UNIVERSITÉ DU LUXEMBOURG

PhD-FSTM-2022-094
The Faculty of Science, Technology and Medicine

DISSERTATION

Defence held on 09/09/2022 in Esch-sur-Alzette

to obtain the degree of

DOCTEUR DE L'UNIVERSITÉ DU LUXEMBOURG

EN BIOLOGIE

by

Céline BARLIER
Born on 23 June 1994 in Colmar (France)

SINGLE CELL BASED COMPUTATIONAL APPROACHES TO UNRAVEL DYSREGULATIONS IN DISEASES

Dissertation defence committee

Dr. Antonio del Sol, dissertation supervisor
Professor, Université du Luxembourg

Dr. Ernest Arenas
Professor, Karolinska Institutet

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

Dr. Urko Martínez Marigorta
CIC bioGUNE

Dr. Thomas Sauter
Professor, Université du Luxembourg

Affidavit

I hereby confirm that the PhD thesis entitled “SINGLE CELL BASED COMPUTATIONAL APPROACHES TO UNRAVEL DYSREGULATIONS IN DISEASES” has been written independently and without any other sources than cited.

Luxembourg, 25/07/2022
Céline Barlier

Acknowledgements

First of all, I would like to thank my PhD supervisor Prof. Dr. Antonio del Sol for the opportunity he gave me when accepting me as a PhD candidate in his research group. I would like to thank him for his help and the guidance he provided me during my PhD studies. I have learnt to think critically, to become more independent in my research, and I improved my skills and knowledge in the field thanks to the training I had during these years.

Then, I would like to thank my CET members, Prof. Dr. Jens Schwamborn and Prof. Dr. Ernest Arenas for their yearly follow-up, the scientific discussions and valuable feedbacks they gave me to improve my research projects. I would also like to thank the Fond National de la Recherche Luxembourg for funding my research and all the colleagues of my DTU (PARK-QC) for all the exciting scientific exchanges we had. Moreover, I would like to thank my collaborators, especially Prof. Dr. Juan Anguita and Diego Barriales for their valuable contribution in one of my research projects.

Finally, I would like to thank Dr. Srikanth Ravichandran and Dr. Sascha Jung for their co-supervision of my PhD projects. Thank you so much for all the time you took to guide me, provide me with valuable guidance and suggestions, for all the stimulating scientific discussions, and most of all for your advice to help me grow as a researcher. I would also like to thank all my CBG colleagues, especially Meztli Matadamas, Mariana Ribeiro and Menglin Zheng, for all the help and support they provided me during these years.

Dedication

This thesis is dedicated to:

My parents, sister and grand-mothers *for their endless support;*

The memory of ***my grand-father***

Table of Contents

List of Figures

Introduction

Manuscript 1

Manuscript 2

Manuscript 3

List of Abbreviations

AD: Alzheimer's disease

BP: Biological process

ChIP-seq: Chromatin Immunoprecipitation sequencing

CPM: Count per million

DNA: Deoxyribonucleic acid

DE: Differential expression

DEG: Differentially expressed gene

DREAM: Dialogue on reverse engineering assessment and methods

GRN: Gene regulatory network

GS: Gold standard

GSEA: Gene Set Enrichment Analysis

GWAS: Genome wide association study

Hi-C: High-throughput chromatin conformation capture

HPC: High performance computing

iPSCs: induced pluripotent stem cells

MRTF: Master regulator transcription factor

RNA: Ribonucleic acid

mRNA: Messenger ribonucleic acid

PBMC: Peripheral blood mononuclear cell

PD: Parkinson's disease

PPV: Positive predictive value

scATAC-seq: single cell assay for transposase-accessible chromatin sequencing

SNP: Single nucleotide polymorphism

scRNA-seq: single cell assay for ribonucleic acid sequencing

TF: Transcription factor

TPM: Transcript per million

t-SNE: t-distributed stochastic neighbor embedding

UMAP: Uniform manifold approximation and projection

UMI: Unique molecular identifier

Summary

The characterization of cells escaping the physiological landscape, the understanding of pathological mechanisms, and the identification of novel targets for new therapeutic strategies are part of the main aims of computational disease modelling. The accurate characterization of cell identity and identification of key transcription factors (TFs) for cell conversion holds great promises to revert disease states towards healthy ones. Moreover, the characterization of the Gene Regulatory Network (GRN) is crucial to better understand impaired regulatory mechanisms and identify potential targets for disease treatment. To date, several computational methods have been implemented to tackle the aforementioned aims. First, some methods were developed to characterize cell identity, including the identification of cell identity genes. However, these computational methods solely rely on tissue samples, usually composed of a mixture of cell classification (e.g., cell types, subtypes) which hinders the accurate capture of identity genes. Moreover, they categorize genes as being expressed or non-expressed, and hence discard intermediate levels of expression which have been shown to be involved in the functional outcome of the cells. Further, current methods rely on genome-wide or highly variable genes to identify subtle differences such as cell states. However, these approaches do not accurately decipher functional cell states neither the genes that characterize them. Finally, several GRN inference methods based on single cell transcriptomics have been developed over the years. However, few of them exploit the single cell multi-OMICS data to infer more comprehensive GRNs, including the interaction between TFs and the enhancers of regulated genes, to provide a better understanding of impaired regulatory mechanisms in disease conditions.

In this thesis, three computational strategies were developed to overcome the limitations of current methods and tackle main challenges of systems biology and disease modelling. First, HCellig was implemented to accurately characterize cellular identity. HCellig is based on a hierarchical cell identity composed of three layers including cell type, subtype and phenotype to overcome the mixture of different cell classification that can hinder the capture of identity genes. In addition, HCellig quantifies gene into three levels of expression to provide a more refined functional characterization of the cell identity. The use of HCellig on mouse and human large-scale datasets allowed us to generate two high-resolution cell identity atlases for both organisms. Second, FunPart was developed to decipher functional cell states while capturing the key genes characterizing them by using a feature selection strategy combined with a clustering approach. The application of FunPart on a large

compendium of mouse infection datasets generated a *Catalogus Immune Muris* comprising all the functional cell states identified and the key genes defining their state. In particular, these genes could be candidate immunomodulator as we demonstrated for *Zfp591*, a previously unknown transcription factor modulating macrophages response to *Salmonella* infection. Lastly, we designed RNetDys, a systematic multi-OMICS pipeline to infer regulatory interactions mediated by TFs and enhancers of regulated genes for specific cell (sub)types or states and identify candidate impaired regulatory interactions in diseases due to single nucleotide polymorphisms (SNPs). We showed that RNetDys overcome current approaches to infer cell (sub)type specific GRN and validated the relevance of captured impaired interactions across five diseases.

In summary, the three computational methods proposed in this thesis cover the cell identity and gene regulatory mechanisms aspects, in physiological and pathological conditions. Together, they will contribute to a better understanding of cells escaping the physiological landscape, a more accurate characterization of pathological cells states and dysregulated regulatory mechanisms, and the identification of candidate genes to design novel therapeutic strategies to treat diseases.

1 Introduction

1.1 Disease modelling to guide new therapeutic approaches

1.1.1 Diseases and systems biology

Diseases result from abnormal modifications in the function or structure of a tissue, organ or group of organs. They can be roughly grouped as those resulting from genetic factors or environmental factors (Antony *et al.*, 2012). In particular, diseases resulting from genetic factors can range from single causal factors (monogenic disease) to polygenic or multifactorial diseases (Weatherall, 2000; Antonarakis and Beckmann, 2006; Visscher *et al.*, 2021). Nevertheless, the combination of both genetic and environmental factors has been reported to impact the onset and progression of most diseases (Knip *et al.*, 2005; Antony *et al.*, 2012). In that regard, multifactorial or complex diseases, such as Parkinson's disease (PD) and epilepsy, are those for which the interplay of environmental factors and several genes is believed to influence their progression (Ottman *et al.*, 1996; Warner and Schapira, 2003). For instance, the multifactorial nature of PD has been demonstrated by the identification of several PD-related genes (e.g. *SCNA, LRRK2, DJ-1*), the characterization of diverse genetic risk factors and the study of some environmental factors such as cigarette smoking and caffeine consumption that could alter the risk of PD development (Pérez-Tur, 2006; Kouli *et al.*, 2018).

The prevalence of many complex diseases, such as diabetes and cardiovascular diseases, has dramatically increased in the last few years (Mardinoglu and Nielsen, 2016). Moreover, a considerable number of diseases still lack of effective medical treatments to prevent, treat and cure them (Kiser and Pronovost, 2009; Cummings *et al.*, 2021; Hansson, 2021). Therefore, there is a need for new therapeutic approaches that would allow the detection, prevention and treatment of diseases. However, the development of new therapeutic strategies requires a deep understanding of the cellular heterogeneity and underlying molecular mechanisms involved (Gitler *et al.*, 2017; Schett *et al.*, 2021; Mortada *et al.*, 2021). In that regard, systems biology is an active and evolving multidisciplinary field of research that includes computational modelling and wet-lab expertise, to pave the way towards new therapeutic approaches and personalized medicine (Wolkenhauer *et al.*, 2013; Gabhann *et al.*, 2010; van Kampen and Moerland, 2016). In particular, disease modelling using computational approaches is an active research field of systems biology that aims at

developing computational models to study different aspects of diseases. These models aspire at providing a valuable guidance for experimental and clinical setups to develop strategies that detect, prevent and/or treat diseases.

1.1.2 Computational approaches to guide treatments and personalized medicine

Computational modelling methods aim at developing models based on assumptions and data evidences to provide explanations and insights into a scientific problem that can then tested or refined using further investigations involving experimental validations (Barh *et al.*, 2020). In particular, models can be categorized in two main categories depending on their general aim, with descriptive models intending at providing explanations for an observation, and predictive models aiming at predicting the result of novel observations (Motta and Pappalardo, 2013). The development of a model is an iterative process in which it is common to use additional observations for refinement purposes. In addition, a descriptive model or the predictions obtained from a predictive model can be validated using *in vitro* or *in vivo* experimental strategies (Kitano, 2002). Over the years, several computational models have been implemented to study a wide spectrum of diseases and get a better understanding of their cellular and molecular complexity. For instance, methods have been developed to study the characteristics of diseases and identify candidate genes involved in diseases using different models and approaches (Gill *et al.*, 2014). Notably, computational modelling methods contributed to the discovery of heterogeneity and complexity of Alzheimer's disease (AD) and PD for which the notion that they are fundamentally governed by amyloid-β, tau, and α-synuclein proteins has been challenged (Lam *et al.*, 2020).

The findings and insights provided by computational methods help the development of novel therapeutic approaches and personalized medicine strategies (Figure 1). In particular, models at the cellular and molecular levels such as Gene Regulatory Network (GRN) based methods holds great promise to predict key transcription factors (TFs) for cellular conversion that can be applied for cell-based therapies (del Sol and Jung, 2021). In that regard, one main goal of regenerative medicine is the replacement of damaged cells by healthy and functional ones using cell transplantation strategies (Edgar *et al.*, 2020). The guidance provided by computational methods greatly contributed to the stem cell engineering, allowing the reprogramming or differentiation of cells toward the target cells of interest (Cahan *et al.*, 2021). For instance, induced pluripotent stem cells (iPSCs) are used to produce functionally mature dopaminergic neurons to treat PD, characterized by the loss of dopaminergic neurons

in the substantia nigra. Notably, the first clinical trial to treat PD using iPSCs has been initiated by Shinya Yamanka in 2018 (Aly, 2020). In addition, computational models are powerful tools to advance personalized medicine by optimizing outcomes of patients based on their unique disease features and biological properties (Figure 1). The generation of patient-specific models held great promises to monitor diseases and open new venues for personalized healthcare (Chen and Snyder, 2012).

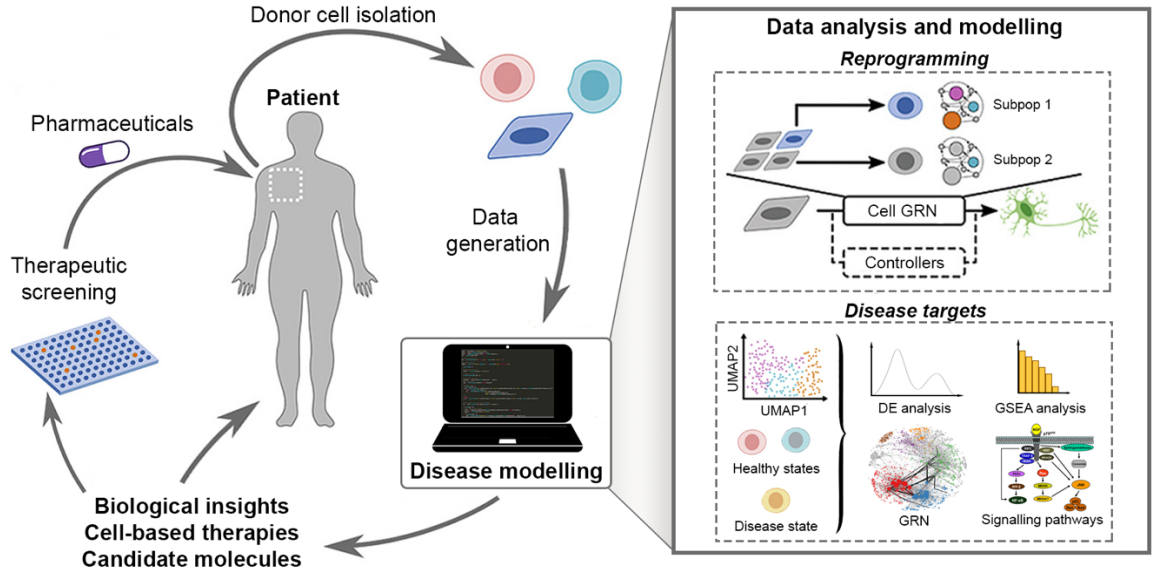

Figure 1. Computational disease modelling contribution for personalized medicine.
Figure modified from (Wang *et al.*, 2020), from (Gu *et al.*, 2012) for the GRN and (Niewiadomska *et al.*, 2011) for the signalling pathway. It shows the interplay of disease modelling to contribute to the development of novel therapeutic strategies applied to personalized medicine.

Over the years, several studies have been focusing on different aspects of diseases using computational biology approaches to dissect cellular heterogeneity and shed the light towards the composition of biological systems (Satija *et al.*, 2015; Butler *et al.*, 2018). Moreover, many approaches focused on deciphering their molecular complexity to provide mechanistic insights on the processes involved in diseases progression or to identify candidate genes that could be used as therapeutic targets for disease treatment (Szabo *et al.*, 2019; De Luca *et al.*, 2020). Notably, efforts have been made to develop computational methods driving the discovery of cellular heterogeneity, identifying molecules for cell phenotype conversion and providing insights of the underlying mechanisms leading to a disease state (Figure 2) (Hassan *et al.*, 2018; Jenner *et al.*, 2020; Collin *et al.*, 2022; Pappalardo *et al.*, 2016; Ford Versypt, 2021). However, despite recent efforts and valuable contributions to develop computational systems biology strategies aiming for therapeutic or

clinical applications, several challenges remain to be solved (Ma and Lim, 2021; Cha and Lee, 2020; Zhao *et al.*, 2020).

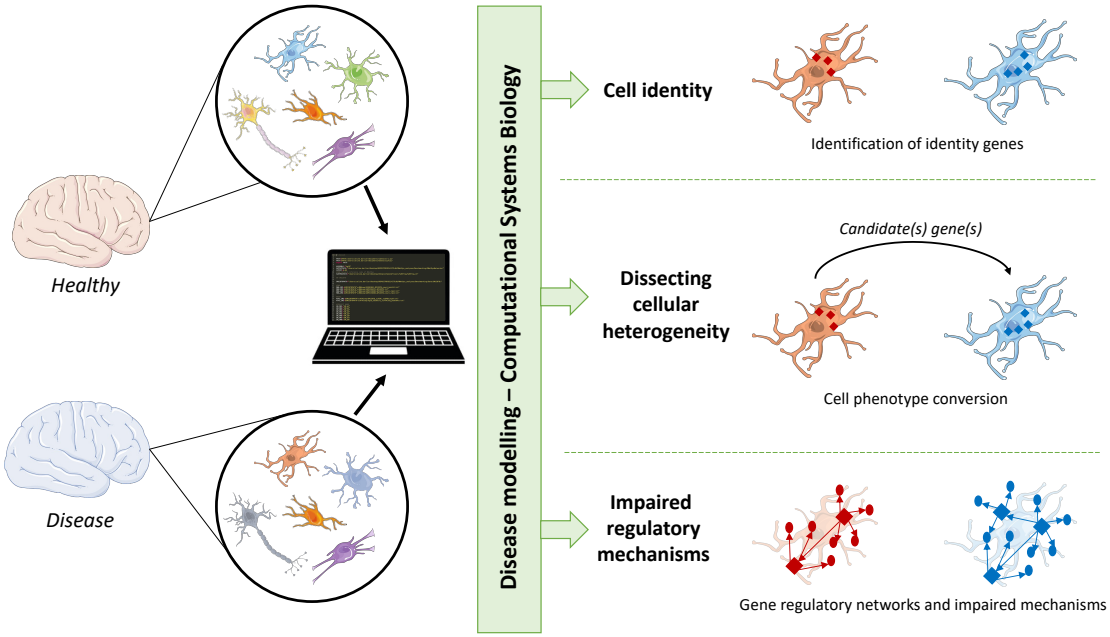

Figure 2. Computational models for disease modelling.
This figure summarizes some aspects tackled by research projects exploiting computational models to study diseases. The development of computational systems biology approaches focused on different aspects such as the comprehensive characterization of cell identity, the dissection of cellular heterogeneity and identification of candidate genes for cell phenotype conversion, and the study of dysregulated regulatory mechanisms.

1.1.3 Development of single cell-based technologies

For years, research studies have been relying on bulk-sequencing technologies allowing the measurement of features such as genes quantification across pool of cell populations (Li and Wang, 2021). However, these averaged measurements obscured the discovery of tissues composition, cell-to-cell variability and rare cell (sub)populations that could be involved in diseases (Wang and Navin, 2015). The emergence and fast development of single-cell based technologies led to the generation of different types of omics data, such as single cell RNA-seq and single cell ATAC-seq, that allowed large-scale and refined measurements at the cell level. The high-resolution of this data enhanced the dissection of cells heterogeneity and molecular complexity of mechanisms involved at different biological scales (Wang and Bodovitz, 2010; Trapnell, 2015; Papalexi and Satija, 2017; Lähnemann *et al.*, 2020).

Single-cell technologies uncovered a considerable number of previously unknown cell (sub)types throughout the generation of organism-wide cell atlases (Rozenblatt-Rosen *et al.*,

2017; The Tabula Muris Consortium *et al.*, 2018; Zhang *et al.*, 2021; The Tabula Sapiens Consortium *et al.*, 2022). In addition, this high-resolution data enhanced the discovery and characterization of novel cellular states (Trapnell, 2015). The creation of a comprehensive landscape of cell phenotypes would contribute to the systematic identification of cells that cross physiological bounds towards pathological states (Morris *et al.*, 2019; Szabo *et al.*, 2019). Moreover, single-cell data contributed to our understanding of cell fates and gene regulatory mechanisms by providing an unprecedent molecular resolution at the cell level (Perkel, 2021). The development of single cell OMICS data allowed to link different features to decipher the complexity of multicellular organisms and the underlying mechanisms driving physiological and pathological processes. In particular, the combination of epigenomics and transcriptomics helped the study of lineage determination and mechanisms involved in the development of diseases (Ogbeide *et al.*, 2022). However, the precision in features measurement provided by single cell technologies raised several challenges (Potter, 2018; Cha and Lee, 2020; Lähnemann *et al.*, 2020). Notably, the sparsity and important variability of single cell data hinders the accurate detection of relevant features, complex gene patterns and discovery of new cell (sub)types or states (Kiselev, Tallulah S. Andrews, *et al.*, 2019; Lähnemann *et al.*, 2020).

1.2 Characterization of cell identity

1.2.1 Cellular identity and destabilization in disease state

Multicellular organisms are composed of highly heterogeneous cells organized in different layers to form complex entities such as tissues and organs. For a long time, cells were classified based on diverse features including their location, morphology or interactions with other cells (Arendt *et al.*, 2016; Morris *et al.*, 2019). However, the emergence of single-cell based technologies allowed for a more precise and refined measurement of cell features that uncovered the wide complexity of biological systems and showed the limitation of the previous classification system. Indeed, the generation of organism-wide cell atlases provided more insights into the cellular heterogeneity (The Tabula Muris Consortium *et al.*, 2018; Zhang *et al.*, 2021; The Tabula Sapiens Consortium *et al.*, 2022). For instance, the Tabula Sapiens is a single cell transcriptomics atlas reporting the gene expression profiles for 475 cell (sub)types across 24 human tissues. This atlas allowed for the discovery of shared and tissue-specific properties across cell types such as the macrophages, a cell type shared across tissues but displaying subtle differences in genes expression that are tissue-specific (The Tabula Sapiens Consortium *et al.*, 2022).

Cells originate from different lineages and acquire part of their identity during the developmental process, guided by cell fate determinants, in which pluripotent cells differentiate to give rise to more specialized cells such as cell types or cell subtypes (Mayor, 2019; Belmonte-Mateos and Pujades, 2022). In addition, cells express different sets of genes depending on their micro-environment and the functions they have to perform, leading to different phenotypes. Indeed, the most refined level of resolution for cellular heterogeneity is the cell state level for which the same cell (sub)type could respond differently to perturbations and hence display a variety of phenotypes (Dueck *et al.*, 2016; Nimmo *et al.*, 2015). Single cell RNA-seq technologies greatly contributed to the dissection of cellular heterogeneity by leveraging the high-resolution of gene expression patterns displayed by individual cells (Choi and Kim, 2019). The hematopoietic system has been widely studied to better understand hematopoiesis and uncover the wide diversity of cell types and subtypes differentiating from hematopoietic stem cells (Watcham *et al.*, 2019; Dolgalev and Tikhonova, 2021). Indeed, the study of the hematopoietic cell landscape using single-cell technologies shaped, modified and extended the hematopoietic development tree (Watcham *et al.*, 2019). Notably, hematopoietic progenitor cells were found to be in a continuous transcriptional landscape branching into seven fates including erythroid, basophilic, megakaryocytic, lymphocytic, dendritic, monocytic and granulocytic neutrophil lineages (Tusi *et al.*, 2018). In addition, immune cells have been shown to display a wide diversity of phenotypes during immune responses, highlighting their dynamic and plasticity (Satija and Shalek, 2014; Gause *et al.*, 2020). In particular, the binary classification of M1 and M2 macrophages, with M1 macrophages displaying pro-inflammatory properties and M2 macrophages displaying anti-inflammatory properties, has been questioned by the discovery of the wide spectrum of macrophages polarization states (Kim and Nair, 2019; Liu *et al.*, 2020). Therefore, the identification and molecular characterization of more subtle differences such as rare cell (sub)populations or cell states still remains elusive (Nguyen *et al.*, 2018; Andreatta *et al.*, 2021).

The comprehensive characterization of cells identities in the organism cellular landscape (e.g., human) would allow the identification of cells displaying non-physiological features and potentially going toward disease-related states (Morris *et al.*, 2019). Indeed, the maintenance of cellular identity is crucial to conserve the homeostasis and integrity of the organism. Cell identity is maintained by a set of genes, named identity genes, that ensure the

physiological properties of the cells such as their functions (Xia *et al.*, 2020; Kim *et al.*, 2021). In that regard, identity genes are defined as a combination of unique genes specifically expressed to characterize and maintain cell identity. The loss or perturbation of identity genes can lead to the destabilization or disruption of cell identity, which has been shown to be associated with pathological processes and involved in several diseases (Ikeda *et al.*, 2018; Brumbaugh *et al.*, 2019; Budday *et al.*, 2015). For instance, the identity of human dopaminergic neurons was shown to be destabilized in response to diverse PD related stress factors (Fernandes *et al.*, 2020). Another example is the loss of β-cell identity which has been shown to be involved in diabetic phenotypes (Mostafa *et al.*, 2020). Therefore, it is required to accurately characterize cell identity by deciphering identity genes to have a comprehensive understanding of the cellular landscape heterogeneity and distinguish physiological features from pathological ones.

1.2.2 Deciphering identity genes and its limitations

The accurate identification of identity genes to characterize cell identity remains a central challenge in biology (Morris *et al.*, 2019). Several efforts have been made in this direction and diverse computational methods aiming at identifying such genes based on single-cell transcriptomics data have been developed and used in the past few years (Stuart *et al.*, 2019; Wang *et al.*, 2019; Delaney *et al.*, 2019). Notably, Seurat is a well-established pipeline for single cell RNA-seq datasets analysis composed of several features such as the quality control of the data, normalization, dimensionality reduction, visualization and identification of identity genes based on differential expression (DE) using a Wilcoxon test by default (Satija *et al.*, 2015). The use of DE analysis methods allows to discover differentially expressed genes (DEGs) that have a significant quantitative change in their expression between different conditions or group of cells (Mou *et al.*, 2020). DEGs found to be uniquely up-regulated in one condition or one cell (sub)type have been used as markers as this property reflects their specificity to characterize the condition or cell (sub)type (Cliff *et al.*, 2004; Squair *et al.*, 2021). In addition, other computational methods relying on different strategies have been implemented such as scMarker that uses information theory principles to identify markers for cell types (Wang *et al.*, 2019).

Existing computational methods have several limitations that hinders the accurate capture of identity genes to characterize cell identity (Figure 3). First, they do not account for the underlying biological complexity of cells classified in a hierarchy composed of cell types,

subtypes and phenotypes. Indeed, the identity genes identification highly relies on the biological environment in which cells are studied accordingly with their hierarchical classification. Nevertheless, these methods identify such genes by performing comparison of gene expression profiles between a target cell population with other cell populations in given tissues (Figure 3). Whereas these tissues do not necessarily contain all representative cell populations, they are also usually composed of a mixture of different cell types, subtypes and phenotypes, which hinders the accurate identification of the target cell population identity genes.

