et al., 2021; Bryois et al., 2022). In addition, recent studies have shown that specialized group of cells, including cell types, subtypes and phenotypes, are not equally involved in diseases (Nathan et al., 2022; Kamath et al., 2022). However, current approaches have been mainly focused on cell types, lacking ability to identify dysregulated mechanisms at deeper levels of resolution. RNetDys is a systematic multi-OMICS pipeline to decipher cell (sub)type or state specific regulatory interactions impaired due to SNPs in diseases. This pipeline exploits the high-resolution of single cell to infer a comprehensive regulatory landscape, leveraged to identify impairment due to SNPs. We applied RNetDys to five disease cases and showed that cell (sub)types specific regulatory mechanisms were not equally impaired, suggesting their differential involvement in the studied diseases. Moreover, we validated the relevance of some impaired regulatory mechanisms using GWAS and eQTL data (Landrum et al., 2018; Bryois et al., 2022). In that regard, we provided additional mechanistic insights into the regulatory mechanisms dysregulated and identified the main TF regulators involved. Notably, the presented analysis was performed

using SNPs retrieved from ClinVar, but RNetDys could be of great use to provide valuable regulatory mechanistic insights by using SNPs derived from genotyping studies. In the present study, we were able to predict known and unreported cell (sub)type specific SNP-gene interactions, hence showing how our pipeline could facilitate the discovery of regulatory impairments. To conclude, we foresee RNetDys to be a valuable tool to comprehensively identify cell (sub)type specific regulatory mechanisms impaired due to SNPs and aid the development of strategies for therapeutic intervention in diseases.

Data and Material availability

RNetDys is a pipeline publicly available at https://github.com/BarlierC/RNetDys.git.

The repository of generated regulatory networks, results and scripts used in this study are available at https://gitlab.com/C.Barlier/RNetDys_analyses.

Acknowledgements

The authors thank Dr. Patrick May for the valuable feedback and insights provided for this project. The benchmarking of the state-of-the-art GRNs method and data processing was performed using the HPC facilities of the University of Luxembourg (https://hpc.uni.lu).

Author contributions

C.B. implemented RNetDys, collected and processed the data, performed the benchmarking of the GRNs, generated the cell (sub)type specific GRNs, collected the disease-related SNPs, performed the data analysis and wrote the manuscript, M.M.R. collected and processed the data, extracted the healthy cell (sub)types datasets, performed the validation of impaired interactions, the data analysis and wrote the manuscript, S.J. supervised the computational work, A.d.S conceived the idea and supervised the project.

Funding

C.B. is supported by funding from the Luxembourg National Research Fund (FNR) within PARK-QC DTU (PRIDE17/12244779/PARK-QC). M.M.R. is supported by Fonds National de la Recherche Luxembourg (C17/BM/11662681). S.J. is supported by the Spanish Ministry of Science and Innovation MCIN/AEI (PID2020-118605RB-I00).

Conflict of Interest

None declared.

References

Aibar,S. et al. (2017) SCENIC: single-cell regulatory network inference and clustering. Nat Methods, 14, 1083–1086.

Akhlaghipour,I. et al. (2022) Single-nucleotide polymorphisms as important risk factors of diabetes among Middle East population. Hum Genomics, 16, 11.

Amberger,J.S. et al. (2019) OMIM.org: leveraging knowledge across phenotype–gene relationships. Nucleic Acids Res, 47, D1038–D1043.

Ament,S.A. et al. (2018) Transcriptional regulatory networks underlying gene expression changes in Huntington's disease. Mol Syst Biol, 14.

Andersson,R. et al. (2014) An atlas of active enhancers across human cell types and tissues. Nature, 507, 455–461.

Annese,V. et al. (2015) Metalloproteinase-9 contributes to inflammatory glia activation and nigrostriatal pathway degeneration in both mouse and monkey models of 1-methyl-4-phenyl-1,2,3,6-tetrahydropyridine (MPTP)-induced Parkinsonism. Brain Struct Funct, 220, 703–727.

Bakker,O.B. et al. (2021) Linking common and rare disease genetics through gene regulatory networks Genetic and Genomic Medicine.

Bryois,J. et al. (2022) Cell-type-specific cis-eQTLs in eight human brain cell types identify novel risk genes for psychiatric and neurological disorders. Nat Neurosci, 25, 1104–1112.

Chan,T.E. et al. (2017) Gene Regulatory Network Inference from Single-Cell Data Using Multivariate Information Measures. Cell Systems, 5, 251-267.e3.

Chiou,J. et al. (2021) Interpreting type 1 diabetes risk with genetics and single-cell epigenomics. Nature, 594, 398–402.

Choi,D.H. et al. (2011) Role of Matrix Metalloproteinase 3-mediated α-Synuclein Cleavage in Dopaminergic Cell Death. J Biol Chem, 286, 14168.

Claringbould,A. and Zaugg,J.B. (2021) Enhancers in disease: molecular basis and emerging treatment strategies. Trends in Molecular Medicine, 27, 1060–1073.

Cubillos-Angulo,J.M. et al. (2020) In silico transcriptional analysis of mRNA and miRNA reveals unique biosignatures that characterizes different types of diabetes. PLoS One, 15, e0239061.

Cui, X., Yang, Y., & Yan, A. (2022). MiR-654-3p Constrains Proliferation, Invasion, and Migration of Sinonasal Squamous Cell Carcinoma via CREB1/PSEN1 Regulatory Axis. Frontiers in Genetics, 12.

David,Y. et al. (2009) Astrocytic dysfunction in epileptogenesis: consequence of altered potassium and glutamate homeostasis? J Neurosci, 29, 10588–10599.

Desai,M.K. et al. (2011) An Alzheimer's Disease-relevant Presenilin-1 Mutation Augments Amyloid-beta-induced Oligodendrocyte Dysfunction. Glia, 59, 627.

Dewachter,I. et al. (2002) Neuronal deficiency of presenilin 1 inhibits amyloid plaque formation and corrects hippocampal long-term potentiation but not a cognitive defect of amyloid precursor protein [V717I] transgenic mice. J Neurosci, 22, 3445–3453.

Donkels,C. et al. (2020) Oligodendrocyte lineage and myelination are compromised in the gray matter of focal cortical dysplasia type IIa. Epilepsia, 61, 171–184.

E. Vorontsov,I. et al. (2015) PERFECTOS-APE - Predicting Regulatory Functional Effect of SNPs by Approximate P-value Estimation: In, Proceedings of the International Conference on Bioinformatics Models, Methods and Algorithms. SCITEPRESS - Science and Technology Publications, Lisbon, Portugal, pp. 102–108.

Emmert-Streib,F. et al. (2014) Gene regulatory networks and their applications: understanding biological and medical problems in terms of networks. Front. Cell Dev. Biol., 2.

Fakhoury,M. (2018) Microglia and Astrocytes in Alzheimer's Disease: Implications for Therapy. Curr Neuropharmacol, 16, 508–518.

Farh,K.K.H. et al. (2014) Genetic and epigenetic fine mapping of causal autoimmune disease variants. Nature 2014 518:7539, 518, 337–343.

Fishilevich,S. et al. (2017) GeneHancer: genome-wide integration of enhancers and target genes in GeneCards. Database, 2017.

Frost,G.R. and Li,Y.M. (2017) The role of astrocytes in amyloid production and Alzheimer's disease. Open Biol, 7, 170228.

Gazal,S. et al. (2022) Combining SNP-to-gene linking strategies to identify disease genes and assess disease omnigenicity. Nat Genet, 54, 827–836.

Hao,Y. et al. (2021) Integrated analysis of multimodal single-cell data. Cell, 184, 3573-3587.e29.

Heinz,S. et al. (2010) Simple Combinations of Lineage-Determining Transcription Factors Prime cis-Regulatory Elements Required for Macrophage and B Cell Identities. Molecular Cell, 38, 576–589.

Herzig,S. et al. (2001) CREB regulates hepatic gluconeogenesis through the coactivator PGC-1. Nature, 413, 179–183.

Hiramoto,M. et al. (2015) Comparative analysis of type 2 diabetes-associated SNP alleles identifies allele-specific DNA-binding proteins for the KCNQ1 locus. International Journal of Molecular Medicine, 36, 222–230.

Huynh-Thu,V.A. et al. (2010) Inferring Regulatory Networks from Expression Data Using Tree-Based Methods. PLoS ONE, 5, e12776.

Jansen,L.A. et al. (2014) Glial localization of antiquitin: implications for pyridoxine-dependent epilepsy. Ann Neurol, 75, 22–32.

Jeng,X.J. et al. (2020) Effective SNP ranking improves the performance of eQTL mapping. Genet Epidemiol, 44, 611–619.

Kamath,T. et al. (2022) Single-cell genomic profiling of human dopamine neurons identifies a population that selectively degenerates in Parkinson's disease. Nat Neurosci, 25, 588–595.

Kim,S. (2015) ppcor: An R Package for a Fast Calculation to Semi-partial Correlation Coefficients. Communications for Statistical Applications and Methods, 22, 665–674.

Landrum,M.J. et al. (2018) ClinVar: improving access to variant interpretations and supporting evidence. Nucleic Acids Res, 46, D1062–D1067.

Latchman,D.S. (2011) Transcriptional Gene Regulation in Eukaryotes. In, John Wiley & Sons, Ltd (ed), eLS. Wiley.

Lestón Pinilla,L. et al. (2021) Hypoxia Signaling in Parkinson's Disease: There Is Use in Asking "What HIF?" Biology, 10, 723.

Luo,Y. et al. (2015) Alterations in hippocampal myelin and oligodendrocyte precursor cells during epileptogenesis. Brain Res, 1627, 154–164.

Malpartida,A.B. et al. (2021) Mitochondrial Dysfunction and Mitophagy in Parkinson's Disease: From Mechanism to Therapy. Trends Biochem Sci, 46, 329–343.

Mashili,F. et al. (2013) Constitutive STAT3 phosphorylation contributes to skeletal muscle insulin resistance in type 2 diabetes. Diabetes, 62, 457–465.

Matsui,T. et al. (2007) Expression of APP pathway mRNAs and proteins in Alzheimer's disease. Brain Res, 1161, 116–123.

Mei,S. et al. (2017) Cistrome Data Browser: a data portal for ChIP-Seq and chromatin accessibility data in human and mouse. Nucleic Acids Res., 45, D658–D662.

Muratore,C.R. et al. (2017) Cell-type Dependent Alzheimer's Disease Phenotypes: Probing the Biology of Selective Neuronal Vulnerability. Stem Cell Reports, 9, 1868–1884.

Nathan,A. et al. (2022) Single-cell eQTL models reveal dynamic T cell state dependence of disease loci. Nature.

Nica,A.C. and Dermitzakis,E.T. (2013) Expression quantitative trait loci: present and future. Phil. Trans. R. Soc. B, 368, 20120362.

Oki,S. et al. (2018) ChIP -Atlas: a data-mining suite powered by full integration of public Ch IP -seq data. EMBO Rep, 19.

Palop,J.J. and Mucke,L. (2010) Amyloid-beta-induced neuronal dysfunction in Alzheimer's disease: from synapses toward neural networks. Nat Neurosci, 13, 812–818.

Pliner,H.A. et al. (2018) Cicero Predicts cis-Regulatory DNA Interactions from Single-Cell Chromatin Accessibility Data. Mol Cell, 71, 858-871.e8.

Quinlan,A.R. and Hall,I.M. (2010) BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics, 26, 841–842.

Rao,S. et al. (2021) Editing GWAS: experimental approaches to dissect and exploit disease-associated genetic variation. Genome Med, 13, 41.

Saarimäki-Vire,J. et al. (2017) An Activating STAT3 Mutation Causes Neonatal Diabetes through Premature Induction of Pancreatic Differentiation. Cell Rep, 19, 281–294.

Shaywitz,A.J. and Greenberg,M.E. (1999) CREB: a stimulus-induced transcription factor activated by a diverse array of extracellular signals. Annu Rev Biochem, 68, 821–861.

Stuart,T. et al. (2020) Multimodal single-cell chromatin analysis with Signac Genomics.

Sung,J.Y. et al. (2005) Proteolytic cleavage of extracellular secreted {alpha}-synuclein via matrix metalloproteinases. J Biol Chem, 280, 25216–25224.

Tanaka,M. et al. (2012) GABRB3, Epilepsy, and Neurodevelopment. In, Noebels,J.L. et al. (eds), Jasper's Basic Mechanisms of the Epilepsies. National Center for Biotechnology Information (US), Bethesda (MD).

Toral-Rios,D. et al. (2020) Activation of STAT3 Regulates Reactive Astrogliosis and Neuronal Death Induced by AβO Neurotoxicity. Int J Mol Sci, 21, 1–29.

Uddin,F. et al. (2020) CRISPR Gene Therapy: Applications, Limitations, and Implications for the Future. Front Oncol, 10, 1387.

Visscher,P.M. et al. (2017) 10 Years of GWAS Discovery: Biology, Function, and Translation. The American Journal of Human Genetics, 101, 5–22.

Wang,G. et al. (2020) Advances in Understanding CREB Signaling-Mediated Regulation of the Pathogenesis and Progression of Epilepsy. Clinical Neurology and Neurosurgery, 196, 106018.

Yang,D. et al. (2018) 3DIV: A 3D-genome Interaction Viewer and database. Nucleic Acids Research, 46, D52–D57.

Yin,K.J. et al. (2006) Matrix metalloproteinases expressed by astrocytes mediate extracellular amyloid-beta peptide catabolism. J Neurosci, 26, 10939–10948.

Yu,F. et al. (2022) Variant to function mapping at single-cell resolution through network propagation. Nature Biotechnology.

Zhao,Y. et al. (2022) Accelerated aging-related transcriptome alterations in neurovascular unit cells in the brain of Alzheimer's disease. Front Aging Neurosci, 14, 937.

Zhu,X. et al. (2012) Decreased CREB levels suppress epilepsy. Neurobiol Dis, 45, 253–263.

Tables

Table 1. TF regulators involved in impaired regulatory mechanisms.

DISEASE	CELL (SUB)TYPE	RANKED TFS*
AD	Astrocyte	MXI1, STAT3
	Excitatory neuron	CREB1, USF2, MXI1
	Inhibitory neuron	CREB1, MXI1, STAT3
	Microglia	CREB1, USF2, MXI1, IKZF1
	Oligodendrocyte	CREB1, MXI1
	OPCs	CREB1, MXI1, ETV1
EPI	Astrocyte	MXI1, STAT3, BCL6, ZFX, RXRA
	Excitatory neuron	CREB1, MXI1
	Inhibitory neuron	CREB1, STAT3, STAT1, MXI1
	Microglia	CREB1, MXI1
	Oligodendrocyte	CREB1
	OPCs	CREB1, BCL6, MXI1, STAT1, ETV1
PD	Astrocyte	MXI1, BCL6
	Dopaminergic neuron	STAT3
	Excitatory neuron	MXI1, CREB1
	Oligodendrocyte	MXI1
	OPCs	BCL6, MXI1, ETV1
T1D	Alpha cell	STAT3, STAT1, RXRA
	Beta cell	STAT3, CREB1
	Delta cell	STAT3, CREB1
T2D	Alpha cell	STAT3, RXRA, STAT1, CREB1, ATF2, EHF
	Beta cell	CREB1, STAT1, STAT3, PDX1, ETS1, ATF2, RXRA, MXI1
	Delta cell	CREB1, STAT1, STAT3, PDX1, ETV1, EHF, ATF2
	Gamma cell	STAT3, CREB1, STAT1, ETV1, EHF, ATF2

* TFs are ranked by their order of importance in the detected impaired regulatory mechanisms.

Figures

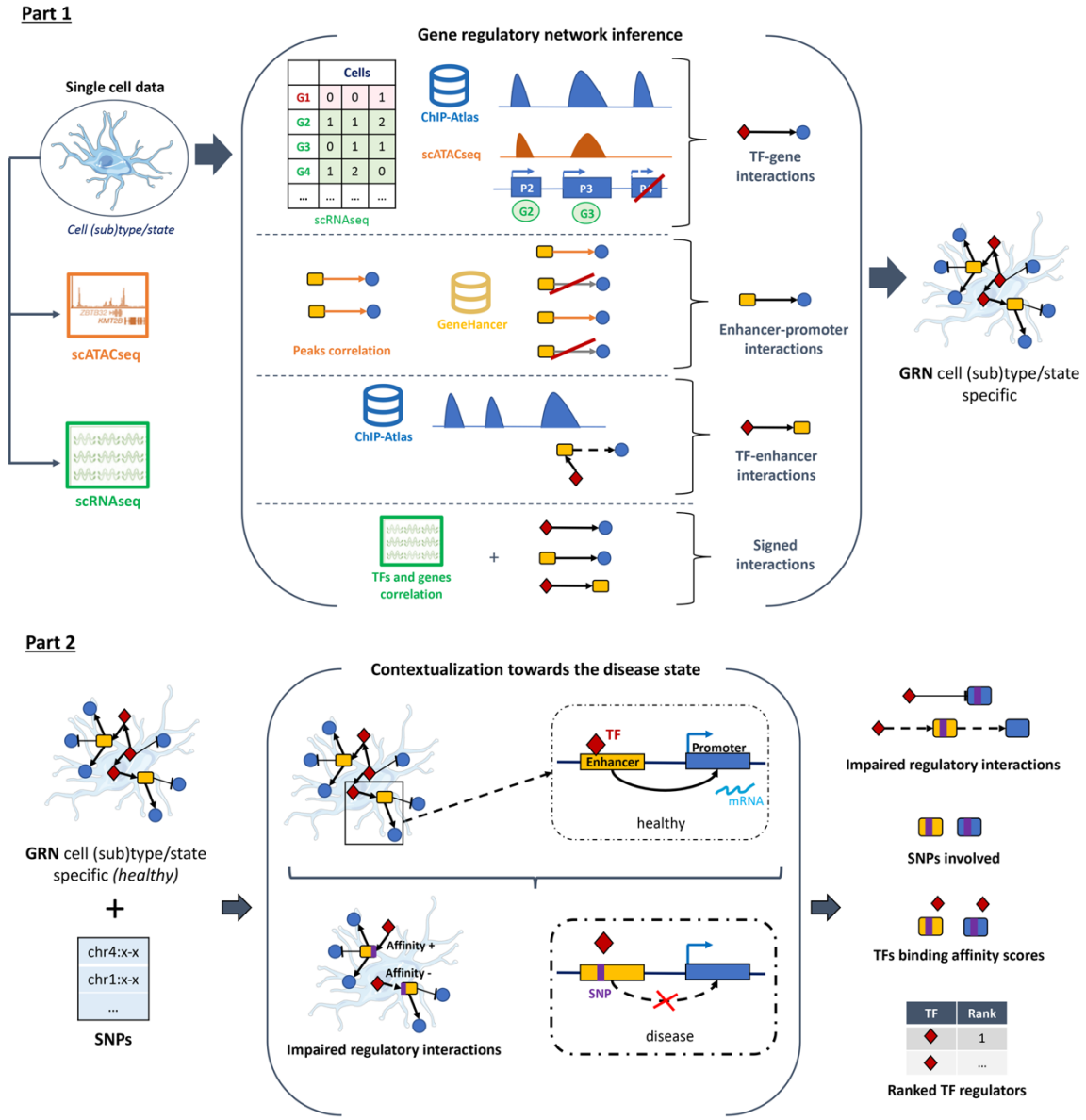


Fig. 1. General workflow of RNetDys to decipher regulatory dysregulation in diseases. RNetDys is composed of two main parts including (1) the cell (sub)type specific GRN inference using scRNA-seq, scATAC-seq and prior-knowledge, and (2) the identification of candidates impaired regulatory interactions using the GRN and a list of SNPs, followed by the TF-binding affinity analysis. The part one provides the cell (sub)type or state specific GRN describing the regulatory interactions mediated by TFs and enhancers of regulated genes. The part two provides the list of candidate impaired regulatory interactions in the cell (sub)types, the SNPs that were mapped to these interactions and the TFs for which the binding ability might be impaired and regulatory TFs ranked based on their importance in the impairments.

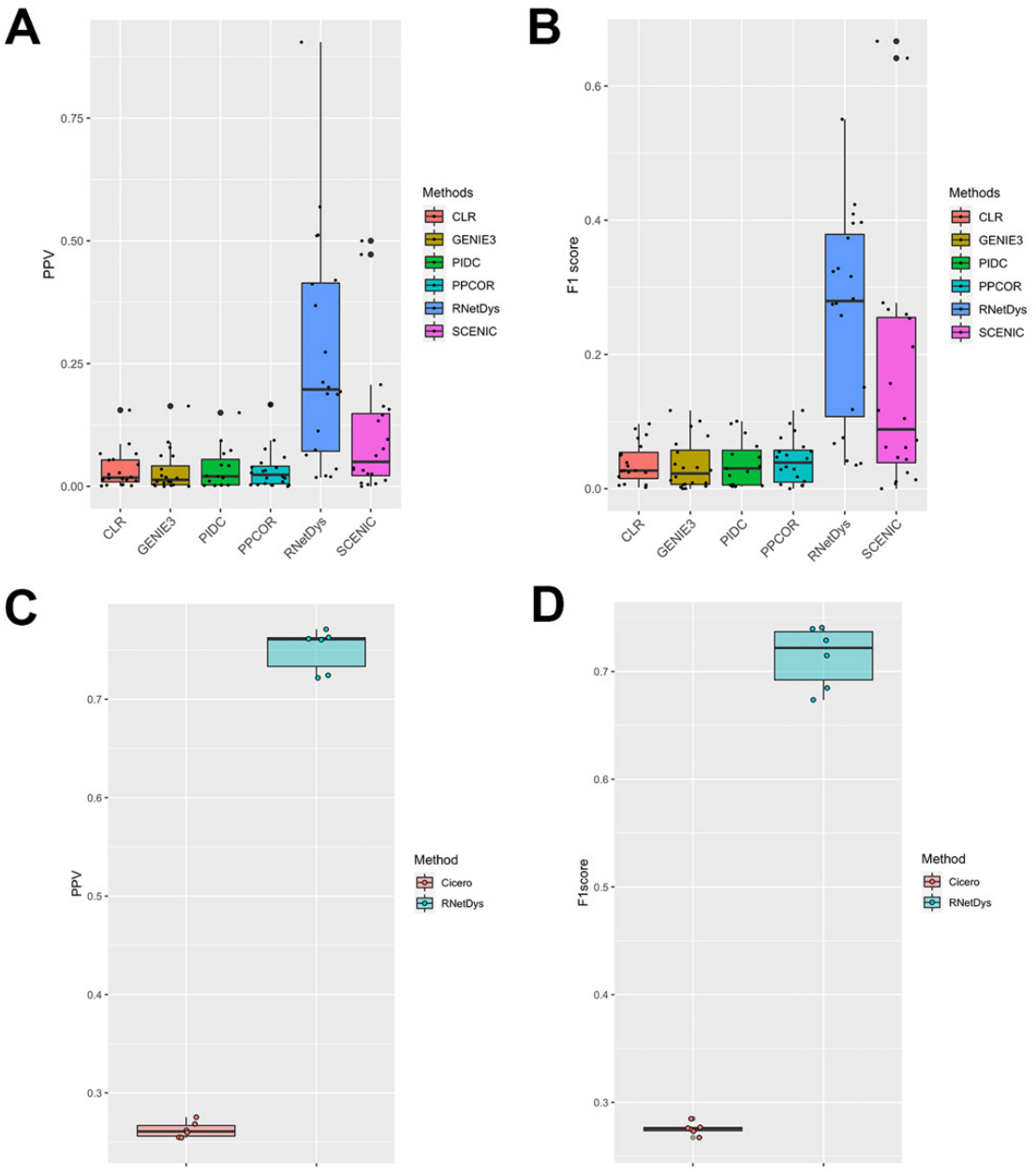


Fig. 2. Performances of RNetDys and comparison to other methods. (A, B) TF-promoter regulatory interactions performances assessed using (A) the PPV and (B) the F1-score metrics. Performances were assessed for RNetDys, state-of-the-art methods and metrics on 20 datasets from six human cell lines. (C, D) Enhancer-promoters regulatory interactions performance assessment using (C) the PPV and (D) the F1-score metrics. Performances were assessed for RNetDys and Cicero on 6 scATAC-seq datasets from three human cell lines.