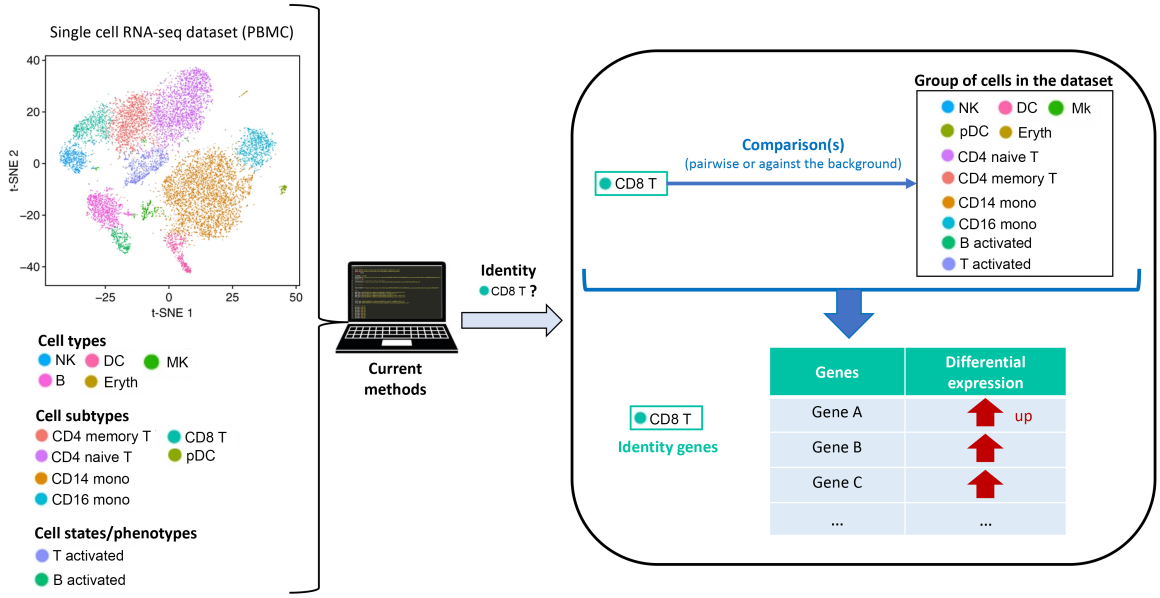

Figure 3. Limitations of current methods to characterize cellular identity.
The t-SNE was modified from (Butler *et al.*, 2018). This figure shows the concept behind current computational method to identify identity genes of a target cell (sub)population. For instance, the identity genes of TCD8+ cells are identified by performing pairwise comparisons with all other group of cells or by performing a comparison against all of them grouped together as a background. From this comparison based on differential expression, identify genes are captured as up-regulated in the target cell (sub)population. Of note, down-regulated genes that could correspond to negative markers can also be identified but are not shown in the figure. NK: natural killer, DC: dendritic cell, eryth: erythrocyte, MK: megakaryocyte, pDC: plasmacytoid DC, mono: monocyte.

Moreover, current methods rely on a Boolean approach of gene expression that identifies whether a gene is expressed or not expressed in a specific cell population. However, it has been shown that the same gene with different levels of expression can lead to different functional outcomes (Bigas and Espinosa, 2012; Shats *et al.*, 2017; Huang, Yang, George W Ye, *et al.*, 2021) and, hence their approach is too stringent to capture such subtle differences. For instance, *E2F1* expression levels were shown to be critical in the control of cell fates, with a low level promoting cell proliferation, an intermediate level driving the mitotic cell cycle arrest and a high one promoting apoptosis (Shats *et al.*, 2017). In addition, it have been

shown that Notch targets and receptors are found at different levels of expression in hematopoietic cell types and impact on their lineages fate (Sandy and Maillard, 2009; Huang, Yang, George W. Ye, *et al.*, 2021).

1.3 Identification of disease-related functional states and genes

1.3.1 Functional cell states identification and its limitations

Cell identity is defined by a set of genes that characterize the specific features, such as specific functions, displayed by the cell. Whereas the functional specialization of cell (sub)types arose during the developmental process, it is further shaped by external signals. Indeed, in response to various stimuli, the same cell (sub)type can exhibit diverse phenotypes defined by specific molecular and functional features, hence corresponding to different functional cell states (Morris *et al.*, 2019; Masuda *et al.*, 2020). A compendium of computational methods based on single-cell transcriptomics data have been implemented in the past few years to identify cell (sub)populations using clustering-based approaches (Andrews and Hemberg, 2018). Clustering methods aim at grouping cells that share similar expression patterns to identify cell (sub)populations that could correspond to cell types, subtypes and/or states. They can be divided into two major categories comprising unsupervised clustering approaches that solely rely on the data and supervised clustering ones that use prior-knowledge to guide de grouping of cells (Abdelaal *et al.*, 2019; Kiselev, Tallulah S Andrews, *et al.*, 2019; Sun *et al.*, 2022). Firstly, a wide range of unsupervised clustering methods have been implement and commonly used such as K-means (Kiselev *et al.*, 2017), hierarchical clustering (Guo *et al.*, 2015), density based clustering (Januzaj *et al.*, 2004) or graph-based clustering (Satija *et al.*, 2015). In that regard, the standard method is the k-means algorithm which identifies k centroids, corresponding to cluster centers, and assign each cell to the closest centroid. In addition, the hierarchical clustering is another widely used algorithm that combines cells into larger groups (agglomerative) or divides group of cells into smaller ones. In particular, Seurat, a state-of-the-art pipeline to analyze single-cell data (Satija *et al.*, 2015), builds a shared-nearest-neighbors graph to connect cells and applies the Louvain community detection algorithm to detect strongly connected communities that corresponds to cluster of cells. These clustering approaches are widely used for the discovery of novel cell (sub)populations, but they often miss cell states displaying subtle changes and require additional analyses to annotate and characterize them. Identified clusters corresponding to cell (sub)types can be manually annotated based on

expert knowledge well-defined markers (X. Zhang *et al.*, 2019) or with the use of computational methods providing systematic approaches (Sun *et al.*, 2022). Notably, iterative clustering approaches have been developed to identify sub-clusters, aiming at a better identification of rare cell subpopulations or states but prone to over-clustering (Miao *et al.*, 2020). Secondly, supervised clustering approaches have been developed to overcome the manual annotation of cells by identifying group of cells based on reference datasets or set of defined markers (A. W. Zhang *et al.*, 2019; Pliner *et al.*, 2019; Lee and Hemberg, 2019). However, supervised approaches are limited to the prior-knowledge provided and are then unable to discover new cell (sub)populations.

1.3.2 Identification of relevant features to characterize cell states

Current computational methods to resolve cellular heterogeneity primarily focus on genome-wide gene expression patterns to identify cluster of cells (Andrews and Hemberg, 2018). However, the use of genome-wide gene expression patterns usually obscured the detection of cell clusters distinguished by subtle differences. To overcome this limitation, diverse approaches were implemented to pre-select the most relevant features to improve the cell partitioning (Xie *et al.*, 2019; Yang *et al.*, 2021). In that regard, one standard feature selection strategy has been the detection of highly variable genes (HVGs), implemented in Seurat (Satija *et al.*, 2015), consisting of genes having the highest variability across cells to leverage the capacity of clustering methods to better account for subtle differences (Yip *et al.*, 2019). Once the cell (sub)clusters are identified, it is usually required to characterize them and interpret their biological meaning (Figure 4) (Kiselev, Tallulah S. Andrews, *et al.*, 2019). One standard strategy consists of the identification of DEGs between the cell (sub)populations identified, followed by functional enrichment analyses to find biological processes and/or pathways over-represented to guide the biological interpretation (Figure 4) (Luecken and Theis, 2019). In particular, gene set enrichment analysis (GSEA) aims at identifying genes significantly enriched in specific annotations of interest to guide the functional interpretation (Reimand *et al.*, 2019). Of note, widely used annotations are the Gene Ontology (Ashburner *et al.*, 2000), composed of biological processes, molecular functions and cellular component, as well as pathways such as KEGG pathways (Wixon, 2001; Kanehisa *et al.*, 2017). Nevertheless, the functional characterization of identified cell (sub)populations, such as cell states, highly relies on the obtained clusters which has been shown to not be always reliable (Andrews and Hemberg, 2018). Therefore, existing

computational methods lack functional relevance when aiming at detecting distinct functional cell states.

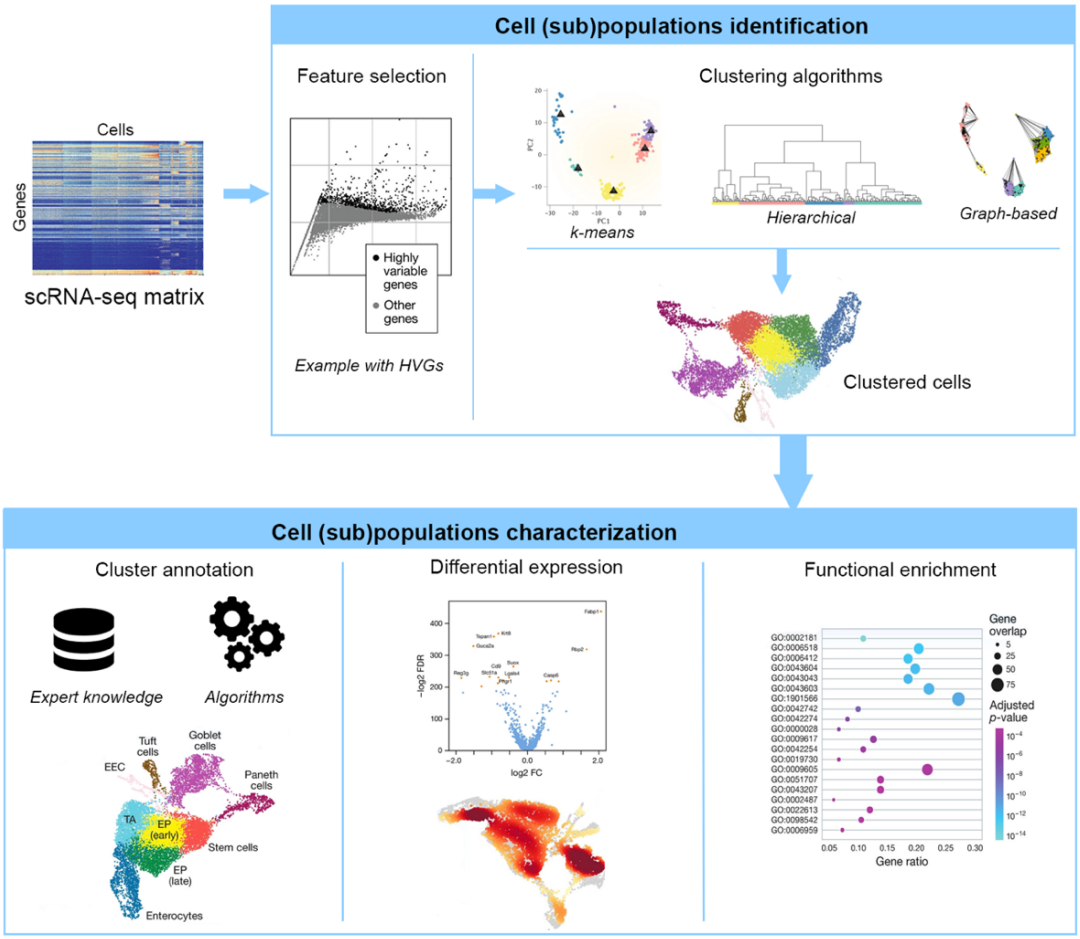

Figure 4. General workflow to identify and characterize cell (sub)populations.
Figure modified from (Luecken and Theis, 2019) and (Kiselev, Tallulah S. Andrews, *et al.*, 2019). The general workflow to identify cluster of cells consists of the feature selection to select the most informative genes to then perform the clustering to identify cell (sub)populations. Then, the characterization of these cell (sub)populations is usually performed by manual or automatic annotation, the identification of differentially expressed genes and functional enrichment analysis.

1.3.3 Cell states conversion

The cell states conversion or transition is a biological process happening in physiological conditions to maintain the organism integrity in response to different stimuli. For instance, neurogenic niches are composed of neuronal stem cells in active or quiescent state to ensure tissue maintenance (Codega *et al.*, 2014). Another example is the transitioning states of immune cells, such as quiescent or active T cells, in response to inflammatory signals (Andreatta *et al.*, 2021; Hua and Thompson, 2001). However, cells can display non-physiological expression patterns in response to stimuli and potentially undergo toward pathologic states due to important dysregulations (Schwartz *et al.*, 2013; Prinz and Priller,

2017). The accurate characterization of cellular states is required to allow the accurate identification of candidate genes that could be used to induce conversion between cell states. Especially, some cell conversion strategies aim at inducing the transition between cell states to promote changes under different conditions, as for instance the conversion of a disease state towards a healthy one. In that regard, several studies aimed at identifying candidate genes that could be used to modulate or convert cellular states to pave the way towards new therapeutic approaches applied to diseases treatment (Kwon and Koh, 2020; Gyun Jee Song, 2017). For instance, it has been shown that the conversion between pancreatic endocrine cell states was a promising strategy to recover the β cell mass for diabetes (Wei *et al.*, 2022). In addition, immunomodulator candidates were identified using a single-cell based approach in the case of acute myeloid leukemia (Guo *et al.*, 2021). Indeed, single-cell transcriptomics based analyses are a valuable strategy to identify candidate genes that could be targeted using a variety of experimental techniques such as viral vectors (Miyamoto *et al.*, 2018) or guide RNA (Liu *et al.*, 2018) to overexpress or repress the specific candidate genes(s). In addition, chemical compounds specifically targeting candidate gene(s) can contribute to the drug discovery and development field (Ebrahimi, 2016; Liu *et al.*, 2016; Li and Ding, 2010). Whereas computational-based predictions are not directly applicable for clinical setups, they provide a valuable guidance for experimental investigation and contribute to the development of novel therapies aiming at preventing or treating diseases.

1.4 Deciphering impaired regulatory mechanisms in diseases

1.4.1 Gene Regulatory Networks inference and limitations

Gene regulation constitutes a fundamental biological process involving mechanisms that activate and repress genes to specify the gene expression profile of cells and hence their identity (Almeida *et al.*, 2021). This process generates diverse gene expression patterns leading to a high cellular heterogeneity for which cells have different sets of proteins to ensure their identity and functionality. It is composed of complex mechanisms in which molecular regulators interact following internal and external signals sent by the (micro)environment (Bahrami and Drabløs, 2016). The gene expression in eukaryotes involves several steps that can be regulated from the DNA availability to the translation in proteins (Wray *et al.*, 2003; Cooper, 2000) (Figure 5). Indeed, a gene can be regulated at each step of the regulatory process. First, the chromatin accessibility can be modulated from a compact to an open structure in order to make enhancer and/or promoter regions of genes

accessible (Klemm *et al.*, 2019). Then, the transcriptional process, a key point of the regulatory process, in which TFs bind to specific parts of the DNA such as enhancer and/or promoter regions to initiate or repress the transcription of a gene (Spitz and Furlong, 2012). Finally, the transcribed RNAs are processed via splicing for which the same pre-mRNA produced can lead to different mRNAs (Nilsen and Graveley, 2010). Then these mRNAs are translated into proteins that can undergo several post-translational modifications that may affect their activity including ubiquitylation, methylation, phosphorylation or acetylation (Wang *et al.*, 2013).

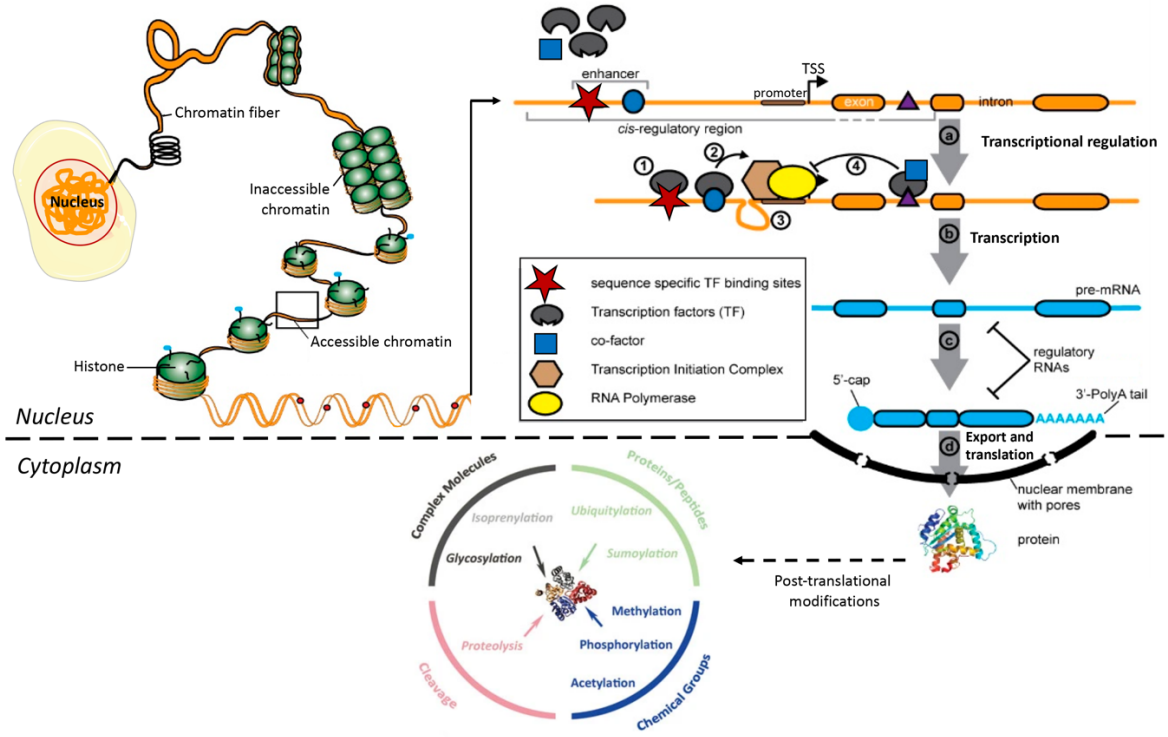

Figure 5. Gene regulation mechanisms in eukaryotes.
Figure modified from (Buchberger et al., 2019), the post-translational modifications picture was taken from (Wang et al., 2013). This figure summarizes the main regulatory mechanisms covering the gene regulation to the set of proteins expressed in the cell. The transcriptional regulation takes place in the nucleus, once the chromatin is open TF binding sites are accessible and TFs can bind to enhancer or promoter regions to enhance or repress the transcription of a specific gene. Once the mRNA is transcribed, it is exported in the cytoplasm to be translated into amino acids and form the protein. Several post-translational modifications such as acetylation or methylation can then modify the protein structure and/or its function.

Reliable and fast inference of large-scale GRNs from transcriptomics data is a long-standing challenge and is crucial for understanding key biological processes such as differentiation and reprogramming (Marbach *et al.*, 2010). Over the years, diverse computational methods have been proposed to infer GRNs from transcriptomics data, especially during the DREAM

challenges (Meyer and Saez-Rodriguez, 2021). These methods aim at reconstructing the GRN that reflects the underlying mechanisms regulating cell expression patterns. Early methods for bulk gene expression data were based on capturing changes in average gene expression profiles as a function of time or perturbations (Sima *et al.*, 2009; Huynh-Thu *et al.*, 2010; Margolin *et al.*, 2006). However, bulk gene expression data can often obscure true biological signals due to averaging of expression over all cells in a given sample (Pratapa *et al.*, 2020; Chen *et al.*, 2019). In this regard, advances in single cell RNA-sequencing has led to development of different kinds of computational methods that leverage the high-resolution gene expression profiling of individual cells and overcomes the major limitations of bulk sequencing (Efremova and Teichmann, 2020). One of the key challenges in GRN inference involves the accurate prediction of regulatory relationships between TFs and their target genes from their expression patterns. Putative regulatory relationships between genes can be detected by simple correlation analysis, or through more advanced measures like mutual information or partial information decomposition that detect statistical dependency between pairs of genes (Chan *et al.*, 2017; Aibar *et al.*, 2017; Nguyen *et al.*, 2021). Notably, SCENIC uses a two steps approach by first using GENIE3, a method using random forests that detects regulatory relationships based on covariation (Huynh-Thu *et al.*, 2010), to infer the regulatory interactions between genes based on scRNA-seq data. It then performs a TF-motif enrichment analysis using RcisTarget to refine the predictions and identify putative TF-targets regulatory interactions (Aibar *et al.*, 2017). In particular, most of the existing GRN inference methods exclusively relies on scRNA-seq data, widely available and rapidly expanding (Chen *et al.*, 2019; Mercatelli *et al.*, 2020).

Computational methods inferring networks from single cell gene expression data still poorly perform (Chen and Mar, 2018; Pratapa *et al.*, 2020). The increasing generation of chromatin immunoprecipitation sequencing (ChIP-seq) data greatly contributed to the understanding of the transcriptional regulatory landscape by providing TF-binding evidence supporting TF-genes regulatory interactions (Mei *et al.*, 2017; Oki *et al.*, 2018). In that regard, GRN inference methods that solely rely on scRNA-seq can predict regulatory interactions among TFs and genes (Wray *et al.*, 2003). However, these methods are not designed to model the direct regulatory interactions involving enhancers, and hence the regulatory mechanistic insights provided remains limited. Indeed, it is well described that genes are regulated in time and space by the interplay between enhancers and promoters to define specific expression patterns (Dao and Spicuglia, 2018). Therefore, another key challenge to model

genes regulation is the prediction of enhancer-promoter regulatory interactions. Particularly, some methods have been implemented to predict regulatory interactions between enhancers and promoters using chromatin physical interactions bulk data such as Hi-C or CTCF ChIP-seq (Hariprakash and Ferrari, 2019; Belokopytova *et al.*, 2020). In addition, valuable resources have been created such as GeneHancer, a comprehensive database reporting known human enhancers and their connected genes (Fishilevich *et al.*, 2017). Notably, computational approaches relying on bulk data are strongly limited regarding their applicability to uncover cell (sub)type or state specific enhancer-promoter regulatory interactions that would require the high-resolution of single cell technologies. In that regard, Cicero, a single-cell cis-regulatory network method relying on scATAC-seq data was developed to exploit the high-resolution of single cell data by identifying co-accessible pairs of DNA elements, and hence connect regulatory elements such as enhancers to their putative target genes (Pliner *et al.*, 2018). Whereas the described methods are valuable to uncover regulatory relationships involving enhancers and key elements impacted in disease conditions (Claringbould and Zaugg, 2021), they partially model the regulatory machinery as the interplay with TFs remains missing. Therefore, GRN inference methods based on single-OMICS data remain limited to model comprehensive regulatory interactions between the key elements involved in gene regulation.

1.4.2 Multi-OMICS approaches for better mechanistic insights

In order to address the mechanistic limitations of GRN inference methods based on single-OMICS approaches, such as scRNA-seq only, strategies based on multi-OMICS data were implemented (Hu *et al.*, 2020). In that regard, combinative or integrative approaches have been developed based on multi-OMICS data over the past few years to account for gene expression and genomics information, such as chromatin accessibility and/or histone modifications (Zarayeneh *et al.*, 2017; Jung *et al.*, 2021). For instance, IRENE, a systematic GRN inference method that integrates diverse OMICS data including gene expression, chromatin accessibility, histone modification, ChIP-seq, and protein–protein interaction data was developed to predict cell-type specific core GRNs (Jung *et al.*, 2021). Whereas efforts have been made in integrating bulk-based multi-OMICS data to predict more comprehensive GRNs that cover a larger part of the complex regulatory machinery, future direction of development should integrate different single-cell layers to more accurately depict regulatory mechanisms underlying disease and biological processes (Figure 6) (Hu *et al.*, 2020). In that regard, scGRNom, a computational pipeline combining bulk and single cell

multi-OMICs, was developed to infer tissue and cell type specific regulatory interactions involving TFs, genes and enhancers using Hi-C data, single cell transcriptomics and/or chromatin accessibility data (Jin *et al.*, 2021).

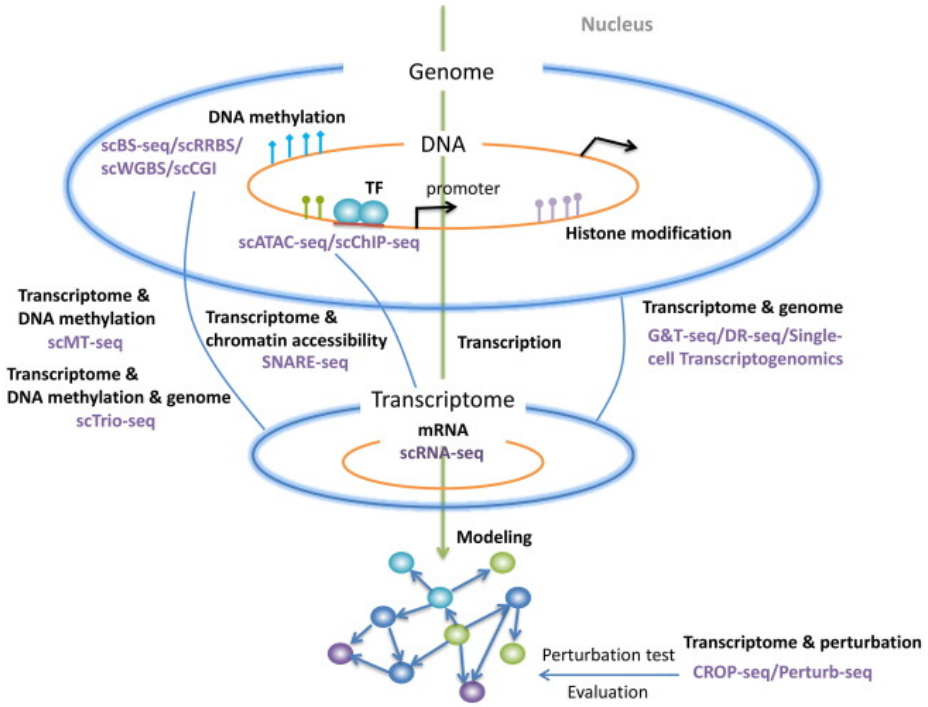

Figure 6. Multi-OMICS integration for more accurate GRNs prediction.
Figure from (Hu *et al.*, 2020). It shows the different single-cell technologies ranging from the DNA to the mRNA that could be used for single cell-based multi-OMICS GRN inference approaches to investigate gene regulatory mechanisms.

Furthermore, the exploitation of single cell modalities to decipher regulatory mechanisms of heterogeneous cell (sub)populations still remains a challenge, mainly due to the lack of single cell sequencing techniques or datasets (Bravo González-Blas *et al.*, 2020). Recently, efforts have been made to integrate single cell layers and provide a more comprehensive understanding of the regulatory mechanisms landscape (Kartha *et al.*, 2021; Boix *et al.*, 2021; Lyu *et al.*, 2021). Notably, EpiMap, a map of the human epigenome has been generated and used to compile a comprehensive view of the human genes regulation across tissues and cell lines that describe gene regulatory regions, their upstream regulators and specific targets (Boix *et al.*, 2021). In addition, the integration of scRNA-seq and scATAC-seq data allowed for the generation of a comprehensive *cis*-regulatory landscape for immune responses across cell types, time and different stimuli (Kartha *et al.*, 2021). Recently, DIRECT-NET, a GRN inference method based on matched scRNA-seq and scATAC-seq has been developed to model the regulatory relationships between key elements involved in

genes regulation including TFs, genes and enhancers (Zhang *et al.*, 2022). Therefore, the future direction to uncover cell heterogeneity at the transcriptional regulatory level will be based on the integration or combination of different single cell layers (Hu *et al.*, 2020) (Figure 6).

1.4.3 Exploiting the gene regulatory network information

GRN models are powerful tools to unveil the fundamentals of cells heterogeneity and functionality (Liu *et al.*, 2019). They provide a guidance for the resolution of several biological and biomedical questions (Emmert-Streib *et al.*, 2014). Indeed, GRNs provide a molecular map that can be used to derive novel hypotheses about these mechanisms and their implications. Standard GRNs are weighted and directed graphs in which source nodes (e.g., TFs) are regulating target nodes (e.g., genes) with a certain degree of confidence or strength (weight) (Aibar *et al.*, 2017). In addition, these graphs can be signed to provide information about the type of interaction which could be an activation or a repression. Graph theory approaches can then be applied to exploit the structure and topology of the graph to identify particularities or features of interest (Koutrouli *et al.*, 2020) (Figure 7).

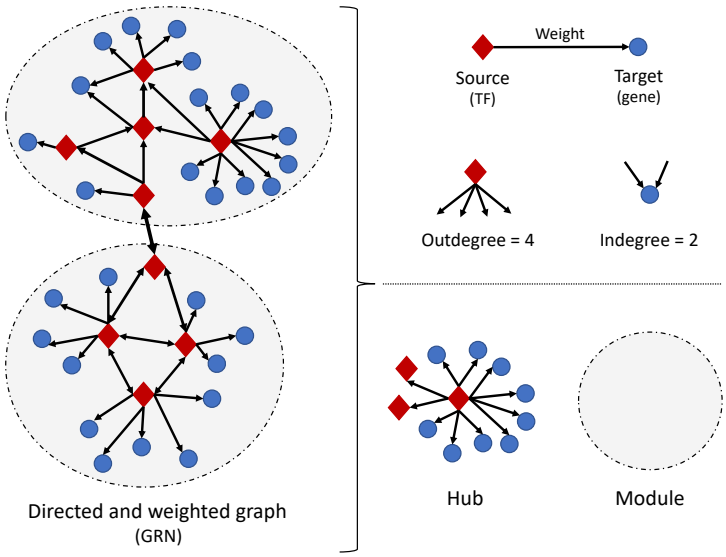

Figure 7. Overview of graph theory to exploit GRNs.
Gene regulatory networks in which TFs are sources (red) and genes are targets (blue). A non-extensive representation of the main graph properties is presented: hubs in which a TF has a high outdegree and modules that contains highly connected nodes.

In particular, highly connected genes, named hubs, have been of particular interest to identify main regulators of a network. They can be identified based on the indegree of a node, corresponding to the number of other nodes regulating it, and its outdegree, corresponding

to the number of targets (Figure 7). Indeed, the main regulators are usually hubs with a high outdegree compared to the other nodes of the network (Wolf *et al.*, 2021). Moreover, regulatory modules can be detected based on the network topology by identifying clusters of nodes highly connected (Song *et al.*, 2017) (Figure 7). These regulatory modules have been shown to provide functional insights into the biological processes involved (Manners *et al.*, 2016). In addition, other topological properties such as specific motifs are of specific interest to understand the regulatory mechanisms involved (Zhang and Zhang, 2013).

GRNs have been widely exploited to identify optimal TFs candidate based on the network topology for cell conversion strategies such as reprogramming into a cell type of interest (Hartmann *et al.*, 2019). Master regulator TFs (MRTFs) can be captured by identifying TFs acting like hubs in the GRN, and hence are the main regulators of a group of genes. These MRTFs have most likely an important effect on the gene regulation if their expression is perturbed, which has been shown as a promising strategy for cell conversions (Wild and Tosh, 2021). Moreover, GRNs have been used to study diseases and provide transcriptional mechanisms insights into the dysregulations involved (Iacono *et al.*, 2019). One standard strategy consists of performing a comparative analysis between the GRN in healthy condition and in disease one (Singh *et al.*, 2018). Such comparison unveils the changes in regulatory interactions and allows to identify the regulations involved in the disease, the genes dysregulated and potential targets for disease treatment (Weighill *et al.*, 2021). For instance, the generation of a GRN around *LRRK2* in PD guided the discovery of *RGS2* as a modulator of *LRRK2* activity that could be used as a therapeutic target to interfere with neurodegeneration in PD patients having the *LRKK2* mutations (Dusonchet *et al.*, 2014).

1.4.4 Identification of dysregulated mechanisms in diseases

The disruption of gene regulation is an important contributor to diseases (Tong Ihn Lee, 2013). Indeed, the impairment in gene expression levels above or below certain thresholds can lead to significant impacts on the phenotype of cells and lead to a wide diversity of diseases (Matharu and Ahituv, 2020). It has been shown that mutations in cis-regulatory elements such as enhancers are key drivers of the alteration of gene regulation in diseases (Epstein, 2009; Claringbould and Zaugg, 2021). In addition, genome-wide association studies (GWAS) have been widely used to discover genomic loci containing SNPs associated with disease-related phenotypes and systematically investigate disease-related molecular mechanisms (Visscher *et al.*, 2017). In particular, the majority of SNPs have been

shown to lie in non-coding regions such as enhancers, for which the regulatory mechanisms remain unresolved (Ward and Kellis, 2012; Claringbould and Zaugg, 2021). Nevertheless, the identification and study of expression quantitative trait loci (eQTLs) uncovered the effect of variants on gene expression to provide a better understanding of their implication in diseases (Nica and Dermitzakis, 2013). Recent efforts were made to provide more insights into the functions of disease variants by building enhancer-gene landscapes for cell types and studying their relationships with SNPs (Kikuchi *et al.*, 2019; Chen *et al.*, 2021; Võsa *et al.*, 2021; Nasser *et al.*, 2021). For instance, genome-wide maps containing millions of enhancer-gene connections were generated to highlight the function of variants related to inflammatory bowel disease (Nasser *et al.*, 2021). Notably, future directions for a more comprehensive view of the complexity of dysregulated mechanisms involved in diseases lie into the use of single cell strategies to dissect the heterogeneity of eQTLs effects across cells, such cell state-dependent eQTLs effects (Nathan *et al.*, 2022).

Data-driven computational methods to infer regulatory interactions in healthy and disease conditions provided a valuable approach to study dysregulations of transcriptional regulatory mechanisms (Emmert-Streib *et al.*, 2014). Indeed, the detailed GRN of disease-relevant cell (sub)types or states is required to translate risk-variants into mechanistic insights (Chiou *et al.*, 2021). Moreover, the fine-mapping of SNPs to regulatory networks has been used to aid the discovery of core disease genes and downstream impacts to provide cell type and disease specific insights (Broekema *et al.*, 2020). In addition, the prioritization of SNPs falling into regulatory regions has been widely performed using computational analysis of the TF binding sites or motifs (Maurano *et al.*, 2015; Broekema *et al.*, 2020). Whereas cell-type specific impairment has been widely studied over the years (Watanabe *et al.*, 2019; Doostparast Torshizi *et al.*, 2020; Bryois *et al.*, 2021; Wong *et al.*, 2021), the specificity of cell subtype and state impairment as well as their implication in diseases remains undetermined (Figure 8). In that regard, a recent study uncovered for the first time a specific subpopulation of dopaminergic neurons that selectively degenerate in PD, demonstrating the importance of dissecting cell (sub)populations heterogeneity to discover specific impairment in diseases (Kamath *et al.*, 2022). In addition, a method to characterize complex trait and disease relevant genetic associations has been implemented to study cell (sub)types, states and trajectories, hence putting emphasis on the importance to leverage the single-cell resolution to unveil the complexity of diseases (Yu *et al.*, 2022).

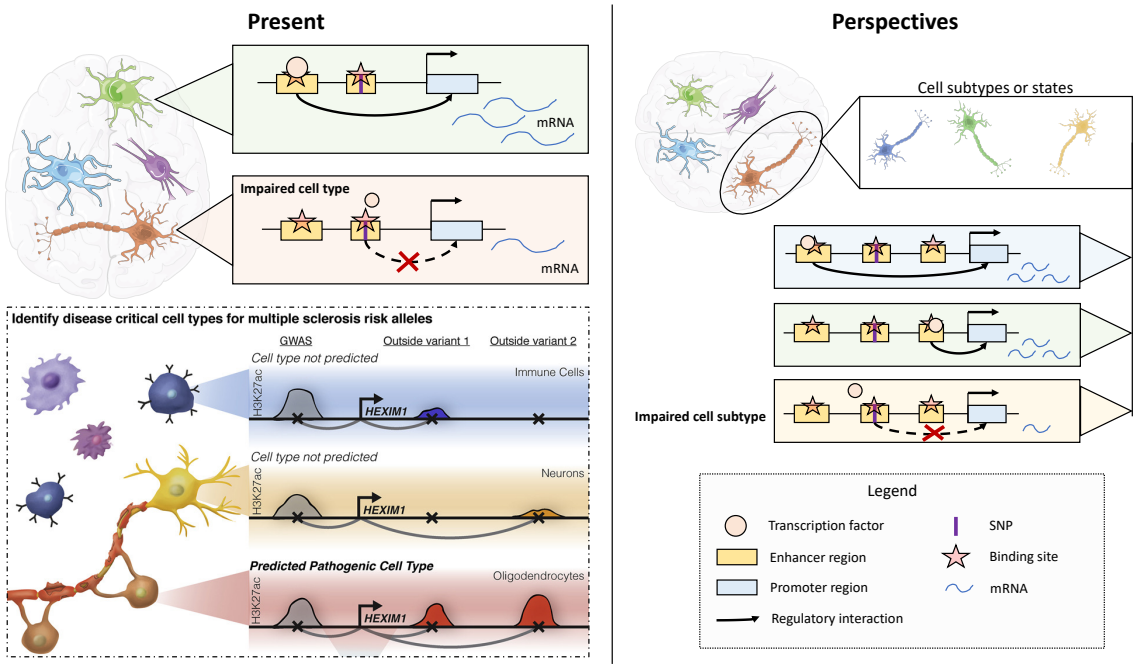

Figure 8. Perspectives to uncover cell (sub)populations specific impairment in diseases. Bottom left picture was taken from (Factor *et al.*, 2020). This figure summarizes the current studies focusing on uncovering cell-type specific impairment or involvement in specific diseases (left panel) and recent studies that started to focus on the characterization of cell subtypes or states specific impairment (right panel).

2 Scope and aims of thesis

2.1 Scope

This PhD thesis is focused on disease modelling using single cell-based computational approaches to pave the way towards the development of new therapeutic strategies. Computational biology applied to disease modelling is an active research field of systems biology aiming at developing computational methods to study and understand diseases. In that regard, the aspects presented in this thesis include the characterization of cell identity, the dissection of cellular heterogeneity, the identification of potential candidate molecules for cellular conversion, and the capture of impaired regulatory mechanisms in diseases. The emergence and fast development of single-cell based technologies allowed an unprecedented resolution of the cell features measurement and highly contributed to the development of novel computational methods that tackle different aspects of disease modelling. However, despites several efforts that have been made in the past few years, several challenges remain to be solved and many limitations need to be overcome.

2.2 Aims

This PhD project aimed at developing three computational methods to address different aspects of disease modelling to guide the development of novel therapeutic strategies. Taken together, these methods aim at contributing to a better understanding and characterization of cellular heterogeneity in physiological and pathological conditions, and at providing additional insights into impaired regulatory mechanisms in diseases.

Aim 1: **Development of a computational method to characterize cell identity and accurately capture identity genes.** In this study, we focused on the characterization of cell identity by capturing identity genes for any cell type, subtype and phenotype. We aimed at overcoming the main limitations of current methods that do not account for the underlying biological complexity of cells (e.g., mixture of cell types, subtypes and phenotypes) and also drastically categorize genes as being expressed or not, hence discarding genes at medium level of expression. Moreover, several studies have shown that the levels of expression of genes can lead to different functional outcome of the cells. Therefore, we sought at developing a computational method that can be applied to accurately characterize the identity of any cell type, subtype or phenotype. In addition, we aimed at generating high-resolution cell identity atlases to complete the existent knowledge and provide a comprehensive identity landscape that could be used to uncover cells displaying non-physiological features.

Aim 2: **Implementation of a computational method to dissect functional heterogeneity and decipher the key genes characterizing the functional cell states identified.** In this research study, we focused on the development of a systematic approach to dissect functional heterogeneity across cells in physiological and non-physiological conditions. We aimed to overcome the current limitations of computational methods that do not put emphasis on the subtle differences between cell states and do not provide an accurate functional characterization. Therefore, we sought at developing a method to accurately decipher functional cell states, the genes characterizing them as well as providing insights into the biological processes in which they are involved. In addition, we aimed at dissecting the functional heterogeneity of mouse immune cells in different type of infections to generate a large-scale catalog of candidate immunomodulators to pave the way towards the development of novel immunotherapies strategies.

Aim 3: **Development of a computational pipeline to infer comprehensive cell (sub)types specific GRNs and identify impaired regulatory mechanisms due to SNPs in diseases.** With this study, we sought at developing a multi-OMICS pipeline to infer comprehensive cell (sub)type or state specific GRNs and systematically identify impaired regulatory mechanisms due to disease-related SNPs. Notably, we focused on the inference of GRNs describing the regulatory interactions mediated by TFs and enhancers of regulated genes by combining scRNA-seq, scATAC-seq, ChIP-seq and prior-knowledge data. We aimed at providing the scientific community with a user-friendly pipeline that exploit the GRN information to identify cell (sub)type or state specific regulatory interactions impaired by disease-related SNPs to provide better regulatory mechanistic insights for the disease.

2.3 Originality

The three computational methods presented in this thesis are addressing different challenges of disease modelling. They were developed to overcome the main limitations of existing approaches to characterize cell identity, decipher functional cell states, and infer comprehensive cell (sub)type or state GRNs to identify impaired regulatory interactions due to SNPs in diseases. Therefore, these methods and the findings of this PhD thesis are a valuable resource to have a better understanding of the cellular heterogeneity and complexity in physiological and pathological conditions. Furthermore, these outcomes provide a guidance for the development of novel therapeutic strategies.

3 Materials and methods

Materials and methods details are presented in the results section of this thesis for the three manuscripts (sections 4.1 to 4.3). A brief summary is described below for each of them.

3.1 Characterization of cell identity

In *"Quantification of gene level to characterize hierarchical cell identity"* (section 4.1), we developed HCellig, a hierarchical cell identity-based computational method that quantifies gene expression into three levels, including low, medium and high, to capture identity genes of any cell type, subtype and phenotype. The implemented hierarchical cell identity model was composed of three hierarchical layers including cell type, subtype and phenotype. These three layers were used as a reference background to quantify the gene expression levels and capture identity genes of a target cell (sub)population. Notably, the gene expression quantification strategy was a single-cell based implementation of RefBool that uses a reference background dataset to quantify genes in three levels of expression including low, medium and high, using bulk RNA-seq data (Jung *et al.*, 2017). Briefly, the reference background datasets were generated by first cleaning and normalizing the data using scTransform (Hafemeister and Satija, 2019), and then scaling each gene using its maximum value. Then, a bootstrapping approach was used to sample the background so that each cell population was equally represented to derive lower and upper thresholds distributions for each gene by solving an optimization problem. Finally, we identified genes displaying a bimodal pattern by using kernel density to compute the number of modes of the distributions (Statisticat, LLC., 2021) and selected genes having a distribution with two modes. Using a query cell (sub)population and the appropriate background dataset (e.g., cell type, subtype or phenotype), genes from the query were first normalized and scaled using the background information to make them comparable. Then, gene expression levels were quantified into a discretized matrix by computing p-values and q-values for each gene using the background thresholds distribution to categorize them as significantly low, medium or high. Notably, genes for which no significance could be determined were classified as non-significant, as the quantification level could not be determined with confidence. The general expression level of each gene in the query (sub)population was determined by first identifying genes significantly not expressed, and then distinguishing medium and high level of expression by computing their frequency across cells. Finally, identity genes of the query cell (sub)population were identified as genes being expressed with a high level and medium

level, under the condition that the gene was displaying a bimodal pattern in the background. Indeed, a gene expressed at a medium level in the query cell (sub)population and having a bimodal pattern in the background reflects a unique medium pattern for the cell (sub)population.

We compiled reference background datasets for each hierarchical layer using the Tabula Muris (The Tabula Muris Consortium *et al.*, 2018) and Tabula Sapiens datasets (The Tabula Sapiens Consortium and Quake, 2021) by manually curating the annotations and classifying them as cell type, subtype and phenotype. We then applied HCellig to all available cell types, subtypes and phenotypes to generate a high-resolution cell identity atlas for mouse and human independently. In addition, we applied our method to the mouse neuronal landscape (La Manno *et al.*, 2021) to characterize the identity of neurons, neuron subtypes and neuron phenotypes, corresponding to neuron subtypes located in different brain regions. Finally, we performed functional enrichment analyses (Wu *et al.*, 2021) and found extensive literature-based evidences to showcase the functional relevance of the captured identity genes.

3.2 Identification of disease-related functional states and genes

In *"A Catalogus Immune Muris of the mouse immune responses to diverse pathogens"* (section 4.2), we developed FunPart, a network-based method combined with a recursive hierarchical clustering approach to decipher functional cell states and capture the functional gene modules characterizing them. Briefly, FunPart was based on a recursive feature selecting strategy combined with a hierarchical clustering approach to first select the most relevant set of genes and then perform the binary clustering of the cells. The identification of the functional gene modules was done by first building a correlation network between genes while keeping significant edges. Then, cliques of genes positively correlated together were identified and, pairs of cliques negatively correlated were selected. Finally, a functional enrichment (Wu *et al.*, 2021) for each pair of cliques was performed and candidate pairs of cliques were selected if both cliques, named modules of genes, were found enriched in specialized biological processes. The best candidate was then identified by ranking the modules of genes by their enrichment score, and it was then used to perform the binary splitting of the cells using the hierarchical clustering approach. This strategy was performed recursively over the groups of cells until no more module of functional genes was identified, in which case the algorithm stopped as it had reached to functional homogeneity of the identified cell states. The output of the method was composed of all the functional cell states

identified, the module of genes characterizing each of them, and the biological processes in which these gene modules were enriched.

FunPart was applied to a large compendium of mouse single-cell RNA-seq datasets composed of six immune cell types in the context of twelve different pathogens including virus, bacteria, fungi and parasites. A large atlas of immune functional states and key functional genes was generated from this analysis. Moreover, we compared FunPart performances to Seurat (Stuart *et al.*, 2019), using the default parameters, and assessed their ability to identify homogeneous functional cell states. Finally, we performed an experimental validation of two TFs (*Stat1* and *Zfp597*) belonging to gene modules negatively correlated and characterizing two distinct functional states of macrophages under *Salmonella* infection. The experiments were performed using shRNAs (Moore *et al.*, 2010) to silence the two TFs independently, and the survival of *Salmonella* in each condition was assessed based on the hypothesis made from the analysis of the two macrophage states.

3.3 Deciphering impaired regulatory mechanisms in diseases

In "*RNetDys: regulatory network inference to identify impaired interactions in diseases*" (section 4.3), we developed RNetDys, a multi-OMICS pipeline to infer comprehensive cell (sub)type or state specific GRNs and systematically identify impaired regulatory mechanisms due to SNPs in diseases. Notably, the combination of scRNA-seq, scATAC-seq, ChIP-seq and prior-knowledge data allowed the inference of GRNs describing the regulatory interactions mediated by TFs and enhancers of regulated genes. The pipeline was divided into two main parts composed of (i) the cell (sub)type or state specific GRN inference in healthy condition and (ii) the contextualization towards the disease state to identify impaired regulatory interactions due to SNPs. First, the GRN inference was performed by inferring TF-gene regulatory interactions using scRNA-seq data to select genes conserved at least in 50% of the cells from the cell (sub)type, and open promoter regions as well as binding TFs were identified by intersecting scATAC-seq and ChIP-seq data (Oki *et al.*, 2018). Then, we inferred the enhancer-promoter interactions by computing a scATAC-seq peak correlation analysis to identify significant correlations between promoter and enhancer regions that were then intersected with the GeneHancer database (Fishilevich *et al.*, 2017). We then predicted TF-enhancers interactions by intersecting the ChIP-seq and scATAC-seq data. Finally, each regulatory interaction of the GRN was signed based on correlation to distinguish activation from repression. Second, using the cell

(sub)type specific GRN, we contextualized the network towards the disease state and identified candidate impaired interactions by mapping the SNPs to TF binding sites of enhancer and promoter regions. We then performed a TF binding affinity analysis for TFs involved in the impaired interactions. The list of impaired regulatory interactions was then refined by selecting the ones involving at least one TF with impaired affinity. Finally, we ranked the TF regulators mediating the regulatory impairment by their degree of importance using the network topology, their impaired binding affinity score and the MAF score of each SNP involved.

We assessed the performances of RNetDys in inferring cell (sub)type specific GRNs by benchmarking two types of interactions. First, we benchmarked the capacity of our approach to accurately predict cell (sub)type specific TF-gene regulatory interactions, and compared its performances to state-of-the-art methods including ppcor (Kim, 2015), CLR (Faith *et al.*, 2007), GENIE3 (Huynh-Thu *et al.*, 2010), PIDC (Chan *et al.*, 2017) and SCENIC (Aibar *et al.*, 2017). Then, we assessed the accuracy of RNetDys to accurately predict cell (sub)type specific enhancer-promoter regulatory interactions compared to Cicero (Pliner *et al.*, 2018). Of note, state-of-the-art GRN inference methods solely relying on scRNA-seq did not infer such interactions and hence were not included in this comparison. The precision (PPV) and F1-score were used to assess the performances of each method, using human cell line specific ChIP-seq GS for the TF-gene interactions and pcHi-C GS for the enhancer-promoter ones. Finally, we applied RNetDys to infer cell (sub)type specific GRNs from human brain and pancreas tissues and identify impaired regulatory interactions due to disease-related SNPs. We collected disease-related 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). We validated the relevance of the predicted impaired regulatory interactions using GWAS, eQTL and literature-based evidences.

4 Results

4.1 Characterization of cell identity

4.1.1 Preface

In this study entitled "*Quantifying gene expression to characterize hierarchical cell identity*", we tackled one of the main challenges in systems biology consisting of the accurate characterization of cell identity. Current methods can identify identity genes by comparing gene expression of mixed cell (sub)populations, usually composed of different cell types, subtypes and phenotypes, which hinders the accurate characterization of the target cell (sub)population identity. Moreover, they do not distinguish between genes expressed at high or medium levels, important to accurately determine the cell functional identity as shown in several studies. To overcome these limitations, we present HCellig, a computational method that relies on the hierarchical organization of cell identity in three hierarchical layers including cell type, subtype and phenotype, to quantify gene expression levels into low, medium or high and capture identity genes.

We made two novel contributions with this study. First, we developed a computational method to accurately capture identity genes and pre-compiled large-scale hierarchical background datasets to allow the identification of identity genes for any cell type, subtype or phenotype. Second, using HCellig we generated a high-resolution identity atlas composed of several cell types, subtypes and phenotypes for mouse and human. Furthermore, we showed the functional relevance of the captured identity genes and their importance for the cell identity. Finally, we applied HCellig to study the mouse neuronal landscape identity and highlighted the brain-region dependence of some identity genes, especially the ones expressed at a medium level. In summary, this study generated a high-resolution characterization of cell identity, while putting emphasis on the importance of the identity genes expressed at a medium level, which have been poorly studied so far. Moreover, HCellig will be of great use to pave the way towards a more accurate characterization of cell identity, especially with the ongoing generation of new organism-wide scRNA-seq data.