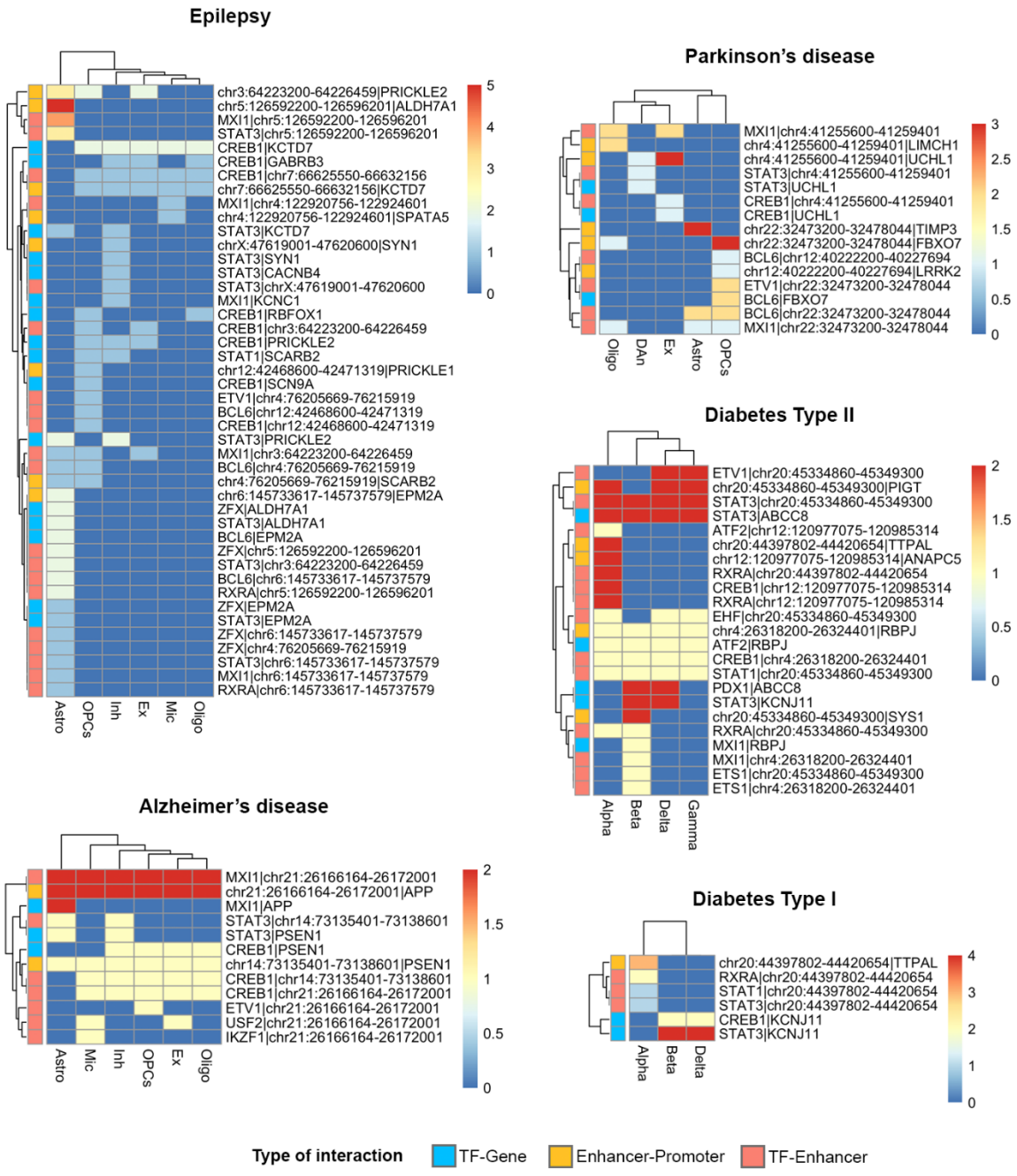


Fig. 3. Cell (sub)type differential regulatory impairment in diseases. Heatmaps showing the distribution of impaired interactions due to disease-related-SNPs across cell (sub)types for Alzheimer's disease (AD), Parkinson's disease (PD), Epilepsy (EPI), Diabetes type I (T1D) and type II (T2D). The colors of the heatmap represent the number of SNPs impacting the regulatory interactions. Astro: astrocytes, Ex: excitatory neurons, Inh: inhibitory neurons, Mic: microglia, Oligo: oligodendrocytes, OPCs: oligodendrocyte progenitors, DAn: dopaminergic neurons.

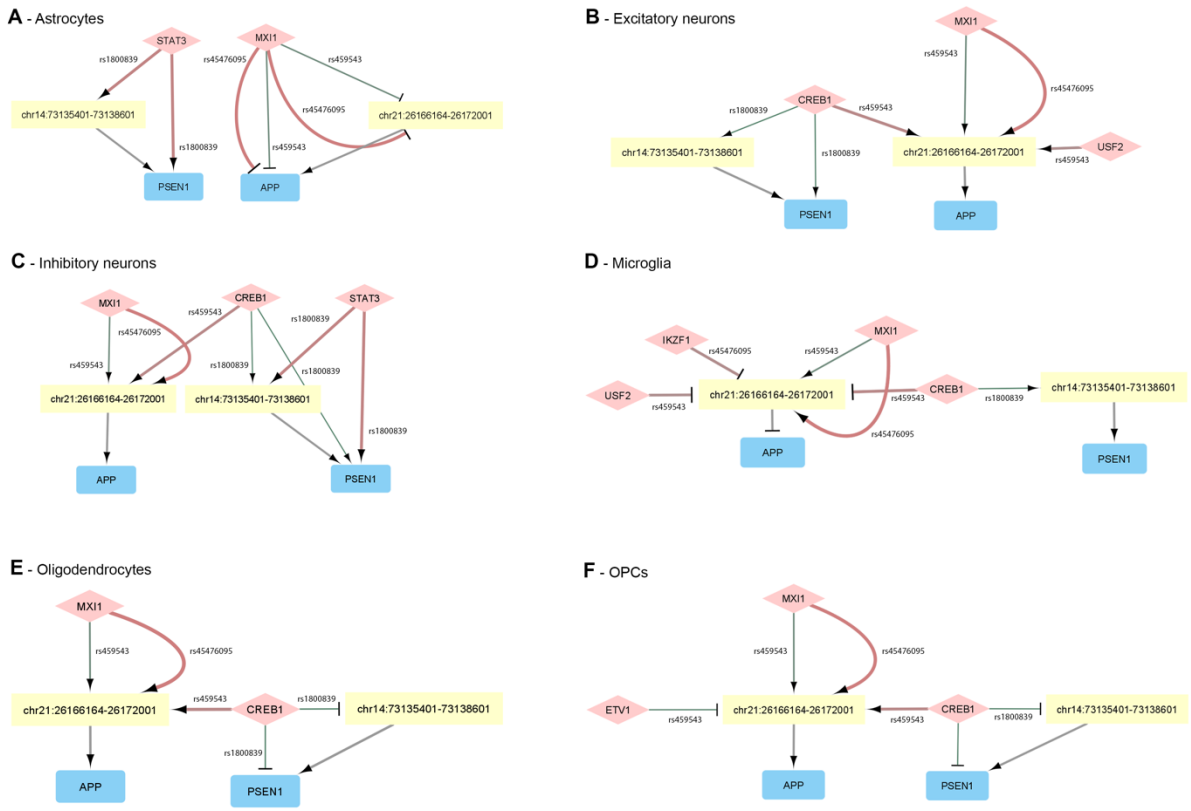


Fig. 4. Cell (sub)type specific regulatory impairment in AD. Network visualization of impaired regulatory interactions for (A) astrocytes, (B) excitatory neurons, (C) inhibitory neurons, (D) microglia, (E) oligodendrocytes and (F) OPCs. TFs are represented as diamonds in light red, enhancers as yellow rectangles and genes in blue rectangles. Arrows represent activations and T edges represent repressions. The weight of edges from TFs correspond to the strength of the impairment, with the thinnest translating a weaker binding affinity and a large edge being a strong increase in binding affinity. The color of the edges from TFs represents the log2FC with green being a decreased affinity and red an increased one.

Supplementary Information

Supplementary Methods

RNetDys workflow

Cell (sub)type and state specific GRN inference

The GRN inference part of RNetDys relies on the combination of multi-OMICS data including single cell datasets (scRNA-seq and scATAC-seq) and prior-knowledge (ChIP-seq and GeneHancer). First, a quality control is performed on the scRNA-seq and scATAC-seq in which any rows (gene or peaks) or columns (cells) having a sum of zero is removed from further analyses. Then, the following steps are computed to infer the cell (sub)type or state specific regulatory interactions:

(1) TF-Genes interactions: First, using the scRNA-seq data, we pre-selected genes conserved at least in 50% of the cells for candidate interactions. Indeed, we consider genes expressed in the majority of the cells to be representative in the specific cell (sub)type. In addition, from the scATAC-seq peaks matrix, coordinates are extracted to identify accessible promoter regions. Notably, a gene promoter region was identified from the ChIP-seq collected from ChIP-Atlas (Oki et al., 2018), using HOMER (Heinz et al., 2010) annotations by filtering peaks related to gene types annotated as protein coding, and defined as a region between 1500bp upstream and 500bp downstream. A promoter is considered as accessible if its gene has been considered as expressed (conserved at least in 50% of the cells) and at least one ATAC peak is overlapping. The overlap between promoter regions and the peaks coordinates was performed using BEDTools (Quinlan and Hall, 2010) with the parameter -f = 0.48 in reciprocal mode (-r). We identified the overlap parameter f = 0.48 as being the one with the highest probability to capture a real cell (sub)type accessible promoter region. The procedure used to select 0.48 is described in "Identification of accessible gene promoter regions" of the Supplementary Methods. Finally, the resulting overlapping between promoter regions and chromatin accessibility allow us to predict the cell (sub)type or state specific TF-Genes interactions.

(2) Enhancer-Promoters interactions: First, we identified open enhancer regions by intersecting the ChIP-seq data and the scATAC peaks coordinates using BEDTools

with the parameter -F 1.0 selecting open enhancer if 100% of the region is accessible. Then, we splitted the scATAC peaks matrix such that one matrix contains accessible promoter regions, obtained previously, and the other one accessible enhancer regions. We then computed the correlation between the two matrices, using the Pearson metric with the propagate R package (Andrej-Nikolai Spiess, 2018) that requires few computational resources to perform correlation of large matrices. Z-scores and corresponding p-values using a one-sided test on a normal distribution is performed for each pairwise correlation generated. Then, a Benjamini-Hochberg multiple test correction was performed on the computed p-values. The network was generated by selecting enhancer regions as sources, and promoter regions as targets, filtering the edges such as p-adjusted value < 0.05 and keeping promoters for which genes were found in the TF-Genes network. Notably, only positive correlation could be find as being significant as a negative correlation between accessibility peaks translate an absence of interaction between enhancers and promoters. We then retrieved the genes corresponding to the promoter regions using the ChIP-seq data used by RNetDys. Finally, the enhancer-promoter correlation network is intersected with all GeneHancer (Fishilevich et al., 2017) reported connections.

(3) TF-Enhancers interactions: First, enhancers present in the Enhancer-Promoter network are selected. They are then intersected with the ChIP-seq data, using bedtools and -F 1.0, such as if 100% of the TF peak fell inside the enhancer region, then this TF is interacting with the enhancer.

All the interactions of the comprehensive network were then signed based on the scRNA-seq dataset using the Pearson correlation metric between TFs and genes. For TF-Genes interactions, the correlation value defined the sign of the interactions such as positive correlations are most likely activation whereas negative ones are most likely repression. Then, signs for Enhancer-Promoter interactions were determined by computing the sum of correlation values for the TFs binding to the enhancer regulating the specific promoter/gene with the correlation corresponding to the TF-gene relationship (Figure S1) such as:

$$corV_{E_a \rightarrow G_b} = \sum_x corV_{TF_x \rightarrow G_b}$$

With corV: correlation value, TF: transcription factor, E: enhancer, G: gene

Finally, signs for TF-Enhancers were computed by summing, for each TF binding to the enhancer, the TF-genes relationship correlation values for each gene/promoter regulated by the enhancer (Figure S1) such as:

$$corV_{TF_a \to E_b} = \sum_x corV_{TF_a \to G_x}$$

With corV: correlation value, TF: transcription factor, E: enhancer, G: gene

Contextualization towards the disease state to identify candidate impaired interactions

Based on a GRN from a healthy cell (sub)type or state, the regulatory network was contextualized towards the disease condition of interest based on a list of SNPs. First, promoter regions coordinate for which a TF binding site has been identified is retrieved from the ChIP-seq data. Then, provided SNPs are mapped to these regions and enhancer regions of the GRN using bedtools under the condition that the SNP falls exactly inside one of the regions (parameter -F 1). This step allows the identification of candidate impaired regulatory interactions, including TF-genes and enhancer-promoters, for the specific cell (sub)type. Finally, a TF binding affinity analysis is performed on the SNP impacted regions. The fasta sequences for impacted enhancer and promoter regions were retrieved from genome.ucsc.edu accordingly with the genome assembly, 50bp upstream and downstream were selected from the SNP position and the SNP [ref/alt] alleles were added to the sequence. Then, we used PERFECTOS-APE (E. Vorontsov et al., 2015) to perform the TF motif binding affinity analysis for each SNP on each region found to be involved in regulation. Then, using the cell (sub)type specific GRN, TFs that were binding specifically on the impaired promoter or enhancer were retrieved as well as their dysregulated affinity score. Notably, we used PERFECTOS-APE with the following modified parameters: --pvalue-cutoff 0.05 --fold-change-cutoff 2. Finally, we ranked the TFs to prioritize the regulators that are impaired due to SNPs and hence are most likely to play a role in the dysregulations observed in the disease condition. The rank of each TF regulator was computed as follow:

$$Rank_{TF} = RE \times \frac{NG}{RE} \times \left(\sum |AI|_i^r \times \left(MAF_i^r \times \sum MAF^r\right)\right)$$

With RE: number of regulatory elements regulated by the TF, NG: number of downstream genes across RE, AI: binding affinity impairment log2FC, i: SNPs, r: regulatory element.

Identification of accessible gene promoter regions

We intersected ChIP-seq peaks related to gene promoter regions with ATAC peaks from scATAC-seq data to identify accessible cell (sub)type promoter regions using bedtools. In order to define the best threshold to use for the overlapping between the ChIP and ATAC peaks, we collected ChIP-seq from ChIP-ATLAS and compiled four human cell line specific ChIP-seq gold standards (BJ, GM12878, H1 ESC and K-562). We then used all the ChIP-seq collected from ChIP-ATLAS (specific) and considered a ChIP peak to be a true positive (TP) if it was found in the cell line specific GS and a false positive (FP) if it was not found in the GS. We computed the percentage of overlaps between ATAC peaks and TPs or FPs ChIP-peaks independently. Then, we computed the delta probability distribution such as: ecdf(TPs overlap) - ecdf (FPs overlap), and selected the highest point = 0.48. Indeed, 0.48 corresponded to the reciprocal threshold for which the probability to capture a TP (cell (sub)type specific ChIP peak) was the highest and was used as default by the RNetDys (Fig. S8).

Generation of the cell (sub)type specific GRNs in healthy condition

We collected scRNA-seq and scATAC-seq data from human pancreas and brain tissues (Table S2). The scRNA-seq datasets were processed using Seurat v4 (Hao et al., 2021) and, the gene expression and peaks matrices for each cell (sub)type were extracted for each tissue using Signac (Stuart et al., 2020). Annotations were used from their original studies for all tissues.

- Pancreas: we performed the peak calling with MACS2 (-q 0.05 --call-summits) for each cell (sub)type and the peak matrices were extracted for the cell (sub)types having a corresponding scRNA-seq matrix by using the FeatureMatrix function provided by Signac. We then used Seurat to extract all the cell (sub)type scRNA-seq matrices.
- Brain: several datasets were collected to match scRNA-seq and scATAC-seq data in order to extract cell (sub)types and states for different brain regions (Table S3). scATAC-seq fragment files were obtained after request to the authors and the general peaks matrix as well as metadata were retrieved from the public repository of their study (Corces et al., 2020). Each brain region-related scATAC-seq cell (sub)types clusters were annotated using Signac and Seurat with their matched scRNA-seq dataset (Table S3), whereas the cell type annotations were kept from the original

study (Corces et al., 2020). We performed the peak calling with MACS2 (-q 0.05 --call-summits) for each cell (sub)type in each brain region. The peak matrices were extracted for the cell (sub)types having a corresponding scRNA-seq matrix by using the FeatureMatrix function provided by Signac. We then used Seurat to extract all the cell (sub)type scRNA-seq matrices. First, we processed the frontal cortex data, imputed the dropouts using MAGIC due to the high rate of zeros (van Dijk et al., 2018) and used the annotations provided by the authors to extract the cell (sub)types (Lake et al., 2018). Of note, excitatory subtypes were merged as excitatory neurons and inhibitory ones as inhibitory neurons to match with the scATAC-seq. Then, we extracted the cell (sub)types of the substantia nigra for healthy patients while keeping the annotations provided by the authors (Smajić et al., 2022).

Each cell (sub)type GRN was generated using the extracted scRNA-seq and scATAC-seq datasets with the GRN inference part of RNetDys using the default parameters.

GRN inference benchmarking and comparison to state-of-the-art

We first assessed the performances of RNetDys to capture cell (sub)type specific TF-Gene interactions and compared to state-of-the-art methods including CLR (Zhang et al., 2016), GENIE3 (Huynh-Thu et al., 2010), SCENIC (Aibar et al., 2017), PIDC (Chan et al., 2017) and ppcor (Kim, 2015). All methods were used with default parameters to infer the TF-Genes networks and applied to 20 single cell RNA-seq datasets collected from six human cell lines (A549, Jurkat, K-562, GM12878, H1 ESC, BJ). Of note, only genes expressed at least in 50% of the cells for each scRNA-seq dataset were provided to the methods to be consistent for the comparison with RNetDys. In addition, predicted (un)directed GRNs were formatted to obtain TF-gene networks by filtering the Source (regulator) such that it contains any human TFs or co-TFs reported in Animal TFDB (accessed on the 08/04/2022)(Hu et al., 2019). Notably, due to large computational resources or a running time higher than two days, five networks could not be generated, including scRNA-seq datasets of one K562, one GM12878 and three H1-ESCs. RNetDys was used with default parameters on the 20 scRNA-seq datasets and scATAC-seq datasets retrieved for each of the six human cell lines (Table S1). We benchmarked the inferred networks against cell line specific GS standard networks compiled from the Cistrome database and computed the precision (PPV) and accuracy (F1-score). Of note, more than one network was generated by RNetDys for each scRNA-seq dataset used for other methods, depending on the number of scATAC-seq datasets. We hence

computed the median PPV and F1 score over the networks to have one metric by scRNA-seq, as we had for each state-of-the-art method. We then assessed the performances of RNetDys in capturing cell (sub)type specific enhancer-promoter regulatory interactions. State-of-the-art methods used for the TF-Gene benchmarking did not account for enhancers, as they solely relied on scRNA-seq, and hence we performed a comparison using Cicero (Pliner et al., 2018), a widely used strategy to identify co-accessibility between regulatory regions based on scATAC-seq. We applied RNetDys on twelve combinations of scRNA-seq and scATAC-seq datasets for three human cell lines (Table S1) for which we could compile reliable cell line specific gold standard networks from 3DIV database (GM12878, H1 ESC, BJ/IMR90). We used Cicero on the scATAC-seq datasets using default parameters and annotated the enhancer and promoter regions using the ChIP-seq leveraged by RNetDys. Notably, not significance score was provided on the interactions and hence, accordingly with Cicero guideline we selected interactions with a co-accessibility score greater than zero. Finally, we benchmarked the predicted networks against the human cell line specific GS networks to compute the PPV and F1-scores.

Compilation of the gold standard networks

We compiled two types of GS networks, both directed, to assess the performances and validate the specificity in identifying cell (sub)type specific regulatory interactions:

(1) TF-Genes GS networks: for each human cell line, we collected high quality ChIP-seq data specific to the cell line from Cistrome (Mei et al., 2017). The highest quality was defined as peak data passing all the quality control available in Cistrome.
(2) Enhancer-promoter GS networks: for each human cell line, we collected Promoter Capture Hi-C data from 3DIV (Yang et al., 2018) database. We then filtered the GS networks to retain enhancers found in GeneHancer and gene promoter regions defined in the ChIP-seq data retrieved from ChIP-Atlas using BEDTools (Quinlan and Hall, 2010).

Cell (sub)type specific regulatory mechanisms impaired in diseases

We performed a general study of cell (sub)type specific impairment in diseases by using prior-knowledge SNPs to validate the relevance of the captured interactions. We first collected single nucleotide variants from ClinVar (Landrum et al., 2018) and extracted SNPs such as the SNV was found at least in 1% of the global population (MAF >= 0.01). Of note, MAF scores were retrieved for each SNV using BioMart R package and the 'hsapiens_snp'

dataset. Then, we extracted the SNPs for each disease by selecting the ones that have been reported as being related to the disease in ClinVar and, we performed a systematic extraction using regex with the disease name as pattern. Finally, for each cell (sub)type and each disease, we applied RNetDys using the cell (sub)type GRN and the list of SNPs to capture candidate impaired regulatory interactions, TF binding impairment information and the ranked regulators. Notably, SNPs related to AD were mapped to the brain cortex networks whereas SNPs related to PD were mapped to the midbrain networks.

Supplementary References

Aibar,S. et al. (2017) SCENIC: single-cell regulatory network inference and clustering. Nat Methods, 14, 1083–1086.

Andrej-Nikolai Spiess (2018) R Package 'propagate'.

Chan,T.E. et al. (2017) Gene Regulatory Network Inference from Single-Cell Data Using Multivariate Information Measures. Cell Systems, 5, 251-267.e3.

Corces,M.R. et al. (2020) Single-cell epigenomic analyses implicate candidate causal variants at inherited risk loci for Alzheimer's and Parkinson's diseases. Nat Genet, 52, 1158–1168.

van Dijk,D. et al. (2018) Recovering Gene Interactions from Single-Cell Data Using Data Diffusion. Cell, 174, 716-729.e27.

E. Vorontsov,I. et al. (2015) PERFECTOS-APE - Predicting Regulatory Functional Effect of SNPs by Approximate P-value Estimation: In, Proceedings of the International Conference on Bioinformatics Models, Methods and Algorithms. SCITEPRESS - Science and and Technology Publications, Lisbon, Portugal, pp. 102–108.

Fishilevich,S. et al. (2017) GeneHancer: genome-wide integration of enhancers and target genes in GeneCards. Database, 2017.

Hao,Y. et al. (2021) Integrated analysis of multimodal single-cell data. Cell, 184, 3573-3587.e29.

Heinz,S. et al. (2010) Simple Combinations of Lineage-Determining Transcription Factors Prime cis-Regulatory Elements Required for Macrophage and B Cell Identities. Molecular Cell, 38, 576–589.

Hu,H. et al. (2019) AnimalTFDB 3.0: a comprehensive resource for annotation and prediction of animal transcription factors. Nucleic Acids Research, 47, D33–D38.

Huynh-Thu,V.A. et al. (2010) Inferring Regulatory Networks from Expression Data Using Tree-Based Methods. PLoS ONE, 5, e12776.

Kim,S. (2015) ppcor: An R Package for a Fast Calculation to Semi-partial Correlation Coefficients. Communications for Statistical Applications and Methods, 22, 665–674.

Lake,B.B. et al. (2018) Integrative single-cell analysis of transcriptional and epigenetic states in the human adult brain. Nat Biotechnol, 36, 70–80.

Landrum,M.J. et al. (2018) ClinVar: improving access to variant interpretations and supporting evidence. Nucleic Acids Res, 46, D1062–D1067.

Mei,S. et al. (2017) Cistrome Data Browser: a data portal for ChIP-Seq and chromatin accessibility data in human and mouse. Nucleic Acids Res., 45, D658–D662.

Oki,S. et al. (2018) Ch IP -Atlas: a data-mining suite powered by full integration of public Ch IP -seq data. EMBO Rep, 19.

Pliner,H.A. et al. (2018) Cicero Predicts cis-Regulatory DNA Interactions from Single-Cell Chromatin Accessibility Data. Mol Cell, 71, 858-871.e8.

Quinlan,A.R. and Hall,I.M. (2010) BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics, 26, 841–842.

Smajić,S. et al. (2022) Single-cell sequencing of human midbrain reveals glial activation and a Parkinson-specific neuronal state. Brain, 145, 964–978.

Stuart,T. et al. (2020) Multimodal single-cell chromatin analysis with Signac Genomics.

Yang,D. et al. (2018) 3DIV: A 3D-genome Interaction Viewer and database. Nucleic Acids Research, 46, D52–D57.

Zhang,L. et al. (2016) Reconstructing directed gene regulatory network by only gene expression data. BMC Genomics, 17 Suppl 4, 430.

Supplementary Figures

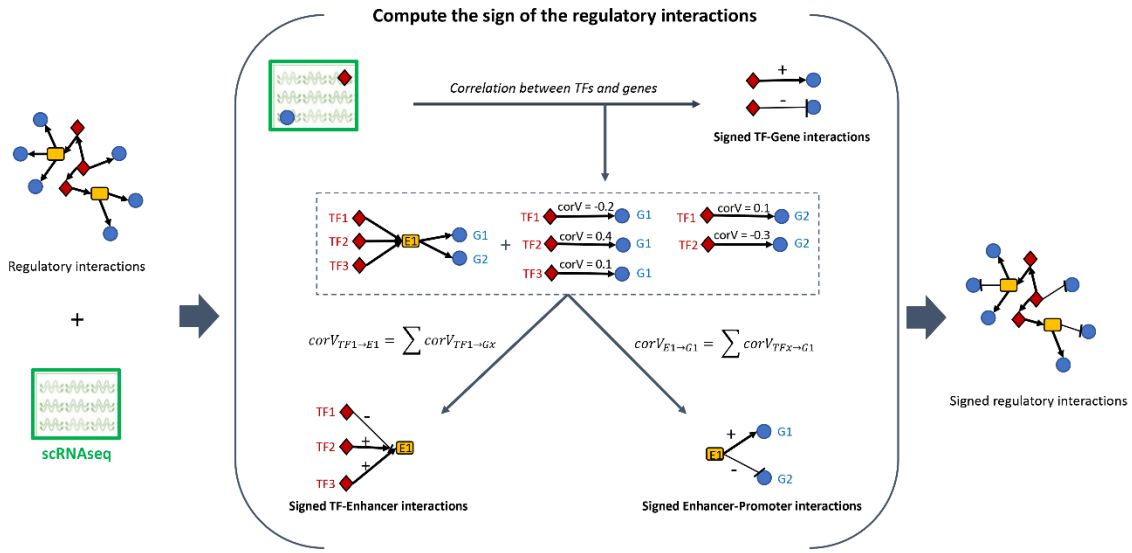


Fig. S1. Strategy to compute the sign of the regulatory interactions. The scRNA-seq dataset is used to compute the correlation between the TFs and genes of the GRN. TF-Gene interactions are directly signed using the correlation values. Enhancer-Promoter interactions are signed by summing the correlation values between the TFs binding to the enhancer and the regulated gene/promoter. TF-Enhancer interactions are signed by computing for each TF the sum of the correlation values between the TF and the genes regulated by the enhancer.