Contribution: I implemented the computational method, performed the stability analysis of the algorithm, collected and pre-processed the human data, manually curated the annotations for the mouse, human and brain datasets, compiled the hierarchical backgrounds, generated the hierarchical cell identity atlases for mouse and human, and wrote the manuscript.

4.1.2 Manuscript

Quantifying gene expression to characterize hierarchical cell identity

Céline Barlier[1], Kartikeya Singh[1], Sascha Jung[2], Antonio del Sol[1,2,3]

1 Computational Biology Group, Luxembourg Centre for Systems Biomedicine (LCSB), University of Luxembourg, L-4362 Esch-sur-Alzette, Luxembourg

2 Computational Biology Group, CIC bioGUNE-BRTA (Basque Research and Technology Alliance), Derio, Spain 48160

3 Ikerbasque, Basque Foundation for Science, Bilbao, Bizkaia, 48012. Spain

Abstract

Cellular identity, which reflects functional specialization of cells, depends on the biological context. In this regard, we present a hierarchical cell identity-based method which considers three hierarchical layers (cell types, subtypes and phenotypes), that quantifies different levels of gene expression to accurately identify cell identity genes (HCellig). We generated a high-resolution cell identity atlas for human and mouse, and showed the functional relevance of identity genes expressed at different levels.

Main text

Multicellular organisms are composed of diverse cells which display different morphologies and specialized functions depending on their hierarchical classification of cell type, subtype or phenotype. The characterization of cellular identity, including the identity genes remains to be a central challenge in biology. Indeed, cell identity is initially acquired during development and then shaped by the micro-environment to perform specific functions[1,2]. The identification of identity genes remains a challenge as it highly relies on the biological context in which cells are characterized according to their hierarchical classification as cell types, subtypes or phenotypes[3]. Current computational methods are able to identify identity genes by comparing gene expression profiles of the target cell population with other cell populations in given tissues[4,5]. Nevertheless, these tissue samples do not always contain all representative cell populations and are usually composed of a mixture of different cell types, subtypes and phenotypes, which hinders the accurate identification of the target cell population identity genes. For example, in order to identify the cell subtype identity genes of a dopaminergic neuron the proper comparison should be carried out with respect to the complete set of other neuronal subtypes (e.g., glutamatergic, serotoninergic) without including non-neuronal (sub)types. In addition, these methods categorize genes as expressed

or non-expressed and hence discard genes displaying medium expression level. However, it has been shown that varying expression levels can lead to diverse cellular functions[6,7].

Here, we present HCellig, a hierarchical cell identity-based method, which considers three hierarchical layers (cell types, subtypes and phenotypes), to quantify gene expression levels for the detection of the identity genes at each of these hierarchical layers (Fig. 1A, Supplementary Fig. S1). In particular, the quantification of gene expression considers three levels (low, medium and high) to characterize cell identity genes. Moreover, the cell type level is composed of all cell types of the organism, whereas the subtype level comprises every subtype of a cell type, and the phenotype level considers the tissue type or condition specificity for a particular cell subtype (Fig. 1A). For this purpose, we built an extensive background database for each hierarchical layer by compiling scRNA-seq data for both mouse[8] and human[9] cell atlases (Supplementary Tables S1, S2). In order to extend the current knowledge of cell identity, we applied HCellig on a large-scale repository of cell types, subtypes, and phenotypes to generate a high-resolution cell identity atlas for each organism (Fig. 1B, Supplementary Fig. S3 and Tables S3, S4). Clustering cell types based on their identity genes showed that cell types which belong to the same broad categories were grouped together, reflecting their functional similarity (Fig. 1B, Supplementary Fig. S2). Furthermore, HCellig identified known markers for specific cell types and cell subtypes. For example, as expected the high expression of CD3 was able to identify T cells in both human and mouse (Supplementary Table S5). Moreover, CD8 expressing T cells were distinguishable as either cytotoxic or memory based on the medium expression of identity genes for these subtypes including *TBX21, STK10* and *ZBTB7A*, which have not yet been fully elucidated (Fig. 1C, Supplementary Table S4). We then showed the functional relevance of detected identity genes, particularly those expressed at a medium level. In this regard, an enrichment analysis indicated that these identity genes participate in biological processes which are consistent with the cell (sub)type function. For instance, identity genes of monocytes were enriched for immune-related processes, such as phagocytosis and regulation of effector immune processes[10] (Supplementary Fig. S4 and Table S6). In addition, we observed that T cell identity genes expressed at the medium level were enriched in functions involved in immune responses such as cytokine production and response to stimulus, supporting their functional relevance[11] (Fig. 1C). In particular, we found that the identity gene *FYN* is typically expressed at the medium level amongst T cells, however amongst specific subtypes the expression level is high (Supplementary Tables S4, S7).

Indeed, some studies have shown that the expression levels of *FYN* have an impact on T cell activation signals[12,13], supporting the accuracy of our method to find functionally relevant cell identity genes. However, we noticed that most of the genes expressed at the medium level were poorly studied for the cell subtype-specific context (Supplementary Table S7).

Finally, we applied HCellig to perform a case study of the mouse neuronal landscape (Fig. 2A, Supplementary Table S8). At the cell type level, we identified neuron identity genes expressed at a medium level known to be involved in neuronal functions (Supplementary Table S9). Moreover, we found that some identity genes ought to be expressed at the medium level, and changes in their expression have been shown to lead to dysregulation of the neuronal functionality (Supplementary Table S9). For instance, low and high expression levels of the gene *Clasp2* dysregulate the neuronal polarity and synaptic function, suggesting its importance as an identity gene expressed at the medium level for neurons[14]. Similarly, Dscaml1 low and high expression levels lead to impaired migration or self-avoidance defects in neurons[15] (Supplementary Table S9). Moreover, we studied the variation of gene levels across the different neuronal subtypes (Fig. 2B, Supplementary Fig. S4). Interestingly, we found *Pbx1* to be an identity gene highly expressed in dopaminergic neurons, while it was not an identity gene for other neurotransmitter neurons but found at a medium level (Fig. 2B, Supplementary Fig. S4). Indeed, *Pbx1* is known to be involved in maturation of dopaminergic neurons from neuroblasts and represses other cell fates, supporting its high level in dopaminergic neurons[16]. Finally, we observed variation of identity genes expressed at the medium level specific for subtypes across brain regions. This shows the region-specificity of these medium expressed genes for the neuronal subtypes (Fig. 2C, Supplementary Fig. S5).

In conclusion, we developed HCellig, a hierarchical cell identity-based method, that quantifies gene expression in low, medium and high levels, to capture cell identity genes for any cell type, subtype and phenotype. We generated a cell identity atlas for mouse and human, the latter being more complete due to the extensiveness of the data. Moreover, we provided background datasets that can be used to quantify gene expression levels to detect identity genes of any new cell type, subtype or phenotype. Furthermore, we showed the functional relevance of deciphered identity genes, highlighting the lack of knowledge for those medium identity genes. Indeed, studies cataloging cells have been only focusing on highly or lowly expressed genes[1,5], hence missing valuable data when it comes to genes

expressed at a medium level that could lead to important functional differences[6,7]. In summary, we expect the cell identity atlases, as well as the implemented method HCellig, to be of great use to pave the way towards a more accurate characterization of functional cellular identity.

Online Method

HCellig workflow

We developed HCellig, a hierarchical cell identity-based method, that quantifies gene expression in low, medium and high levels, to capture cell identity genes for any cell type, subtype and phenotype at each hierarchical level. The quantification strategy is based on RefBool[17] and was adapted for single-cell UMI data. HCellig takes as an input the single cell UMI matrix of a cell type, subtype or phenotype and the background data based on the hierarchical level. The quantification strategy is composed of two main parts: (1) the construction of a background at one of the three hierarchical levels, (2) the gene quantification of a query cell type, cell subtype or phenotype. In fact, the background construction is optional as we built pre-compiled backgrounds for mouse and human, generated in this study, to quantify gene expression levels and identify identity genes of a query cell population.

Background construction

The compilation of a background was the first step of the method and was later used to quantify gene expression of a query cell population, including cell type, cell subtype and phenotype. The strategy to build the background at the cell type level was different from the one at the cell subtype and phenotype level as it included two layers of information to process the thresholds. Regardless of the level of hierarchy considered, the first step was to clean the data. We removed all genes expressed in less than 10 cells, cells with no gene expression, as well as low quality cells using the strategy provided in the *Scuttle* R package. In addition, cell populations with less than 50 cells were removed from further analyses to assess the threshold distribution properly. Then, the single cell matrix was normalized using scTransform[18], which considers the batch correction which is a necessary step especially for cell type backgrounds that might contain different datasets. In fact, normalization factors used for each gene were saved as a component of the background that were used for gene quantification. After normalization, the general scaling factors were retrieved and saved for each gene as being the maximum gene expression value found in the normalized data. Then,

the next step consisted into the generation of lower and upper-threshold distributions using optimization functions combined with a bootstrapping approach:

1. Sampling of the data: this step differed between the different hierarchical levels. In the case of a cell type background, two layers of information composed by the tissue and cell type were used, whereas for the subtype or phenotype level, only the cell population was used to sample the data. For each gene, cells with non-zero values were kept under the condition that at least 10 cells are found. If less than 10 cells with non-zero value are identified, the gene was removed from the analysis. This step aimed to represent equally each cell population in the background. Therefore, 100 cells from each population were sampled with replacement from the normalized data.
2. Scaling of the sampled data: each gene was scaled using its maximum expression value to have a range from 0 to 1, required for the threshold computation.
3. Lower and upper-thresholds identification: for each gene, thresholds were derived using two step functions by solving the optimization problem of maximizing the area over the step function for the left tail and maximizing the area under the step function for the right tail of the empirical cumulative distribution function[17] (Supplementary Fig. S1).

The computation of 1000 bootstraps of the three previous steps resulted into the background threshold distributions, composed of upper and lower thresholds for each gene. Finally, genes with bimodal expression pattern were identified by calculating the number of modes of the distribution based on kernel density using the *LaplacesDemon* R package as follow:

1. We performed 100 bootstraps of background sampling and scaling to obtain the background distribution of each gene.
2. The number of modes was computed. If the gene distribution had 2 modes, it was considered to be bimodal (*is.bimodal()* function of the R package).

These two steps were repeated three times and bimodal genes consistently identified were considered as being truly bimodal in the background. In summary, the background construction provided for each gene the following outputs that are used for the gene level quantification and gene identity identification including: (1) the normalization factor, (2) the scaling factor, (3) the thresholds distribution and (4) if the gene was bimodal.

Gene level quantification

HCellig quantified the gene expression of a query cell population (cell type, cell subtype or phenotype) into three levels of expression: low, medium and high compared to a specific

background. As for the background construction, data was cleaned using the same strategy, each gene was normalized and scaled using the corresponding factors from the background. Then, based on the derived threshold distribution described previously and the genes expression of the query cell population, the method computed p-values for each gene to determine if the gene was significantly high or not by performing a one-sided test against the null hypothesis for which the gene is not significantly high. In addition, q-values were derived and a Benjamini Hochberg multiple testing correction was performed. Furthermore, HCellig identified an intermediate level of expression defined as values significantly greater than the lower-thresholds and significantly lower than the higher thresholds. The output was a discretized matrix of quantified gene levels consisting of low, medium, high expression levels (Supplementary Fig. S1). Genes with non-significant p-values (default to $p < 0.05$) were assigned as non-significant, as the quantification level cannot be determined with confidence. Finally, HCellig identified the overall level of expression of each gene in the query cell population. Based on the generated discretized matrix, for each gene we calculated the frequency of the three levels. Then, we computed a Z2-score threshold based on a binomial distribution to distinguish significantly expressed genes (high or medium) from the ones lowly expressed. Finally, high and medium levels of expression were distinguished based on their maximum frequency in the query cell population.

Identification of identity genes

HCellig identified the identity genes of the cell population (cell type, cell subtype or phenotype). We defined as identity genes those highly expressed genes as well as genes expressed at a medium level with a bimodal pattern in the background. Indeed, if a gene was found to be at a medium expression level in the cell population while it displayed a bimodal pattern in the background, we considered this gene to be part of the identity.

Pre-compiled backgrounds construction

We collected organism-wide UMI data for Human, from the Tabula Sapiens study[9], and Mouse, from the Tabula Muris study[8]. We performed a manual curation of the original metadata to classify cell types and cell subtypes accordingly for each tissue provided. In addition, due to the large size of the Tabula Sapiens data, we performed a downsampling while limiting the loss of information as follows: each cell subtype of each cell type of each tissue was downsampled to 200 cells. In case the number of cells was smaller than 200 cells, we kept the original number of cells for the specific cell subtype. The downsampling of the

Tabula Sapiens allowed us to have about 62 200 cells compared to about 450 000 cells from the original data, while conserving all the cell (sub)types across tissues. For each dataset, a quality control was performed: all non-expressed genes were removed and cells containing less than 2500 UMIs were removed from further analyses. We applied HCellig to pre-compile backgrounds for both organisms at each hierarchical level: cell type, cell subtype and phenotype. In fact, we could not build phenotype backgrounds for Mouse as no subtypes with at least 50 cells could be found in at least two different tissues. The different backgrounds were processed as follow for each organism:

- Cell type backgrounds: the tissue and cell type information were provided to the algorithm, hence allowing a sampling based on two layers of information to build the background thresholds.
- Cell subtype backgrounds: for each cell type independently, the cell subtypes were provided to the algorithm to build the background thresholds.
- Phenotype backgrounds: for each cell subtype independently, the specific tissues in which the cell subtype could be found were used in order to compile the background thresholds.

Large-scale repository of hierarchical cell identity for mouse and human

We built a large-scale identity atlas for mouse and human by quantifying gene expression and identifying key identity genes for each cell population available including cell types, cell subtypes and phenotypes. Indeed, few subtypes and no phenotypes for Mouse could be included due to the limited annotations and number of cells in the dataset. The gene quantification and key identity genes identification was performed for each cell population with at least 10 cells accordingly with the hierarchical cell identity model: (1) the cell type background was used for every cell types, (2) cell subtype backgrounds were used accordingly with the subtype considered (e.g., the T cell background was used for T regulatory cells) and (3) human phenotype backgrounds were selected in accordance with the phenotype to analyze (e.g., the classical monocyte background was used for the classical monocytes of the spleen). Furthermore, as the original Human data provides Ensembl IDs, we identified their corresponding Gene Symbol using the *AnnotationDbi* R package with the *org.Hs.eg.db* database version 3.12.0 and reported them accordingly in Supplementary Table S4. The hierarchical UMAPs were generated by computing for each corresponding level a discrete matrix with all identity genes identified in rows, the corresponding cell populations

in columns and the gene expression levels as value. In the discretized matrix, high level was equal to 1, medium level to 0.5, low level to 0 and non-significant or not found in the cell population to -0.5. For the cell type level UMAP visualization, cell types were grouped by broad categories displayed with different shapes whereas the different cell types were shown with gradient of colors.

Stability assessment of HCellig

We validated the stability of HCellig when different backgrounds were created using the same dataset as input (Supplementary Fig. S6). Indeed, as the method relies on a bootstrapping approach, we ensured its stability across runs and performed 10 runs using available mouse and human data at the subtype level (Supplementary Tables S1, S2). First, we assessed the stability of the background thresholds using a Kolmogorov-Smirnov test between the threshold distributions of each run across all genes for each subtype level background, by considering the thresholds to be stable if p-value < 0.05. Then, we computed the stability of the p-values, used to quantify the gene expression, using a Wilcoxon test between the p-values of each run for each cell subtype across all genes. The p-values thresholds were considered stable if p-value < 0.05. Finally, we verified the stability of the gene quantification into three expression levels by comparing for each cell subtype across each run (1) the significant genes identified and (2) their level of expression. The stability was measured by computing a ratio of common predictions versus all predictions, between pairwise runs, with a ratio of 1 reflecting a stability of 100%. Finally, we assessed the average stability of bimodal genes identified across the 10 runs for each subtype background by computing the ratio of common bimodal genes versus all bimodal genes between pairwise runs and computed the average by calculating the median value for each subtype background.

Literature-based validation of the generated repository of identity genes

We ensured that in our large-scale cell identity atlases, HCellig captured experimentally reported cell type and cell subtypes markers. Of note, phenotypes were discarded due to the lack of cell subtype tissue specific available information in literature. We used the Cell Marker database[19] as a reference to collect experimentally validated markers in normal conditions and reported the ones matching with HCellig identity genes identified for each cell type and cell subtype, for mouse and human. In addition, we performed a literature search of medium identity genes, guided by the functional enrichment, for human T-cells

and four well studied subtypes (Treg, TCD4+ memory, TCD8+ memory, T cytotoxic) to support their functional relevance.

Functional enrichment of large-scale cell identity genes

We validated that the cell identity genes captured at each hierarchical level reflected the functional features of the specific cell types, cell subtypes and phenotypes analyzed. In order to validate the relevance of the captured identity genes, we performed a functional enrichment for every cell population on all identity genes and independently only on mediumly expressed identity genes. We used the *ClusterProfiler* R package[20] to carry out an over-representation analysis of biological processes (BP). The universe was respectively set to all sequenced genes for mouse or human, BP categories tested were limited to categories containing 5 to 500 genes hence removing broad processes, a Benjamini-Hochberg multiple correction was performed and enriched BP with a p adjusted value < 0.05 were selected. Of note, gene symbols were used to perform the functional enrichment of the human cell population. For both organisms and each hierarchical level, we removed shared BPs between cell populations of the same level, hence keeping the specific ones, and selected the top 10 most enriched unique BPs, based on the GeneRatio provided by *ClusterProfiler,* for each cell population.

Application on the mouse brain

We utilized a comprehensive mouse brain atlas data[21] to perform a deep case study. A quality control as well as a down-sampling was performed on the raw data, using the same strategy as the one used for the Tabula Sapiens to conserve as much information as possible. In addition, a manual curation of the neuron annotation was performed to fit the hierarchical identity concept. We applied HCellig to quantify gene expression of neuronal cells at each hierarchical level including cell type, cell subtypes and phenotype. For the cell type level, we used the background constructed from Tabula Muris to quantify the neuronal gene expression. Then, based on the curated metadata, we selected neuronal subtypes including dopaminergic, GABAergic, glutamatergic, serotonergic, glycinergic, sensory and motor neurons to build cell subtype level backgrounds to accordingly quantify the gene expression of these neuronal subtypes and capture identity genes for each one of them. Finally, the neuronal subtypes were further divided based on the major brain region including Forebrain, Midbrain and Hindbrain. The region based neuronal subtypes data was used to construct the phenotype level background for five of the neuronal subtypes. Indeed, phenotype

backgrounds for dopaminergic and motor neurons were not generated due to the lack of data. We then performed pairwise comparisons of gene expression levels across the neuron subtypes to identify genes displaying different levels depending on the neuron subtype. Finally, we performed an analysis at the phenotype level by merging medium identity genes of the neuronal subtypes across the three major brain regions. Genes with strong variation were ranked based on the Euclidean distance between the distribution of gene level quantification values.

References

1. Morris, S. A. The evolving concept of cell identity in the single cell era. *Development* **146**, dev169748 (2019).
2. Wagner, A., Regev, A. & Yosef, N. Revealing the vectors of cellular identity with single-cell genomics. *Nat Biotechnol* **34**, 1145–1160 (2016).
3. Osumi-Sutherland, D. *et al.* Cell type ontologies of the Human Cell Atlas. *Nat Cell Biol* **23**, 1129–1135 (2021).
4. Hao, Y. *et al.* Integrated analysis of multimodal single-cell data. *Cell* **184**, 3573-3587.e29 (2021).
5. Kim, H. J., Tam, P. P. L. & Yang, P. Defining cell identity beyond the premise of differential gene expression. *Cell Regen* **10**, 20 (2021).
6. Shats, I. *et al.* Expression level is a key determinant of E2F1-mediated cell fate. *Cell Death Differ.* **24**, 626–637 (2017).
7. Huang, C., Yang, D., Ye, G. W., Powell, C. A. & Guo, P. Vascular Notch Signaling in Stress Hematopoiesis. *Front. Cell Dev. Biol.* **0**, (2021).
8. The Tabula Muris Consortium *et al.* Single-cell transcriptomics of 20 mouse organs creates a Tabula Muris. *Nature* **562**, 367–372 (2018).
9. The Tabula Sapiens Consortium & Quake, S. R. *The Tabula Sapiens: a multiple organ single cell transcriptomic atlas of humans*. http://biorxiv.org/lookup/doi/10.1101/2021.07.19.452956 (2021) doi:10.1101/2021.07.19.452956.
10. Dale, D. C., Boxer, L. & Liles, W. C. The phagocytes: neutrophils and monocytes. *Blood* **112**, 935–945 (2008).
11. Krummel, M. F., Bartumeus, F. & Gérard, A. T cell migration, search strategies and mechanisms. *Nat Rev Immunol* **16**, 193–201 (2016).
12. Filby, A. *et al.* Fyn Regulates the Duration of TCR Engagement Needed for Commitment to Effector Function. *J Immunol* **179**, 4635–4644 (2007).
13. Palacios, E. H. & Weiss, A. Function of the Src-family kinases, Lck and Fyn, in T-cell development and activation. *Oncogene* **23**, 7990–8000 (2004).
14. Beffert, U. *et al.* Microtubule Plus-End Tracking Protein CLASP2 Regulates Neuronal Polarity and Synaptic Function. *Journal of Neuroscience* **32**, 13906–13916 (2012).
15. Sachse, S. M. *et al.* Nuclear import of the DSCAM -cytoplasmic domain drives signaling capable of inhibiting synapse formation. *EMBO J* **38**, (2019).

16. Villaescusa, J. C. et al. A PBX1 transcriptional network controls dopaminergic neuron development and is impaired in Parkinson's disease. EMBO J 35, 1963–1978 (2016).
17. Jung, S., Hartmann, A. & del Sol, A. RefBool: a reference-based algorithm for discretizing gene expression data. *Bioinformatics* **33**, 1953–1962 (2017).
18. Hafemeister, C. & Satija, R. Normalization and variance stabilization of single-cell RNA-seq data using regularized negative binomial regression. *Genome Biol* **20**, 296 (2019).
19. Zhang, X. *et al.* CellMarker: a manually curated resource of cell markers in human and mouse. *Nucleic Acids Research* **47**, D721–D728 (2019).
20. Wu, T. *et al.* clusterProfiler 4.0: A universal enrichment tool for interpreting omics data. *The Innovation* **2**, 100141 (2021).
21. La Manno, G. *et al.* Molecular architecture of the developing mouse brain. *Nature* **596**, 92–96 (2021).

Data availability

The raw data used in this study are publicly available and accessible in the cited papers. The pre-processed data can be provided by sending a request to the authors. The pre-compiled backgrounds generated with HCellig for mouse and human are available at: https://gitlab.com/C.Barlier/HCellig_backgrounds. The large-scale hierarchical cell identity atlases and the hierarchical cell identity of neurons generated are provided in Supplementary Tables S3, S4 and S8 and are available at: https://gitlab.com/C.Barlier/HCI. All gene quantification is available at https://gitlab.com/C.Barlier/HCellig_analyses.

Code availability

HCellig is an R package available at: https://github.com/BarlierC/HCellig.

All scripts for this study are available at: https://gitlab.com/C.Barlier/HCellig_analyses.

Acknowledgements

C.B. is supported by funding from the Luxembourg National Research Fund (FNR) within PARK-QC DTU (PRIDE17/12244779/PARK-QC). K.S. is supported by the Luxembourg National Research Fund (FNR) under the PRIDE program (Project code: 11012546) within the NextImmune DTU. The analyses were carried out using the HPC facilities of the University of Luxembourg (https://hpc.uni.lu). We thank Chrysovalantou Kalaitzidou for the pre-processing of the Tabula Muris data. We thank Prof. Ernest Arenas for the scientific exchanges we had for the brain analysis.

Authors information

Affiliations

Computational Biology Group, Luxembourg Centre for Systems Biomedicine (LCSB), University of Luxembourg, L-4362 Esch-sur-Alzette, Luxembourg
Céline Barlier, Kartikeya Singh, Antonio del Sol

Computational Biology Group, CIC bioGUNE-BRTA (Basque Research and Technology Alliance), Derio, Spain 48160
Sascha Jung, Antonio del Sol

Ikerbasque, Basque Foundation for Science, Bilbao, Bizkaia, 48012. Spain
Antonio del Sol

Contributions

C.B. developed the method, collected the human data, manually curated the mouse, human and brain annotations, compiled the backgrounds, created the hierarchical cell identity repository, performed the stability analysis and wrote the manuscript; K.S. collected the brain data, performed the corresponding analysis and drafted the manuscript; S.J. supervised the computational work; A.d.S supervised the project, conceived the idea and wrote the manuscript.

Corresponding author

Correspondence to Antonio de Sol.

Ethics declarations

Competing interest

The authors declare no competing interest.

Figures

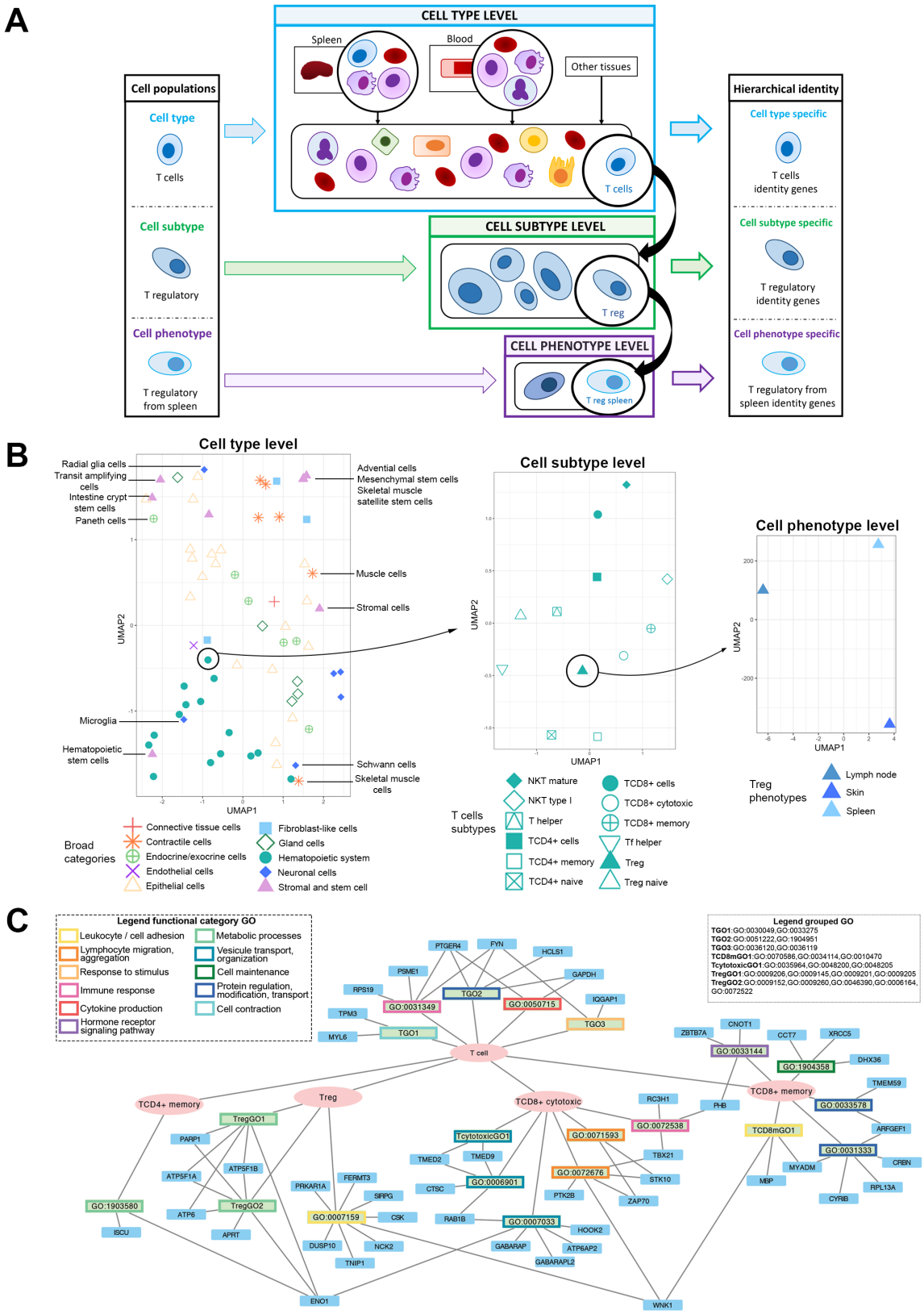

Figure 1. Hierarchical cell identity concept, repository and validation.

(A) Hierarchical cell identity concept to capture refined identity genes. (B) Hierarchical cell identity landscape for Human cell types, T cell subtypes and Treg phenotypes. (C) Functional relevance of medium identity genes for human T cells and four related subtypes.

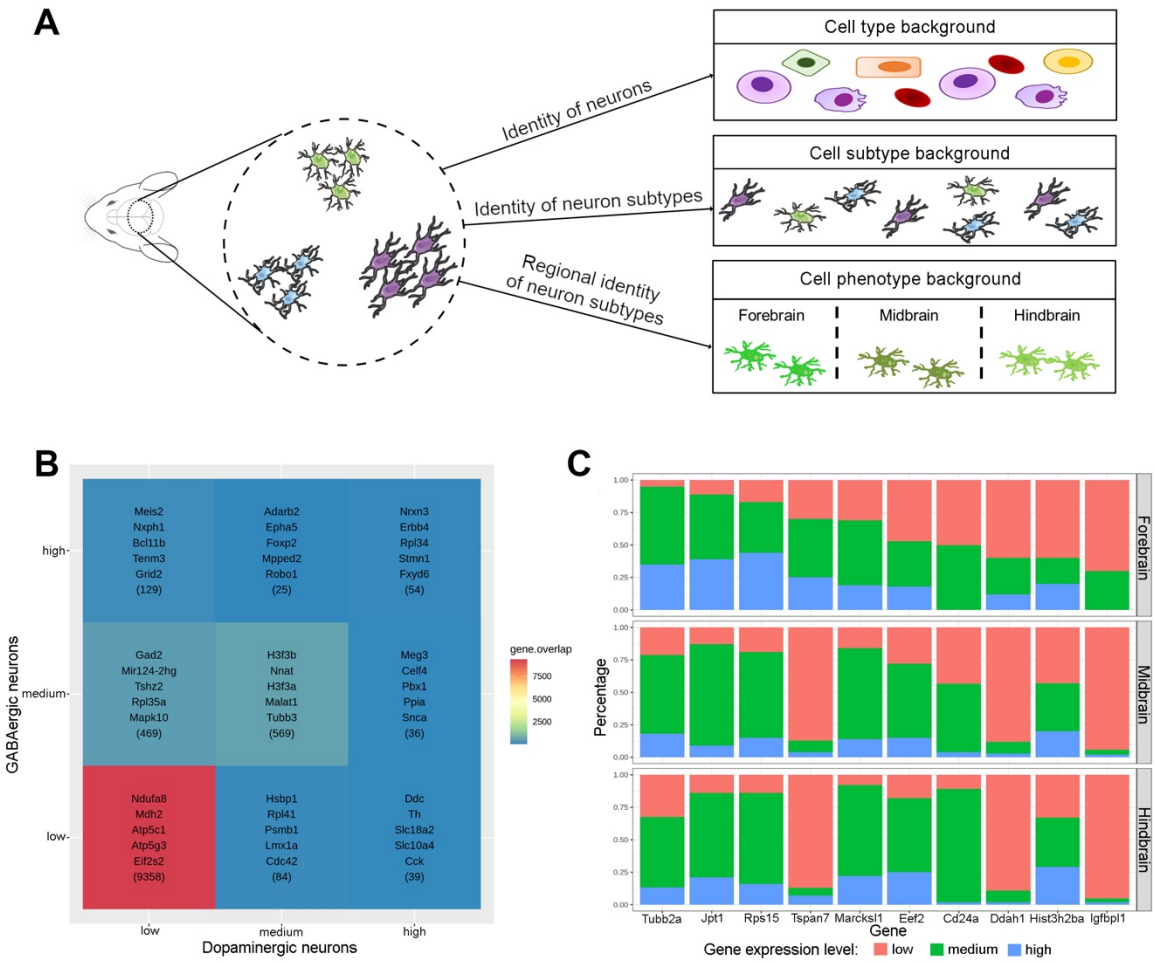

Figure 2. Characterization of hierarchical brain cell identity.

(A) Hierarchical identity concept applied to the mouse brain. (B) GABAergic and Dopaminergic neurons gene quantification comparison. (C) Brain region variation of gene expression levels for Serotoninergic neurons.

4.1.3 Supplementary Information

The supplementary tables S3, S4, S6 and S8 can be found in the GitHub repository of the project, due to their size and complexity: https://gitlab.com/C.Barlier/HCellig_analyses.

Supplementary Tables legends:

Table S1. Composition of the hierarchical layers for Mouse.

Table S2. Composition of the hierarchical layers for Human.

Table S3. Cell identity genes for mouse cell types and cell subtypes.

Table S4. Cell identity genes for human cell types, cell subtypes and phenotypes.

Table S5. Literature based validation of known markers.

Table S6. Functional enrichment of identity genes for each hierarchical level.

Table S7. Literature support for medium identity genes of T cells and related subtypes.

Table S8. Hierarchical cell identity genes for the brain application.

Table S9. Literature support for medium identity genes in the brain study case.

Supplementary Figures:

Figure S1. General workflow of HCellig.

Figure S2. Landscape of mouse cell types.

Figure S3. Functional enrichment of hierarchical identity for human monocytes.

Figure S4. Genes level of dopaminergic neurons compared to other neuronal subtypes.

Figure S5. Glycinergic and Sensory medium identity genes across brain regions.

Figure S6. HCellig stability of thresholds and predictions.

Table S1. Composition of the hierarchical layers for Mouse.

Mouse Cell Type Background	
Cell types	
basal epithelial cell	
basophil	
B cell	
chondrocyte	
endothelial cell	
epithelial cell	
erythroblast	
fibroblast	
granulocyte	
hematopoietic precursor cell	
hepatocyte	
keratinocyte	
luminal epithelial cell	
luminal progenitor	
macrophage	
mesangial cell	
mesenchymal stem cell	
monocyte	
NK cell	
podocyte	
skeletal muscle satellite cell	
stromal cell	
T cell	
Thymocyte	
umbrella cell	
2 Mouse Cell Subtype Backgrounds	
Cell types	**Cell subtypes**
mesenchymal stem cell	mesenchymal cell Car3+
	mesenchymal cell Scara5+
T cell	T cell CD4+
	T cell CD8+

Table S2. Composition of the hierarchical layers for Human.

Human Cell Type Background
Cell types
acinar cell
adventitial cell
aerocyte
B cell
basal cell
basophil
beta cell
cardiomyocyte
ciliated cell
club cell
common myeloid progenitor
dendritic cell
ductal cell
endothelial cell
enterocyte
enteroendocrine cell
epithelial cell
erythrocyte
erythroid progenitor cell
eye photoreceptor cell
fibroblast
goblet cell
granulocyte

hematopoietic stem cell
hepatocyte
intestinal crypt stem cell
ionocyte
keratinocyte
lacrimal gland functional unit cell
limbal stem cell
luminal epithelial cell
macrophage
mast cell
melanocyte
mesenchymal stem cell
microglial cell
monocyte
mucus secreting cell
Muller cell
muscle cell
myoepithelial cell
myometrial cell
natural killer cell
neutrophil
pancreatic stellate cell
paneth cell
pericyte
platelet
pneumocyte
radial glial cell
skeletal muscle satellite stem cell
smooth muscle cell
stromal cell
surface ectodermal cell
T cell
tendon cell
thymocyte
transit amplifying cell
tuft cell
urothelial cell

9 Human Cell Subtype Backgrounds

Cell types	Cell subtypes
B cell	memory b cell
	plasma cell
dendritic cell	CD141 myeloid dendritic cell
	CD1C myeloid dendritic cell
	plasmacytoid dendritic cell
endothelial cell	capillary endothelial cell
	endothelial cell of artery
	lymphatic endothelial
	vein endothelial cell
fibroblast	keratocyte
	myofibroblast cell
monocyte	classical monocyte
	intermediate monocyte
	non-classical monocyte
muscle cell	fast muscle cell
	slow muscle cell
pneumocyte	type i pneumocyte
	type ii pneumocyte
T cell	mature NK T cell
	regulatory t cell
	TCD4 alpha/beta memory
	TCD4 helper
	TCD8 alpha/beta cytotoxic
	TCD8 alpha/beta memory
	DN1 thymic pro-T cell

thymocyte	DN3 thymocyte
24 Human Cell Phenotype Backgrounds	
Cell subtypes	**Tissues**
capillary endothelial cell	bladder organ
	lung
	muscle tissue
	thymus
	tongue
CD1C myeloid dendritic cell	lymph node
	skin of body
classical monocyte	blood
	lung
	lymph node
	spleen
endothelial cell of artery	lung
	mammary gland
	muscle tissue
	thymus
	vasculature
fast muscle cell	muscle tissue
	thymus
immature enterocyte	large intestine
	small intestine
intermediate monocyte	lung
	lymph node
	spleen
lymphatic endothelial	bladder organ
	muscle tissue
	saliva-secreting gland
	thymus
	uterus
mature enterocyte	large intestine
	small intestine
mature NK T cell	adipose tissue
	bladder organ
	blood
	bone marrow
	kidney
	liver
	lung
	lymph node
	prostate gland
	saliva-secreting gland
	skin of body
	spleen
	thymus
	vasculature
memory b cell	blood
	bone marrow
	lymph node
	saliva-secreting gland
	spleen
	thymus
myofibroblast cell	adipose tissue
	bladder organ
naive b cell	blood
	bone marrow
	lymph node
	saliva-secreting gland
	spleen
	thymus
plasma cell	bladder organ
	blood

	bone marrow
	large intestine
	lung
	lymph node
	mammary gland
	pancreas
	saliva-secreting gland
	small intestine
	spleen
	thymus
	trachea
regulatory t cell	lymph node
	skin of body
	spleen
TCD4 alpha/beta	bone marrow
	large intestine
	lung
	lymph node
	muscle tissue
	skin of body
	small intestine
	spleen
	trachea
TCD4 alpha/beta memory	blood
	lymph node
	skin of body
	spleen
TCD4 alpha/beta naive	blood
	lymph node
TCD4 helper	kidney
	saliva-secreting gland
	skin of body
	thymus
TCD8 alpha/beta	blood
	bone marrow
	kidney
	large intestine
	lung
	lymph node
	prostate gland
	saliva-secreting gland
	skin of body
	small intestine
	spleen
	thymus
	trachea
TCD8 alpha/beta cytotoxic	skin of body
	thymus
TCD8 alpha/beta memory	lymph node
	spleen
type I NKT cell	blood
	lymph node
	spleen
vein endothelial cell	bladder organ
	lung
	mammary gland
	thymus
	tongue

Table S5. Literature based validation of known markers.

HUMAN		
Hierarchical Level	**Cell (sub)population**	**Identity genes found in CellMarker**
CELL TYPE	Hepatocyte	*CPS1,ABCC2,HNF4A,ARG1,CYP3A4,ALB*
	Limbal stem cell	*TP63,SOD2,KRT15,KRT14*
	Endothelial cell	*ICAM1,FLT1,ENG,ICAM2,VWF,PTPRB,NECTIN2,PLVAP,EMCN ,THBD,CDH5,AQP1,ECSCR,PECAM1*
	Hematopoietic stem cell	*CD38,PROM1,PTPRC*
	Fibroblast	*VIM,PDGFRB,PDGFRA*
	Pericyte	*MCAM,ACTA2,PDGFRB,CSPG4,PECAM1*
	Mesenchymal stem cell	*VIM,CD44,ZBTB16,CD81,PDGFRA,ITGB1,VCAM1,BSG,MME*
	Smooth muscle cell	*MCAM,ACTA2,DES*
	Myoepithelial cell	*ACTN1,CNN1,ACTN4,BHLHE40,ITGB1,S100A1,KRT14*
	Cardiomyocyte	*ACTN1,TNNT2,TNNI3,VCAM1,MYH6*
	Macrophage	*TFRC,FCGR2B,PTPRC,ICAM1,LYZ,CD83,IL1RN,FCGR2A,ITGB 2,CD14,CD163,HLA-DQB1,CSF1R,HLA-DRB1,HLA-DQA1,HLA-DRB5,FCGR3A,HLA-DMA,HLA-DRA,AIF1,HLA-DPB1,HLA-DPA1,MRC1*
	B cell	*CD74,POU2F2,PTPRC,CD79A,MS4A1,HLA-DRB1,HLA-DRA,IGHM*
	Microglial cell	*AIF1*
	Stromal cell	*VIM,CD44,NT5E,ITGB1,ANPEP,GREM1,CD34*
	Epithelial cell	*VIM,CDH1,KLF6,TJP1,KRT18,EPCAM,KRT7,PIP,CLDN1,CTNN B1,KRT8,KRT19,KRT13,MUC16,MUC1,KRT3*
	Intestinal crypt stem cell	*CD24*
	Keratinocyte	*CD44,ALDH3A2,ITGA6,ALDH3A1,ALDH2,ALDH3B2,ALDH9A1, ALDH7A1,ALDH1A1,KRT5,KRT14,SPRR2A*
	Erythrocyte	*GYPA*
	Neutrophil	*PTPRC,CEACAM8,LCN2,FCGR3B,MNDA,CXCL8,ITGAM,CD14, FPR1,FCGR3A,CD24*
	Neuron	*MAP2*
	Monocyte	*TNFRSF1B,PTPRC,LYZ,CD36,FCGR2A,S100A8,ITGB2,MNDA,C D52,CD14,CD163,SELL,HLA-DRB1,FCGR3A,HLA-DRA,PECAM1*
	T cell	*PTPRC,CD69,CD2,CD3G,CD3D,IL7R,CD7,CD3E*
	Granulocyte	*PTPRC,FUT4,HLA-DRA*
	Basophil	*CD69,HLA-DRA*
	Dendritic cell	*CD83,CD86,CD1A,CD1C,CD14,THBD,HLA-DQB1,HLA-DRB1,HLA-DQA1,CLEC9A,HLA-DMA,HLA-DRA,HLA-DPB1,HLA-DPA1,HLA-DQA2,HLA-DMB*
	Platelet	*ITGA2B,SELP,PECAM1*
	Basal cell	*CD151,KRT5,KRT14,S100A6*
	Natural killer cell	*PTPRC,KLRD1*
	Mast cell	*KIT,SLC18A2,FCER1A*
	Beta cell	*FXYD2,NKX6-1,HEPACAM2,INS*
	Mesothelial cell	*UPK3B*
	Eye photoreceptor cell	*CRX,RCVRN,RHO*
	Urothelial cell	*DHRS2,UPK1B,NECTIN4,S100P*
	Transit amplifying cell	*FABP5*
	Luminal epithelial cell	*KRT18,KRT19,MUC1*
	Alpha cell	*GCG*
	Muller cell	*GLUL*
	Club cell	*SCGB1A1*
	Erythroid progenitor cell	*TFRC*
CELL SUBTYPE	Myofibroblast cell	*VIM,ACTA2,FN1*
	Type II pneumocyte	*CD44,PGC*
	TCD4 alpha/beta naive	*CCR7,SELL,CD3E*
	T helper follicular	*CCR7,ICOS,SELL,CD3E*
	Regulatory T cell	*IKZF2,FOXP3,IL2RA,ENTPD1,CTLA4,CD3D,TIGIT,TNFRSF18, CD3E*
	TCD4 alpha/beta	*CD3E,LTB*
	Intermediate monocyte	*CEACAM8,CD14,FCGR3A*

	Classical monocyte	*CD14*
	Non-classical monocyte	*FCGR3A*
	Plasmablast	*CD38*
	Plasma cell	*CD38,TNFRSF17,SDC1,CD27*
	CD1C myeloid dendritic cell	*ITGAX,CD1C*
	CD141 myeloid dendritic cell	*ITGAX,CD1C*
	Lymphatic endothelial	*FLT4,PROX1,PDPN*
	Mature NKT cell	*GZMB,KLRD1,FCGR3A*
	TCD8 alpha/beta	*NKG7,CD8A,CD3E*
	TCD8 alpha/beta cytotoxic	*CD8A,CD3E*
	Plasmacytoid dendritic cell	*THBD,CLEC4C*
MOUSE		
Hierarchical Level	**Cell (sub)population**	**Identity genes found in CellMarker**
CELL TYPE	Basal epithelial cell	*Cd24a,Itga6*
	Basophil	*Mcpt8*
	B cell	*Cd79a,Ms4a1,Cd24a,Ptprc*
	Cardiomyocyte	*Tnnt2,Ryr2,Actc1,Nppa,Tnnc1,Myh6,Atp2a2,Actn2*
	Dendritic cell	*Cd74,H2-Ab1*
	Endothelial cell	*Egfl7,Fabp4,Cdh5,Pecam1,Eng,Emcn,Epas1,Plvap,Tie1,Cd34,Ednrb,Lyve1*
	Epithelial cell	*Ly6a,Cd24a,Epcam*
	Erythroblast	*Tfrc*
	Fibroblast	*Gsn,Sparc,Vim,Fstl1,Mmp2,Fbln2,Col3a1,Col1a2*
	Granulocyte	*Itgam*
	Hematopoietic precursor cell	*Cd47,Cd48,Kit,Cd34*
	Hepatocyte	*Alb*
	Luminal epithelial cell	*Cd24a*
	Macrophage	*Cd74,H2-Ab1,Lyz1,Lgals3,S100a4,Csf1r,S100a10,Fcgr3,S100a9,S100a8*
	Mesenchymal stem cell	*Ly6a,Cd34,Thy1,Itgb1,Pdgfra,Vcam1*
	Monocyte	*Cd48,Itgam*
	Skeletal muscle satellite cell	*Cav1,Cdh15*
	Smooth muscle cell	*Rgs5*
	Stromal cell	*Cd34*
	T cell	*Cd3d,Thy1,Cd2,Cd3e,Cd5,Cd8a*
	Thymocyte	*Cd8a,Cd4,Cd5*
CELL SUBTYPE	TCD4 cell	*Ptprc*
	Mesenchymal cell Car3+	*Car3*
	Mesenchymal cell Scara5+	*Scara5*

Table S7. Literature support for medium identity genes of T cells and related subtypes.

Hierarchical Level	Cell (sub)population	Gene Symbol	PMID	Comment
CELL TYPE	T cells	*MYL6*	25770220	
		TPM3	https://www.jimmunol.org/content/206/1_Supplement/14.05	
		RPS19	28228558	Antitumor immune reponses
		PSME1	9189757	
		PTGER4	22544928	PA28 subunit
			22544928	Support for level variation
		FYN	15489916	
			7594580	Support for the medium level identity
		HCLS1	30537294	
		GAPDH		
		IQGAP1	22573807	

CELL SUBTYPE	TCD4+ memory	*ISCU*	34880854	T cells in general (cell type)
		ENO1	32709897	T cells in general (cell type)
	Treg	*PARP1*	23977081	
		ATP5F1B	20686167	Importance of Adenosine for Treg
		ATP5F1A		
		ATP6		
		APRT		
		ENO1	32709897	T cells in general (cell type)
		PRKAR1A	24007532	
		FERMT3	30187863	T cells in general (cell type)
		SIRPG	18524990	T-cell transendothelial migration
		CSK	26302204	T cells in general (cell type)
		NCK2	20709959	T cells in general (cell type)
		TNIP1	20181891	
		DUSP10	22387553	Described for TCD8+ supressor T cells
		WNK1	27400149	T cells in general (cell type)
	TCD8+ cytotoxic	*TMED2*		
		TMED9		
		CTSC		
		RAB1B	31375559	Indirect validation: response to viral infection
		GABARAP	31632966	T cells in general (cell type)
		GABARAPL2		
		ENO1	32709897	T cells in general (cell type)
		ATP6AP2		
		HOOK2		
		PTK2B	20688918	
		ZAP70	24596147	
		STK10		
		WNK1	27400149	T cells in general (cell type)
		TBX21	29488879	T cells (cell type) general program, involvement toward TCD8+ (memory)
		RC3H1	34879274	
		PHB	18086671	T cells in general (cell type)
	TCD8+ memory	*WNK1*	27400149	T cells in general (cell type)
		MBP	12067310	Study in disease case (MS)
		MYADM		
		CYRIB		
		RPL13A	32005148	T cells in general (cell type)
		CRBN	https://ashpublications.org/blood/article/126/23/3440/90809/Genetic-Ablation-of-Cereblon-CRBN-Increases-Long	
		ARFGEF1		
		TMEM59		

		DHX36		
		XRCC5		
		CCT7	33268369	T cells in general (cell type)
		CNOT1	34349771	Thymocyte to T cell transition
		ZBTB7A	34349770	T cells development program
		PHB	18086671	T cells in general (cell type)

Table S9. Literature support for medium identity genes in the brain study case.

CELL TYPE		
Medium Identity Gene	**PMID**	**Comment**
Ncam1	32632143	
Ulk2	29099309	
Nf1	31234911	
Camk2d	22612808	
Klf7	15964824	Required at medium level for neurons
Klf7	11336497	Required at medium level for neurons
Clasp2	28285824	Required at medium level for neurons
Clasp2	23035100	Required at medium level for neurons
Epha7	24707048	Required at medium level for neurons
Dscaml1	33585465	Required at medium level for neurons
Dscaml1	30745319	Required at medium level for neurons
Fzd3	34414184	Required at medium level for neurons
Fzd3	26939553	Required at medium level for neurons
CELL SUBTYPE		
Medium Identity Gene	**PMID**	**Comment**
Tubb3	22159867	
Chl1	23949217	
Stk11	30333724	
Id4	15882580	
Id4	31552825	
Cux2	20510857	
Sema6a	22685427	
Ache	15136152	
Ache	31031601	
Ndel1	22114287	
Nr3c1	33715314	
Nr3c1	32547368	

Supplementary Figure 1

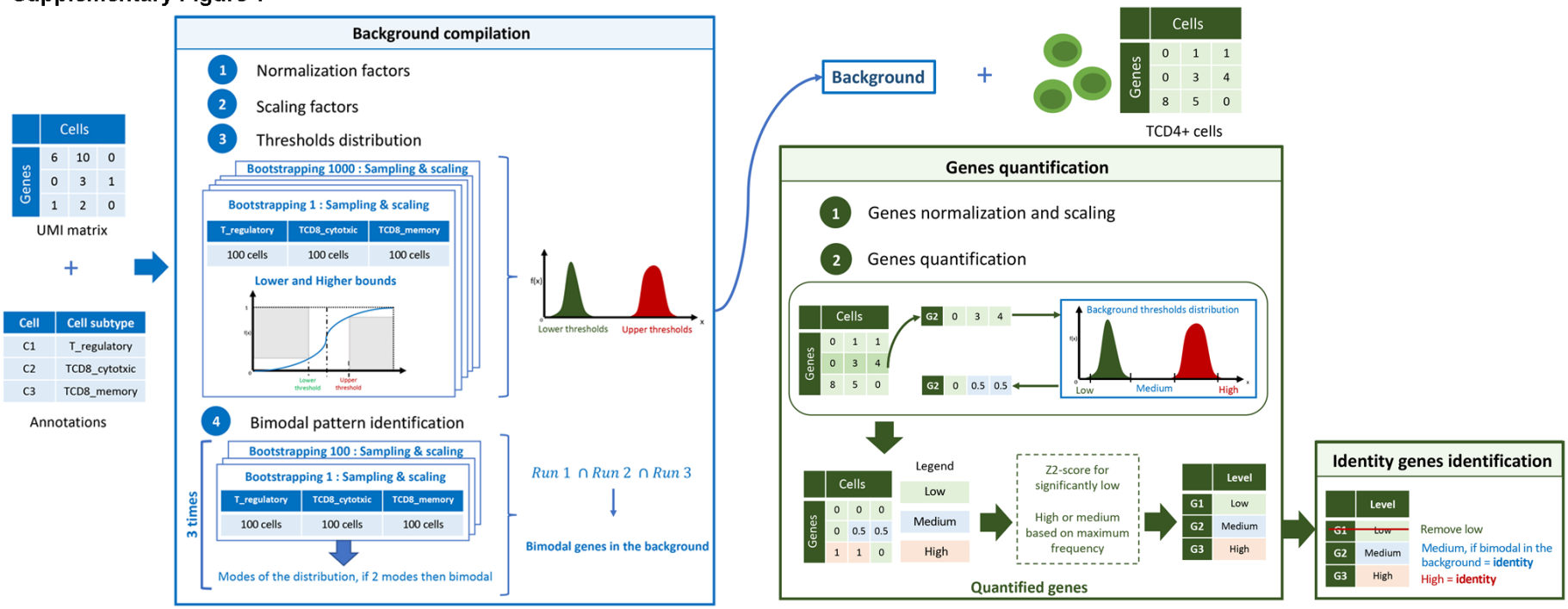

Figure S1. General workflow of HCellig.