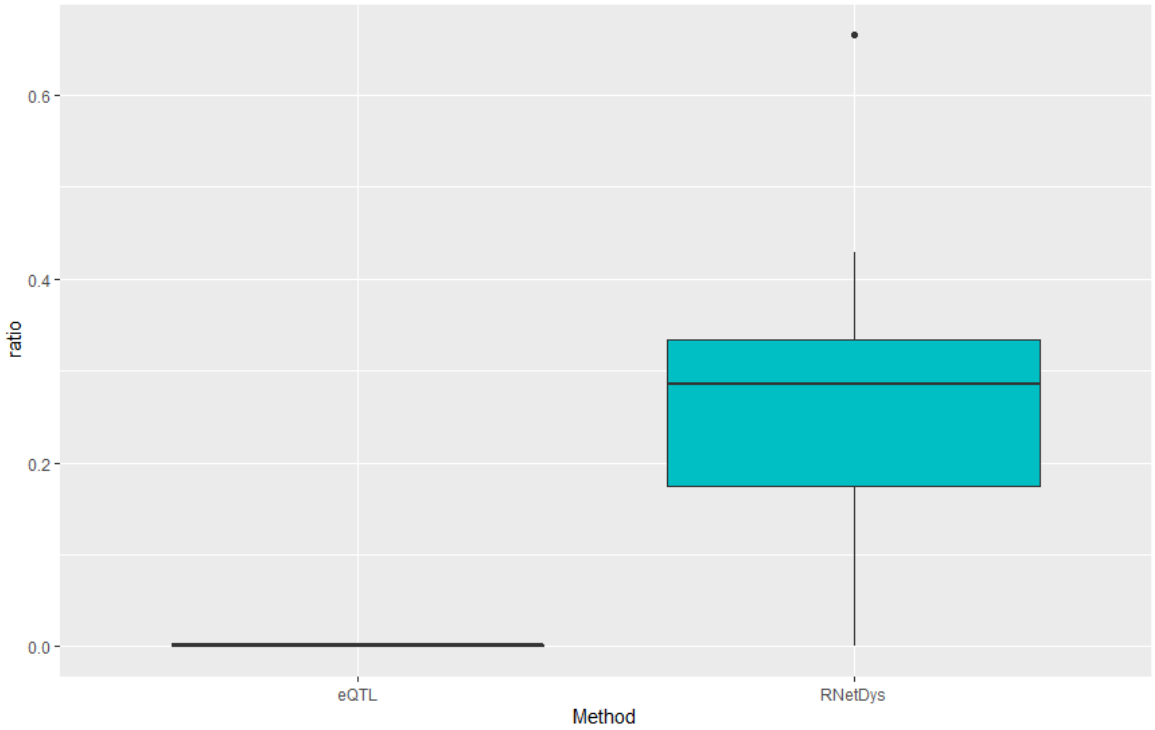


Fig. S2. Comparison of the precision in the identification of gene-disease interactions between eQTL and RNetDys. Ratio for the captured genes reported as linked to the disease according to OMIM is represented in y axis. Each boxplot represents ratios across all cell (sub)types for AD, EPI, PD, T1D and T2D.

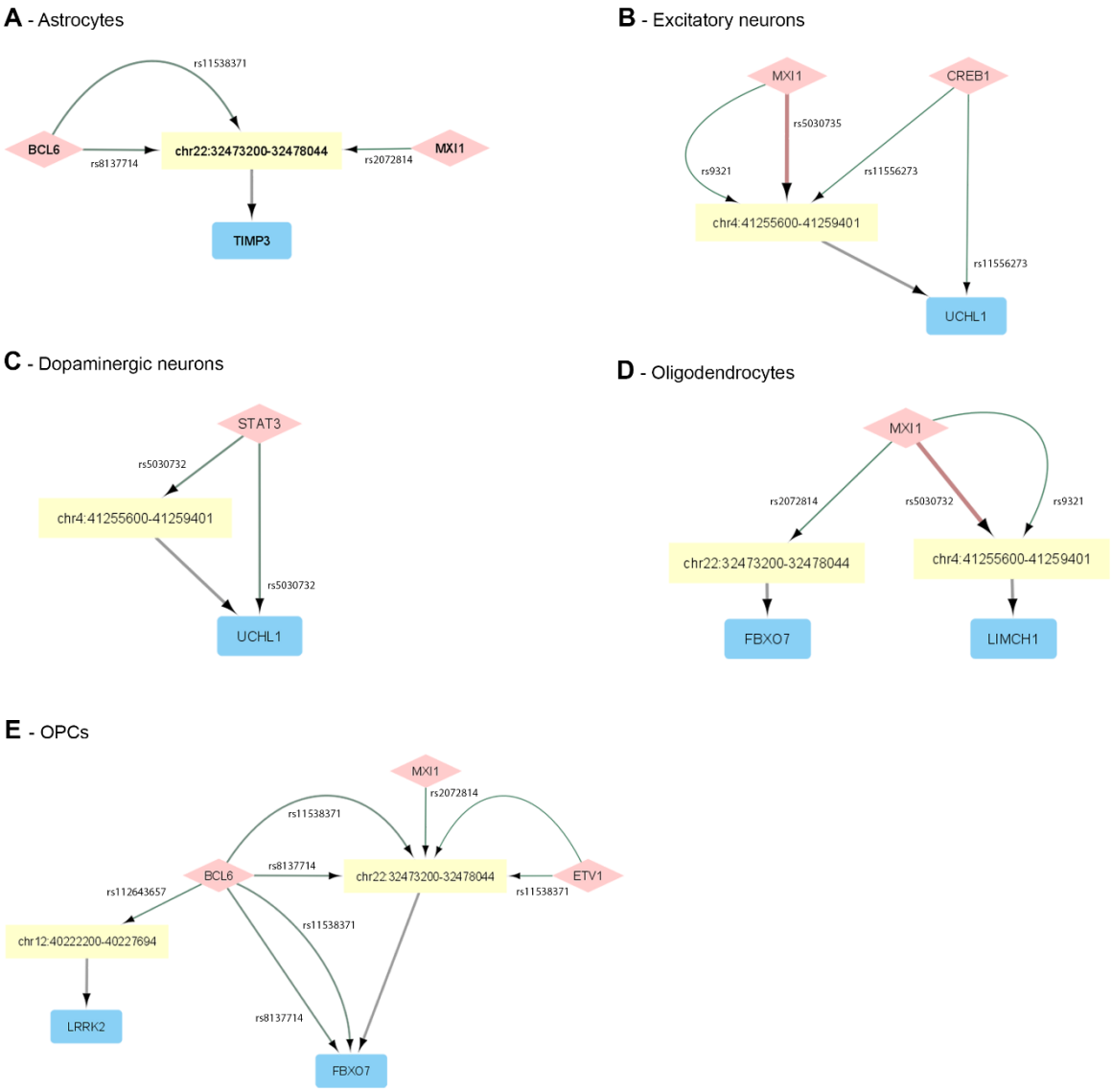


Fig. S3. Cell (sub)type specific regulatory impairment in PD. Network visualization of impaired regulatory interactions for (A) astrocytes, (B) excitatory neurons, (C) dopaminergic neurons, (D) oligodendrocytes and (E) OPCs. TFs are represented as diamond in light red, enhancers as yellow rectangles and genes in blue rectangles. Arrows represent activations. The weight of edges from TFs correspond to the strength of the impairment, with the thinnest translating a strong lack of binding affinity and a large edge being a strong increase in binding affinity. The color of the edges from TFs represents the log2FC with green being a decreased affinity and red an increased one.

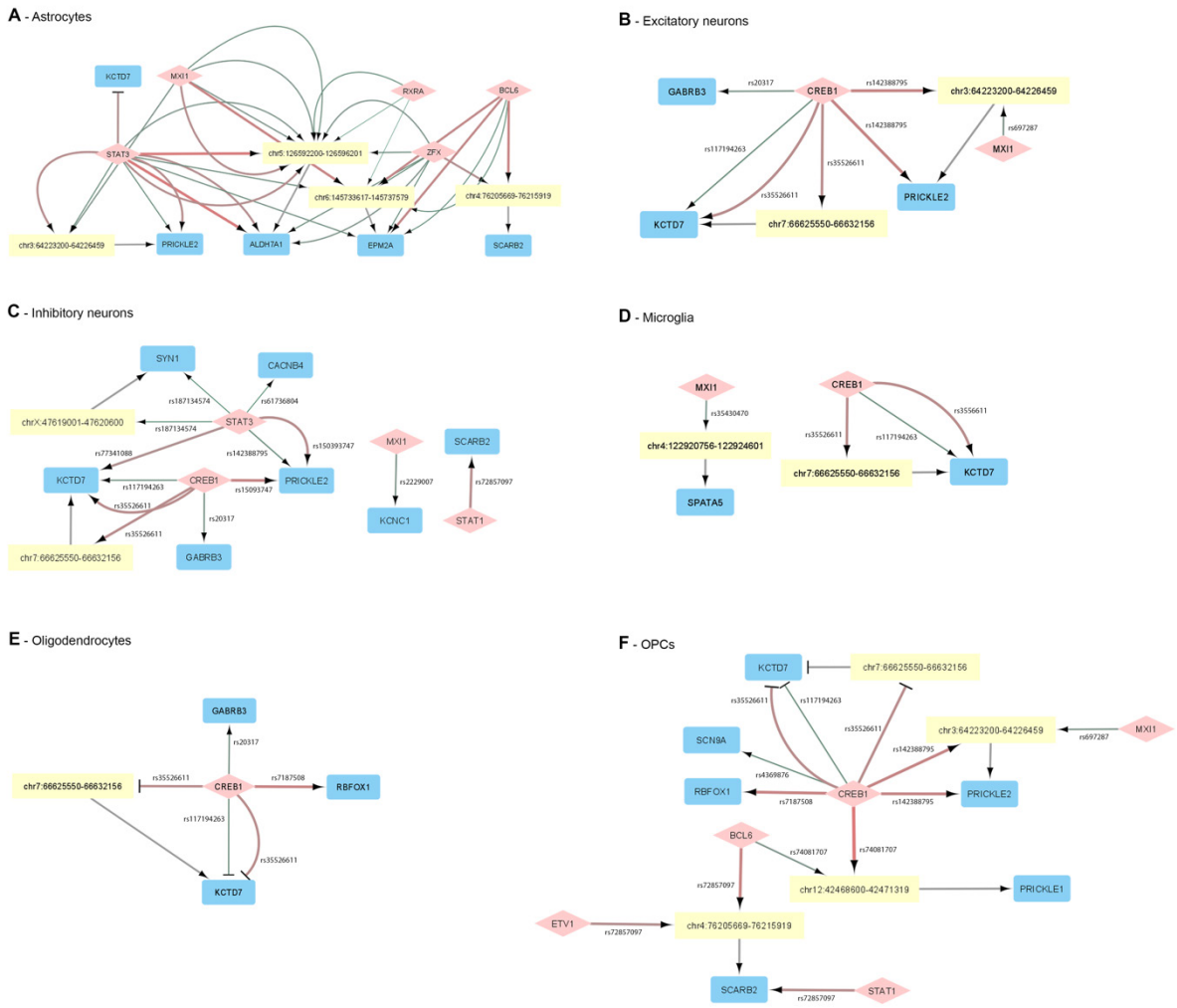


Fig. S4. Cell (sub)type specific regulatory impairment in EPI. Network visualization of impaired regulatory interactions for (A) astrocytes, (B) excitatory neurons, (C) inhibitory neurons, (D) microglia, (E) oligodendrocytes and (F) OPCs. TFs are represented as diamond in light red, enhancers as yellow rectangles and genes in blue rectangles. Arrows represent activations and T edges represent repressions. The weight of edges from TFs correspond to the strength of the impairment, with the thinnest translating a strong lack of binding affinity and a large edge being a strong increase in binding affinity. The color of the edges from TFs represents the log2FC with green being a decreased affinity and red an increased one.

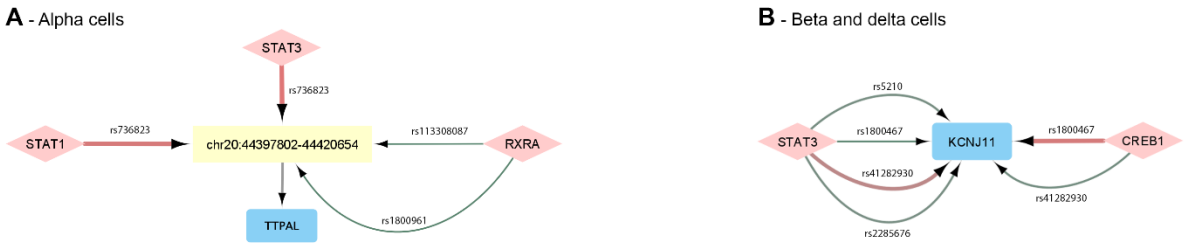


Fig. S5. Cell type specific impairment in T1D. Network visualization of impaired regulatory interactions for (A) alpha cells and (B) beta and delta cells. TFs are represented as diamond in light red, enhancers as yellow rectangles and genes in blue rectangles. Arrows represent activations. The weight of edges from TFs correspond to the strength of the impairment, with the thinnest translating a strong lack of binding affinity and a large edge being a strong increase in binding affinity. The color of the edges from TFs represents the log2FC with green being a decreased affinity and red an increased one.

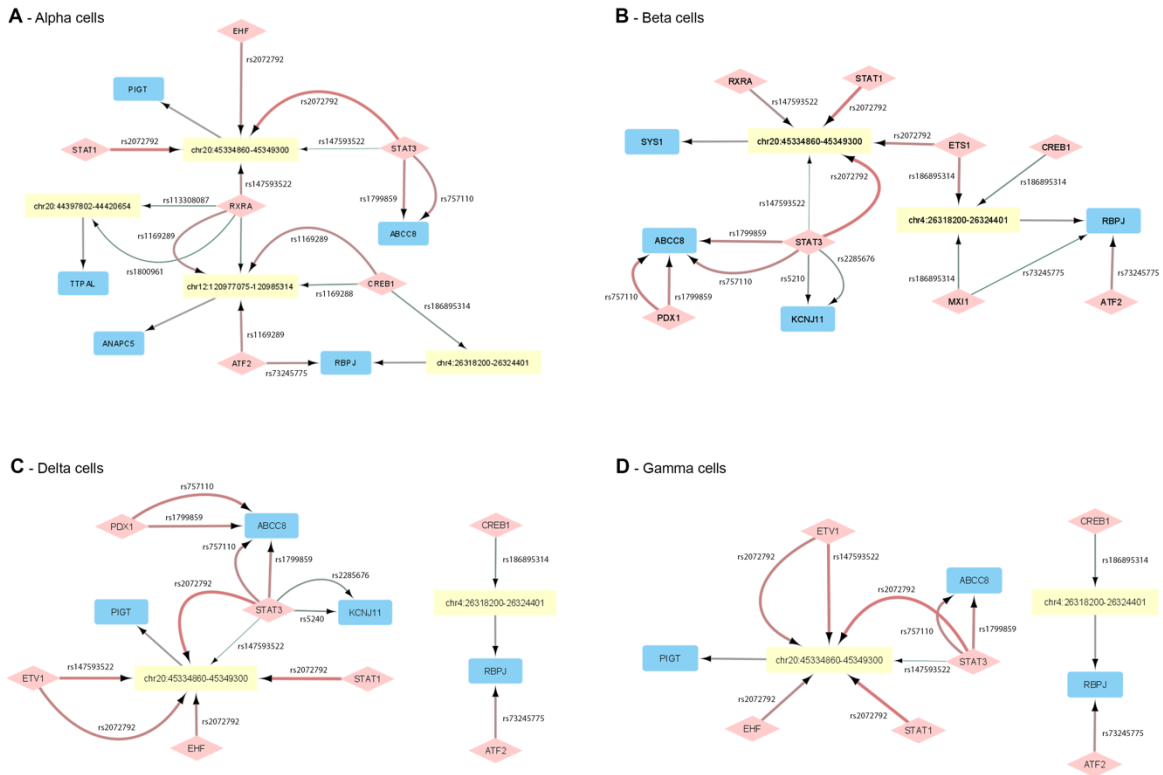


Fig. S6. Cell type specific impairment in T2D. Network visualization of impaired regulatory interactions for (A) alpha cells, (B) beta cells, (C) delta cells and (D) gamma cells. TFs are represented as diamond in light red, enhancers as yellow rectangles and genes in blue rectangles. Arrows represent activations. The weight of edges from TFs correspond to the strength of the

impairment, with the thinnest translating a strong lack of binding affinity and a large edge being a strong increase in binding affinity. The color of the edges from TFs represents the log2FC with green being a decreased affinity and red an increased one.

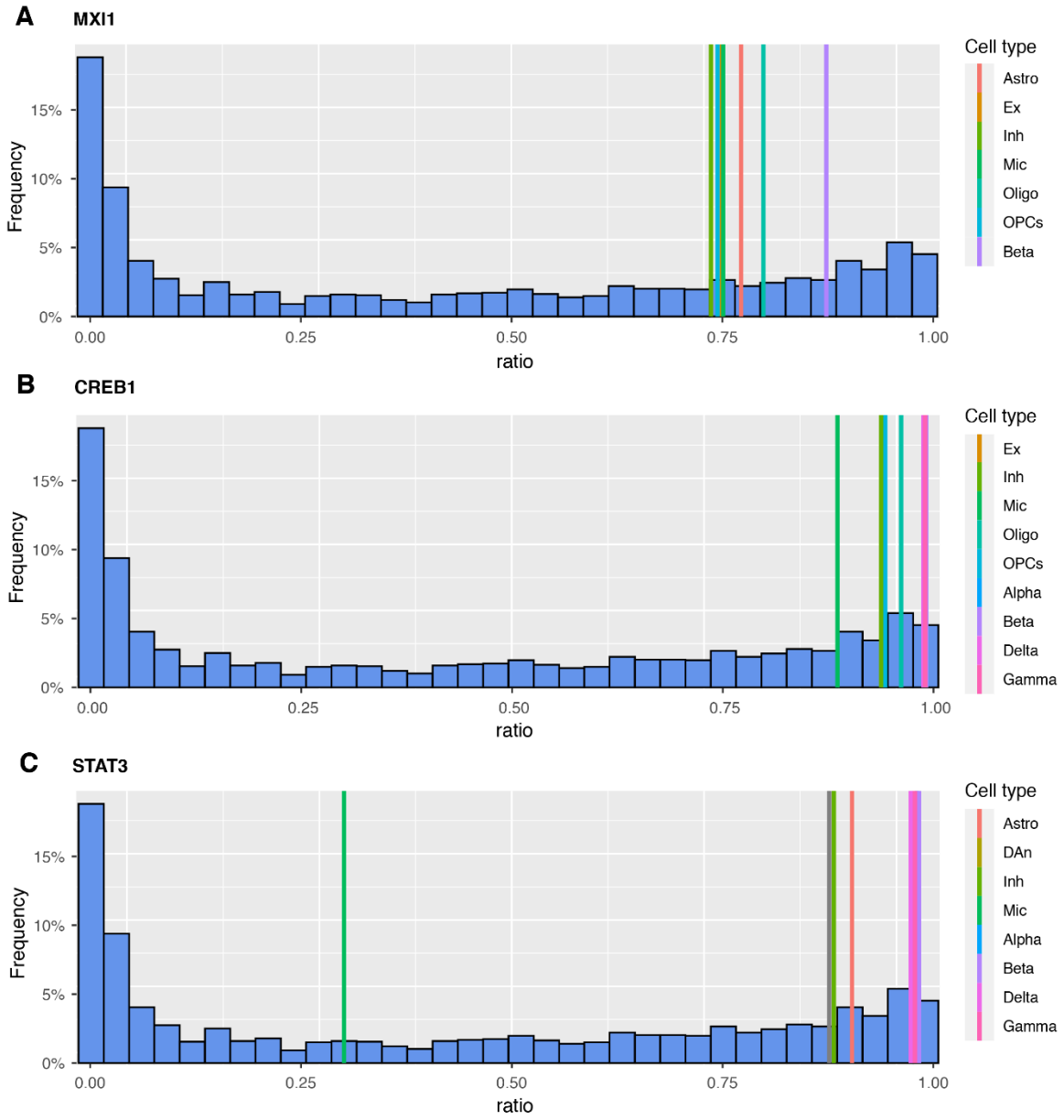


Fig. S7. Distribution of the outdegree ratio for specific TFs across cell (sub)types. Histogram showing the frequency of outdegree ratios across all cell (sub)types for three TFs. The outdegree ratio of (A) MXI1, (B) CREB1, and (C) STAT3 in each specific cell (sub)type is represented by coloured vertical lines in the histograms. Astro: astrocytes, Ex: excitatory neurons, DAn: dopaminergic neurons, Inh: inhibitory neurons, Mic: microglia, Oligo: oligodendrocytes, OPCs: oligodendrocyte progenitors, Alpha: alpha cells, Beta: beta cells, Delta: delta cells, Gamma: gamma cells.

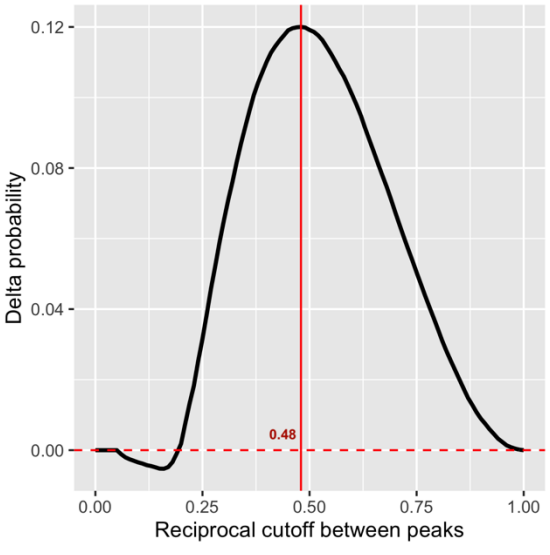


Fig. S8. Threshold selection to define accessibility of promoter regions. Delta probability between true positives and false positives. The peak of the distribution, equal to 0.48, corresponds to the highest probability to capture a true accessible promoter region in the cell (sub)type.

Supplementary Tables

Table S1. Single cell datasets used for validation and comparison

Accession Number	Cell line	Type of data	TF-Promoter benchmarking	Enhancer-Promoter benchmarking
GSE100344	BJ	scRNA-seq	X	X
GSE113415	BJ	scRNA-seq	X	X
GSE160910	BJ	scRNA-seq	X	X
GSE166935	BJ	scRNA-seq	X	X
scOpen*	BJ	scATAC-seq	X	X
GSE99172	BJ	scATAC-seq	X	X
GSE81861	GM12878	scRNA-seq	X	X
GSM3596321	GM12878	scRNA-seq	X	X
GSM4156602	GM12878	scRNA-seq	X	X
GSM4156603	GM12878	scRNA-seq	X	X
scOpen*	GM12878	scATAC-seq	X	X
GSE99172	GM12878	scATAC-seq	X	X
GSE64016	H1-ESC	scRNA-seq	X	X
GSE75748	H1-ESC	scRNA-seq	X	X
GSE81861	H1-ESC	scRNA-seq	X	X
GSM5534158	H1-ESC	scRNA-seq	X	X
scOpen*	H1-ESC	scATAC-seq	X	X
GSE99172	H1-ESC	scATAC-seq	X	X
GSE81861	A549	scRNA-seq	X	
GSM3271042	A549	scRNA-seq	X	
GSM3271043	A549	scATAC-seq	X	
GSM4224433	A549	scATAC-seq	X	

GSE105451	Jurkat	scRNA-seq	X
10x platform**	Jurkat	scRNA-seq	X
GSE107816	Jurkat	scATAC-seq	X
GSE81861	K562	scRNA-seq	X
GSE90063	K562	scRNA-seq	X
GSE113415	K562	scRNA-seq	X
GSM1599500	K562	scRNA-seq	X
scOpen*	K562	scATAC-seq	X
GSE99172	K562	scATAC-seq	X

*scOpen: https://github.com/CostaLab/scopen-reproducibility

**10x platform: https://www.10xgenomics.com/resources/datasets/jurkat-cells-1-standard-1-1-0

Table S2. Collected datasets to generate healthy cell (sub)type GRNs.

System	Accession	Type of data
Pancreas	GSE85241	scRNA-seq
	GSM558939	scATAC-seq
Brain	GSE157783 (Healthy)	scRNA-seq
	GSE97942	scRNA-seq
	GSE147672	scATAC-seq

Table S3. Matching of the scRNA-seq and scATAC-seq brain datasets.

scATAC-seq Brain Regions	scRNA-seq Brain Region Matched	Brain region abbreviation
Substantia Nigra	Human Midbrain (GSE157783, Healthy)	SUNI
Middle Frontal Gyrus	Frontal Cortex (GSE97942)	MDFG

Table S4. Literature-based validation of the predicted impaired regulatory interactions.