This figure presents the workflow of HCellig using the subtype level of hierarchy as an example. The method first builds a background data using an UMI matrix and cell annotations to identify for each gene the normalization and scaling factors, the upper-bound threshold distribution and if the gene is bimodal or not. Using the generated background and the UMI matrix of a specific cell subtype (e.g. T cell CD4+), it quantifies gene expression accordingly with the background threshold distribution into three levels of expression: low, medium and high. Finally, identity genes for the subtype are identified by selecting genes expressed at a high or medium level, under the condition the gene is bimodal in the background for the medium level.

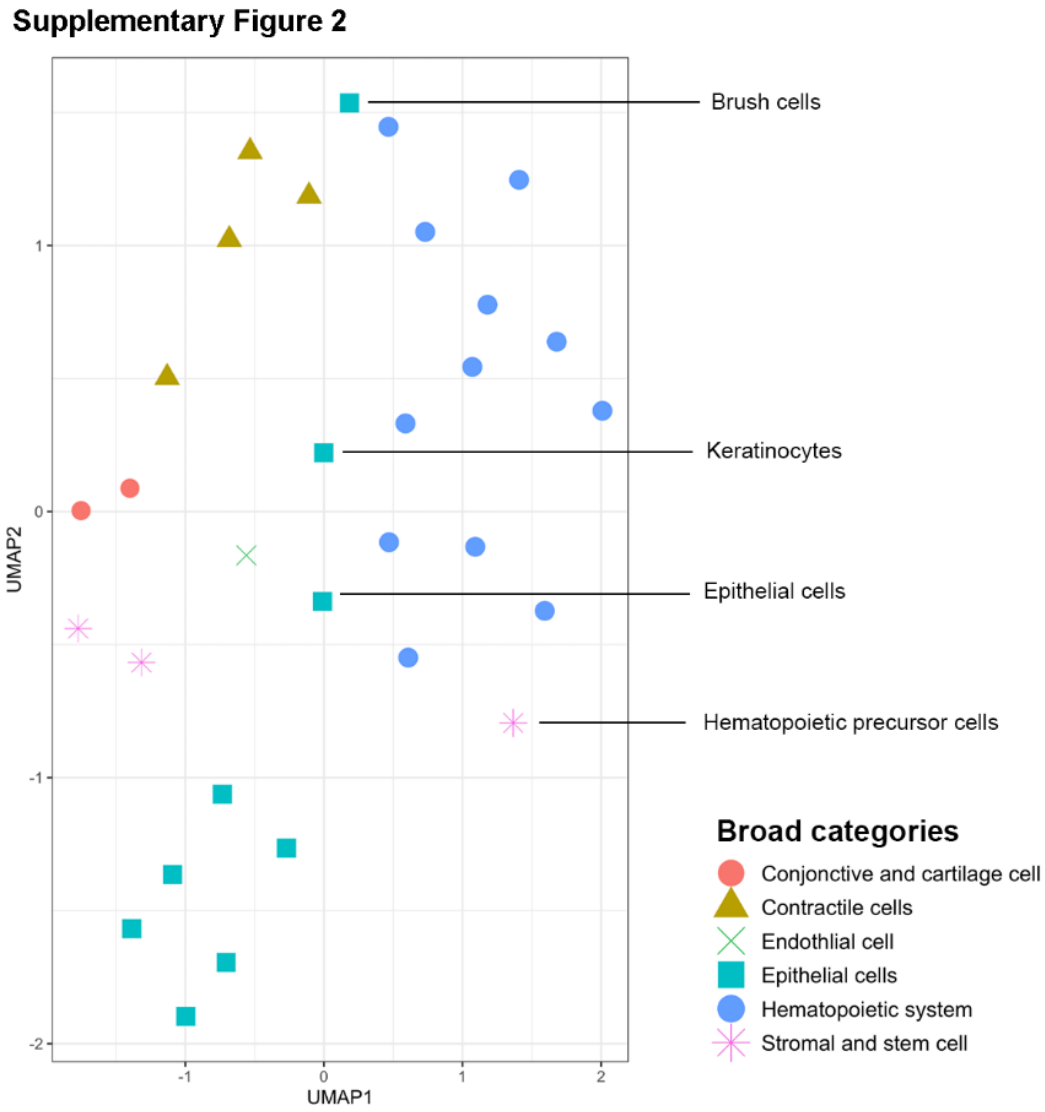

Figure S2. Landscape of mouse cell types.
UMAP of cell types generated using high and medium identity genes identified with HCellig.

Supplementary Figure 3

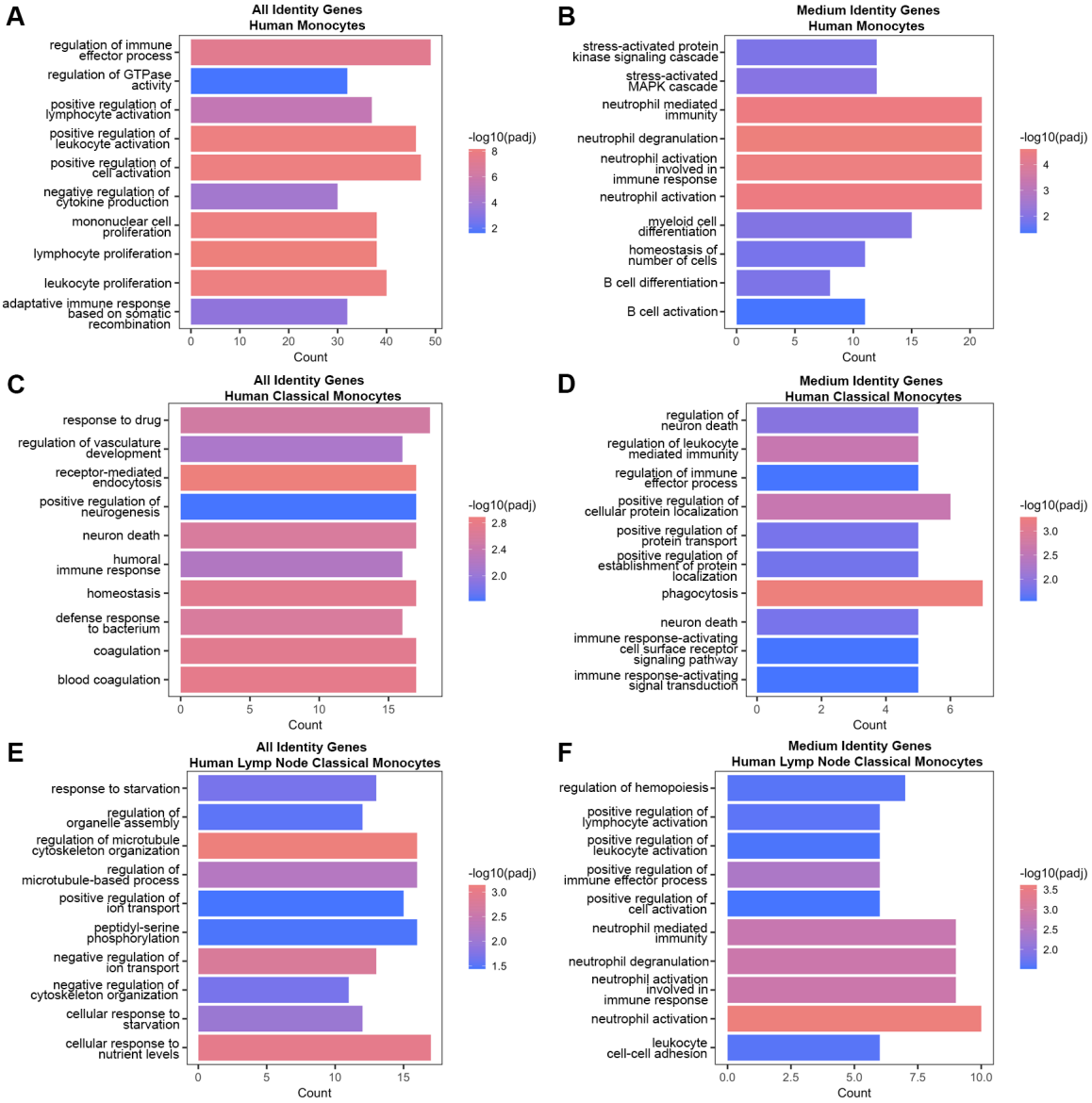

Figure S3. Functional enrichment of hierarchical identity for human monocytes.
Top 10 unique BPs for (A) all identity genes and (B) only medium ones of human monocytes, (C) all identity genes and (D) only medium ones of human classical monocytes, (E) all identity genes and (F) only medium ones of human classical monocytes from lymph nodes.

Supplementary Figure 4

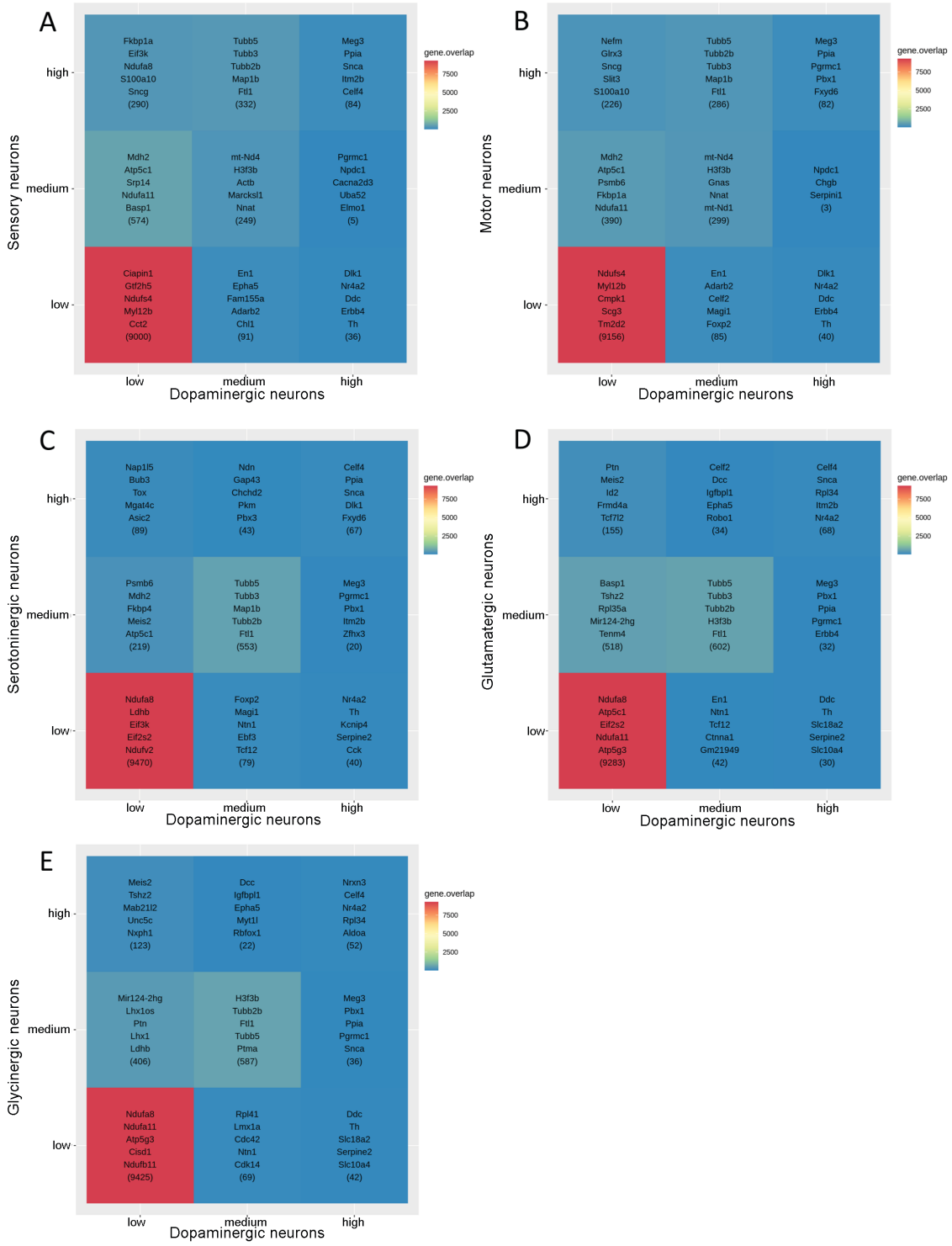

Figure S4. Genes level of dopaminergic neurons compared to other neuronal subtypes. Dopaminergic neurons compared to (A) sensory neurons, (B) motor neurons, (C) serotoninergic neurons, (D) glutamatergic neurons and (E) glycinergic neurons. Heatmaps display the pairwise comparison of neuronal subtypes of their gene expression levels.

Supplementary figure 5

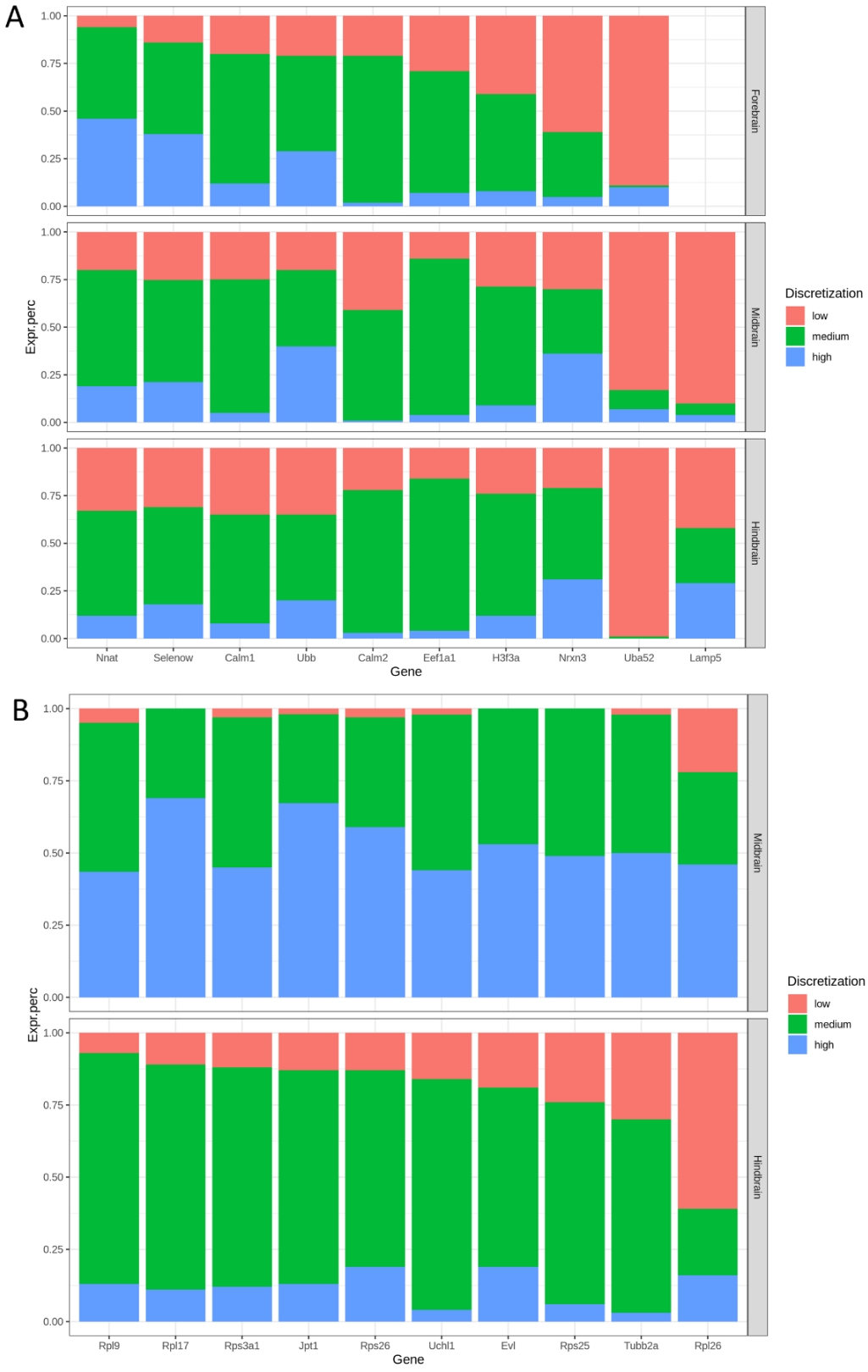

Figure S5. Glycinergic and Sensory medium identity genes across brain regions. Barplots for the functionally relevant genes of (A) glycinergic and (B) sensory neurons. This figure shows the variation of gene levels across the three major brain regions studied.

Supplementary Figure 6

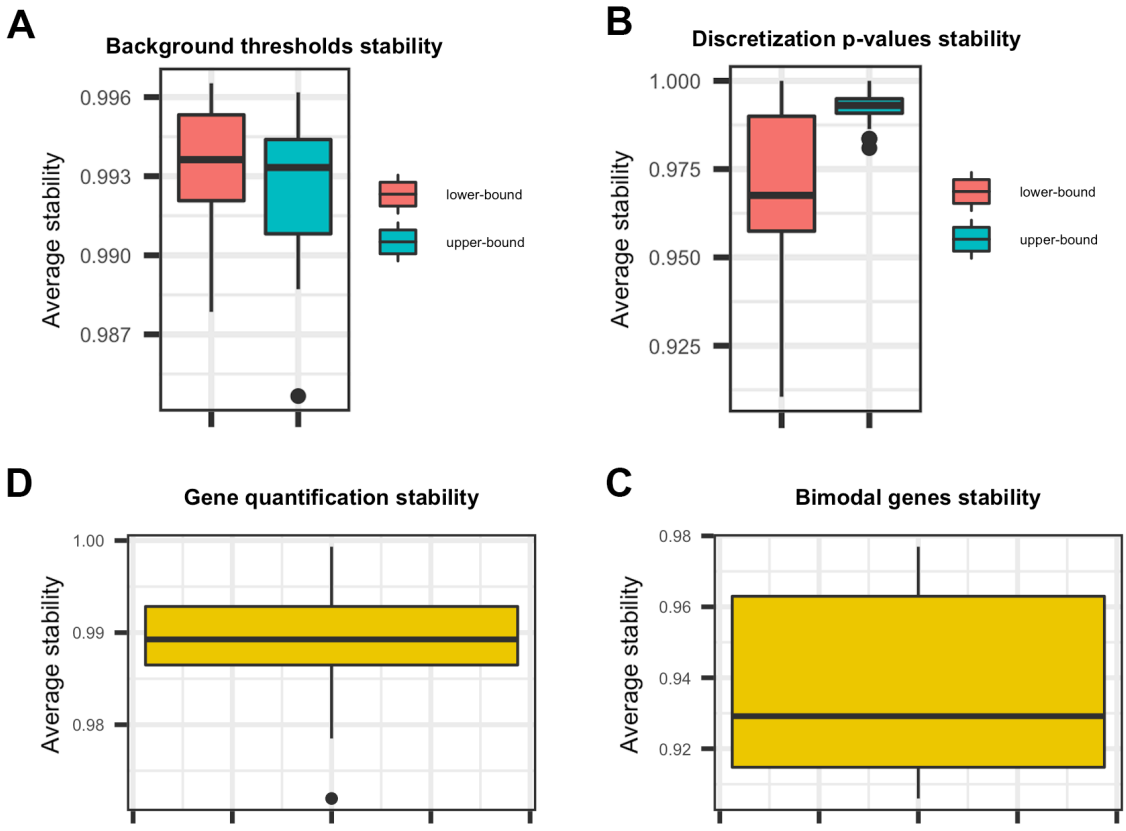

Figure S6. HCellig stability of thresholds and predictions.
(A) Background thresholds average stability across 10 runs using all human and mouse subtype backgrounds. (B) p-value thresholds average stability across 10 runs for all cell subtypes discretized using its corresponding subtype background. (C) Stability of the gene quantification for each cell subtype. (D) Average stability of the bimodal genes identified across 10 runs for all human and mouse subtype backgrounds.

4.2 Identification of disease-related functional states and genes

4.2.1 Preface

This manuscript entitled "*A Catalogus Immune Muris of the mouse immune responses to diverse pathogens*" has been published in Cell Death and Disease in August 2021 and is accessible with the DOI: 10.1038/s41419-021-04075-y. The paper is under a CC BY license and the accepted version of the manuscript is presented in this thesis. The supplementary methods and figures are shown in this thesis, but supplementary tables are accessible online.

In this study, we present a *Catalogus Immune Muris***,** a valuable resource of functional immune cell states for designing novel immunomodulatory strategies. Indeed, discerning the functional states of immune cells and their transcriptional characterization is pivotal for the development of immunomodulatory therapeutic strategies. However, the development of such therapies based on the reprogramming of functional states is significantly impeded by the incomplete knowledge about the functional cell states established in response to pathogens and their characterization. We made two novel contributions with this study. First, we developed FunPart, a computational method to decipher functional cell states in diverse conditions and identify the genes characterizing their states. We showed that genes identified are functionally relevant for the deciphered cell state by manually collecting literature evidence. Moreover, we showed that our method accurately detects functional cell states compared to current state-of-the-art methods. Second, we built a *Catalogus Immune Muris* by applying FunPart to 114 single-cell datasets composed of six immune cell types in the context of twelve viral, bacterial, fungal and parasite infections. We demonstrated how the resource can be exploited to modulate the cellular response to pathogens in the context of macrophages infected by *Salmonella enterica* Serovar Typhimurium. Indeed, we identified a previously unreported TF, *Zfp597*, as a functionally relevant gene of a macrophage cell state and showed that its inhibition significantly increases their phagocytic activity, and hence results in a significant decrease in surviving bacteria.

Contribution: I implemented the computational method, collected and processed the data, performed the analyses, and wrote the manuscript.

4.2.2 Published paper

Title: A *Catalogus Immune Muris* of the mouse immune responses to diverse pathogens

Running title: An atlas of the mouse immune response to pathogens

Celine Barlier[1], Diego Barriales[2], Alexey Samosyuk[3], Sascha Jung[4], Srikanth Ravichandran[1], Yulia A. Medvedeva[3,5,6], Juan Anguita[2,7] and Antonio del Sol[1,4,7,‡]

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

[2] Inflammation and Macrophage Plasticity laboratory, CIC bioGUNE-BRTA (Basque Research and Technology Alliance), Derio, Spain 48160

[3] Department of Biological and Medical Physics, Moscow Institute of Physics and Technology, Dolgoprudny, Russian Federation

[4] Computational Biology Group, CIC bioGUNE-BRTA (Basque Research and Technology Alliance), Derio, Spain 48160

[5] Institute of Bioengineering, Research Center of Biotechnology, Russian Academy of Science, Moscow, Russian Federation

[6] Department of Computational Biology, Vavilov Institute of General Genetics, Russian Academy of Science, Moscow, Russian Federation

[7] Ikerbasque, Basque Foundation for Science, Bilbao, Bizkaia, 48012. Spain

[‡]To whom correspondence should be addressed

Email: Antonio.delsol@uni.lu

Abstract

Immunomodulation strategies are crucial for several biomedical applications. However, the immune system is highly heterogeneous and its functional responses to infections remains elusive. Indeed, the characterization of immune response particularities to different pathogens is needed to identify immunomodulatory candidates. To address this issue, we compiled a comprehensive map of functional immune cell states of mouse in response to 12 pathogens. To create this atlas, we developed a single-cell-based computational method that partitions heterogeneous cell types into functionally distinct states and simultaneously identifies modules of functionally relevant genes characterizing them. We identified 295 functional states using 114 datasets of six immune cell types, creating a Catalogus Immune Muris. As a result, we found common as well as pathogen-specific functional states and experimentally characterized the function of an unknown macrophage cell state that modulates the response to Salmonella Typhimurium infection. Thus, we expect our Catalogus Immune Muris to be an important resource for studies aiming at discovering new immunomodulatory candidates.