PD						
Source (TF or enhancer)	Gene	RSID	Cell (sub)pop	GWAS		Cell type specific e-QTL*
				SNP Linked to gene	PMID	SNP Linked to gene
chr22:32473200-32478044	TIMP3	rs11538371	Astro			x
chr22:32473200-32478044	TIMP3	rs2072814	Astro			x
chr22:32473200-32478044	TIMP3	rs8137714	Astro			x
chr4:41255600-41259401	UCHL1	rs5030732	DAn	x	28253266, 25370916, 22839974	x
STAT3	UCHL1	rs5030732	DAn	x		x
NFKB1, STAT3	PRKAG2	rs117728810	DAn	x		x
NFKB1, STAT3	PRKAG2	rs66628686	DAn	x		x

STAT3	PRKAG2	rs77902041	DAn	x		x
chr4:41255600-41259401	UCHL1	rs11556273	Ex	x		x
chr4:41255600-41259401	UCHL1	rs5030732	Ex	x		x
chr4:41255600-41259401	UCHL1	rs9321	Ex	x		x
CREB1	UCHL1	rs11556273	Ex	x		x
chr22:32473200-32478044	FBXO7	rs2072814	Oligo	x		x
chr4:41255600-41259401	LIMCH1	rs5030732	Oligo			x
chr4:41255600-41259401	LIMCH1	rs9321	Oligo			x
chr22:32473200-32478044	FBXO7	rs11538371	OPCs	x		x
BCL6	FBXO7	rs11538371	OPCs	x		x
chr22:32473200-32478044	FBXO7	rs2072814	OPCs	x		x
chr22:32473200-32478044	FBXO7	rs8137714	OPCs	x	18513678	x
BCL6	FBXO7	rs8137714	OPCs	x		x
chr12:40222200-40227694	LRRK2	rs112643657	OPCs	x		
AD						
Source	**Target**	**RSID**	**Pop**	**GWAS**		**Cell type specific e-QTL***
				Linked to gene	**PMID**	**Linked to gene**
chr14:73135401-73138601	PSEN1	rs1800839	Astro	x	28821390, 11389157	x
STAT3	PSEN1	rs1800839	Astro	x		x
chr21:26166164-26172001	APP	rs45476095	Astro	x	21654062	
MXI1	APP	rs45476095	Astro	x		
chr14:73135401-73138601	APP	rs459543	Astro	x		
MXI1	APP	rs459543	Astro	x	21654062, 16685645	
chr14:73135401-73138601	PSEN1	rs1800839	Ex	x	28821390, 11389157	x
CREB1	PSEN1	rs1800839	Ex	x		x
chr21:26166164-26172001	APP	rs45476095	Ex	x	21654062	
chr21:26166164-26172001	APP	rs459543	Ex	x	21654062, 16685645	
chr14:73135401-73138601	PSEN1	rs1800839	Inh	x	28821390, 11389157	
CREB1, STAT3	PSEN1	rs1800839	Inh	x		
chr21:26166164-26172001	APP	rs45476095	Inh	x	21654062	
chr21:26166164-26172001	APP	rs459543	Inh	x	21654062, 16685645	
chr21:26166164-26172001	APP	rs1800839	Mic			
chr21:26166164-26172001	APP	rs45476095	Mic	x	21654062	
chr14:73135401-73138601	APP	rs459543	Mic	x	21654062, 16685645	
CREB1	PSEN1	rs1800839	Oligo	x	28821390, 11389157	

chr21:26166164-26172001	APP	rs45476095	Oligo	x	21654062	
chr14:73135401-73138601	APP	rs459543	Oligo	x	21654062, 16685645	
chr14:73135401-73138601	PSEN1	rs1800839	OPCs	x		x
CREB1	PSEN1	rs1800839	OPCs	x	28821390, 11389157	x
chr21:26166164-26172001	APP	rs45476095	OPCs	x		
chr21:26166164-26172001	APP	rs459543	OPCs	x		

EPI						
Source	**Target**	**RSID**	**Pop**	**GWAS**		**Cell type specific e-QTL***
				Linked to gene	**PMID**	**Linked to gene**
chr5:126592200-126596201	ALDH7A1	rs144272515	Astro	x		x
ZFX	ALDH7A1	rs144272515	Astro	x		x
chr3:64223200-64226459	PRICKLE2	rs697287	Astro	x		x
chr3:64223200-64226459	PRICKLE2	rs900641	Astro			
chr3:64223200-64226459	PRICKLE2	rs142388795	Astro	x		
STAT3	PRICKLE2	rs142388795	Astro	x		
chr5:126592200-126596201	ALDH7A1	rs146562077	Astro	x		
STAT3	ALDH7A1	rs146562077	Astro	x		
chr3:64223200-64226459	PRICKLE2	rs150393747	Astro	x		
STAT3	PRICKLE2	rs150393747	Astro	x		
chr6:145733617-145737579	EPM2A	rs2235482	Astro	x		
BCL6, STAT3, ZFX	EPM2A	rs2235482	Astro	x		
chr6:145733617-145737579	EPM2A	rs374338349	Astro	x	11735300	
BCL6	EPM2A	rs374338349	Astro	x		
chr5:126592200-126596201	ALDH7A1	rs60720055	Astro	x		
chr5:126592200-126596201	ALDH7A1	rs72857097	Astro			
STAT3	KCTD7	rs77341088	Astro	x		
chr5:126592200-126596201	ALDH7A1	rs900640	Astro	x		
STAT3	ALDH7A1	rs900640	Astro	x		
ZFX	ALDH7A1	rs900640	Astro	x		
chr3:64223200-64226459	PRICKLE2	rs697287	Ex	x		x
CREB1	GABRB3	rs20317	Ex	x	30074174, 24999380, 25025424	x
CREB1	KCTD7	rs117194263	Ex	x		
chr3:64223200-64226459	PRICKLE2	rs142388795	Ex	x		
CREB1	PRICKLE2	rs142388795	Ex	x		
chr7:66625550-66632156	KCTD7	rs35526611	Ex	x		
CREB1	KCTD7	rs35526611	Ex	x		

CREB1	GABRB3	rs20317	Inh	x	30074174, 24999380, 25025424	x
CREB1	KCTD7	rs117194263	Inh	x		
CREB1, STAT3	PRICKLE2	rs142388795	Inh	x		
STAT3	PRICKLE2	rs150393747	Inh	x		
MXI1	KCNC1	rs2229007	Inh	x		
chr7:66625550-66632156	KCTD7	rs35526611	Inh	x		
CREB1	KCTD7	rs35526611	Inh	x		
STAT3	CACNB4	rs61736804	Inh	x		
STAT1	SCARB2	rs72857097	Inh	x		
STAT3	KCTD7	rs77341088	Inh	x		
chrX:47619001-47620600	SYN1	rs187134574	Inh	x		No data on chrX
STAT3	SYN1	rs187134574	Inh	x		No data on chrX
CREB1	KCTD7	rs117194263	Mic	x		
chr4:122920756-122924601	SPATA5	rs35430470	Mic	x		
chr7:66625550-66632156	KCTD7	rs35526611	Mic	x		
CREB1	KCTD7	rs35526611	Mic	x		
CREB1	KCTD7	rs117194263	Oligo	x		
CREB1	GABRB3	rs20317	Oligo	x	30074174, 24999380, 25025424	
chr7:66625550-66632156	KCTD7	rs35526611	Oligo	x		
CREB1	KCTD7	rs35526611	Oligo	x		
CREB1	RBFOX1	rs7187508	Oligo	x		
chr3:64223200-64226459	PRICKLE2	rs697287	OPCs	x		x
CREB1	KCTD7	rs117194263	OPCs	x		
chr3:64223200-64226459	PRICKLE2	rs142388795	OPCs	x		
CREB1	PRICKLE2	rs142388795	OPCs	x		
chr7:66625550-66632156	KCTD7	rs35526611	OPCs	x		
CREB1	KCTD7	rs35526611	OPCs	x		
CREB1	SCN9A	rs4369876	OPCs	x	23292638, 21698661	
CREB1	RBFOX1	rs7187508	OPCs	x		
chr4:76205669-76215919	SCARB2	rs72857097	OPCs	x		
STAT1	SCARB2	rs72857097	OPCs	x		
chr12:42468600-42471319	PRICKLE1	rs74081707	OPCs	x		

T1D

Source	Target	RSID	Pop	GWAS		Cell type specific e-QTL
				Linked to gene	PMID	Linked to gene
chr20:44397802-44420654	TTPAL	rs113308087	Alpha			No data
chr20:44397802-44420654	TTPAL	rs1800961	Alpha			

chr20:44397802-44420654	TTPAL	rs736823	Alpha			
CREB1, STAT3	KCNJ11	rs1800467	Beta	x	25733456, 26937418, 25247988	
STAT3	KCNJ11	rs2285676	Beta	x	32930968, 29903275, 27249660	
CREB1, STAT3	KCNJ11	rs41282930	Beta	x	25247988, 22289434, 15115830	
STAT3	KCNJ11	rs5210	Beta	x	32693412, 33101408, 30641791	
CREB1, STAT3	KCNJ11	rs1800467	Delta	x	25733456, 26937418, 25247988	
STAT3	KCNJ11	rs2285676	Delta	x	32930968, 29903275, 27249660	
CREB1, STAT3	KCNJ11	rs41282930	Delta	x	25247988, 22289434, 15115830	
STAT3	KCNJ11	rs5210	Delta	x	32693412, 33101408, 30641791	
T2D						
Source	**Target**	**RSID**	**Pop**	**GWAS**		**Cell type specific e-QTL**
				Linked to gene	**PMID**	**Linked to gene**
chr20:44397802-44420654	TTPAL	rs113308087	Alpha			
chr20:44397802-44420654	TTPAL	rs1169288	Alpha			
chr12:120977075-120985314	ANAPC5	rs1169289	Alpha			
chr20:45334860-45349300	PIGT	rs147593522	Alpha			
STAT3	ABCC8	rs1799859	Alpha	x	28587604, 26740944	
chr20:44397802-44420654	TTPAL	rs1800961	Alpha			
chr4:26318200-26324401	RBPJ	rs186895314	Alpha	x		
chr20:44397802-44420654	TTPAL	rs2072792	Alpha			No data
ATF2	RBPJ	rs73245775	Alpha	x		
STAT3	ABCC8	rs757110	Alpha	x	32660410, 32468916, 32930968	
chr20:45334860-45349300	SYS1	rs147593522	Beta			
PDX1, STAT3	ABCC8	rs1799859	Beta	x	28587604, 26740944	
chr4:26318200-26324401	RBPJ	rs186895314	Beta	x		
chr20:45334860-45349300	SYS1	rs2072792	Beta			

*https://zenodo.org/record/6104982#.Yq2eUy0RryY

5. Discussion and Perspectives

The human brain is a complex organ that the research community has been striving to understand. Deciphering the intricate relationships among the multiple brain cell types and the underlying molecular mechanisms can significantly contribute to the development of more effective treatments for brain diseases.

The emergence of single-cell technologies paved the way for a more in-depth characterization of the brain's cellular heterogeneity by unveiling previously undefined cellular subtypes [81,83,309]. Computational approaches have been leveraging the high resolutions of this data to uncover new biological insights that can further demonstrate the crucial role these specialized cell subtypes play in the development of the most common brain diseases [76,83–85,310]. Identifying cellular identity TFs at the subtype level has been one of the focuses of these algorithms due to the potential importance of these TFs in improving the development of regenerative medicine applications. Being able to successfully generate a cellular subpopulation of choice can have vital applications in different fields, namely CRT, disease modelling, and drug discovery [311]. Furthermore, identifying disease-specific or disease-prevalent cellular populations has the potential to be used as a diagnostics and prognosis tool, providing key information to better guide the prescription of more adequate and effective treatments to patients [312–314].

In a world where the incidence of brain diseases is rising, being able to characterize the regulatory mechanisms behind the development of these disorders has become crucial for detecting the most suitable therapeutical targets and for the development of successful clinical applications [315]. Understanding how modifications at the genome level (e.g., mutations, SNPs) can impair the regulatory landscape and drive healthy cellular populations towards a pathological state can further improve the development of personalized gene therapies [316].

This dissertation addresses two open challenges in the development of clinical applications for brain diseases. On the one hand, we leverage the determination of cellular identity TFs to develop novel computer-guided experimental strategies with potential to advance CRT and to identify potential biomarkers that can provide a more accurate disease prognosis. On the other hand, we delve into disease modelling to identify impaired

regulatory interactions due to the presence of SNPs, unveiling promising therapeutical targets.

5.1. Regulating pioneer factors to improve cellular conversion

Cellular identity is strongly regulated by specific TFs that drive the expression of gene programs which control characteristic functions of a given cellular system [199,212–214]. Identifying these identity TFs would aid the characterization of any cellular (sub)population, improve the efficiency of cellular conversion protocols, and obtain functional cell (sub)types for regenerative medicine. Initial computational approaches relied on bulk RNA-seq data to characterize identity TFs at the cell type level which contributed to the development of successful cellular conversion strategies [152,153]. However, determining identity TFs at the cell subtype level relies on the identification of more subtle transcriptomic differences between target cell subpopulations and the cellular background in which they reside. Recent developments in gene expression profiling at single-cell level provided high resolution datasets that reflect this diversity between cellular subpopulations. Novel computational methods based on scRNA-seq enable an accurate characterization of identity TFs, potentially improving cellular conversion at the subpopulation level [154,200].

Due to their key role in controlling cell identity, ectopic expression of these identity TFs has been the most widely used strategy to convert between cell (sub)types [259,269,271]. However, the generated cells do not completely recapitulate the functionalities and maturity features of their *in vivo* counterparts, hindering their potential use in CRT [216,317,318]. One of the major reasons behind this partial reprogramming is an incomplete reorganization of epigenetic marks [216]. If the epigenetic profile of the starting cellular population is not favorable to the binding of the identity TFs, the generation of fully converted cell (sub)types might be compromised [228].

In section 4.2 of this dissertation, we address this issue by developing TransSynW, a network-free, single-cell based computational approach that determines cell conversion TFs by identifying transcriptional regulatory cores. Based on a previously published method, TransSynW uses transcriptional synergy to identify cell (sub)population specific TFs and non-specifically expressed PFs [201]. PFs have a unique ability to bind to closed areas of the chromatin and promote the expression of their target genes [229]. Moreover, PFs bind and activate specific enhancers, prompting chromatin opening mechanisms and, consequently,

gene expression [233–235]. Due to these characteristic features, PFs have been extensively described as one of the main determinants of cellular differentiation and used successfully to convert between cellular identities [230–232]. It has been proposed that TFs can act synergistically to induce the expression of cell identity specific gene expression programs by binding to the same enhancer regions [203,205,206]. Indeed, some of the TFs that have been described to act synergistically, namely *GATA2*, *ISL1*, *POU5F1*, and *SOX2*, have been classified as PFs [202–204,229]. By combining these findings, we demonstrate that the measurement of synergy can be applied to determine not only identity TFs, but also PFs.

TransSynW also identifies marker genes for each target cellular (sub)population, enabling researchers to validate the accuracy of their experimental conversion protocol. To illustrate the applicability of our method, we applied TransSynW to several cellular systems, including cell types, subtypes, and phenotypic states, and performed an *in silico* validation. By crosschecking our predicted TFs and markers with literature evidence, we show that our computational tool well-recapitulated most of the previously described marker genes and TFs used to successfully generate the analyzed cellular (sub)populations. Moreover, we used Metacore, a manually curated database of molecular interactions to further validate the newly identified conversion TFs and markers. We were able to observe that our novel conversion TFs have been described to interact with known conversion factors and marker genes. These findings support the biological relevance of our method in identifying conversion TFs for a wide range of cellular (sub)populations with potential applications in regenerative medicine.

Further experimental validation of our method could be done to support our findings. For instance, the generation of fully functional DANs has great potential clinical applications in CRT [258]. Our method identified, among others, *NPAS4* and *PBX1* as novel TFs to obtain DANs. *NPAS4* has been reported to play a neuroprotective role in DANs upon exposure of these cells to a nigrostriatal toxin [319]. *PBX1* has been described as a key PF in controlling the neurogenic process behind the generation of DANs [320,321]. Regulating the expression of these TFs could improve the quality of the generated neurons and their applications in novel regenerative medicine applications.

TransSynW relies on synergy-based measurement to identify conversion TFs. TransSynW identified *ISL1* and *PHOX2A* as capable of generating oculomotor neurons. Indeed, these TFs have been shown to form complexes and act synergistically to generate

motor neurons [203]. Nonetheless, additional experiments could be performed to demonstrate the broad applicability of TransSynW. For instance, the generation of ChIP-seq data for the TFs identified by our computational approach could help revealing if these factors bind to the same regulatory elements. This strategy has been used to show the co-occupancy of the same promoter regions by *NANOG*, *POU5F1*, and *SOX2* in human ESCs [202]. Moreover, co-immunoprecipitation followed by Western-Blot could be used to capture a TF and identify its interacting proteins using specific antibodies. This methodology has been used to detect protein-protein interactions between *ID3* and *PAX5* [322]. Another alternative would be to use Förster resonant energy transfer technique to detect cooperative binding between two TFs by using specific fluorescent proteins and appropriate readouts [323].

Another distinctive feature of TransSynW is the prioritization of PFs to promote epigenetic modifications that can facilitate the expression of the target cellular (sub)population gene program. To examine to which extent PFs remodel the epigenetic landscape, we could generate scATAC-seq data and perform single-cell BS-seq on the generated cellular (sub)populations to examine the changes in chromatin accessibility and DNA methylation, respectively [324].

TransSynW relies on a precompiled list of TFs to determine cell conversion factors for any population identified in scRNA-seq data. We chose AnimalTFDB, currently the most complete TF database, to obtain a comprehensive and up to date list of human and mouse TFs [325]. Due to the lack of precompiled resources, we manually curated a list of PFs based on extensive literature research. Therefore, to ensure the accuracy of our method, periodical literature searches should be performed to keep these precompiled lists up to date.

Besides scRNA-seq data, we could expand our method and make use of publicly available ChIP-seq data, such as ChIP-Atlas, to complement our results [162,173]. Once we identify the conversion TFs, we could use this prior knowledge to determine the regulatory regions binded by the identified TFs. In addition, this would provide mechanistic insights to help elucidating how the TFs regulate cellular identity, including their putative synergistic activity. Recently, ChIP-Atlas was updated to contain not only ChIP-seq, but also ATAC-seq data [173]. Therefore, we could also leverage the available bulk ATAC-seq data to have a general overview of the chromatin conformation at our starting cell population at the cell type level. Then, we could examine if the identified open chromatin areas match the binding regions of our predicted conversion TFs. With this, we could *a priori* identify which

epigenetic barriers we might need to address in order to facilitate the regulation of our target conversion TFs. In summary, adding these layers of epigenetic information would make our results more accurate and robust.

Finally, another major roadblock in the translation of cellular conversion protocols to clinical applications is the functionality and similarity of the generated cells with those observed in the healthy tissue [311]. Recent efforts have focused on pinpointing genes that are responsible for regulating different maturation stages of a given cell (sub)type to attempt to overcome this issue [326]. We could extend our method to contain a database of genes that have been previously reported to be involved in the maturation process of specific cell (sub)types by text mining published literature. In addition to our current output, we would provide the user with a list of cell (sub)type specific maturation factors to aid the generated cells to attain the desired phenotype. These maturation factors could be ranked by their impact on network topology, using the scRNA-seq data provided by the user to generate the regulatory network and contextualize the collected maturation factors [301,327]. Adding these functionalities to our method could prompt the development of improved cellular conversion protocols and accelerate clinical applications.

5.2. Generating neuronal populations for regenerative medicine

Direct reprogramming is one of the most promising approaches in regenerative medicine. This method allows for *in situ* conversion between fully differentiated cell types which has major implications in the development of potential clinical applications [257,259,271]. Cellular conversion targets one of the hallmarks of neurodegeneration, that is the loss of neurons and their function. For this reason, cellular conversion strategies have become one of the major focuses of CRT [258,328].

It has been shown that replacing the lost DANs in PD patients can alleviate the associated motor symptoms [329]. Moreover, astrogliosis has been described to occur upon the presence of α-synuclein aggregates in PD [330,331]. During this process, reactive astrocytes rewire their molecular landscape and undergo tightly regulated functional and morphological changes, such as cellular proliferation and hypertrophy [48]. Some studies have shown that astrocytes close to a lesion site can display stem cell-like features similar to the ones observed in radial glia cells, and that blocking *Notch* signaling in the striatum induces a dormant neurogenic program in astrocytes [332,333]. Therefore, direct reprogramming of

astrocytes is a promising strategy to generate DANs for CRT [258]. Current direct reprogramming protocols have mostly relied on the overexpression of specific TFs to successfully generate DANs [271,334,335]. However, these approaches have low efficiency and generate DANs with different levels of functionality [336].

In section 4.3 of this dissertation, we developed direct reprogramming protocols to convert human astrocytes into specific DAN subtypes (hDA0, hDA1, and hDA2). As mentioned in the previous section, one of the major roadblocks hindering the success of cellular conversion strategies comprises the lack of binding of conversion TFs to the regulatory regions of their target genes [216]. This obstacle is particularly relevant in direct reprogramming protocols since the chromatin conformation of the starting cell population might be incompatible with the binding of conversion factors [216,228]. Therefore, we applied TransSynW, the computational tool presented in section 4.2 of this dissertation, to identify the most suitable conversion TFs for our experimental goals [154]. This computational method addresses the above-mentioned issues by prioritizing PFs in each set of conversion factors. The combination of specific TFs with PFs has been extensively shown to promote conversion between somatic cell types [245,259,260,271]. Notably, the overexpression of *ASCL1*, an established PF, has been shown to promote accumulation of non-CG methylation, a unique characteristic of the neuronal epigenome [337]. Taken together, this literature evidence supports the applicability of this computational platform to our experimental design.

We applied TransSynW to determine the necessary conversion TFs to generate DANs subtypes using a direct and a sequential reprogramming protocol. To overexpress the identified conversion TFs, we adapted a previously established CRISPR-dCas9 platform to simultaneously activate the expression of specific TFs from their endogenous locus [338]. We observed that it was not possible to activate the expression of some of these conversion factors, even upon treatment with epigenetic modifiers. To examine whether closed chromatin modifications are acting as barriers to the activation of gene expression, we could perform ATAC-seq on hIA-dCas9 to characterize its global profile of chromatin accessibility. To identify the mechanisms behind the putative close conformation of the promoter regions of our target TFs, we could perform ChIP-seq to assess the presence of histone modifications associated with gene repression, such H3K9me3 and H3K27me3, and BS-seq to map DNA methylation patterns [339,340].

The implemented CRISPR-dCas9 platform used here shifts the recruitment of regulatory effectors from the dCas9 to the RNA scaffold, allowing for different regulatory functions at distinct loci [338]. Hence, upon characterization of the chromatin barriers we are facing, we could use this CRISPR-dCas9 system to induce targeted epigenetic modifications in each of our target TFs. For instance, we could replace the current activation domains in the RNA scaffold by Tet1, a 5-methylcytosine hydroxylase involved in DNA demethylation, and design gRNAs for the promoters of the TFs which expression was identified to be impaired by this mechanism [264,341]. On the other hand, we could promote the expression of p300, a histone acetyltransferase that induces the activation of enhancers as well as remodeling of the nucleosome, and target its action to specific genomic loci [342–344].

Regardless of the constrains mentioned above, we were able to validate gRNAs that induce the overexpression of the conversion TFs in the hDA0 and hDA1 combination. By applying these conversion factors in our direct reprogramming protocol, we were able to obtain neuron-like, TUBB3-positive cells. The generated cells did not present immunoreactivity for TH. Similarly, another study using CRISPR activation (CRISPRa) to directly reprogram mouse astrocytes into DANs *in vivo* also reported lack of TH expression on their generated induced neurons [265]. They suggest that the lower levels of expression induced by the CRISPRa system, when compared to the ectopic expression protocols, may shift the dopaminergic fate towards a GABAergic phenotype, which they verify by immunocytochemistry [265,335]. To examine if this stands for our protocols, we could examine the presence of GAD65/67, a GABAergic marker, and evaluate the electrophysiological properties of the generated cells.

To be able to perform the direct and sequential protocol to obtain the hDA2 subtype, we constructed an inducible lentiviral expression system encoding the cDNA sequences of our target TFs. Here, we observed that the sequential protocol generates cells with a neuronal-like morphology, but neither of the protocols produced cells expressing TH. This is in contrast with our positive control (NeAL218 [271]), which generated TH-positive cells with neuronal morphology but no specific cell subtype identity. This result suggests that a combination of NeAL218 with the predicted TFs to generate hDA2 may be required to achieve cells of this cell subtype.

All the attempted protocols are based on the overexpression of specific conversion factors. However, it has been shown that downregulation of specific genes also induces

cellular conversion. A pioneer study showed that suppressing *PTBP1* in human and mouse astrocytes leads to the generation of DANs that can reinnervate the nigrostriatal pathway *in vivo* [345]. Moreover, the depletion of this RNA binding protein in a PD mouse model apparently led not only to the reconstruction of the nigrostriatal circuit, but also to the restoration of the levels of dopamine and to the alleviation of motor symptoms. The proposed mechanism for this contested study [346–349] is that *PTBP1* depletion results in an increase of the expression levels of miR-124 [345,350]. This leads to the activation of the neuronal gene program through inhibition of the transcriptional repressor *REST*, facilitating neuronal conversion. However, it remains unclear how tissue-specific neuronal phenotypes are specified.

Future studies could leverage the high flexibility of the CRISPR-dCas9 system used in our experiments to simultaneously induce the expression of TFs and downregulation of specific genes, such as *PTBP1*, and potentially improve the outcome of our conversion protocols. We could replace the current activation domains in the RNA scaffold by KRAB, a transcriptional repression domain, and design gRNAs for the promoter of *PTBP1* [351–353]. In addition, TransSynW's conversion TFs are ranked by fold-change by comparing the levels of expression of these TFs with the ones in the starting cellular (sub)population [154]. By using this computational tool, we could obtain more information about TFs that could be downregulated to generate the target cell (sub)types.

It has been shown that during the direct reprogramming of human fibroblasts into motor neurons, overexpressing miR-124 first erases the fibroblast identity at both transcriptional and epigenetic level, and only induces a neuronal fate afterwards [354]. Therefore, besides forcing the expression of the desired cellular identity, we could also erase the identity of the starting cell type. To achieve this, we could determine the identity TFs of our initial cellular population using TransSynW and downregulate these TFs using either CRISPR inactivation or RNA interference systems [351,355].

Once the difficulties faced during the experimental setup are addressed, the generated subtypes could be characterized by scRNA-seq. To maximize the number of input cells for scRNA-seq, we could use a multiplexing technology to label each of our protocols individually and then pool all the samples together before library preparation [136]. By profiling these cell subpopulations at the transcriptomic level, we would provide a new in-depth layer of scRNA-seq data that we could use to identify additional molecular features relevant to

these cell subtypes. Moreover, we could compare the gene expression profile of the generated cell subpopulations to the newly identified DAN subpopulation that specifically degenerates during the development of PD for potential clinical applications [83].

We could further characterize these cellular subtypes using functional assays and *in vivo* studies. To assess the functionality of the generate cells, we could evaluate the presence of not only TH, but also its colocalization with ALDH1A1 and DDC, other specific markers of DANs [271]. Additionally, evaluating the presence of MAP2 and SYN1, markers of mature neurons, and the transport channels KCNJ6 and SLC6A3 would provide a full characterization of the functionality and maturity of the generated cells. Additionally, evaluating the ability of the generated cells to release dopamine could be quantified by high performance liquid chromatography and their electrophysiological signature (e.g., sag rectification, pace-maker activity) by calcium imaging assays [265,271].

To translate our findings to an *in vivo* setup, we first have to ensure that the cellular conversion is only taking place in astrocytes. To achieve this, we could change our expression system from LVs to adeno-associated viruses (AAVs) by using the recombinant AAV2/5 serotype, which has been shown to have a tropism for astrocytes [265,356,357]. Alternatively, we could condition the expression of our conversion factors by placing them under the control of the GFAP promoter, an astrocyte-specific marker, using Cre-Lox recombination [271]. These strategies could be used simultaneous when using a CRISPRa system [265]. A recent *in vivo* study used the recombinant AAV2/5 serotype to deliver gRNAs and the activator sequence specifically to astrocytes and placed dCas9 expression under the control of a GFAP-Cre activation system [265]. The disadvantage of this system is that the expression levels were variable, and astrocytes did not convert to DANs but rather GABAergic neurons, despite using TFs capable of making DANs *in vitro*. Alternative strategies will need to be explored in the future.

The most interesting application of our findings would be to evaluate the capacity of our protocol to replenish the lost DAN population in a PD model. To examine this, we could use adult mice unilaterally lesioned with 6-hydroxydopamine (6-OHDA), which induces the degeneration of DANs in the ventral midbrain [358]. Besides the functional assays mentioned previously, we would need to examine the recovery of the motor function in 6-OHDA-lesioned mice to evaluate the functional success of our *in vivo* studies. Behavioral tests, such as drug-induced and spontaneous rotation, gait analysis, limb usage, and axial symmetry are

some of the assays that could be performed to evaluate the rescue of motor symptoms [265,321,359].

When performing direct reprogramming protocols, there is the inherent risk to deplete the initial cell population. Since astrocytes are the major metabolic supporters of neurons, depleting this cell population could undermine the applicability of this strategy in clinical trials [336]. However, it has been shown that A1 reactive astrocytes are characterized by the upregulation of complement cascade genes and by the secretion of neurotoxins that lead to the destruction of synapses and consequent neuronal death [51,360]. Several studies have profiled the transcriptome of this cell subtype and identified potential marker genes of this cellular population, namely *C3* and *SERPING1* [51,361,362]. Indeed, *SERPING1* upregulation has been associated with DANs death in a PD mouse model [363].

Although further studies would be necessary to clarify if this gene is uniquely expressed on A1 astrocytes, this or other potential marker genes could be used to selectively target A1 astrocytes with our cellular conversion protocols. Briefly, we could condition the expression of the conversion TFs by placing them under the promoter of a marker gene of A1 astrocytes and limit their activation to this cellular subpopulation. This approach would prevent the depletion of the astrocyte population at the injection site, opening the door to the development of pioneer CRTs in which a deleterious cell (A1 astrocyte) would then be converted into a cell lost by disease (DANs). These strategies would thus simultaneously prevent further damage and promote repair by replacing the cells damaged by neurodegeneration, with immediate applications to PD.

5.3. Leveraging cellular lineage to determine prognosis biomarkers

Along the differentiation process, cells progressively commit to a specific lineage and start specializing until they acquire a fully differentiated and functional state. However, these mature cells still retain some plasticity and can, upon specific stimuli, dedifferentiate and attain a new cellular identity [227,364]. This process has been described to occur under physiological conditions, such as tissue repair, but also during the development of several disorders, including neurodegenerative diseases and tumor formation [365–369]. Specifically, tumorigenesis has been one of the major focuses of the study of cellular plasticity. During this process, fully differentiated cells lose their identity and start to gain stem-like properties, such as the ability to self-renew and generate mature cell types of any type of tissue [370,371].

However, it has been shown that the developmental origin of these stem-like cells has an impact on their properties and determines some distinct molecular characteristics, particularly in GBM [372,373].

GBM is the most common and severe type of glioma with poor prognosis and low life expectancy [374]. Based on the combination of different layers of NGS data, it has been possible to classify GBMs into three different subtypes: i) classical, ii) proneuronal, and iii) mesenchymal [375,376]. Among these subtypes, the mesenchymal subtype presents the most aggressive and therapy resistant features [376]. However, the cellular lineage of origin of this subtype and its associated molecular mechanisms have not been elucidated [377,378]. Identifying those could provide important biomarkers and therapeutical targets for the development of more successful clinical applications.

In section 4.4 of this dissertation, we profiled the transcriptome of low-grade gliomas and GBM cells, including progenitors, from several patients to study their cellular lineage patterns. We see that around half of the analyzed GBM progenitors present molecular features similar to the Rgl lineage, as previously described [379–381]. The remaining GBM progenitors presented a molecular profile analogous to the PeriV lineage, which was uniquely identified in GBM samples when compared to the low-grade gliomas. Furthermore, when compared to the TCGA classification of GBM subtypes, the molecular profile of tumors derived from the PeriV lineage shared most of its features with the mesenchymal subtype [382]. Indeed, patients with GBMs derived from the PeriV lineage presented significantly lower survival than the ones with a Rgl origin.

In our study, most of the tumor cells derived from the PeriV lineage were identified as pericytes and vascular leptomeningeal cells. These cell types are normally located near the blood vessels and the endfeet of astrocytes and have been described to have a role in the remodeling of the brain's vasculature [383,384]. In particular, pericytes have been described to be involved in regulating blood flow and maintaining the BBB [385,386]. Recent studies identified a novel subtype of PeriV cells that presents stem-like characteristics and can give rise to pericytes and mesenchymal cells in the adult brain [387–389]. The combination of these features with the inherent angiogenic role of the PeriV cell lineage can explain the increased severity of GBM with this origin.

Identifying the cell lineage of origin in GBM, specifically by stratifying them in Rgl and PeriV origin, can have profound implications in the prognosis of the disease and in the

prescription of adequate treatments. To address this, we identified *PROX1* and *FOXC1* as specifically expressed in Rgl and PeriV cell lineages, respectively, and validated their expression in patient-derived xenografts. In summary, our results identified the PeriV cellular lineage as uniquely present in GBM, characterized the molecular profile of this cell lineage as very similar to the mesenchymal subtype, and determined *PROX1* and *FOXC1* as potential biomarkers for disease prognosis.

PeriV cells can be found not only in the brain, but also across almost all tissues in our body. Given our findings associating cells of this origin with a poorer prognosis in GBM patients, it would be interesting to investigate the presence of this cell lineage in other types of tumors. Pericytes have been extensively described as regulators of tumor microenvironment, by contributing to higher permeabilization of blood vessels, promoting tumor progression, and the ability to evade immunosurveillance [390–392]. Therefore, finding cells derived from the PeriV lineage in other types of tumors could be an opportunity to find common therapeutical targets and enable the translation of effective therapies across tumor types. It has been shown that breast and high-grade ovarian tumors have common molecular signatures, namely inactivation of *BRCA1* and *BRCA2*, mutations in the *ATM* gene, and *RB1* loss [382,393,394]. These studies serve as a proof of concept that different cancer types can have a shared etiology and consequently similar therapeutical applications.

In this study, we focused on identifying specifically expressed TFs in GBM derived from PeriV origin. However, specific epigenetic marks, namely DNA methylation patterns, have also been shown to play a role in GBM development [395,396]. Methylation of the *MGMT* gene has been shown to be correlated with a better prognosis in patients with GBM [397,398]. Complementing current treatments with temozolomide, an alkylating compound that methylates DNA, has been the preferential therapeutical approach to significantly improve patients' prognosis [399,400]. Furthermore, the presence of a specific DNA hypermethylation profile, named glioma CpG island methylator phenotype, has been correlated with a better survival of GBM patients [401].

PFs have been shown to act as chromatin remodelers by regulating DNA methylation across different types of cancer [402]. The overexpression of *ASCL1* in GBM cells has been shown to reduce the stem-like features and tumorigenicity of the tumor [272]. Other epigenetic marks, such as histone modifications and chromatin conformation, have also been shown to play a fundamental role in the development of GBM [396,403]. Using epigenomic profiling

methods, such as BS-seq to map DNA methylation patterns and ChIP-seq to profile the histone acetylation and methylation profiles, could unveil unique epigenetic modifications in PeriV lineage GBM [339,340]. Performing these studies could not only identify biomarkers for GBM patient's stratification, but also provide novel therapeutical targets for the development of improved treatments.

5.4. Deciphering the role of disease-associated single nucleotide polymorphisms in the impairment of regulatory interactions

The regulation of gene expression is a complex process that involves fine-tuning of both transcriptomic and epigenetic mechanisms which activate or repress genes upon stimuli [217,218]. To activate gene expression, chromatin remodeling mechanisms promote an open conformation by controlling the accessibility of specific regulatory elements, such as promoters and enhancers [217]. Then, TFs bind to these available regions and recruit the transcriptional machinery, containing RNA polymerase II, initiating gene transcription [184]. Dysregulation of these mechanisms due to the impairment of specific regulatory interactions can lead to modifications in gene expression that might lead to disease onset [218].

Mutations in DNA sequences have been extensively described to be connected to disease pathogenesis in cancer, brain diseases, and diabetes [218,404,405]. SNPs are the most common variants in our genome and have been described to affect gene expression by changing TF binding affinity, mRNA stability, and by disrupting protein binding domains [406,407]. SNPs have been extensively characterized as risk factors of disease onset and several methods, such as GWAS and eQTL analysis, were developed to associate these variants to disease causing genes [408–411].

Most of the profiled SNPs are in non-coding regions of the DNA, especially enhancers [287]. These have been described to act in a cell type specific manner to control regulatory mechanisms essential for activation of gene expression profiles [412]. Recent studies have focused on characterizing cell type specific regulatory landscapes and illustrating how SNPs affect cell type specific biological functions [175,413]. However, these methods do not provide mechanistic insights regarding specific regulatory interactions that are impaired due to the presence of disease-related SNPs. Therefore, characterizing the impact of these nucleotide variations on the regulatory landscape would allow for a better understanding of

the pathological mechanisms behind the disease and open doors for the development of novel gene therapy applications [316].

In section 4.5 of this dissertation, we presented RNetDys, a computational pipeline based on multi-omics data, that identifies regulatory mechanisms impaired due to disease-associated SNPs. This systematic approach builds cell (sub)type specific GRNs based on the combination of scRNA-seq, scATAC-seq, and prior knowledge data, including ChIP-seq and GeneHancer database [162,174]. We showed that the use of multi-omics data improves the accuracy of our method in predicting regulatory interactions when compared to state-of-the-art approaches. Furthermore, RNetDys can be applied to any cellular system characterized in both scRNA-seq and scATAC-seq datasets, independently of whether the information is captured on the same cell.

RNetDys relies on prior knowledge to infer the GRNs, therefore, only previously described regulatory mechanisms mediated by TFs and enhancers of regulated genes can be predicted as impaired by the presence of SNPs. In that regard, we included ChIP-Atlas and GeneHancer databases to provide complete and up to date information about TF binding sites and promoter-enhancer interactions [162,174]. To ensure the accuracy and consistency of our systematic pipeline, a regular literature search to evaluate whether these resources can be complemented with novel findings should be performed. GeneHancer is a manually curated database of exclusively human enhancer interactions [174]. Therefore, our pipeline can only be applied to human data. To address this shortcoming, RNetDys could be extended to also include EnhancerDB, a resource that compiles enhancer interactions reported in mouse data [414]. Our method considers only cis-regulatory interactions, the most common regulatory mechanism controlling activation of gene expression [415]. Although scarcer, trans-regulatory interactions have also been described to have a stronger effect on enhancers when compared to promoters. Once the necessary computational resources are available, it would be valuable to include trans-regulatory interactions in our methodology to provide further insights on enhancer-specific regulatory mechanisms.

Histone methylation (H3K4me1 and H3K4me3) and acetylation (H3K27ac) are specific epigenetic marks that can be used to evaluate if an enhancer is active [416]. We could not integrate this information in our pipeline since the profiling of these chromatin signatures at the single-cell level is still incomplete [417]. Instead, we assume that an enhancer is active if

it is accessible, at least one of its binding TFs is expressed, and one of its regulated genes is also expressed [77,412] .

RNetDys uses inferred networks to identify specific dysregulated interactions due to the presence of any SNPs. This pipeline identifies impaired mechanistic interactions mediated by TFs and enhancers of regulated genes. It also provides additional information about the type of the regulatory interaction (activation or repression), the impact of SNPs on the binding affinity of each TF, and a list of TFs ranked by their effect on the impairments.

Our method relies solely on TF binding sites reported in ChIP-seq data to provide additional mechanistic insights behind the identified dysregulated interactions. Besides keeping our ChIP-seq data up to date, we could also expand our pipeline to include an algorithm that predicts the generation of novel TF binding sites due to the presence of a given SNP [418].

RNetDys was applied to diseases known to have a strong genetic component, such as AD, PD, EPI, and diabetes. Here, we were able to accurately identify dysregulated interactions associated with several disease-associated SNPs and provide further insights on the mechanisms behind these impairments. Based on these results, we show that RNetDys is a systematic pipeline that can leverage multi-omics data to decipher cell (sub)type specific mechanisms underlaying impaired regulatory interactions due to disease-associated SNPs. For this application, we used SNPs reported in ClinVar, a database with associations between human SNPs and phenotypes [419]. However, our pipeline could also be applied to patient-specific SNPs extracted from genotyping data which would provide a more contextualized and targeted analysis of the involvement of these SNPs in dysregulated mechanisms.

Due to the high variability of the human genome, characterizing the impact of a single SNP in the regulatory landscape and in the onset of a disease is still difficult. Mice strains have extremely low genetic variance. Thus, a possible experimental validation of our method could consist of *in vivo* experiments using mice models. Based on CRISPR technology, we could insert a single SNP previously identified by our method to impair key regulatory mechanisms and characterize their impact on the onset of the respective disease [420].

5.5. Outlook

The successful application of regenerative medicine strategies into clinical practice highly depends on the generation of high quality and functional cell (sub)types. Currently, the low efficiency of cellular conversion protocols as well as the lack of molecular and functional features in the generated cells *in vitro* is one of the major roadblocks to clinical translation of these approaches. Moreover, studying and characterizing the molecular prolife of specific cellular populations could accelerate the identification of robust biomarkers that can be used as therapeutical targets for clinical applications. The lack of disease modelling approaches that can provide further mechanistic insights regarding the impact of SNPs on disease onset and development has been delaying the identification of suitable targets for gene therapy. Leveraging single cell technology and its combination with other omics data is a promising strategy to advance the development of pioneer methods that can address these issues.

The aims of this PhD dissertation were to develop single-cell based computational and experimental strategies that would improve cellular conversion protocols by addressing epigenetic constrains and achieving control over subtype specification. We also aimed at identifying cellular lineage-specific TFs to provide potential prognosis biomarkers and decipher the role of disease-associated SNPs in the impairment of regulatory mechanisms. This dissertation provides the following contributions to the scientific research community:

- **Implementation of a computational platform that identifies novel cellular conversion TFs for the generation of cell (sub)types**: TransSynW is single-cell based computational method that leverages transcriptomic information to identify conversion TFs for any cellular population identified by scRNA-seq. This method prioritizes PF among the determined conversion factors to address epigenetic constrains that are often related with the low efficiency of current cellular conversion protocols. We were able to well-recapitulate known conversion TFs in different cellular systems and predict novel TFs that have the potential to improve the outcome of cellular conversion protocols. We confirmed the biological relevance of our findings by cross-referencing our novel TFs with a manually curated database of molecular interactions. TransSynW is hosted in a user-friendly web interface that is freely available at https://transsynw.lcsb.uni.lu/.

- **Identification of marker genes for any cellular (sub)population characterized in scRNA-seq**: TransSynW further exploits scRNA-seq data to identify marker genes for any target cell (sub)type. We were able to well-recapitulate known markers in different cellular systems and predict previously unknown genes that allow researchers to evaluate the performance their cellular conversion protocols more precisely.
- **Implementation of direct reprogramming strategies to convert human astrocytes into TUBB3-positive cells based on the endogenous regulation of novel conversion TFs**: We adapted a previously established CRISPR-dCas9 system to promote the cellular conversion of human astrocytes by overexpression of specific conversion TFs. Lentiviral-based deliver of gRNAs and transcriptional activators resulted on the generation of TUBB3-positive cells. Optimization of TF overexpression and targeting epigenetic mechanisms can improve the maturation level of the generated cells.
- **Development of a two-step reprogramming approach to generate neuronal-like cells from human astrocytes**: We developed a novel direct reprogramming approach that consists of the initial overexpression of TFs that promote the conversion of human astrocytes into DANs. This step is followed by upregulation of TFs that induce the specialization of DANs into the target subtype. The application of this approach resulted in cells with a neuronal morphology and further optimization can unlock the potential of this protocol in generating cellular subtypes with high efficiency and functionality.
- **Identification of potential lineage-specific biomarkers in GBM with implications in disease prognosis**: Based on the transcriptomic profile of GBM and low-grade glioma cells, we determined that the PeriV lineage is specifically present in GBM samples. We identified *PROX1* and *FOXC1* as TFs specifically expressed in GBM derived from Rgl and PeriV lineage, respectively, and validated their expression in patient-derived xenografts.
- **Development of a multi-omics approach that infers GRNs specific for any cell (sub)type**: RNetDys is a systematic pipeline that leverages the combination of single cell transcriptomic and epigenomic data with prior knowledge information to comprehensively characterize the regulatory landscape for any cell (sub)population. This method systematically identifies TFs and enhancers of the regulated genes involved in specific cell (sub)type regulatory mechanisms.

RNetDys is a user-friendly pipeline that is freely available at https://github.com/BarlierC/RNetDys.

- **Implementation of a systematic pipeline that identifies impaired regulatory interactions related with disease-associated SNPs**: Given a list of disease-associated SNPs, RNetDys identifies the corresponding impaired regulatory interactions at the cell (sub)type level. This pipeline provides additional mechanistic insights for each of the dysregulated interactions by leveraging the GRN information.

In summary, the findings described in this PhD dissertation leverage single cell data to develop novel computer guided experimental strategies with potential applications in regenerative medicine and reveal therapeutical targets and complex molecular mechanisms associated to the onset of brain diseases.

6. References

1. Akinrodoye, M. A. & Lui, F. *Neuroanatomy, Somatic Nervous System*. *StatPearls* (StatPearls Publishing, 2020).
2. Waxenbaum, J. A. & Varacallo, M. *Anatomy, Autonomic Nervous System*. *StatPearls* (StatPearls Publishing, 2019).
3. Betts, J. G. *et al.* Anatomy of the Nervous System. in *Anatomy & Physiology* (OpenStax, 2013).
4. Ackerman, S. Major Structures and Functions of the Brain. in *Discovering the Brain* (National Academies Press (US), 1992).
5. Shipp, S. Structure and function of the cerebral cortex. *Current Biology* vol. 17 Preprint at https://doi.org/10.1016/j.cub.2007.03.044 (2007).
6. Basinger, H. & Hogg, J. P. *Neuroanatomy, Brainstem*. *StatPearls* (StatPearls Publishing, 2019).
7. Haines, D. E. & Mihailoff, G. A. *Fundamental neuroscience for basic and clinical applications*. (Elsevier, 2017).
8. Caminero, F. & Cascella, M. Neuroanatomy, Mesencephalon Midbrain. *StatPearls* (2021).
9. Sonne, J. & Beato, M. R. *Neuroanatomy, Substantia Nigra*. *StatPearls* (StatPearls Publishing, 2018).
10. Fabbri, M. *et al.* Substantia Nigra Neuromelanin as an Imaging Biomarker of Disease Progression in Parkinson's Disease. *J Parkinsons Dis* **7**, 491–501 (2017).
11. Hodge, G. K. & Butcher, L. L. Pars compacta of the substantia nigra modulates motor activity but is not involved importantly in regulating food and water intake. *Naunyn Schmiedebergs Arch Pharmacol* **313**, 51–67 (1980).
12. Fields, R. D. & Stevens-Graham, B. *New Insights into Neuron-Glia Communication*. *Science* vol. 298 www.sciencemag.org/cgi/content/full/298/5593/556/DC1, (2002).
13. Purves, D. *et al.* Nerve Cells. in *Neuroscience* (Sinauer Associates, 2001).
14. Singh, K. K. & Tsai, L. H. MicroTUB(B3)ules and Brain Development. *Cell* vol. 140 30–32 Preprint at https://doi.org/10.1016/j.cell.2009.12.038 (2010).
15. Menezes, J. R. L. & Luskin, M. B. Expression of neuron-specific tubulin defines a novel population in the proliferative layers of the developing telencephalon. *Journal of Neuroscience* **14**, 5399–5416 (1994).
16. Pereda, A. E. Electrical synapses and their functional interactions with chemical synapses. *Nat Rev Neurosci* **15**, 250–263 (2014).
17. Sheffler, Z. M. & Pillarisetty, L. S. *Physiology, Neurotransmitters*. *StatPearls* (StatPearls Publishing, 2019).
18. Zhou, Y. & Danbolt, N. C. Glutamate as a neurotransmitter in the healthy brain. *Journal of Neural Transmission* vol. 121 799–817 Preprint at https://doi.org/10.1007/s00702-014-1180-8 (2014).
19. Lau, A. & Tymianski, M. Glutamate receptors, neurotoxicity and neurodegeneration. *Pflugers Arch* **460**, 525–542 (2010).

20. Bergman, H., Wichmann, T., Karmon, B. & DeLong, M. R. The primate subthalamic nucleus. II. Neuronal activity in the MPTP model of parkinsonism. *J Neurophysiol* **72**, 507–520 (1994).

21. Hynd, M. R., Scott, H. L. & Dodd, P. R. Glutamate-mediated excitotoxicity and neurodegeneration in Alzheimer's disease. *Neurochemistry International* vol. 45 583–595 Preprint at https://doi.org/10.1016/j.neuint.2004.03.007 (2004).

22. Repici, M. & Giorgini, F. DJ-1 in Parkinson's disease: Clinical insights and therapeutic perspectives. *Journal of Clinical Medicine* vol. 8 Preprint at https://doi.org/10.3390/jcm8091377 (2019).

23. Dong, X. X., Wang, Y. & Qin, Z. H. Molecular mechanisms of excitotoxicity and their relevance to pathogenesis of neurodegenerative diseases. *Acta Pharmacologica Sinica* vol. 30 379–387 Preprint at https://doi.org/10.1038/aps.2009.24 (2009).

24. Meredith, G. E., Totterdell, S., Beales, M. & Meshul, C. K. Impaired glutamate homeostasis and programmed cell death in a chronic MPTP mouse model of Parkinson's disease. *Exp Neurol* **219**, 334–340 (2009).

25. Sonsalla, P. K., Albers, D. S. & Zeevalk, G. D. Role of glutamate in neurodegeneration of dopamine neurons in several animal models of parkinsonism. in *Amino Acids* vol. 14 69–74 (Amino Acids, 1998).

26. Treiman, D. M. GABAergic mechanisms in epilepsy. in *Epilepsia* vol. 42 8–12 (Epilepsia, 2001).

27. Chagnac-Amitai, Y. & Connors, B. W. Horizontal spread of synchronized activity in neocortex and its control by GABA-mediated inhibition. *J Neurophysiol* **61**, 747–758 (1989).

28. During, M. J., Ryder, K. M. & Spencer, D. D. Hippocampal GABA transporter function in temporal-lobe epileps. *Nature* **376**, 174–177 (1995).

29. Grunewald, R. A. *et al.* Effects of vigabatrin on partial seizures and cognitive function. *J Neurol Neurosurg Psychiatry* **57**, 1057–1063 (1994).

30. Franco, R., Reyes-Resina, I. & Navarro, G. Dopamine in health and disease: Much more than a neurotransmitter. *Biomedicines* vol. 9 1–13 Preprint at https://doi.org/10.3390/biomedicines9020109 (2021).

31. Hornykiewicz, O. The discovery of dopamine deficiency in the parkinsonian brain. in *Journal of Neural Transmission, Supplement* 9–15 (Springer Wien, 2006). doi:10.1007/978-3-211-45295-0_3.

32. Tambasco, N., Romoli, M. & Calabresi, P. Levodopa in Parkinson's Disease: Current Status and Future Developments. *Curr Neuropharmacol* **16**, 1239–1252 (2017).

33. Lodish, H. *et al.* Nerve Cells. in *Molecular Cell Biology* (W. H. Freeman, 2013).

34. Sofroniew, M. v. & Vinters, H. v. Astrocytes: Biology and pathology. *Acta Neuropathologica* vol. 119 7–35 Preprint at https://doi.org/10.1007/s00401-009-0619-8 (2010).