Introduction

The immune response to pathogens, such as viruses, bacteria, or fungi, is a complex process involving multiple immune and nonimmune cell types[1,2]. Although transcriptional changes of these cells in response to pathogens have been studied for decades, the development of sensitive analytical techniques such as single-cell RNA sequencing (scRNAseq) only now

enables the identification and functional characterization of cellular subpopulations in response to different stimuli. Thus, heterogeneous subpopulations can be identified by specialized transcriptional profiles that determine their identity and govern their interactions with invading pathogens[3-6]. Recent studies utilizing various pathogens have shown that complex transcriptional variability in macrophages govern their divergent response against individual invasive agents[7,8]. For instance, in the case of Salmonella enterica Serovar Typhimurium, the interplay between the bacteria and macrophages triggers two different scenarios in which some cells are polarized to anti-inflammatory response whereas others display an inflammatory output[9]. Moreover, a subsequent study was able to identify two distinct cellular states that are responsible for a bimodal type I interferon response[10]. However, most of these studies focus on a single pathogen, making them unable to decipher common and distinct cellular states established in response to different infections. To date, only a few meta-analyses exist that aim at identifying common and unique patterns of the immune response to pathogens[11]. Nevertheless, these studies are based on the average response across a population of cells or tissues, making them unable to detect functionally distinct subpopulations. Moreover, the number of pathogens considered in these studies remains limited, which impedes more general conclusions regarding the cellular response to different types of infectious agents.

To date, several functional states of immune cells, such as macrophages, natural killer, and T cells, have been identified and characterized[12-15]. In general, discerning the functional states of immune cells and their transcriptional characterization is pivotal for the development of immunomodulatory therapeutic strategies. For instance, previous studies demonstrated the beneficial effect of reprogramming the macrophage polarization state to promote tumor suppression or alleviate autoimmunity in encephalomyelitis[16,17]. However, the development of new immunomodulatory therapies based on the reprogramming of functional states is significantly impeded by the incomplete knowledge about the functional cell states established in response to pathogens and their characterization.

To address this challenge, we collected 114 single-cell datasets of six immune cell types in the context of 12 viral, bacterial, fungal, and parasite infections, and developed a computational method for identifying functional immune cell states in response to these pathogens, creating a Catalogus Immune Muris. We believe it will serve as a valuable resource of functional immune cell states to devise novel immunomodulatory strategies.

Materials and Methods

Data collection, processing and annotation

We collected 114 single-cell datasets composed of 6 immune cell types and 12 pathogens (Table S1). Raw data (accession numbers: PRJEB14043, E-MTAB-3857, and E-MTAB-4388) were processed using state-of-the-art pipelines[18]. Smart-seq data were subjected to a quality control step using fastqc, reads were mapped to the mm10 genome using STAR aligner and the count matrix were obtained using featureCounts tool. A similar workflow was applied for UMI-based data, adding the demultiplexing step and replacing the counting tool by umi-tool.

Datasets composed of several cell types were clustered using Seurat pipeline with default parameters, manually annotated and extracted. Cells were annotated using prior knowledge and CIPR web tool with default parameters[19]. Only the cells annotated with a good confidence were extracted and used to build the resource.

Functional partitioning algorithm

In order to reliably identify and characterize functionally relevant cell states, we developed a network-based approach combined with a recursive hierarchical clustering named FunPart. The algorithm is composed of four main parts: (1) cleaning and normalization of the data, (2) network-based approach to identify set of genes strongly correlated, (3) functional characterization of the set of genes using manually annotated immune modules by Singhania et al.[11], and (4) recursive unsupervised hierarchical clustering to perform the splits. Each step is detailed in the Supplementary Information. A dataset for which no module is found is considered to be functionally homogenous and corresponds to one functional cell state.

Validations and comparison with the state-of-the-art

We first aimed at validating our method at two levels: (1) the relevance of genes belonging to the detected functional modules, and (2) the relevance of the predicted cell states. We collected literature evidences for some of the main TFs identified in each module focusing on evidences of the immune process identified for macrophages. Next, we aimed at comparing our method with Seurat, a state-of-the-art method[20]. Seurat and FunPart were used with default parameters for the 17 macrophages datasets. We assessed the functional relevance of predicted clusters by both methods and computed a score reflecting the

precision of each method in identifying real or artificial functional heterogeneity per dataset (Supplementary Information).

Characterization of functional cell states

FunPart provides gene modules characterizing the predicted functional cell states as well as the specific immune process in which they are enriched. In order to have an additional layer of information, we aimed at identifying known markers to further characterize these cell states. Immune cell type markers were collected from the CellMarker database by considering experimentally validated evidences only[21]. We performed feature selection using the Boruta algorithm[22], a wrapper built around the random forest classification algorithm, to determine the importance of markers in classifying each cell states. Boruta was used in classification mode with default parameters for each cell state, details are provided in Supplementary Information. Fold changes and cell expression ratios were then computed for each cell states markers extracted by the algorithm (Supplementary Information).

Metadata analysis

Data integration was performed for each dataset using the standard workflow of Seurat (Supplementary Information). Cell states were then aggregated across datasets for each cell type by following a hierarchical clustering approach: (1) Each dataset was first normalized individually by the third quantile to overcome the different types of expression values present in the different datasets (TPM, CPM, UMI and counts), (2) The median expression of each gene in each cell state was calculated, (3) Euclidean distance was then used to build the dendrogram reflecting the similarity between states, and (4) the dendrogram was splitted at a height corresponding to the seventh quantile of the heights distribution. The aggregated states were then embedded into the computed UMAP for visualization and analyses.

Mice and bacteria

C57Bl/6 (B6) mice were purchased from Charles River Laboratories and bred in the Animal Facility at CIC bioGUNE. All the assays performed were approved by the competent authority (Diputación de Bizkaia) under European and Spanish directives. CIC bioGUNE is accredited by AAALAC Intl.

Salmonella enterica subsp. enterica serovar Typhimurium SL1344 (German Collection of Microorganisms and Cell Cultures, Leibniz, DE) was grown in Luria Bertani medium (Sigma–Aldrich) without antibiotics.

Cell culture and gene silencing

Bone-marrow-derived macrophages (BMMs) were generated from 6–12-week-old B6 mice, as previously described[23]. Low-passaged HEK293FT cells were cultured in DMEM containing 10% FBS and 1% penicillin-streptomycin.

Lentiviral particles containing shRNA targeting *Zfp597* (TRCN0000215620, TRCN0000179758, TRCN0000245367, Sigma–Aldrich) and *Stat1* (TRCN0000235839) were generated using a third-generation lentivirus vector with a conditional packaging system[24,25]. *Zfp597*-silencing in BMMs was conducted by co-infection with lentiviral particles containing the three silencing constructs whereas for Stat1 one single construct was used. Lentiviral particles were added at days 3 and 5 of the differentiation process in the presence of 8 μg/ml protamine sulfate (Sigma–Aldrich). Controls were infected with lentiviral particles containing the empty vector, PLKO.1. BMMs derived from three independent mice were used in each silencing assay.

Salmonella survival in murine macrophages

S. typhimurium was grown from a diluted (1:50) overnight inoculum until they reached an O.D. = 0.6. BMMs were infected following the protocol by Avraham et al.[10] at an m.o.i. of 10. In the experiments using shSTAT1 cells, 100 ng/ml of recombinant murine IFNγ was added at the same time than the bacteria. The mixture was centrifuged, incubated for 30 min, washed twice, and further incubated in the presence of 50 μg/ml gentamicin for 1 h. Macrophages were then washed and lysed in medium containing 0.1% Triton X-100. Cell lysates were centrifuged and resuspended in 1 ml of LB broth. Serial 1:10 dilutions were plated on LB-agar plates to determine the number of live intracellular bacteria per condition.

Real time PCR

Total RNA was isolated using the NucleoSpin® RNA kit (Macherey-Nagel) and reverse transcribed with M-MLV reverse transcriptase (Thermo Fisher Scientific). Real-time PCR was performed using the PerfeCTa SYBR Green SuperMix low ROX (Quantabio) on a ViiA 7^{TM} Real-Time PCR System (Thermo Fisher Scientific). Fold induction of *Zfp597* was calculated relative to *Rpl19* whereas *Stat1* was compared to *Actb* by using the $2^{-\Delta\Delta Ct}$ method. Standard curves of all primers were performed by testing serial dilutions of cDNA-experimental samples obtaining an average of 100% ±5% efficiency. Correlation between target and housekeeping genes was assessed by standard curve comparisons (*Zfp597-Rpl19*

slope 0.*0194* / *Stat1-Actin* slope 0.0188). Details about the primers used can be found as Supplementary Information.

Statistics

Three independent mice were used in each silencing assay. Data normality assumption was first validated using the Shapiro-Wilk test and variances between groups were analyzed using an F-test. Statistical difference between the two groups (control versus silenced assay) was then computed using a paired Student t-test. Results with a p value less than 0.05 were considered as being significant.

Results

Identification and characterization of functional immune cell states

In order to create an atlas of functional immune cell states, we developed FunPart, a single-cell-based computational method that partitions heterogeneous cell types into functionally distinct states and simultaneously identifies modules of functionally relevant genes that characterize them. Starting from a population of cells belonging to the same cell type, the method partitions them into two subpopulations by searching for modules that are (i) exclusively expressed in one subpopulation and (ii) composed of co-expressed TFs belonging to the same immunological process. This procedure is recursively repeated until no functionally relevant modules, associated to new subpopulations, can be found (Fig.1A).

To demonstrate the ability of this method to detect functional immune cell states, we collected 17 macrophage datasets corresponding to the infection with eight different pathogens profiled at different timepoints (Table S1). Application of our proposed method to these datasets revealed the presence of 9 M1-like, 13 M2-like cell states, and 14 middle range states expressing simultaneously some M1-like and M2-like markers[12] (Fig. S1). Moreover, literature evidences were found for every immune process and pathway reported by FunPart for the 12 intermediate genes modules, used to distinguish groups of functional states and 26 terminal gene modules, characterizing each individual state (Fig. 1B, C, Table S2). Next, we aimed at demonstrating that current clustering tools are unable to identify subtle functional differences and applied Seurat[20,26], a widely used state-of-the-art clustering method, to each of the datasets. As expected, the subpopulations obtained are vastly different, with FunPart identifying 46% of functionally enriched ones compared to 33% for Seurat across the 17 datasets (Fig. S2A,B). Furthermore, FunPart distinguishes more accurately functional homogeneity and heterogeneity with 67% and 43% of true positives,

respectively, compared to 25% and 22% for Seurat (Fig. S2C). In summary, FunPart identifies immune cell states more reliably and with an increased resolution compared to state-of-the-art methods.

295 functional immune cell states create a *Catalogus Immune Muris*

After validating our approach for detecting functional cell states, we collected 114 single-cell RNA-seq datasets of B cells, T cells, natural killer (NK) cells, macrophages, monocytes, and dendritic cells (DCs) in the context of 12 viral, bacterial, fungal, and parasitic pathogens (Table S1). For each cell type we obtained data for six to nine pathogens across three to six tissues (Fig. 2A, B). Application of our method to these datasets resulted in the detection of 295 functional cell states in total, thus, creating a *Catalogus Immune Muris* (Fig.2C, Table S3). On average, we identified 2.26 cell states per dataset and cell type, with NK cells and B cells having the lowest (average:1.06 and 1.07, respectively) and T cells having the highest (average: 4.45) functional heterogeneity. The low levels of functional heterogeneity in B cells are expected as their primary function is antibody secretion. Only in the context of lymphocytic choriomeningitis (LCMV), B cells exist in two distinct states characterized by two TFs modules composed of *Irf2*, *Rere*, *Sp140* for the first and *Irf5*, *Tcf25*, *Tcf4* for the second state, respectively (Fig. 3A, B, C). Moreover, *Irf5* is known to play a role in B cell differentiation[27] whereas *Irf2* is known to regulate B cell proliferation and antibody production[28], suggesting differences in the maturation stage of these cells. On the contrary, T cells exist in multiple cell states upon infection with various pathogens, such as LCMV, Influenza, and Salmonella Typhimurium. These are characterized by a marked difference in processes linked to stress response, inflammation and oxidative phosphorylation (Fig. S3). Interestingly, these processes are known to be involved in the functional diversity of T cells, more specifically by playing a role in their differentiation, activation, and function[29,30]. Finally, we extracted known cell markers to further characterize the identified functional cell states (Fig.3D, Table S4, S5). We found that combination of broad markers (e.g., CD3 for T cells) and specific markers (e.g., *Tlr9* for DCs) was important to classify the functional cell states, regardless of their relative expression (Table S4, Fig. S4). Finally, we further characterized functional states by identifying the expression of the extracted known cell markers for each functional state (Fig.3D, Table S5). Interestingly, we observed few diversity in markers signatures for B and NK cell states whereas specific signature patterns were found for macrophages and T cells (Fig.3D).

Exploiting TF modules for modulating the inflammatory response

Due to the enrichment of TF modules distinguishing different cell states in immune cell processes, we hypothesized that the *Catalogus Immune Muris* can be exploited to modulate the inflammatory response to pathogens by perturbing the TFs characteristic of different states. In order to provide support to this hypothesis, we selected the macrophage response to *Salmonella enterica* Serovar Typhimurium[10] due to a characteristic temporal change in macrophage states during the infection. In particular, while only a single macrophage state can be detected 2.5 h after the infection, heterogeneity rapidly increases after 4 h (three states) and diminishes again after 8 h (two states) (Fig. 4A). By focusing on the two macrophage states detected 8 h after the infection, we found the first state to be characterized by the module containing *Irf7*, *Hmga1*, *Zfp275*, and *Stat1* (Fig. 4B) that has been previously shown to initiate the inflammatory response to pathogens in an interferon gamma dependent manner[10]. In contrast, the second state is characterized by a module composed of *Zfp597*, *Zbtb38*, and *Zfp180* (Fig. 4C), but lacks a functional characterization. Enrichment of these TFs and their co-expressed targets showed their involvement in RNA and DNA processes as well as pathways such as janus kinase (JNK) signal transduction (Table S2). Indeed, previous studies highlighted the importance of kinase activity in response to bacterial infection and the interference of pathogens with kinase-mediated phosphorylation as a beneficial strategy for bacterial survival, replication and dissemination[31,32]. Thus, we hypothesized that macrophages exhibiting the second cell state are not responding to Salmonella infection due to kinase-mediated phosphorylation of proviral signaling pathways. We sought to validate this hypothesis by knockdown of *Zfp597* as this TF had the strongest co-expression pattern with its targets in the cell state characterized by the gene module. Therefore, we assessed the survival of Salmonella in primary murine bone-marrow-derived macrophages after silencing *Zpf597* with shRNA lentiviral constructs during the differentiation process[23] (Fig. 4D). The results in three independent mice showed that silencing of *Zpf597* resulted in a decreased ability to recover viable bacteria upon 90 min incubation periods demonstrating that *Zpf597* is responsible for preventing the macrophage response to Salmonella infection (Fig. 4E). Thus, the subpopulation characterized by the module involving *Zfp597* is indeed not responding to the pathogen due to the propathogenic effects of *Zfp597* and its inhibition induced a change in cell state. To further support the induced macrophage state change, we employed the same experimental setup to silence *Stat1* and hypothesized that bacterial survival is increased. Indeed, recovery of viable bacteria upon 90 min incubation periods in the presence of IFNγ demonstrated that *Stat1* is a driver

of bacterial clearance (Fig. 4F), which is consistent with previous reports[33,34]. Moreover, we analyzed the expression of both silenced TFs on their respective TF module counterparts in order to determine regulatory relationship between the two modules (Fig. 4G, H). We observed that silencing of *Zfp597* induced a significant increase in *Stat1* expression whereas *Stat1* silencing did not significantly alter *Zfp597* expression (Fig. 4G, H). This suggests a regulatory relationship between the two modules, with *Zfp597* inhibiting the expression of *Stat1*, which belongs to the opposite module.

In summary, the TFs characteristic of the detected cell states could be harnessed to modulate the immune response to pathogens by inducing a transition of cell states.

Integration across pathogens identifies common and unique cell states in time and space

As previously described, a major bottleneck of previous studies is the inability to compare the immune response across pathogens and timepoints. To address this issue, we set out to unify the previously detected cell states across different datasets by combining similar states. As a result, we obtained between 5 and 45 unique states for each cell type. We observed that the majority of functional states is homogeneous although some states display heterogeneous functionalities shared by other states (Fig. S5). Similar to the analysis conducted for individual datasets, NK and B cells have the lowest number of unique states whereas T cells have the highest. Next, we leveraged this integrated collection to identify functional states common and unique in the response to different pathogens. Interestingly, we observed largely distinct responses to different types of pathogens for most of the cell types, underscoring the previously reported predominance of pathogen-specific immune responses[35] (Fig. 5A). Finally, we set out to interrogate the changes in cell states at different timepoints of an infection. We analyzed the *Mycobacterium smegmatis* infection for the six cell types and observed a conserved functional state for T cells, NK cells, and monocytes across the three timepoints, respectively (Fig. 5B). Indeed, no functional diversity is observed for T cells, which are in one conserved state across the 7 days. However, B cells and DCs have conserved and unique states, with the functional diversity of DCs increasing at day 7. We noticed a shift of functional B cell states between the first and second day, mainly characterized by the differential expression of IgD (Fig. (Fig.5C)5C) [36]. Furthermore, we observed that the functional diversity of DCs at day 7 is characterized by three functional states (Fig. 5D) and could reflect differential DCs maturation during the inflammatory response, as reported in previous reports[37]. In addition, the functional state

CS3 is the most different with the expression of *Cd86*, *Cd4*, and especially *Ccl22*, suggesting this state to be actively recruiting other cells, such as invariant NKT or regulatory T cells, in response to the infection[38-41].

Discussion

In this study, we developed FunPart, a single-cell-based computational method to dissect the heterogeneous cellular response of immune cells to pathogens. In particular, this method is conceptually different from traditional clustering methods[42] as it accounts for functional aspects by identifying specific set of genes required to belong to the same immune process. Moreover, the striking difference between our approach and current clustering methodologies can be exemplified in the context of B cell states. Although traditional clustering methods detected 11 memory B cell states in a recent study, only a few states exhibited significant differences[43]. This is in accordance with our observation that B cells do not exhibit a high functional diversity with respect to immune processes. Furthermore, it was not unexpected to identify the largest number of functional states for T cells[44,45]. The differential diversity between B and T cells was observed at the marker expression level, initially used to distinguish cells (sub)types[21], but not fully explanatory of the functional diversity captured. Thus, the main advantage of our approach is that it mainly captures functional rather than transcriptional heterogeneity. Moreover, FunPart provides modules of genes used to identify the functional cell states and the immune processes[11] to which they belong. As a result, we were able to compile a *Catalogus Immune Muris*, the most comprehensive atlas of immune cell states currently available to the research community.

In addition, the *Catalogus Immune Muris* contains a molecular characterization of each state that can be leveraged to design novel immunomodulatory strategies. Here, we showed that the cellular response to *Salmonella* infection can be modulated by inhibiting TFs from identified gene modules by FunPart to enhance or inhibit pathogen clearance. Indeed, as reported in previous studies, we found *Stat1* to be a driver of bacterial clearance[33,34], whereas we identified *Zfp597*, a previously unreported TF, to have propathogenic effects. We showed that perturbation of TFs predicted to be characteristic of two macrophage cell states allows the modulation of their response to the infection by a switch between functional cell states. Moreover, our analysis suggests a regulatory relationship between the two modules where *Zfp597* inhibits the expression of *Stat1*. Therefore, targeting the identified TFs provides a rationale strategy for immunomodulatory therapies[46,47]. Nevertheless, the development of

novel immunomodulatory therapies typically relies on the utilization of drugs and compounds to alter cellular functions[48,49]. In this regard, a limitation of the presented strategy is that it solely considers modules composed of transcription factors that are potentially difficult to target.

Finally, the strategy implemented in FunPart could be of use for deciphering and characterizing functional heterogeneity within cell populations in diverse pathological and physiological conditions. Indeed, our method is not biased by the cell type it analyzes and thus could be applied to any cell type in any tissue or condition. Although FunPart currently identifies modules enriched in immune cell processes, it can be easily adapted to other genesets characteristic of any biological process. For instance, it could be applied to study the functional impairment of cell (sub)types in liver-related diseases[50,51]. Indeed, it is known that the cellular location around the lobule plays an important role for their function[52], however the dysregulations imparing the hepatocytes functions is not well defined[51,53]. The identification and characterization of such functional subtypes could help improving regenerative medicine strategies[54].

In summary, we presented a computational strategy for resolving functional cell states in the context of infections and identifying TFs involved in the maintenance of these states. We expect our approach to be of great utility for deciphering and characterizing functionally distinct cell states in physiological and pathological conditions. Moreover, application of our method to 114 datasets created a *Catalogus Immune Muris*, which we believe to be of great utility in the development of novel immunomodulatory therapies.

References

1 Chaplin DD. Overview of the immune response. *Journal of Allergy and Clinical Immunology* 2010; **125**: S3–S23.

2 Iwasaki A, Medzhitov R. Control of adaptive immunity by the innate immune system. *Nat Immunol* 2015; **16**: 343–353.

3 Lawrence T, Natoli G. Transcriptional regulation of macrophage polarization: enabling diversity with identity. *Nat Rev Immunol* 2011; **11**: 750–761.

4 MacParland SA, Liu JC, Ma X-Z, Innes BT, Bartczak AM, Gage BK *et al.* Single cell RNA sequencing of human liver reveals distinct intrahepatic macrophage populations. *Nat Commun* 2018; **9**: 4383.

5 Mould KJ, Jackson ND, Henson PM, Seibold M, Janssen WJ. Single cell RNA sequencing identifies unique inflammatory airspace macrophage subsets. *JCI Insight* 2019; **4**: e126556.

6 the Immunological Genome Consortium, Gautier EL, Shay T, Miller J, Greter M, Jakubzick C *et al.* Gene-expression profiles and transcriptional regulatory pathways that underlie the identity and diversity of mouse tissue macrophages. *Nat Immunol* 2012; **13**: 1118–1128.
7 Muñoz JF, Gade L, Chow NA, Loparev VN, Juieng P, Berkow EL *et al.* Genomic insights into multidrug-resistance, mating and virulence in Candida auris and related emerging species. *Nat Commun* 2018; **9**: 5346.
8 The International IBD Genetics Consortium (IIBDGC), Jostins L, Ripke S, Weersma RK, Duerr RH, McGovern DP *et al.* Host–microbe interactions have shaped the genetic architecture of inflammatory bowel disease. *Nature* 2012; **491**: 119–124.
9 Saliba A-E, Li L, Westermann AJ, Appenzeller S, Stapels DAC, Schulte LN *et al.* Single-cell RNA-seq ties macrophage polarization to growth rate of intracellular Salmonella. *Nat Microbiol* 2016; **2**: 16206.
10 Avraham R, Haseley N, Brown D, Penaranda C, Jijon HB, Trombetta JJ *et al.* Pathogen Cell-to-Cell Variability Drives Heterogeneity in Host Immune Responses. *Cell* 2015; **162**: 1309–1321.
11 Singhania A, Graham CM, Gabryšová L, Moreira-Teixeira L, Stavropoulos E, Pitt JM *et al.* Transcriptional profiling unveils type I and II interferon networks in blood and tissues across diseases. *Nat Commun* 2019; **10**: 2887.
12 Mantovani A, Sica A, Sozzani S, Allavena P, Vecchi A, Locati M. The chemokine system in diverse forms of macrophage activation and polarization. *Trends in Immunology* 2004; **25**: 677–686.
13 Myers LM, Tal MC, Torrez Dulgeroff LB, Carmody AB, Messer RJ, Gulati G *et al.* A functional subset of CD8+ T cells during chronic exhaustion is defined by SIRPα expression. *Nat Commun* 2019; **10**: 794.
14 Sica A, Mantovani A. Macrophage plasticity and polarization: in vivo veritas. *J Clin Invest* 2012; **122**: 787–795.
15 Siewiera J, Gouilly J, Hocine H-R, Cartron G, Levy C, Al-Daccak R *et al.* Natural cytotoxicity receptor splice variants orchestrate the distinct functions of human natural killer cell subtypes. *Nat Commun* 2015; **6**: 10183.
16 Liu C, Li Y, Yu J, Feng L, Hou S, Liu Y *et al.* Targeting the shift from M1 to M2 macrophages in experimental autoimmune encephalomyelitis mice treated with fasudil. *PLoS One* 2013; **8**: e54841.
17 Zhang F, Parayath NN, Ene CI, Stephan SB, Koehne AL, Coon ME *et al.* Genetic programming of macrophages to perform anti-tumor functions using targeted mRNA nanocarriers. *Nat Commun* 2019; **10**: 3974.
18 Luecken MD, Theis FJ. Current best practices in single-cell RNA-seq analysis: a tutorial. *Mol Syst Biol* 2019; **15**. doi:10.15252/msb.20188746.
19 Ekiz HA, Conley CJ, Stephens WZ, O'Connell RM. CIPR: a web-based R/shiny app and R package to annotate cell clusters in single cell RNA sequencing experiments. *BMC Bioinformatics* 2020; **21**: 191.
20 Butler A, Hoffman P, Smibert P, Papalexi E, Satija R. Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nat Biotechnol* 2018; **36**: 411–420.