35. Pekny, M. & Pekna, M. Astrocyte intermediate filaments in CNS pathologies and regeneration. *Journal of Pathology* vol. 204 428–437 Preprint at https://doi.org/10.1002/path.1645 (2004).

36. Doyle, D. *The Fine Structure of the Nervous System: The Neurons and Supporting Cells*. *Journal of Neurology, Neurosurgery, and Psychiatry* vol. 41 (1978).

37. Ramón y Cajal, S. *Histologie du système nerveux de l'homme & des vertébrés*. *Histologie du système nerveux de l'homme & des vertébrés* (Maloine, 2011). doi:10.5962/bhl.title.48637.

38. Bak, L. K., Schousboe, A. & Waagepetersen, H. S. The glutamate/GABA-glutamine cycle: Aspects of transport, neurotransmitter homeostasis and ammonia transfer. *Journal of Neurochemistry* vol. 98 641–653 Preprint at https://doi.org/10.1111/j.1471-4159.2006.03913.x (2006).

39. Magistretti, P. J. & Allaman, I. Brain energy and metabolism. in *Neuroscience in the 21st Century: From Basic to Clinical, Second Edition* 1879–1909 (Springer New York, 2016). doi:10.1007/978-1-4939-3474-4_56.

40. McKenna, M. C. The glutamate-glutamine cycle is not stoichiometric: Fates of glutamate in brain. *J Neurosci Res* **85**, 3347–3358 (2007).

41. Brown, L. S. *et al.* Pericytes and Neurovascular Function in the Healthy and Diseased Brain. *Front Cell Neurosci* **13**, 282 (2019).

42. Pellerin, L. & Magistretti, P. J. Glutamate uptake into astrocytes stimulates aerobic glycolysis: A mechanism coupling neuronal activity to glucose utilization. *Proc Natl Acad Sci U S A* **91**, 10625–10629 (1994).

43. Herrero-Mendez, A. *et al.* The bioenergetic and antioxidant status of neurons is controlled by continuous degradation of a key glycolytic enzyme by APC/C-Cdh1. *Nat Cell Biol* **11**, 747–752 (2009).

44. Lovatt, D. *et al.* The transcriptome and metabolic gene signature of protoplasmic astrocytes in the adult murine cortex. *Journal of Neuroscience* **27**, 12255–12266 (2007).

45. Ronald Zielke, H., Zielke, C. L. & Baab, P. J. Direct measurement of oxidative metabolism in the living brain by microdialysis: a review. *J Neurochem* **109**, 24–29 (2009).

46. Bouzier-Sore, A. K. *et al.* Competition between glucose and lactate as oxidative energy substrates in both neurons and astrocytes: A comparative NMR study. *European Journal of Neuroscience* **24**, 1687–1694 (2006).

47. Itoh, Y. *et al.* Dichloroacetate effects on glucose and lactate oxidation by neurons and astroglia in vitro and on glucose utilization by brain in vivo. *Proc Natl Acad Sci U S A* **100**, 4879–4884 (2003).

48. Sofroniew, M. v. Astrogliosis. *Cold Spring Harb Perspect Biol* **7**, (2015).

49. Sofroniew, M. v. Molecular dissection of reactive astrogliosis and glial scar formation. *Trends in Neurosciences* vol. 32 638–647 Preprint at https://doi.org/10.1016/j.tins.2009.08.002 (2009).

50. Wanner, I. B. *et al.* Glial scar borders are formed by newly proliferated, elongated astrocytes that interact to corral inflammatory and fibrotic cells via STAT3-dependent mechanisms after spinal cord injury. *Journal of Neuroscience* **33**, 12870–12886 (2013).

51. Liddelow, S. A. *et al.* Neurotoxic reactive astrocytes are induced by activated microglia. *Nature* **541**, 481 (2017).

52. Liddelow, S. A. & Barres, B. A. Reactive Astrocytes: Production, Function, and Therapeutic Potential. *Immunity* **46**, 957–967 (2017).

53. Yun, S. P. *et al.* Block of A1 astrocyte conversion by microglia is neuroprotective in models of Parkinson's disease. *Nat Med* **24**, 931 (2018).

54. Xu, X. *et al.* MFG-E8 reverses microglial-induced neurotoxic astrocyte (A1) via NF-κB and PI3K-Akt pathways. *J Cell Physiol* **234**, 904–914 (2018).

55. Guo, Z. *et al.* In vivo direct reprogramming of reactive glial cells into functional neurons after brain injury and in an Alzheimer's disease model. *Cell Stem Cell* **14**, 188–202 (2014).

56. Chen, Y. *et al.* Functional repair after ischemic injury through high efficiency in situ astrocyte-to-neuron conversion. *bioRxiv* 294967 (2018) doi:10.1101/294967.

57. Zhang, L. *et al.* Reversing Glial Scar Back To Neural Tissue Through NeuroD1-Mediated Astrocyte-To-Neuron Conversion. *bioRxiv* 261438 (2018) doi:10.1101/261438.

58. Tremblay, M. È. *et al.* The role of microglia in the healthy brain. *Journal of Neuroscience* **31**, 16064–16069 (2011).

59. Nimmerjahn, A., Kirchhoff, F. & Helmchen, F. Neuroscience: Resting microglial cells are highly dynamic surveillants of brain parenchyma in vivo. *Science (1979)* **308**, 1314–1318 (2005).

60. Colonna, M. & Butovsky, O. Microglia function in the central nervous system during health and neurodegeneration. *Annual Review of Immunology* vol. 35 441–468 Preprint at https://doi.org/10.1146/annurev-immunol-051116-052358 (2017).

61. Doorn, K. J. *et al.* Increased amoeboid microglial density in the Olfactory Bulb of Parkinson's and Alzheimer's Patients. *Brain Pathology* **24**, 152–165 (2014).

62. Tang, Y. & Le, W. Differential Roles of M1 and M2 Microglia in Neurodegenerative Diseases. *Molecular Neurobiology* vol. 53 1181–1194 Preprint at https://doi.org/10.1007/s12035-014-9070-5 (2016).

63. Keren-Shaul, H. *et al.* A Unique Microglia Type Associated with Restricting Development of Alzheimer's Disease. *Cell* **169**, 1276-1290.e17 (2017).

64. de Biase, L. M. *et al.* Local Cues Establish and Maintain Region-Specific Phenotypes of Basal Ganglia Microglia. *Neuron* **95**, 341-356.e6 (2017).

65. Heneka, M. T. *et al.* Focal glial activation coincides with increased BACE1 activation and precedes amyloid plaque deposition in APP[V717I] transgenic mice. *J Neuroinflammation* **2**, (2005).

66. Zhang, G., Wang, Z., Hu, H., Zhao, M. & Sun, L. Microglia in Alzheimer's Disease: A Target for Therapeutic Intervention. *Frontiers in Cellular Neuroscience* vol. 15 479 Preprint at https://doi.org/10.3389/fncel.2021.749587 (2021).

67. Hunot, S. *et al.* Nitric oxide synthase and neuronal vulnerability in Parkinson's disease. *Neuroscience* **72**, 355–363 (1996).

68. Perry, V. H. Innate inflammation in Parkinson's disease. *Cold Spring Harb Perspect Med* **2**, (2012).

69. Mogi, M. *et al.* Caspase activities and tumor necrosis factor receptor R1 (p55) level are elevated in the substantia nigra from Parkinsonian brain. *J Neural Transm* **107**, 335–341 (2000).

70. Baumann, N. & Pham-Dinh, D. Biology of oligodendrocyte and myelin in the mammalian central nervous system. *Physiological Reviews* vol. 81 871–927 Preprint at https://doi.org/10.1152/physrev.2001.81.2.871 (2001).

71. Snaidero, N. *et al.* Myelin membrane wrapping of CNS axons by PI(3,4,5)P3-dependent polarized growth at the inner tongue. *Cell* **156**, 277–290 (2014).

72. Stadelmann, C., Timmler, S., Barrantes-Freer, A. & Simons, M. Myelin in the central nervous system: Structure, function, and pathology. *Physiol Rev* **99**, 1381–1431 (2019).

73. Filippi, M. *et al.* Multiple sclerosis. *Nat Rev Dis Primers* **4**, 1–27 (2018).

74. Keirstead, H. S. & Blakemore, W. F. Identification of post-mitotic oligodendrocytes incapable of remyelination within the demyelinated adult spinal cord. *J Neuropathol Exp Neurol* **56**, 1191–1201 (1997).

75. Sim, F. J., Zhao, C., Penderis, J. & Franklin, R. J. M. The Age-Related Decrease in CNS Remyelination Efficiency Is Attributable to an Impairment of Both Oligodendrocyte Progenitor Recruitment and Differentiation. *Journal of Neuroscience* **22**, 2451–2459 (2002).

76. Jäkel, S. *et al.* Altered human oligodendrocyte heterogeneity in multiple sclerosis. *Nature* **566**, 543–547 (2019).

77. Alberts, B. *et al. Molecular Biology of the Cell*. *Molecular Biology of the Cell* (W.W. Norton & Company, 2017). doi:10.1201/9781315735368.

78. Sheridan, G. K. & Murphy, K. J. Neuron–glia crosstalk in health and disease: fractalkine and CX3CR1 take centre stage. *Open Biol* **3**, (2013).

79. Morris, S. A. The evolving concept of cell identity in the single cell era. *Development* **146**, (2019).

80. Sunkin, S. M. *et al.* Allen Brain Atlas: An integrated spatio-temporal portal for exploring the central nervous system. *Nucleic Acids Res* **41**, D996 (2013).

81. la Manno, G. *et al.* Molecular Diversity of Midbrain Development in Mouse, Human, and Stem Cells. *Cell* **167**, 566-580.e19 (2016).

82. Callaway, E. M. *et al.* A multimodal cell census and atlas of the mammalian primary motor cortex. *Nature* **598**, 86–102 (2021).

83. Kamath, T. *et al.* Single-cell genomic profiling of human dopamine neurons identifies a population that selectively degenerates in Parkinson's disease. *Nat Neurosci* **25**, 588–595 (2022).

84. Sarkar, S. & Biswas, S. C. Astrocyte subtype-specific approach to Alzheimer's disease treatment. *Neurochemistry International* vol. 145 104956 Preprint at https://doi.org/10.1016/j.neuint.2021.104956 (2021).

85. Masuda, T., Sankowski, R., Staszewski, O. & Prinz, M. Microglia Heterogeneity in the Single-Cell Era. *Cell Reports* vol. 30 1271–1281 Preprint at https://doi.org/10.1016/j.celrep.2020.01.010 (2020).

86. Watson, J. D. & Crick, F. H. C. Molecular structure of nucleic acids: A structure for deoxyribose nucleic acid. *Nature* **171**, 737–738 (1953).

87. Wilkins, M. H. F., Stokes, A. R. & Wilson, H. R. Molecular structure of nucleic acids: Molecular structure of deoxypentose nucleic acids. *Nature* **171**, 738–740 (1953).

88. Franklin, R. E. & Gosling, R. G. Molecular configuration in sodium thymonucleate. *Nature* **171**, 740–741 (1953).

89. Sanger, F., Nicklen, S. & Coulson, A. R. DNA sequencing with chain-terminating inhibitors. *Proc Natl Acad Sci U S A* **74**, 5463–5467 (1977).

90. Shendure, J. *et al.* DNA sequencing at 40: Past, present and future. *Nature* **550**, (2017).

91. Mardis, E. R. Next-generation sequencing platforms. *Annual Review of Analytical Chemistry* **6**, 287–303 (2013).

92. McGinnis, S. & Madden, T. L. BLAST: At the core of a powerful and diverse set of sequence analysis tools. *Nucleic Acids Res* **32**, W20 (2004).

93. Benson, D. A., Karsch-Mizrachi, I., Lipman, D. J., Ostell, J. & Wheeler, D. L. GenBank. *Nucleic Acids Res* **33**, D34 (2005).

94. Levy, S. *et al.* The Diploid Genome Sequence of an Individual Human. *PLoS Biol* **5**, e254 (2007).

95. Tammi, M. T., Arner, E., Kindlund, E. & Andersson, B. Correcting errors in shotgun sequences. *Nucleic Acids Res* **31**, 4663–4672 (2003).

96. Stark, R., Grzelak, M. & Hadfield, J. RNA sequencing: the teenage years. *Nature Reviews Genetics* vol. 20 631–656 Preprint at https://doi.org/10.1038/s41576-019-0150-2 (2019).

97. Kuksin, M. *et al.* Applications of single-cell and bulk RNA sequencing in onco-immunology. *Eur J Cancer* **149**, 193–210 (2021).

98. Biswas, D. *et al.* A clonal expression biomarker associates with lung cancer mortality. *Nat Med* **25**, 1540–1548 (2019).

99. Han, L. O., Li, X. Y., Cao, M. M., Cao, Y. & Zhou, L. H. Development and validation of an individualized diagnostic signature in thyroid cancer. *Cancer Med* **7**, 1135–1140 (2018).

100. Farnham, P. J. Insights from genomic profiling of transcription factors. *Nature Reviews Genetics* vol. 10 605–616 Preprint at https://doi.org/10.1038/nrg2636 (2009).

101. Grandi, F. C., Modi, H., Kampman, L. & Corces, M. R. Chromatin accessibility profiling by ATAC-seq. *Nature Protocols* vol. 17 1518–1552 Preprint at https://doi.org/10.1038/s41596-022-00692-9 (2022).

102. Buenrostro, J. D., Giresi, P. G., Zaba, L. C., Chang, H. Y. & Greenleaf, W. J. Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nat Methods* **10**, 1213–1218 (2013).

103. Ernst, J. *et al.* Mapping and analysis of chromatin state dynamics in nine human cell types. *Nature* **473**, 43–49 (2011).

104. Xie, W. *et al.* Epigenomic analysis of multilineage differentiation of human embryonic stem cells. *Cell* **153**, 1134–1148 (2013).

105. Dirks, R. A. M., Stunnenberg, H. G. & Marks, H. Genome-wide epigenomic profiling for biomarker discovery. *Clinical Epigenetics* vol. 8 Preprint at https://doi.org/10.1186/s13148-016-0284-4 (2016).

106. Rendeiro, A. F. *et al.* Chromatin accessibility maps of chronic lymphocytic leukaemia identify subtype-specific epigenome signatures and transcription regulatory networks. *Nat Commun* **7**, (2016).

107. Elsasser, W. M. Outline of a theory of cellular heterogeneity. *Proc Natl Acad Sci U S A* **81**, 5126–5129 (1984).

108. Altschuler, S. J. & Wu, L. F. Cellular Heterogeneity: Do Differences Make a Difference? *Cell* vol. 141 559–563 Preprint at https://doi.org/10.1016/j.cell.2010.04.033 (2010).

109. Lieberman, B. *et al.* Toward uncharted territory of cellular heterogeneity: advances and applications of single-cell RNA-seq. *J Transl Genet Genom* (2020) doi:10.20517/jtgg.2020.51.

110. Elowitz, M. B., Levine, A. J., Siggia, E. D. & Swain, P. S. Stochastic gene expression in a single cell. *Science (1979)* **297**, 1183–1186 (2002).

111. Kalisky, T., Blainey, P. & Quake, S. R. Genomic analysis at the single-cell level. *Annu Rev Genet* **45**, 431–445 (2011).

112. Kumar, R. M. *et al.* Deconstructing transcriptional heterogeneity in pluripotent stem cells. *Nature* **516**, 56–61 (2014).

113. Choi, Y. H. & Kim, J. K. Dissecting cellular heterogeneity using single-cell RNA sequencing. *Molecules and Cells* vol. 42 189–199 Preprint at https://doi.org/10.14348/molcells.2019.2446 (2019).

114. González-Silva, L., Quevedo, L. & Varela, I. Tumor Functional Heterogeneity Unraveled by scRNA-seq Technologies. *Trends in Cancer* vol. 6 13–19 Preprint at https://doi.org/10.1016/j.trecan.2019.11.010 (2020).

115. Zhang, P. *et al.* Single-cell RNA sequencing to track novel perspectives in HSC heterogeneity. *Stem Cell Research and Therapy* vol. 13 Preprint at https://doi.org/10.1186/s13287-022-02718-1 (2022).

116. Goldman, S. L. *et al.* The impact of heterogeneity on single-cell sequencing. *Frontiers in Genetics* vol. 10 Preprint at https://doi.org/10.3389/fgene.2019.00008 (2019).

117. Tang, F. *et al.* mRNA-Seq whole-transcriptome analysis of a single cell. *Nat Methods* (2009) doi:10.1038/nmeth.1315.

118. Wu, A. R. *et al.* Quantitative assessment of single-cell RNA-sequencing methods. *Nat Methods* (2014) doi:10.1038/nmeth.2694.

119. Tu, A. A. *et al.* TCR sequencing paired with massively parallel 3′ RNA-seq reveals clonotypic T cell signatures. *Nat Immunol* **20**, 1692–1699 (2019).

120. Petropoulos, S. *et al.* Single-Cell RNA-Seq Reveals Lineage and X Chromosome Dynamics in Human Preimplantation Embryos. *Cell* **165**, 1012–1026 (2016).

121. Stubbington, M. J. T. *et al.* T cell fate and clonality inference from single-cell transcriptomes. *Nat Methods* **13**, 329–332 (2016).

122. Shaffer, S. M. *et al.* Rare cell variability and drug-induced reprogramming as a mode of cancer drug resistance. *Nature* **546**, 431–435 (2017).

123. Shalek, A. K. *et al.* Single-cell transcriptomics reveals bimodality in expression and splicing in immune cells. *Nature* **498**, 236–240 (2013).

124. Hwang, B., Lee, J. H. & Bang, D. Single-cell RNA sequencing technologies and bioinformatics pipelines. *Experimental and Molecular Medicine* vol. 50 Preprint at https://doi.org/10.1038/s12276-018-0071-8 (2018).

125. Picelli, S. *et al.* Smart-seq2 for sensitive full-length transcriptome profiling in single cells. *Nat Methods* **10**, 1096–1100 (2013).

126. Sheng, K., Cao, W., Niu, Y., Deng, Q. & Zong, C. Effective detection of variation in single-cell transcriptomes using MATQ-seq. *Nat Methods* **14**, 267–270 (2017).

127. Ziegenhain, C. *et al.* Comparative Analysis of Single-Cell RNA Sequencing Methods. *Mol Cell* **65**, 631-643.e4 (2017).

128. Haque, A., Engel, J., Teichmann, S. A. & Lönnberg, T. A practical guide to single-cell RNA-sequencing for biomedical research and clinical applications. *Genome Medicine* vol. 9 1–12 Preprint at https://doi.org/10.1186/s13073-017-0467-4 (2017).

129. Macosko, E. Z. *et al.* Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. *Cell* **161**, 1202–1214 (2015).

130. Zheng, G. X. Y. *et al.* Massively parallel digital transcriptional profiling of single cells. *Nat Commun* **8**, 1–12 (2017).

131. Klein, A. M. *et al.* Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells. *Cell* **161**, 1187–1201 (2015).

132. Marcus, J. S., Anderson, W. F. & Quake, S. R. Microfluidic single-cell mRNA isolation and analysis. *Anal Chem* **78**, 3084–3089 (2006).

133. Weibel, D. B. & Whitesides, G. M. Applications of microfluidics in chemical biology. *Curr Opin Chem Biol* **10**, 584–591 (2006).

134. Hu, P., Zhang, W., Xin, H. & Deng, G. Single cell isolation and analysis. *Frontiers in Cell and Developmental Biology* vol. 4 116 Preprint at https://doi.org/10.3389/fcell.2016.00116 (2016).

135. Mylka, V. *et al.* Comparative analysis of antibody- and lipid-based multiplexing methods for single-cell RNA-seq. *Genome Biol* **23**, 55 (2022).

136. McGinnis, C. S. *et al.* MULTI-seq: sample multiplexing for single-cell RNA sequencing using lipid-tagged indices. *Nat Methods* **16**, 619–626 (2019).

137. Stoeckius, M. *et al.* Simultaneous epitope and transcriptome measurement in single cells. *Nat Methods* **14**, 865–868 (2017).

138. Kivioja, T. *et al.* Counting absolute numbers of molecules using unique molecular identifiers. *Nat Methods* **9**, 72–74 (2012).

139. Ziegenhain, C. *et al.* Comparative Analysis of Single-Cell RNA Sequencing Methods. *Mol Cell* **65**, 631-643.e4 (2017).

140. Yu, P. & Lin, W. Single-cell Transcriptome Study as Big Data. *Genomics Proteomics Bioinformatics* **14**, 21–30 (2016).

141. Andrews, S. FastQC. (2010).

142. Ilicic, T. *et al.* Classification of low quality cells from single-cell RNA-seq data. *Genome Biol* **17**, 29 (2016).

143. Osorio, D. & Cai, J. J. Systematic determination of the mitochondrial proportion in human and mice tissues for single-cell RNA-sequencing data quality control. *Bioinformatics* **37**, 963 (2021).

144. Vallejos, C. A., Risso, D., Scialdone, A., Dudoit, S. & Marioni, J. C. Normalizing single-cell RNA sequencing data: Challenges and opportunities. *Nature Methods* vol. 14 565–571 Preprint at https://doi.org/10.1038/nmeth.4292 (2017).

145. Hafemeister, C. & Satija, R. Normalization and variance stabilization of single-cell RNA-seq data using regularized negative binomial regression. *Genome Biol* **20**, 296 (2019).

146. Haghverdi, L., Lun, A. T. L., Morgan, M. D. & Marioni, J. C. Batch effects in single-cell RNA-sequencing data are corrected by matching mutual nearest neighbors. *Nat Biotechnol* **36**, 421–427 (2018).

147. Nayak, R. & Hasija, Y. A hitchhiker's guide to single-cell transcriptomics and data analysis pipelines. *Genomics* vol. 113 606–619 Preprint at https://doi.org/10.1016/j.ygeno.2021.01.007 (2021).

148. Chen, G., Ning, B. & Shi, T. Single-Cell RNA-Seq Technologies and Related Computational Data Analysis. *Fronteirs in Genetics* **10**, 1–13 (2019).

149. Townes, F. W., Hicks, S. C., Aryee, M. J. & Irizarry, R. A. Feature selection and dimension reduction for single-cell RNA-Seq based on a multinomial model. *Genome Biol* **20**, 295 (2019).

150. Hao, Y. *et al.* Integrated analysis of multimodal single-cell data. *Cell* **184**, 3573-3587.e29 (2021).

151. Aibar, S. *et al.* SCENIC: Single-cell regulatory network inference and clustering. *Nat Methods* **14**, 1083–1086 (2017).

152. Cahan, P. *et al.* CellNet: Network biology applied to stem cell engineering. *Cell* **158**, 903–915 (2014).

153. Rackham, O. J. L. *et al.* A predictive computational framework for direct reprogramming between human cell types. *Nat Genet* **48**, 331–335 (2016).

154. Ribeiro, M. M., Okawa, S. & del Sol, A. TransSynW: A single-cell RNA-sequencing based web application to guide cell conversion experiments. *Stem Cells Transl Med* **10**, 230–238 (2021).

155. Kiselev, V. Y., Yiu, A. & Hemberg, M. Scmap: Projection of single-cell RNA-seq data across data sets. *Nat Methods* **15**, 359–362 (2018).

156. la Manno, G. *et al.* RNA velocity of single cells. *Nature* **560**, 494–498 (2018).

157. Qiu, X. *et al.* Reversed graph embedding resolves complex single-cell trajectories. *Nat Methods* **14**, 979–982 (2017).

158. Gibney, E. R. & Nolan, C. M. Epigenetics and gene expression. *Heredity* vol. 105 4–13 Preprint at https://doi.org/10.1038/hdy.2010.54 (2010).

159. Li, Z. *et al.* Single-cell RNA-seq and chromatin accessibility profiling decipher the heterogeneity of mouse γδ T cells. *Sci Bull (Beijing)* **67**, 408–426 (2022).

160. Behjati Ardakani, F. *et al.* Integrative analysis of single-cell expression data reveals distinct regulatory states in bidirectional promoters. *Epigenetics Chromatin* **11**, 66 (2018).

161. Mei, S. *et al.* Cistrome Data Browser: A data portal for ChIP-Seq and chromatin accessibility data in human and mouse. *Nucleic Acids Res* **45**, D658–D662 (2017).

162. Oki, S. *et al.* ChIP-Atlas: a data-mining suite powered by full integration of public ChIP-seq data . *EMBO Rep* **19**, (2018).

163. Bickmore, W. A. & van Steensel, B. Genome architecture: Domain organization of interphase chromosomes. *Cell* vol. 152 1270–1284 Preprint at https://doi.org/10.1016/j.cell.2013.02.001 (2013).

164. Fraser, P. & Bickmore, W. Nuclear organization of the genome and the potential for gene regulation. *Nature* vol. 447 413–417 Preprint at https://doi.org/10.1038/nature05916 (2007).

165. Belton, J. M. *et al.* Hi-C: A comprehensive technique to capture the conformation of genomes. *Methods* **58**, 268–276 (2012).

166. Ron, G., Globerson, Y., Moran, D. & Kaplan, T. Promoter-enhancer interactions identified from Hi-C data using probabilistic models and hierarchical topological domains. *Nat Commun* **8**, 1–12 (2017).

167. Hariprakash, J. M. & Ferrari, F. Computational Biology Solutions to Identify Enhancers-target Gene Pairs. *Computational and Structural Biotechnology Journal* vol. 17 821–831 Preprint at https://doi.org/10.1016/j.csbj.2019.06.012 (2019).

168. Belokopytova, P. S., Nuriddinov, M. A., Mozheiko, E. A., Fishman, D. & Fishman, V. Quantitative prediction of enhancer–promoter interactions. *Genome Res* **30**, 72–84 (2020).

169. Thurman, R. E. *et al.* The accessible chromatin landscape of the human genome. *Nature* **489**, 75–82 (2012).

170. Sun, Y., Miao, N. & Sun, T. Detect accessible chromatin using ATAC-sequencing, from principle to applications. *Hereditas* vol. 156 29 Preprint at https://doi.org/10.1186/s41065-019-0105-9 (2019).

171. de la Torre-Ubieta, L. *et al.* The Dynamic Landscape of Open Chromatin during Human Cortical Neurogenesis. *Cell* **172**, 289-304.e18 (2018).

172. Raurell-Vila, H., Ramos-Rodríguez, M. & Pasquali, L. Assay for transposase accessible chromatin (ATAC-Seq) to chart the open chromatin landscape of human Pancreatic Islets. in *Methods in Molecular Biology* vol. 1766 197–208 (Humana Press Inc., 2018).

173. Zou, Z., Ohta, T., Miura, F. & Oki, S. ChIP-Atlas 2021 update: a data-mining suite for exploring epigenomic landscapes by fully integrating ChIP-seq, ATAC-seq and Bisulfite-seq data. *Nucleic Acids Res* **50**, W175–W182 (2022).

174. Fishilevich, S. *et al.* GeneHancer: Genome-wide integration of enhancers and target genes in GeneCards. *Database* **2017**, (2017).

175. Zhang, L., Zhang, J. & Nie, Q. DIRECT-NET: An efficient method to discover cis-regulatory elements and construct regulatory networks from single-cell multiomics data. *Sci Adv* **8**, (2022).

176. Lyu, P. *et al.* Gene regulatory networks controlling temporal patterning, neurogenesis, and cell-fate specification in mammalian retina. *Cell Rep* **37**, (2021).

177. Ali, M. & del Sol, A. Modeling of Cellular Systems: Application in Stem Cell Research and Computational Disease Modeling. in *Theoretical and Applied Aspects of Systems Biology* vol. 27 129–138 (2018).

178. Bolour, H. *Computational modeling of gene regulatory networks — A primer*. *Computational Modeling of Gene Regulatory Networks - A Primer* (Imperial College Press, 2008). doi:10.1142/P567.

179. Ahluwalia, U., Katyal, N. & Deep, S. Models of Protein Folding. *J Proteins Proteom* (2013).

180. Hughey, J. J., Lee, T. K. & Covert, M. W. Computational modeling of mammalian signaling networks. *Wiley Interdisciplinary Reviews: Systems Biology and Medicine* vol. 2 194–209 Preprint at https://doi.org/10.1002/wsbm.52 (2010).

181. Buccitelli, C. & Selbach, M. mRNAs, proteins and the emerging principles of gene expression control. *Nature Reviews Genetics* vol. 21 630–644 Preprint at https://doi.org/10.1038/s41576-020-0258-4 (2020).

182. Berger, S. L. The complex language of chromatin regulation during transcription. *Nature* vol. 447 407–412 Preprint at https://doi.org/10.1038/nature05915 (2007).

183. Kouzarides, T. & Berger, S. Chromatin modifications and their mechanism of action. in *Epigenetics* 191–209 (CSHL Press: New York, 2007).

184. Levine, M. & Tjian, R. Transcription regulation and animal diversity. *Nature* **424**, 147–151 (2003).

185. Chen, K. & Rajewsky, N. The evolution of gene regulation by transcription factors and microRNAs. *Nature Reviews Genetics* vol. 8 93–103 Preprint at https://doi.org/10.1038/nrg1990 (2007).

186. Nilsen, T. W. & Graveley, B. R. Expansion of the eukaryotic proteome by alternative splicing. *Nature* vol. 463 457–463 Preprint at https://doi.org/10.1038/nature08909 (2010).

187. Decker, C. J. & Parker, R. Mechanisms of mRNA degradation in eukaryotes. *Trends in Biochemical Sciences* vol. 19 336–340 Preprint at https://doi.org/10.1016/0968-0004(94)90073-6 (1994).

188. Mercatelli, D., Scalambra, L., Triboli, L., Ray, F. & Giorgi, F. M. Gene regulatory network inference resources: A practical overview. *Biochimica et Biophysica Acta - Gene Regulatory Mechanisms* vol. 1863 194430 Preprint at https://doi.org/10.1016/j.bbagrm.2019.194430 (2020).

189. Koutrouli, M., Karatzas, E., Paez-Espino, D. & Pavlopoulos, G. A. A Guide to Conquer the Biological Network Era Using Graph Theory. *Frontiers in Bioengineering and Biotechnology* vol. 8 34 Preprint at https://doi.org/10.3389/fbioe.2020.00034 (2020).

190. Seo, C. H., Kim, J.-R., Kim, M.-S. & Cho, K.-H. Hub genes with positive feedbacks function as master switches in developmental gene regulatory networks. *Bioinformatics* **25**, 1898–1904 (2009).

191. Okawa, S. & del Sol, A. A general computational approach to predicting synergistic transcriptional cores that determine cell subpopulation identities. *Nucleic Acids Res* **47**, 3333–3343 (2019).

192. Nguyen, H., Tran, D., Tran, B., Pehlivan, B. & Nguyen, T. A comprehensive survey of regulatory network inference methods using single cell RNA sequencing data. *Briefings in Bioinformatics* vol. 22 Preprint at https://doi.org/10.1093/bib/bbaa190 (2021).

193. Hartmann, A. S., Ravichandran, S. & del Sol, A. Modeling Cellular Differentiation and Reprogramming with Gene Regulatory Networks. in *Methods in Molecular Biology* vol. Chapter 2 37–51 (2019).

194. Huynh-Thu, V. A., Irrthum, A., Wehenkel, L. & Geurts, P. Inferring Regulatory Networks from Expression Data Using Tree-Based Methods. *PLoS One* **5**, e12776 (2010).

195. Pratapa, A., Jalihal, A. P., Law, J. N., Bharadwaj, A. & Murali, T. M. Benchmarking algorithms for gene regulatory network inference from single-cell transcriptomic data. *Nat Methods* **17**, 147–154 (2020).

196. Jung, S., Appleton, E., Ali, M., Church, G. M. & del Sol, A. A computer-guided design tool to increase the efficiency of cellular conversions. *Nat Commun* **12**, 1–12 (2021).

197. Hu, X., Hu, Y., Wu, F., Leung, R. W. T. & Qin, J. Integration of single-cell multi-omics for gene regulatory network inference. *Comput Struct Biotechnol J* **18**, 1925–1938 (2020).

198. Zarayeneh, N. *et al.* Integrative Gene Regulatory Network inference using multi-omics data. *Proceedings - 2016 IEEE International Conference on Bioinformatics and Biomedicine, BIBM 2016* 1336–1340 (2017) doi:10.1109/BIBM.2016.7822711.

199. Holmberg, J. & Perlmann, T. Maintaining differentiated cellular identity. *Nature Reviews Genetics* vol. 13 429–439 Preprint at https://doi.org/10.1038/nrg3209 (2012).

200. Chan, T. E., Stumpf, M. P. H. & Babtie, A. C. Gene Regulatory Network Inference from Single-Cell Data Using Multivariate Information Measures. *Cell Syst* **5**, 251-267.e3 (2017).

201. Okawa, S. *et al.* Transcriptional synergy as an emergent property defining cell subpopulation identity enables population shift. *Nat Commun* **9**, 1–10 (2018).

202. Boyer, L. A. *et al.* Core transcriptional regulatory circuitry in human embryonic stem cells. *Cell* **122**, 947–956 (2005).

203. Mazzoni, E. O. *et al.* Synergistic binding of transcription factors to cell-specific enhancers programs motor neuron identity. *Nat Neurosci* **16**, 1219–1227 (2013).

204. Wilson, N. K. *et al.* Combinatorial transcriptional control in blood stem/progenitor cells: Genome-wide analysis of ten major transcriptional regulators. *Cell Stem Cell* **7**, 532–544 (2010).

205. Maniatis, T., Goodbourn, S. & Fischer, J. A. Regulation of inducible and tissue-specific gene expression. *Science (1979)* **236**, 1237–1245 (1987).

206. Slattery, M. *et al.* Cofactor binding evokes latent differences in DNA binding specificity between hox proteins. *Cell* **147**, 1270–1282 (2011).

207. Bell, A. J. *THE CO-INFORMATION LATTICE*.

208. Griffith, V. & Koch, C. Quantifying Synergistic Mutual Information. in 159–190 (2014). doi:10.1007/978-3-642-53734-9_6.

209. Ameri, A. J. & Lewis, Z. A. Shannon entropy as a metric for conditional gene expression in Neurospora crassa. *G3 Genes|Genomes|Genetics* **11**, (2021).

210. Timme, N., Alford, W., Flecker, B. & Beggs, J. M. Synergy, redundancy, and multivariate information measures: An experimentalist's perspective. *Journal of Computational Neuroscience* vol. 36 119–140 Preprint at https://doi.org/10.1007/s10827-013-0458-4 (2014).

211. Mayor, R. Cell fate decisions during development. *Science* vol. 364 937–938 Preprint at https://doi.org/10.1126/science.aax7917 (2019).

212. Lassar, A. B., Paterson, B. M. & Weintraub, H. Transfection of a DNA locus that mediates the conversion of 10T1/2 fibroblasts to myoblasts. *Cell* **47**, 649–656 (1986).

213. Kulessa, H., Frampton, J. & Graf, T. GATA-1 reprograms avian myelomonocytic cell lines into eosinophils, thromboblasts, and erythroblasts. *Genes Dev* **9**, 1250–1262 (1995).

214. Heyworth, C., Pearson, S., May, G. & Enver, T. Transcription factor-mediated lineage switching reveals plasticity in primary committed progenitor cells. *EMBO Journal* **21**, 3770–3781 (2002).

215. Fisher, A. G. Cellular identity and lineage choice. *Nat Immunol* **2**, 977–982 (2002).

216. Basu, A. & Tiwari, V. K. Epigenetic reprogramming of cell identity: lessons from development for regenerative medicine. *Clinical Epigenetics* vol. 13 Preprint at https://doi.org/10.1186/s13148-021-01131-4 (2021).

217. Klemm, S. L., Shipony, Z. & Greenleaf, W. J. Chromatin accessibility and the regulatory epigenome. *Nature Reviews Genetics 2018 20:4* **20**, 207–220 (2019).

218. Lee, T. I. & Young, R. A. Transcriptional regulation and its misregulation in disease. *Cell* **152**, 1237–1251 (2013).

219. Whetstine, J. R. Histone methylation. chemically inert but chromatin dynamic. in *Handbook of Cell Signaling, 2/e* vol. 3 2389–2397 (Elsevier Inc., 2010).

220. Miranda, T. B. & Jones, P. A. DNA methylation: The nuts and bolts of repression. *J Cell Physiol* **213**, 384–390 (2007).

221. Gardiner-Garden, M. & Frommer, M. CpG Islands in vertebrate genomes. *J Mol Biol* **196**, 261–282 (1987).

222. Goldberg, A. D., Allis, C. D. & Bernstein, E. Epigenetics: A Landscape Takes Shape. *Cell* vol. 128 635–638 Preprint at https://doi.org/10.1016/j.cell.2007.02.006 (2007).

223. Venters, B. J. & Pugh, B. F. How eukaryotic genes are transcribed regulation of eukaroytic gene transcription. *Critical Reviews in Biochemistry and Molecular Biology* vol. 44 117–141 Preprint at https://doi.org/10.1080/10409230902858785 (2009).

224. Vaquerizas, J. M., Kummerfeld, S. K., Teichmann, S. A. & Luscombe, N. M. A census of human transcription factors: Function, expression and evolution. *Nature Reviews Genetics* vol. 10 252–263 Preprint at https://doi.org/10.1038/nrg2538 (2009).

225. Graf, T. & Enver, T. Forcing cells to change lineages. *Nature* vol. 462 587–594 Preprint at https://doi.org/10.1038/nature08533 (2009).

226. Yamanaka, S. Induced pluripotent stem cells: Past, present, and future. *Cell Stem Cell* vol. 10 678–684 Preprint at https://doi.org/10.1016/j.stem.2012.05.005 (2012).

227. Takahashi, K. & Yamanaka, S. Induction of Pluripotent Stem Cells from Mouse Embryonic and Adult Fibroblast Cultures by Defined Factors. *Cell* **126**, 663–676 (2006).

228. Gaspar-Maia, A., Alajem, A., Meshorer, E. & Ramalho-Santos, M. Open chromatin in pluripotency and reprogramming. *Nature Reviews Molecular Cell Biology* vol. 12 36–47 Preprint at https://doi.org/10.1038/nrm3036 (2011).

229. Balsalobre, A. & Drouin, J. Pioneer factors as master regulators of the epigenome and cell fate. *Nature Reviews Molecular Cell Biology* Preprint at https://doi.org/10.1038/s41580-022-00464-z (2022).

230. Cirillo, L. A. *et al.* Opening of compacted chromatin by early developmental transcription factors HNF3 (FoxA) and GATA-4. *Mol Cell* **9**, 279–289 (2002).

231. Minderjahn, J. *et al.* Mechanisms governing the pioneering and redistribution capabilities of the non-classical pioneer PU.1. *Nat Commun* **11**, 1–16 (2020).

232. Raposo, A. A. S. F. *et al.* Ascl1 coordinately regulates gene expression and the chromatin landscape during neurogenesis. *Cell Rep* **10**, 1544–1556 (2015).

233. Iwafuchi-Doi, M. *et al.* The Pioneer Transcription Factor FoxA Maintains an Accessible Nucleosome Configuration at Enhancers for Tissue-Specific Gene Activation. *Mol Cell* **62**, 79–91 (2016).

234. Heinz, S. *et al.* Simple Combinations of Lineage-Determining Transcription Factors Prime cis-Regulatory Elements Required for Macrophage and B Cell Identities. *Mol Cell* **38**, 576–589 (2010).

235. van Oevelen, C. *et al.* C/EBPα Activates Pre-existing and de Novo Macrophage Enhancers during Induced Pre-B Cell Transdifferentiation and Myelopoiesis. *Stem Cell Reports* **5**, 232–247 (2015).

236. Dodonova, S. O., Zhu, F., Dienemann, C., Taipale, J. & Cramer, P. Nucleosome-bound SOX2 and SOX11 structures elucidate pioneer factor function. *Nature* **580**, 669–672 (2020).

237. Mayran, A. *et al.* Pioneer factor Pax7 deploys a stable enhancer repertoire for specification of cell fate. *Nat Genet* **50**, 259–269 (2018).

238. Reizel, Y. *et al.* FoxA-dependent demethylation of DNA initiates epigenetic memory of cellular identity. *Dev Cell* **56**, 602-612.e4 (2021).

239. Lupien, M. *et al.* FoxA1 Translates Epigenetic Signatures into Enhancer-Driven Lineage-Specific Transcription. *Cell* **132**, 958–970 (2008).

240. Lee, K. *et al.* FOXA2 Is Required for Enhancer Priming during Pancreatic Differentiation. *Cell Rep* **28**, 382-393.e7 (2019).

241. Sekiya, S. & Suzuki, A. Direct conversion of mouse fibroblasts to hepatocyte-like cells by defined factors. *Nature* **475**, 390–395 (2011).

242. Alver, B. H. *et al.* The SWI/SNF chromatin remodelling complex is required for maintenance of lineage specific enhancers. *Nat Commun* **8**, 1–10 (2017).

243. Neiman, D. *et al.* Islet cells share promoter hypomethylation independently of expression, but exhibit cell-type-specific methylation in enhancers. *Proc Natl Acad Sci U S A* **114**, 13525–13530 (2017).

244. Waddington, C. H. *The strategy of the genes; a discussion of some aspects of theoretical biology*. *Allen & Unwin* (London: George Allen & Unwin, Ltd., 1957).

245. Davis, R. L., Weintraub, H. & Lassar, A. B. Expression of a single transfected cDNA converts fibroblasts to myoblasts. *Cell* **51**, 987–1000 (1987).

246. Vierbuchen, T. *et al.* Direct conversion of fibroblasts to functional neurons by defined factors. *Nature* **463**, 1035–1041 (2010).

247. Pang, Z. P. *et al.* Induction of human neuronal cells by defined transcription factors. *Nature* vol. 476 220–223 Preprint at https://doi.org/10.1038/nature10202 (2011).

248. Weltner, J. *et al.* Human pluripotent reprogramming with CRISPR activators. *Nat Commun* **9**, (2018).

249. Guerrero-Ramirez, G. I., Valdez-Cordoba, C. M., Islas-Cisneros, J. F. & Jiang, V. Computational approaches for predicting key transcription factors in targeted cell reprogramming. *Molecular Medicine Reports* vol. 18 1225–1237 Preprint at https://doi.org/10.3892/mmr.2018.9092 (2018).

250. Okawa, S., Nicklas, S., Zickenrott, S., Schwamborn, J. C. & del Sol, A. A Generalized Gene-Regulatory Network Model of Stem Cell Differentiation for Predicting Lineage Specifiers. *Stem Cell Reports* **7**, 307–315 (2016).

251. Kamaraj, U. S., Gough, J., Polo, J. M., Petretto, E. & Rackham, O. J. L. Computational methods for direct cell conversion. *Cell Cycle* vol. 15 3343–3354 Preprint at https://doi.org/10.1080/15384101.2016.1238119 (2016).

252. Almeida, N. *et al.* Employing core regulatory circuits to define cell identity. *EMBO J* **40**, (2021).

253. Morris, S. A. *et al.* Dissecting engineered cell types and enhancing cell fate conversion via Cellnet. *Cell* **158**, 889–902 (2014).

254. Stumpf, P. S. *et al.* Stem Cell Differentiation as a Non-Markov Stochastic Process. *Cell Syst* **5**, 268-282.e7 (2017).

255. Moignard, V. *et al.* Decoding the regulatory network of early blood development from single-cell gene expression measurements. *Nat Biotechnol* **33**, 269–276 (2015).

256. Fernandes, H. J. R. *et al.* Single-Cell Transcriptomics of Parkinson's Disease Human In Vitro Models Reveals Dopamine Neuron-Specific Stress Responses. *Cell Rep* **33**, 108263 (2020).

257. Qian, L. *et al.* In vivo reprogramming of murine cardiac fibroblasts into induced cardiomyocytes. *Nature* **485**, 593–598 (2012).

258. Arenas, E., Denham, M. & Villaescusa, J. C. How to make a midbrain dopaminergic neuron. *Development* **142**, 1918–1936 (2015).

259. Ieda, M. *et al.* Direct reprogramming of fibroblasts into functional cardiomyocytes by defined factors. *Cell* **142**, 375–386 (2010).

260. Marro, S. *et al.* Direct lineage conversion of terminally differentiated hepatocytes to functional neurons. *Cell Stem Cell* **9**, 374–382 (2011).

261. Dominguez, A. A., Lim, W. A. & Qi, L. S. Beyond editing: Repurposing CRISPR-Cas9 for precision genome regulation and interrogation. *Nature Reviews Molecular Cell Biology* Preprint at https://doi.org/10.1038/nrm.2015.2 (2016).

262. Black, J. B. *et al.* Targeted Epigenetic Remodeling of Endogenous Loci by CRISPR/Cas9-Based Transcriptional Activators Directly Converts Fibroblasts to Neuronal Cells. *Cell Stem Cell* **19**, 406–414 (2016).

263. Wang, C. *et al.* Loss of MyoD Promotes Fate Transdifferentiation of Myoblasts Into Brown Adipocytes. *EBioMedicine* **16**, 212–223 (2017).

264. Liu, X. S. *et al.* Editing DNA Methylation in the Mammalian Genome. *Cell* **167**, 233-247.e17 (2016).

265. Giehrl-Schwab, J. *et al.* Parkinson's disease motor symptoms rescue by CRISPRa-reprogramming astrocytes into GABAergic neurons. *EMBO Mol Med* **14**, (2022).

266. Espinosa Angarica, V. & del Sol, A. Modeling heterogeneity in the pluripotent state: A promising strategy for improving the efficiency and fidelity of stem cell differentiation. *BioEssays* vol. 38 758–768 Preprint at https://doi.org/10.1002/bies.201600103 (2016).

267. Lindvall, O. & Björklund, A. Cell Replacement Therapy: Helping the Brain to Repair Itself. *NeuroRx* vol. 1 379–381 Preprint at https://doi.org/10.1602/neurorx.1.4.379 (2004).

268. del Sol, A. & Jung, S. The Importance of Computational Modeling in Stem Cell Research. *Trends in Biotechnology* vol. 39 126–136 Preprint at https://doi.org/10.1016/j.tibtech.2020.07.006 (2021).

269. Torper, O. *et al.* Generation of induced neurons via direct conversion in vivo. *Proc Natl Acad Sci U S A* **110**, 7038–7043 (2013).

270. Guo, Z. *et al.* In vivo direct reprogramming of reactive glial cells into functional neurons after brain injury and in an Alzheimer's disease model. *Cell Stem Cell* **14**, 188–202 (2014).

271. Rivetti Di Val Cervo, P. *et al.* Induction of functional dopamine neurons from human astrocytes in vitro and mouse astrocytes in a Parkinson's disease model. *Nat Biotechnol* **35**, 444–452 (2017).

272. Park, N. I. *et al.* ASCL1 Reorganizes Chromatin to Direct Neuronal Fate and Suppress Tumorigenicity of Glioblastoma Stem Cells. *Cell Stem Cell* **21**, 209-224.e7 (2017).

273. Brumbaugh, J., Stefano, B. di & Hochedlinger, K. Reprogramming: Identifying the mechanisms that safeguard cell identity. *Development (Cambridge)* vol. 146 Preprint at https://doi.org/10.1242/dev.182170 (2019).

274. Ikeda, T. *et al.* Srf destabilizes cellular identity by suppressing cell-type-specific gene expression programs. *Nat Commun* **9**, 1–15 (2018).

275. Maiti, P., Manna, J., Dunbar, G. L., Maiti, P. & Dunbar, G. L. Current understanding of the molecular mechanisms in Parkinson's disease: Targets for potential treatments. *Translational Neurodegeneration* vol. 6 1–35 Preprint at https://doi.org/10.1186/s40035-017-0099-z (2017).

276. Guo, T. *et al.* Molecular and cellular mechanisms underlying the pathogenesis of Alzheimer's disease. *Molecular Neurodegeneration* vol. 15 1–37 Preprint at https://doi.org/10.1186/s13024-020-00391-7 (2020).

277. Hanahan, D. & Weinberg, R. A. Hallmarks of cancer: The next generation. *Cell* vol. 144 646–674 Preprint at https://doi.org/10.1016/j.cell.2011.02.013 (2011).

278. von Bernhardi, R. Glial cell dysregulation: A new perspective on Alzheimer disease. *Neurotox Res* **12**, 215–232 (2007).

279. Navarro, E. *et al.* Dysregulation of mitochondrial and proteolysosomal genes in Parkinson's disease myeloid cells. *Nat Aging* **1**, 850–863 (2021).

280. Bhatia, D. *et al.* T-cell dysregulation is associated with disease severity in Parkinson's Disease. *J Neuroinflammation* **18**, (2021).

281. Soto, C. & Estrada, L. D. Protein misfolding and neurodegeneration. *Archives of Neurology* vol. 65 184–189 Preprint at https://doi.org/10.1001/archneurol.2007.56 (2008).

282. Michel, P. P., Hirsch, E. C. & Hunot, S. Understanding Dopaminergic Cell Death Pathways in Parkinson Disease. *Neuron* vol. 90 675–691 Preprint at https://doi.org/10.1016/j.neuron.2016.03.038 (2016).

283. Epstein, D. J. Cis-regulatory mutations in human disease. *Briefings in Functional Genomics and Proteomics* vol. 8 310–316 Preprint at https://doi.org/10.1093/bfgp/elp021 (2009).

284. Matharu, N. & Ahituv, N. Modulating gene regulation to treat genetic disorders. *Nature Reviews Drug Discovery* vol. 19 757–775 Preprint at https://doi.org/10.1038/s41573-020-0083-7 (2020).

285. Prabantu, V. M., Naveenkumar, N. & Srinivasan, N. Influence of Disease-Causing Mutations on Protein Structural Networks. *Front Mol Biosci* **7**, 492 (2021).

286. Ward, L. D. & Kellis, M. Interpreting noncoding genetic variation in complex traits and human disease. *Nature Biotechnology* vol. 30 1095–1106 Preprint at https://doi.org/10.1038/nbt.2422 (2012).

287. Claringbould, A. & Zaugg, J. B. Enhancers in disease: molecular basis and emerging treatment strategies. *Trends in Molecular Medicine* vol. 27 1060–1073 Preprint at https://doi.org/10.1016/j.molmed.2021.07.012 (2021).

288. Cheung, V. G. & Spielman, R. S. The genetics of variation in gene expression. *Nature Genetics* vol. 32 522–525 Preprint at https://doi.org/10.1038/ng1036 (2002).

289. Karki, R., Pandya, D., Elston, R. C. & Ferlini, C. Defining 'mutation' and 'polymorphism' in the era of personal genomics. *BMC Medical Genomics* vol. 8 Preprint at https://doi.org/10.1186/s12920-015-0115-z (2015).

290. Uffelmann, E. *et al.* Genome-wide association studies. *Nature Reviews Methods Primers* vol. 1 1–21 Preprint at https://doi.org/10.1038/s43586-021-00056-9 (2021).

291. Visscher, P. M., Yengo, L., Cox, N. J. & Wray, N. R. Discovery and implications of polygenicity of common diseases. *Science* vol. 373 1468–1473 Preprint at https://doi.org/10.1126/science.abi8206 (2021).

292. Soto-Ortolaza, A. I. *et al.* GWAS risk factors in Parkinson's disease: LRRK2 coding variation and genetic interaction with PARK16. *American Journal of Neurodegenerative Diseases* **2**, 287–299 (2013).

293. Gorlov, I., Xiao, X., Mayes, M., Gorlova, O. & Amos, C. SNP eQTL status and eQTL density in the adjacent region of the SNP are associated with its statistical significance in GWA studies. *BMC Genet* **20**, 85 (2019).

294. Nica, A. C. & Dermitzakis, E. T. Expression quantitative trait loci: Present and future. *Philosophical Transactions of the Royal Society B: Biological Sciences* vol. 368 Preprint at https://doi.org/10.1098/rstb.2012.0362 (2013).

295. Yazar, S. *et al.* Single-cell eQTL mapping identifies cell type-specific genetic control of autoimmune disease. *Science (1979)* **376**, (2022).

296. Chen, X. *et al.* Tissue-specific enhancer functional networks for associating distal regulatory regions to disease. *Cell Syst* **12**, 353-362.e6 (2021).

297. Võsa, U. *et al.* Large-scale cis- and trans-eQTL analyses identify thousands of genetic loci and polygenic scores that regulate blood gene expression. *Nat Genet* **53**, 1300–1310 (2021).

298. Nasser, J. *et al.* Genome-wide enhancer maps link risk variants to disease genes. *Nature* **593**, 238–243 (2021).

299. Maurano, M. T. *et al.* Large-scale identification of sequence variants influencing human transcription factor occupancy in vivo. *Nat Genet* **47**, 1393–1401 (2015).

300. Broekema, R. v., Bakker, O. B. & Jonkers, I. H. A practical view of fine-mapping and gene prioritization in the post-genome-wide association era. *Open Biol* **10**, (2020).

301. Iacono, G., Massoni-Badosa, R. & Heyn, H. Single-cell transcriptomics unveils gene regulatory network plasticity. *Genome Biol* **20**, 110 (2019).

302. Singh, A. J., Ramsey, S. A., Filtz, T. M. & Kioussi, C. Differential gene regulatory networks in development and disease. *Cellular and Molecular Life Sciences* vol. 75 1013–1025 Preprint at https://doi.org/10.1007/s00018-017-2679-6 (2018).

303. Weighill, D. *et al.* Gene Targeting in Disease Networks. *Front Genet* **12**, 501 (2021).

304. Dusonchet, J. *et al.* A Parkinson's disease gene regulatory network identifies the signaling protein RGS2 as a modulator of LRRK2 activity and neuronal toxicity. *Hum Mol Genet* **23**, 4887–4905 (2014).

305. Schaid, D. J., Chen, W. & Larson, N. B. From genome-wide associations to candidate causal variants by statistical fine-mapping. *Nature Reviews Genetics* vol. 19 491–504 Preprint at https://doi.org/10.1038/s41576-018-0016-z (2018).

306. Chiou, J. *et al.* Interpreting type 1 diabetes risk with genetics and single-cell epigenomics. *Nature* **594**, 398–402 (2021).

307. Yu, F. *et al.* Variant to function mapping at single-cell resolution through network propagation. *Nat Biotechnol* 1–10 (2022) doi:10.1038/s41587-022-01341-y.

308. D'Alessio, A. C. *et al.* A systematic approach to identify candidate transcription factors that control cell identity. *Stem Cell Reports* **5**, 763–775 (2015).

309. Callaway, E. M. *et al.* A multimodal cell census and atlas of the mammalian primary motor cortex. *Nature 2021 598:7879* **598**, 86–102 (2021).

310. Strzelecka, P. M., Ranzoni, A. M. & Cvejic, A. Dissecting human disease with single-cell omics: application in model systems and in the clinic. *Dis Model Mech* **11**, (2018).

311. Grath, A. & Dai, G. Direct cell reprogramming for tissue engineering and regenerative medicine. *Journal of Biological Engineering* vol. 13 14 Preprint at https://doi.org/10.1186/s13036-019-0144-9 (2019).

312. Hernandez, C. *et al.* Systemic blood immune cell populations as biomarkers for the outcome of immune checkpoint inhibitor therapies. *International Journal of Molecular Sciences* vol. 21 Preprint at https://doi.org/10.3390/ijms21072411 (2020).

313. Sant, G. R., Knopf, K. B. & Albala, D. M. Live-single-cell phenotypic cancer biomarkers-future role in precision oncology? *NPJ Precis Oncol* **1**, 1–7 (2017).

314. Hu, Y. *et al.* Neural network learning defines glioblastoma features to be of neural crest perivascular or radial glia lineages. *Sci Adv* **8**, (2022).

315. Emmert-Streib, F., Dehmer, M. & Haibe-Kains, B. Gene regulatory networks and their applications: Understanding biological and medical problems in terms of networks. *Front Cell Dev Biol* **2**, 38 (2014).

316. Uddin, F., Rudin, C. M. & Sen, T. CRISPR Gene Therapy: Applications, Limitations, and Implications for the Future. *Front Oncol* **10**, 1387 (2020).

317. Ohnuki, M. & Takahashi, K. Present and future challenges of induced pluripotent stem cells. *Philos Trans R Soc Lond B Biol Sci* **370**, (2015).

318. Zhang, Y.-X., Chen, S.-L., Li, Y.-M. & Zheng, Y.-W. Limitations and challenges of direct cell reprogramming in vitro and in vivo. *Histol Histopathol* 18458 (2022) doi:10.14670/HH-18-458.

319. Kasai, S., Li, X., Torii, S., Yasumoto, K. ichi & Sogawa, K. Direct protein–protein interaction between Npas4 and IPAS mutually inhibits their critical roles in neuronal cell survival and death. *Cell Death Discov* **7**, (2021).

320. Hau, A. C. *et al.* Transcriptional cooperation of PBX1 and PAX6 in adult neural progenitor cells. *Sci Rep* **11**, (2021).

321. Villaescusa, J. C. *et al.* A PBX1 transcriptional network controls dopaminergic neuron development and is impaired in Parkinson's disease. *EMBO J* **35**, 1963 (2016).

322. Klenova, E., Chernukhin, I., Inoue, T., Shamsuddin, S. & Norton, J. Immunoprecipitation techniques for the analysis of transcription factor complexes. *Methods* **26**, 254–259 (2002).

323. Margineanu, A. *et al.* Screening for protein-protein interactions using Förster resonance energy transfer (FRET) and fluorescence lifetime imaging microscopy (FLIM). *Scientific Reports 2016 6:1* **6**, 1–17 (2016).

324. Smallwood, S. A. *et al.* Single-cell genome-wide bisulfite sequencing for assessing epigenetic heterogeneity. *Nature Methods 2014 11:8* **11**, 817–820 (2014).

325. Hu, H. *et al.* AnimalTFDB 3.0: a comprehensive resource for annotation and prediction of animal transcription factors. *Nucleic Acids Res* **47**, D33 (2019).

326. Alvarez-Dominguez, J. R. & Melton, D. A. Cell maturation: Hallmarks, triggers, and manipulation. *Cell* **185**, 235–249 (2022).

327. Gonçalves, J. P. *et al.* TFRank: network-based prioritization of regulatory associations underlying transcriptional responses. *Bioinformatics* **27**, 3149–3157 (2011).

328. Hung, C. W., Chen, Y. C., Hsieh, W. L., Chiou, S. H. & Kao, C. L. Ageing and neurodegenerative diseases. *Ageing Research Reviews* vol. 9 S36–S46 Preprint at https://doi.org/10.1016/j.arr.2010.08.006 (2010).

329. Barker, R. A., Barrett, J., Mason, S. L. & Björklund, A. Fetal dopaminergic transplantation trials and the future of neural grafting in Parkinson's disease. *The Lancet Neurology* vol. 12 84–91 Preprint at https://doi.org/10.1016/S1474-4422(12)70295-8 (2013).

330. Gu, X. L. *et al.* Astrocytic expression of Parkinson's disease-related A53T -synuclein causes neurodegeneration in mice. *Mol Brain* **3**, 12 (2010).

331. Lee, H. J. *et al.* Direct transfer of α-synuclein from neuron to astroglia causes inflammatory responses in synucleinopathies. *Journal of Biological Chemistry* **285**, 9262–9272 (2010).

332. Magnusson, J. P. *et al.* A latent neurogenic program in astrocytes regulated by Notch signaling in the mouse. *Science* **346**, 237–241 (2014).

333. Vilpoux, C. *et al.* Astrogliosis and compensatory neurogenesis after the first ethanol binge drinking-like exposure in the adolescent rat. *Alcohol Clin Exp Res* **46**, 207–220 (2022).

334. Addis, R. C. *et al.* Efficient Conversion of Astrocytes to Functional Midbrain Dopaminergic Neurons Using a Single Polycistronic Vector. *PLoS One* **6**, e28719 (2011).

335. Torper, O. *et al.* In Vivo Reprogramming of Striatal NG2 Glia into Functional Neurons that Integrate into Local Host Circuitry. *Cell Rep* **12**, 474–481 (2015).

336. Wang, H., Yang, Y., Liu, J. & Qian, L. Direct cell reprogramming: approaches, mechanisms and progress. *Nature Reviews Molecular Cell Biology 2021 22:6* **22**, 410–424 (2021).

337. Luo, C. *et al.* Global DNA methylation remodeling during direct reprogramming of fibroblasts to neurons. *Elife* **8**, (2019).

338. Zalatan, J. G. *et al.* Engineering complex synthetic transcriptional programs with CRISPR RNA scaffolds. *Cell* **160**, 339–350 (2015).

339. Li, Y. & Tollefsbol, T. O. DNA methylation detection: Bisulfite genomic sequencing analysis. *Methods in Molecular Biology* **791**, 11–21 (2011).

340. O'Geen, H., Echipare, L. & Farnham, P. J. Using ChIP-seq technology to generate high-resolution profiles of histone modifications. *Methods in Molecular Biology* **791**, 265–286 (2011).

341. Baumann, V. *et al.* Targeted removal of epigenetic barriers during transcriptional reprogramming. *Nat Commun* **10**, 1–12 (2019).

342. Rada-Iglesias, A. *et al.* A unique chromatin signature uncovers early developmental enhancers in humans. *Nature* **470**, 279–285 (2011).

343. Delvecchio, M., Gaucher, J., Aguilar-Gurrieri, C., Ortega, E. & Panne, D. Structure of the p300 catalytic core and implications for chromatin targeting and HAT regulation. *Nat Struct Mol Biol* **20**, 1040–1046 (2013).

344. Shrimp, J. H. *et al.* Chemical Control of a CRISPR-Cas9 Acetyltransferase. *ACS Chem Biol* **13**, 455–460 (2018).

345. Qian, H. *et al.* Reversing a model of Parkinson's disease with in situ converted nigral neurons. *Nature 2020 582:7813* **582**, 550–556 (2020).

346. Wang, L. L. *et al.* Revisiting astrocyte to neuron conversion with lineage tracing in vivo. *Cell* **184**, 5465-5481.e16 (2021).

347. Yang, G. *et al.* Ptbp1 knockdown in mouse striatum did not induce astrocyte-to-neuron conversion using HA-tagged labeling system. *bioRxiv* 2022.03.29.486202 (2022) doi:10.1101/2022.03.29.486202.

348. Xie, Y., Zhou, J. & Chen, B. Critical examination of Ptbp1-mediated glia-to-neuron conversion in the mouse retina. *Cell Rep* **39**, 110960 (2022).

349. Chen, W., Zhen, Q., Huang, Q., Ma, S. & Li, M. Repressing PTBP1 fails to convert reactive astrocytes to dopaminergic neurons in a 6-hydroxydopamine mouse model of Parkinson's disease. *Elife* **11**, (2022).

350. Xue, Y. *et al.* Direct Conversion of Fibroblasts to Neurons by Reprogramming PTB-Regulated microRNA Circuits. *Cell* **152**, 82 (2013).

351. Li, Y. & Zhou, L. quan. dCas9 techniques for transcriptional repression in mammalian cells: Progress, applications and challenges. *BioEssays* **43**, 2100086 (2021).

352. Yeo, N. C. *et al.* An enhanced CRISPR repressor for targeted mammalian gene regulation. *Nat Methods* **15**, 611 (2018).

353. Alerasool, N., Segal, D., Lee, H. & Taipale, M. An efficient KRAB domain for CRISPRi applications in human cells. *Nature Methods 2020 17:11* **17**, 1093–1096 (2020).

354. Cates, K. *et al.* Deconstructing Stepwise Fate Conversion of Human Fibroblasts to Neurons by MicroRNAs. *Cell Stem Cell* **28**, 127 (2021).

355. Hough, S. R., Clements, I., Welch, P. J. & Wiederholt, K. A. Differentiation of mouse embryonic stem cells after RNA interference-mediated silencing of OCT4 and Nanog. *Stem Cells* **24**, 1467–1475 (2006).

356. Ortinski, P. I. *et al.* Selective induction of astrocytic gliosis generates deficits in neuronal inhibition. *Nat Neurosci* **13**, 584–591 (2010).

357. Xie, Y., Wang, T., Sun, G. Y. & Ding, S. Specific Disruption of Astrocytic Ca2+ Signaling Pathway in vivo by Adeno-Associated Viral Transduction. *Neuroscience* **170**, 992 (2010).

358. Lundblad, M., Picconi, B., Lindgren, H. & Cenci, M. A. A model of L-DOPA-induced dyskinesia in 6-hydroxydopamine lesioned mice: Relation to motor and cellular parameters of nigrostriatal function. *Neurobiol Dis* **16**, 110–123 (2004).

359. Mirelman, A. *et al.* Arm swing as a potential new prodromal marker of Parkinson's disease. *Mov Disord* **31**, 1527–1534 (2016).

360. Li, T., Chen, X., Zhang, C., Zhang, Y. & Yao, W. An update on reactive astrocytes in chronic pain. *J Neuroinflammation* **16**, 1–13 (2019).

361. Clarke, L. E. *et al.* Normal aging induces A1-like astrocyte reactivity. doi:10.1073/pnas.1800165115.

362. Ugalde, C. L. *et al.* Markers of A1 astrocytes stratify to molecular sub-types in sporadic Creutzfeldt–Jakob disease brain. *Brain Commun* **2**, (2020).

363. Seo, M. H. & Yeo, S. Association of increase in Serping1 level with dopaminergic cell reduction in an MPTP-induced Parkinson's disease mouse model. *Brain Res Bull* **162**, 67–72 (2020).

364. Gurdon, J. B. & Melton, D. A. Nuclear reprogramming in cells. *Science* vol. 322 1811–1815 Preprint at https://doi.org/10.1126/science.1160810 (2008).

365. Southall, T. D., Davidson, C. M., Miller, C., Carr, A. & Brand, A. H. Dedifferentiation of neurons precedes tumor formation in lola mutants. *Dev Cell* **28**, 685–696 (2014).

366. Caldwell, A. B. *et al.* Dedifferentiation and neuronal repression define familial Alzheimer's disease. *Sci Adv* **6**, (2020).

367. Jopling, C., Boue, S. & Belmonte, J. C. I. Dedifferentiation, transdifferentiation and reprogramming: Three routes to regeneration. *Nature Reviews Molecular Cell Biology* vol. 12 79–89 Preprint at https://doi.org/10.1038/nrm3043 (2011).

368. Brockes, J. P. & Kumar, A. Plasticity and reprogramming of differentiated cells in amphibian regeneration. *Nature Reviews Molecular Cell Biology* vol. 3 566–574 Preprint at https://doi.org/10.1038/nrm881 (2002).

369. Burclaff, J. & Mills, J. C. Plasticity of differentiated cells in wound repair and tumorigenesis, part I: stomach and pancreas. *DMM Disease Models and Mechanisms* vol. 11 Preprint at https://doi.org/10.1242/DMM.033373 (2018).

370. Cao, Y. Tumorigenesis as a process of gradual loss of original cell identity and gain of properties of neural precursor/progenitor cells. *Cell and Bioscience* vol. 7 1–14 Preprint at https://doi.org/10.1186/s13578-017-0188-9 (2017).

371. Puri, S., Folias, A. E. & Hebrok, M. Plasticity and dedifferentiation within the pancreas: Development, homeostasis, and disease. *Cell Stem Cell* vol. 16 18–31 Preprint at https://doi.org/10.1016/j.stem.2014.11.001 (2015).

372. Wang, Z. *et al.* Cell Lineage-Based Stratification for Glioblastoma. *Cancer Cell* **38**, 366-379.e8 (2020).

373. Lu, X. *et al.* Cell-lineage controlled epigenetic regulation in glioblastoma stem cells determines functionally distinct subgroups and predicts patient survival. *Nature Communications 2022 13:1* **13**, 1–16 (2022).

374. Dunn, G. P. *et al.* Emerging insights into the molecular and cellular basis of glioblastoma. *Genes Dev* **26**, 756–784 (2012).

375. McLendon, R. *et al.* Comprehensive genomic characterization defines human glioblastoma genes and core pathways. *Nature* **455**, 1061–1068 (2008).

376. Brennan, C. W. *et al.* The somatic genomic landscape of glioblastoma. *Cell* **155**, 462 (2013).

377. Neftel, C. *et al.* An Integrative Model of Cellular States, Plasticity, and Genetics for Glioblastoma. *Cell* **178**, 835-849.e21 (2019).

378. Kim, Y. *et al.* Perspective of mesenchymal transformation in glioblastoma. *Acta Neuropathologica Communications* vol. 9 Preprint at https://doi.org/10.1186/s40478-021-01151-4 (2021).

379. Bhaduri, A. *et al.* Outer Radial Glia-like Cancer Stem Cells Contribute to Heterogeneity of Glioblastoma. *Cell Stem Cell* **26**, 48-63.e6 (2020).

380. Yuan, J. *et al.* Single-cell transcriptome analysis of lineage diversity in high-grade glioma. *Genome Med* **10**, (2018).

381. Weng, Q. *et al.* Single-Cell Transcriptomics Uncovers Glial Progenitor Diversity and Cell Fate Determinants during Development and Gliomagenesis. *Cell Stem Cell* **24**, 707-723.e8 (2019).

382. Weinstein, J. N. *et al.* The cancer genome atlas pan-cancer analysis project. *Nat Genet* **45**, 1113–1120 (2013).

383. Kamouchi, M., Ago, T. & Kitazono, T. Brain pericytes: Emerging concepts and functional roles in brain homeostasis. *Cellular and Molecular Neurobiology* vol. 31 175–193 Preprint at https://doi.org/10.1007/s10571-010-9605-x (2011).

384. Marques, S. *et al.* Oligodendrocyte heterogeneity in the mouse juvenile and adult central nervous system. *Science (1979)* **352**, 1326–1329 (2016).

385. Bell, R. D. *et al.* Pericytes Control Key Neurovascular Functions and Neuronal Phenotype in the Adult Brain and during Brain Aging. *Neuron* **68**, 409–427 (2010).

386. Thanabalasundaram, G., Schneidewind, J., Pieper, C. & Galla, H. J. The impact of pericytes on the blood-brain barrier integrity depends critically on the pericyte differentiation stage. *International Journal of Biochemistry and Cell Biology* **43**, 1284–1293 (2011).

387. Mravic, M. *et al.* From pericytes to perivascular tumours: Correlation between pathology, stem cell biology, and tissue engineering. *International Orthopaedics* vol. 38 1819–1824 Preprint at https://doi.org/10.1007/s00264-014-2295-0 (2014).

388. Ando, K. *et al.* Peri-arterial specification of vascular mural cells from naïve mesenchyme requires notch signaling. *Development (Cambridge)* **146**, (2019).

389. Vanlandewijck, M. *et al.* A molecular atlas of cell types and zonation in the brain vasculature. *Nature* **554**, 475–480 (2018).

390. Sun, R., Kong, X., Qiu, X., Huang, C. & Wong, P. P. The Emerging Roles of Pericytes in Modulating Tumor Microenvironment. *Frontiers in Cell and Developmental Biology* vol. 9 1037 Preprint at https://doi.org/10.3389/fcell.2021.676342 (2021).

391. Caspani, E. M., Crossley, P. H., Redondo-Garcia, C. & Martinez, S. Glioblastoma: A pathogenic crosstalk between tumor cells and pericytes. *PLoS One* **9**, (2014).

392. Navarro, R., Compte, M., Álvarez-Vallina, L. & Sanz, L. Immune regulation by pericytes: Modulating innate and adaptive immunity. *Frontiers in Immunology* vol. 7 Preprint at https://doi.org/10.3389/fimmu.2016.00480 (2016).

393. Koboldt, D. C. *et al.* Comprehensive molecular portraits of human breast tumours. *Nature* **490**, 61–70 (2012).

394. Ciriello, G., Cerami, E., Sander, C. & Schultz, N. Mutual exclusivity analysis identifies oncogenic network modules. *Genome Res* **22**, 398–406 (2012).

395. He, R. *et al.* H3K4 demethylase KDM5B regulates cancer cell identity and epigenetic plasticity. *Oncogene* **41**, 2958–2972 (2022).

396. Romani, M., Pistillo, M. P. & Banelli, B. Epigenetic targeting of glioblastoma. *Frontiers in Oncology* vol. 8 Preprint at https://doi.org/10.3389/fonc.2018.00448 (2018).

397. Hegi, M. E. *et al.* MGMT Gene Silencing and Benefit from Temozolomide in Glioblastoma . *New England Journal of Medicine* **352**, 997–1003 (2005).

398. Roos, W. P. *et al.* Apoptosis in malignant glioma cells triggered by the temozolomide-induced DNA lesion O6-methylguanine. *Oncogene* **26**, 186–197 (2007).

399. Stupp, R. *et al.* Effects of radiotherapy with concomitant and adjuvant temozolomide versus radiotherapy alone on survival in glioblastoma in a randomised phase III study: 5-year analysis of the EORTC-NCIC trial. *Lancet Oncol* **10**, 459–466 (2009).

400. Stupp, R. *et al.* Radiotherapy plus Concomitant and Adjuvant Temozolomide for Glioblastoma. *New England Journal of Medicine* **352**, 987–996 (2005).

401. Noushmehr, H. *et al.* Identification of a CpG Island Methylator Phenotype that Defines a Distinct Subgroup of Glioma. *Cancer Cell* **17**, 510–522 (2010).

402. Lemma, R. B. *et al.* Pioneer transcription factors are associated with the modulation of DNA methylation patterns across cancers. *Epigenetics Chromatin* **15**, 1–19 (2022).

403. Gusyatiner, O. & Hegi, M. E. Glioma epigenetics: From subclassification to novel treatment options. *Seminars in Cancer Biology* vol. 51 50–58 Preprint at https://doi.org/10.1016/j.semcancer.2017.11.010 (2018).

404. Hiramoto, M. *et al.* Comparative analysis of type 2 diabetes-associated SNP alleles identifies allele-specific DNA-binding proteins for the KCNQ1 locus. *Int J Mol Med* **36**, 222–230 (2015).

405. Selvaraj, S. & Piramanayagam, S. Impact of gene mutation in the development of Parkinson's disease. *Genes Dis* **6**, 120 (2019).

406. Shastry, B. S. SNPs in disease gene mapping, medicinal drug development and evolution. *Journal of Human Genetics 2007 52:11* **52**, 871–880 (2007).

407. Kimchi-Sarfaty, C. *et al.* A 'silent' polymorphism in the MDR1 gene changes substrate specificity. *Science* **315**, 525–528 (2007).

408. Gray, I. C., Campbell, D. A. & Spurr, N. K. Single nucleotide polymorphisms as tools in human genetics. *Hum Mol Genet* **9**, 2403–2408 (2000).

409. Visscher, P. M. *et al.* 10 Years of GWAS Discovery: Biology, Function, and Translation. *Am J Hum Genet* **101**, 5–22 (2017).

410. Gazal, S. *et al.* Combining SNP-to-gene linking strategies to identify disease genes and assess disease omnigenicity. *Nature Genetics 2022 54:6* **54**, 827–836 (2022).

411. Bryois, J. *et al.* Cell-type-specific cis-eQTLs in eight human brain cell types identify novel risk genes for psychiatric and neurological disorders. *Nat Neurosci* **25**, 1104–1112 (2022).

412. Andersson, R. *et al.* An atlas of active enhancers across human cell types and tissues. *Nature 2014 507:7493* **507**, 455–461 (2014).

413. Yu, F. *et al.* Variant to function mapping at single-cell resolution through network propagation. *Nature Biotechnology 2022* 1–10 (2022) doi:10.1038/s41587-022-01341-y.

414. Kang, R. *et al.* EnhancerDB: a resource of transcriptional regulation in the context of enhancers. *Database* **2019**, (2019).

415. Mattioli, K. *et al.* Cis and trans effects differentially contribute to the evolution of promoters and enhancers. *Genome Biol* **21**, 1–22 (2020).

416. Spicuglia, S. & Vanhille, L. Chromatin signatures of active enhancers. *Nucleus* **3**, 126 (2012).

417. Clark, S. J., Lee, H. J., Smallwood, S. A., Kelsey, G. & Reik, W. Single-cell epigenomics: Powerful new methods for understanding gene regulation and cell identity. *Genome Biol* **17**, 1–10 (2016).

418. Degtyareva, A. O., Antontseva, E. v. & Merkulova, T. I. Regulatory SNPs: Altered Transcription Factor Binding Sites Implicated in Complex Traits and Diseases. *Int J Mol Sci* **22**, 22 (2021).

419. Landrum, M. J. *et al.* ClinVar: improving access to variant interpretations and supporting evidence. *Nucleic Acids Res* **46**, D1062–D1067 (2018).

420. Gao, P., Dong, X., Wang, Y. & Wei, G. H. Optimized CRISPR/Cas9-mediated single nucleotide mutation in adherent cancer cell lines. *STAR Protoc* **2**, (2021).