21 Zhang X, Lan Y, Xu J, Quan F, Zhao E, Deng C *et al.* CellMarker: a manually curated resource of cell markers in human and mouse. *Nucleic Acids Research* 2019; **47**: D721–D728.
22 Kursa MB, Rudnicki WR. Feature Selection with the **Boruta** Package. *J Stat Soft* 2010; **36**. doi:10.18637/jss.v036.i11.
23 Carreras-González A, Barriales D, Palacios A, Montesinos-Robledo M, Navasa N, Azkargorta M *et al.* Regulation of macrophage activity by surface receptors contained within Borrelia burgdorferi-enriched phagosomal fractions. *PLoS Pathog* 2019; **15**: e1008163.
24 Dull T, Zufferey R, Kelly M, Mandel RJ, Nguyen M, Trono D *et al.* A third-generation lentivirus vector with a conditional packaging system. *J Virol* 1998; **72**: 8463–8471.
25 Rubinson DA, Dillon CP, Kwiatkowski AV, Sievers C, Yang L, Kopinja J *et al.* A lentivirus-based system to functionally silence genes in primary mammalian cells, stem cells and transgenic mice by RNA interference. *Nat Genet* 2003; **33**: 401–406.
26 Stuart T, Butler A, Hoffman P, Hafemeister C, Papalexi E, Mauck WM *et al.* Comprehensive Integration of Single-Cell Data. *Cell* 2019; **177**: 1888-1902.e21.
27 Lien C, Fang C-M, Huso D, Livak F, Lu R, Pitha PM. Critical role of IRF-5 in regulation of B-cell differentiation. *Proceedings of the National Academy of Sciences* 2010; **107**: 4664–4668.
28 Minamino K, Takahara K, Adachi T, Nagaoka K, Iyoda T, Taki S *et al.* IRF-2 regulates B-cell proliferation and antibody production through distinct mechanisms. *International Immunology* 2012; **24**: 573–581.
29 Metidji A, Rieder SA, Glass DD, Cremer I, Punkosdy GA, Shevach EM. IFN-α/β Receptor Signaling Promotes Regulatory T Cell Development and Function under Stress Conditions. *JI* 2015; **194**: 4265–4276.
30 Windt GJW, Pearce EL. Metabolic switching and fuel choice during T-cell differentiation and memory development. *Immunol Rev* 2012; **249**: 27–42.
31 Lane K, Andres-Terre M, Kudo T, Monack DM, Covert MW. Escalating Threat Levels of Bacterial Infection Can Be Discriminated by Distinct MAPK and NF-κB Signaling Dynamics in Single Host Cells. *Cell Systems* 2019; **8**: 183-196.e4.
32 Richter E, Mostertz J, Hochgräfe F. Proteomic discovery of host kinase signaling in bacterial infections. *Prot Clin Appl* 2016; **10**: 994–1010.
33 Durbin JE, Hackenmiller R, Simon MC, Levy DE. Targeted Disruption of the Mouse Stat1 Gene Results in Compromised Innate Immunity to Viral Disease. *Cell* 1996; **84**: 443–450.
34 Meraz MA, White JM, Sheehan KCF, Bach EA, Rodig SJ, Dighe AS *et al.* Targeted Disruption of the Stat1 Gene in Mice Reveals Unexpected Physiologic Specificity in the JAK–STAT Signaling Pathway. *Cell* 1996; **84**: 431–442.
35 Rivera A, Siracusa MC, Yap GS, Gause WC. Innate cell communication kick-starts pathogen-specific immunity. *Nat Immunol* 2016; **17**: 356–363.
36 Geisberger R, Lamers M, Achatz G. The riddle of the dual expression of IgM and IgD. *Immunology* 2006; **0**: 060526021554006-???

37 Domínguez PM, Ardavín C. Differentiation and function of mouse monocyte-derived dendritic cells in steady state and inflammation. *Immunological Reviews* 2010; **234**: 90–104.
38 Rapp M, Wintergerst MWM, Kunz WG, Vetter VK, Knott MML, Lisowski D *et al.* CCL22 controls immunity by promoting regulatory T cell communication with dendritic cells in lymph nodes. *Journal of Experimental Medicine* 2019; **216**: 1170–1181.
39 Vulcano M, Albanesi C, Stoppacciaro A, Bagnati R, D'Amico G, Struyf S *et al.* Dendritic cells as a major source of macrophage-derived chemokine/CCL22 in vitro and in vivo. *Eur J Immunol* 2001; **31**: 812–822.
40 Bischoff L, Alvarez S, Dai DL, Soukhatcheva G, Orban PC, Verchere CB. Cellular Mechanisms of CCL22-Mediated Attenuation of Autoimmune Diabetes. *JI* 2015; **194**: 3054–3064.
41 Bialecki E, Macho Fernandez E, Ivanov S, Paget C, Fontaine J, Rodriguez F *et al.* Spleen-Resident CD4+ and CD4− CD8α− Dendritic Cell Subsets Differ in Their Ability to Prime Invariant Natural Killer T Lymphocytes. *PLoS ONE* 2011; **6**: e26919.
42 Kiselev VY, Andrews TS, Hemberg M. Challenges in unsupervised clustering of single-cell RNA-seq data. *Nat Rev Genet* 2019; **20**: 273–282.
43 King HW, Orban N, Riches JC, Clear AJ, Warnes G, Teichmann SA *et al.* Antibody repertoire and gene expression dynamics of diverse human B cell states during affinity maturation. Immunology, 2020 doi:10.1101/2020.04.28.054775.
44 Newell EW, Sigal N, Bendall SC, Nolan GP, Davis MM. Cytometry by Time-of-Flight Shows Combinatorial Cytokine Expression and Virus-Specific Cell Niches within a Continuum of CD8+ T Cell Phenotypes. *Immunity* 2012; **36**: 142–152.
45 Sallusto F, Lanzavecchia A. Heterogeneity of CD4 $^{+}$ memory T cells: Functional modules for tailored immunity: Highlights. *Eur J Immunol* 2009; **39**: 2076–2082.
46 Lee J-U, Kim L-K, Choi J-M. Revisiting the Concept of Targeting NFAT to Control T Cell Immunity and Autoimmune Diseases. *Front Immunol* 2018; **9**: 2747.
47 Cuadrado A, Manda G, Hassan A, Alcaraz MJ, Barbas C, Daiber A *et al.* Transcription Factor NRF2 as a Therapeutic Target for Chronic Diseases: A Systems Medicine Approach. *Pharmacol Rev* 2018; **70**: 348–383.
48 Iqbal Yatoo Mohd, Hamid Z, Rather I, Nazir QUA, Bhat RA, Ul Haq A *et al.* Immunotherapies and immunomodulatory approaches in clinical trials - a mini review. *Human Vaccines & Immunotherapeutics* 2021; : 1–13.
49 Davis JS, Ferreira D, Paige E, Gedye C, Boyle M. Infectious Complications of Biological and Small Molecule Targeted Immunomodulatory Therapies. *Clin Microbiol Reviews* 2020; **33**: e00035-19, /cmr/33/3/CMR.00035-19.atom.
50 Wen Y, Lambrecht J, Ju C, Tacke F. Hepatic macrophages in liver homeostasis and diseases-diversity, plasticity and therapeutic opportunities. *Cell Mol Immunol* 2021; **18**: 45–56.
51 Albillos A, Lario M, Álvarez-Mon M. Cirrhosis-associated immune dysfunction: Distinctive features and clinical relevance. *Journal of Hepatology* 2014; **61**: 1385–1396.
52 Trefts E, Gannon M, Wasserman DH. The liver. *Current Biology* 2017; **27**: R1147–R1151.

53 Gissen P, Arias IM. Structural and functional hepatocyte polarity and liver disease. *Journal of Hepatology* 2015; **63**: 1023–1037.
54 Bhatia SN, Underhill GH, Zaret KS, Fox IJ. Cell and tissue engineering for liver disease. *Sci Transl Med* 2014; **6**: 245sr2-245sr2.

Acknowledgements

CIC bioGUNE thanks the MCI for the Severo Ochoa Excellence accreditation (SEV-2016-0644).

Conflict of Interest Statement

The authors declare no competing interests.

Author contribution Statement

C.B. collected and processed the data, performed the analyses, developed the R package, the shiny interface, and wrote the manuscript. D.B. performed the experimental validation. A.S. performed the macrophages analysis. S.J. supervised the computational work and wrote the manuscript. S.R. supervised the computational work. Y.A.M. supervised the computational work. J.A. supervised the experimental work and wrote the manuscript. A.d.S. supervised the project, conceived the idea and wrote the manuscript.

Ethics Statement

All the assays performed were approved by the competent authority (Diputación de Bizkaia) under European and Spanish directives. CIC bioGUNE is accredited by AAALAC Intl.

Funding Statement

C.B. is supported by funding from the Luxembourg National Research Fund (FNR) within PARK-QC DTU (PRIDE17/12244779/PARK-QC). D.B. is supported by an FPI fellowship from the Spanish Ministry of Science and Innovation (MCI) (BES-2016-078437). J.A. is supported by a grant from the MCI co-financed with FEDER funds (RTI2018-096494-B-100).

Availability of Data and Materials

The accession number of the datasets used are available in the table S1. The integrated datasets for each cell type are available at: https://gitlab.com/C.Barlier/immunofunmap.git. The maps are available via an interface developed with Shiny at: https://gitlab.com/C.Barlier/immunofunmap.git.

The functional states identification algorithm is an R package named FunPart available at: https://github.com/BarlierC/FunPart.git.

Figures

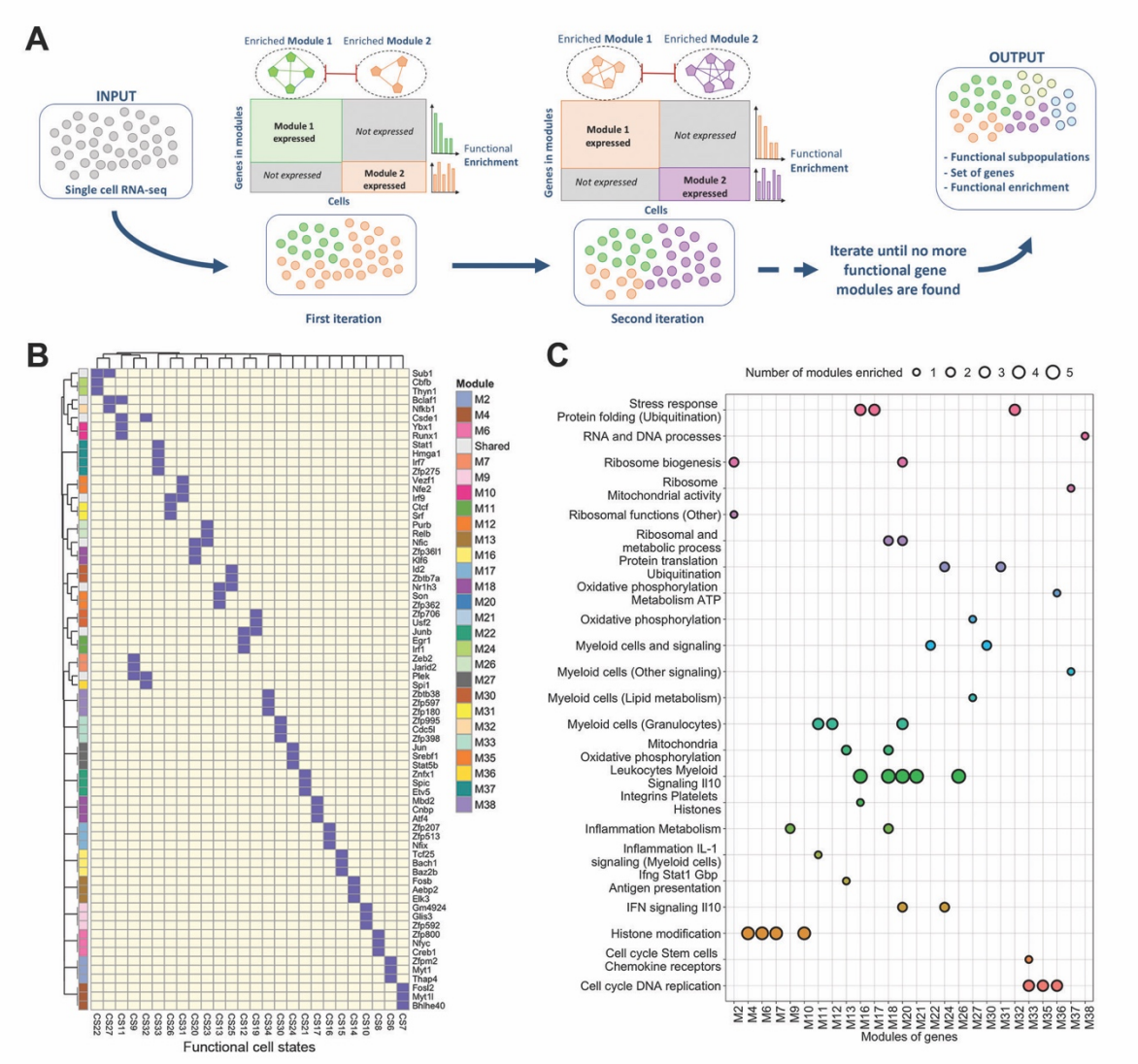

Figure 1. FunPart general workflow and validation.

(A) General workflow of the functional states identification and characterization. The computational method we developed, named FunPart, takes single cell RNA-seq data of one cell type as an input, to identify functional states based on functional modules of genes. The method searches for modules exclusively expressed in one group of cells and belonging to the same immune process. Cells are recursively splitted in two groups until no more functionally relevant modules associated to new states can be found. (B) Binary heatmap of the 26 terminal genes modules identified by FunPart for the macrophages functional cell states CS. Only TFs are displayed. (C) Functional enrichment of these 26 terminal gene modules. Each immune process has a different color, the size of the dots represents the number of gene modules enriched in the specific process. Intermediate gene modules are not displayed.

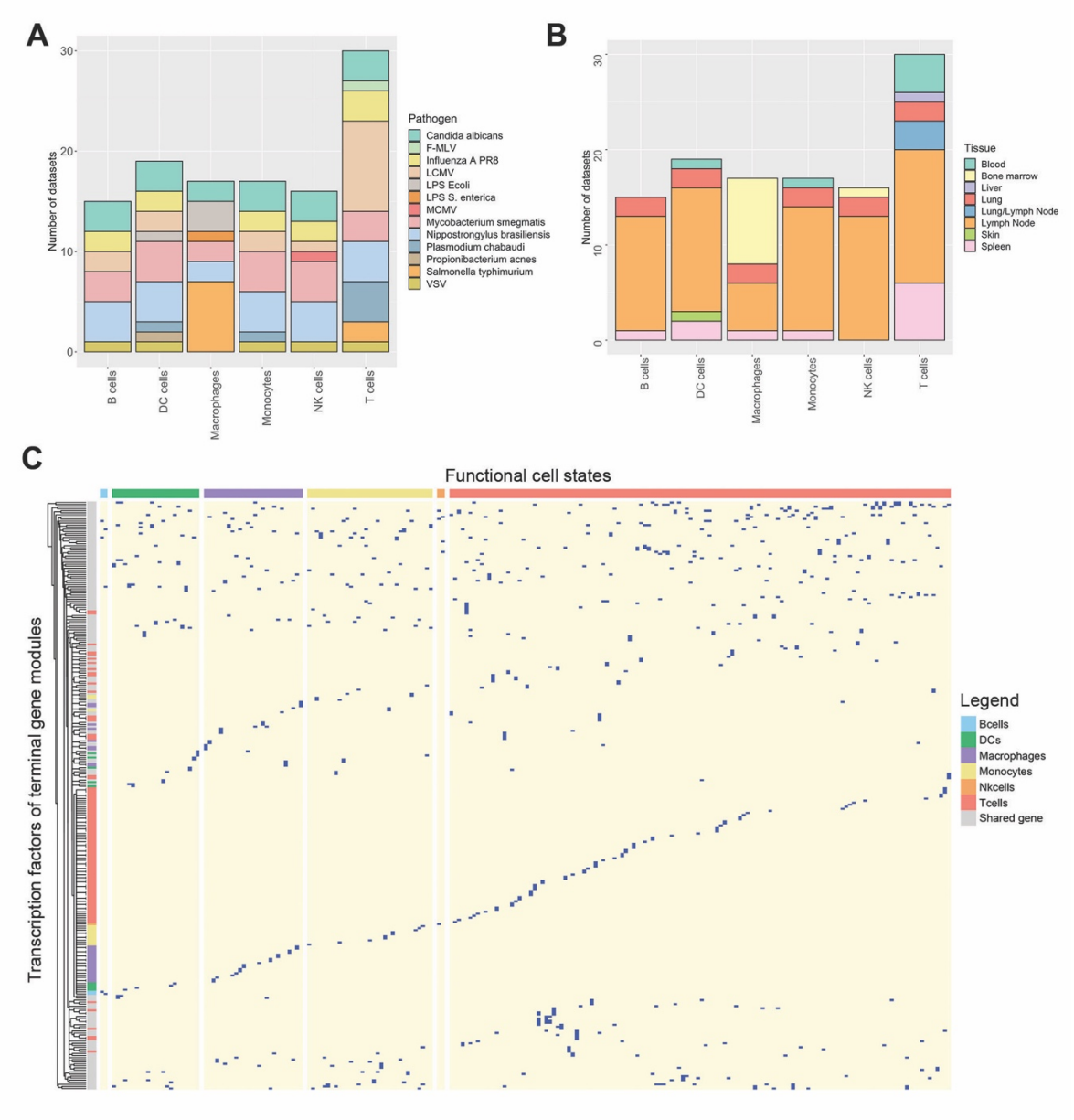

Figure 2. Overview of the Catalogus Immune Muris content.

(A,B) Composition of the Catalogus Immune Muris. Repartition by immune cell type of the (A) nine pathogens and (B) seven tissues across the 114 datasets. (C) Binary heatmap displaying the terminal gene modules identified by FunPart for each functional cell states belonging to one of the six broad immune cell type. Shared genes, colored in grey, correspond to transcription factors found in more than one terminal gene modules. Only TFs of terminal gene modules are displayed.

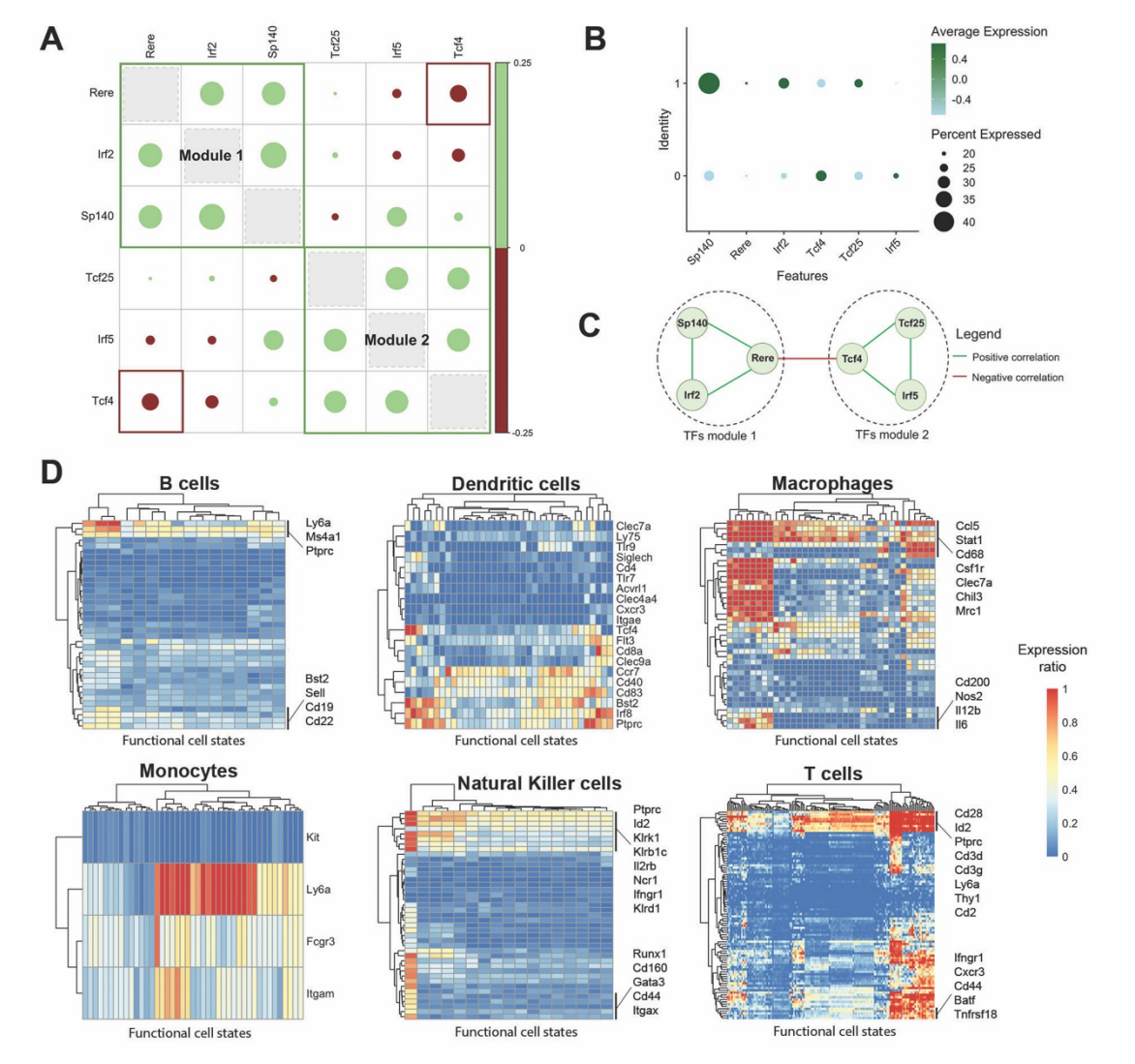

Figure 3. Functional cell states analysis and characterization.

(A) Correlation plot and (B) dotplot of the functional TFs characterizing two B cells functional states in LCMV infection at time point 72h. Colored boxes in (A) indicate correlations considered by the algorithm with green boxes indicating cliques of genes and red boxes the negative correlation considered as significant. (C) Network representation of the significant edges retained by the algorithm for the six TFs shown in (A). Each module consists of a clique of three transcription factors positively correlated together. The negative correlation between the two modules is supported by the interaction between Tcf4 and Rere. (D) Heatmaps showing the expression ratio of the cell markers, extracted using Boruta, for each functional cell state. Identified cell states are in columns and markers in rows. A ratio of one corresponds to the marker being expressed in all cells of the functional cell state whereas a ratio of zero translates to the absence of its expression in the cell state.

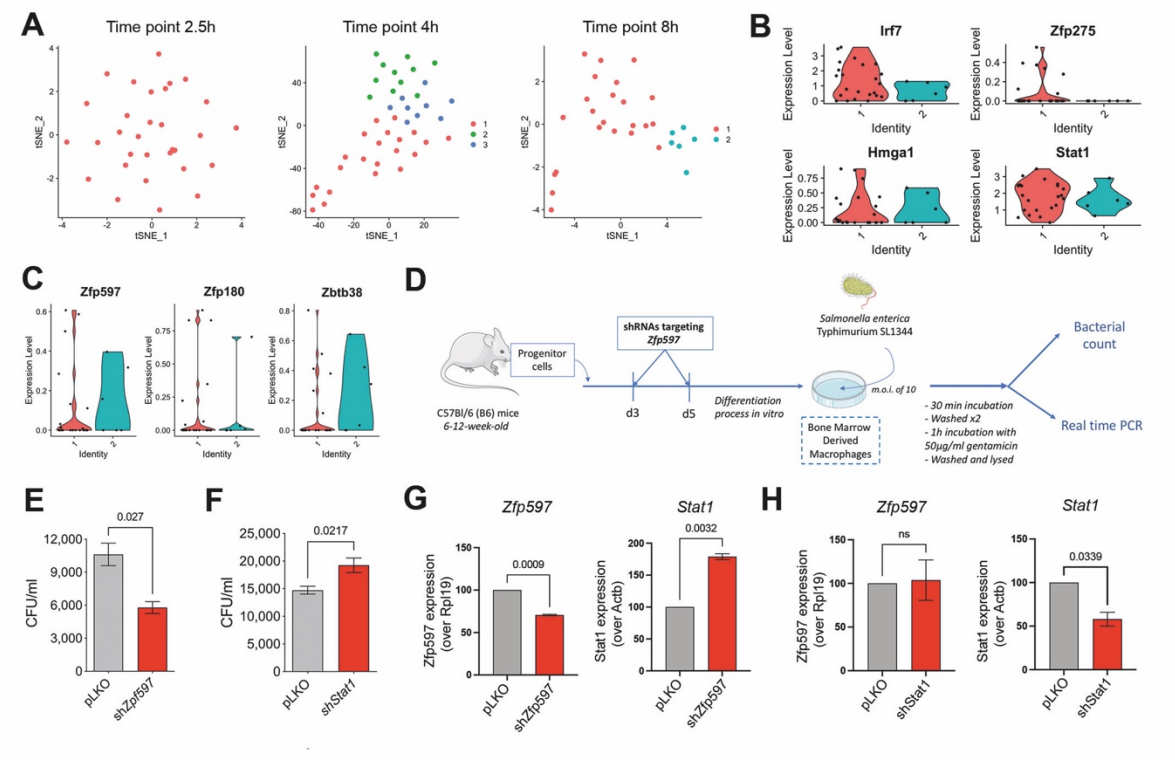

Figure 4. Immunomodulation of macrophage responses and functional states analysis. (A) t-SNE displaying functional states identified by FunPart across three time points for macrophages infected by Salmonella typhimurium. (B,C) Violin plots showing the expression levels for the two functional states identified at time point 8h of (B) the first module composed of Irf7, Zfp275, Hmga1, Stat1 and (C) the second module composed of Zp597, Zfp180, Zbtb38. (D) Summary of the experimental design used to validate Zfp597 and Stat1 as immunomodulators. (E,F,G,H) Differential survival of S. enterica typhimurium in Zfp597-silenced and Stat1-silenced macrophages compared to their respective pLKO controls. (E,F) Colony-forming units recovered from silenced and control-transfected BMMs infected with Salmonella at an m.o.i of 10 for (E) Zfp597 and (F) Stat1. (G,H) Zfp597 and Stat1 gene expression levels in macrophages lentivirally infected with shRNAs targeting the gene or controls (plKO). The results are represented as average ± SE of 3 independent mice per silencing. The p values were calculated by paired Student's t test. A result is considered as significant if its p-value is less than 0.05.

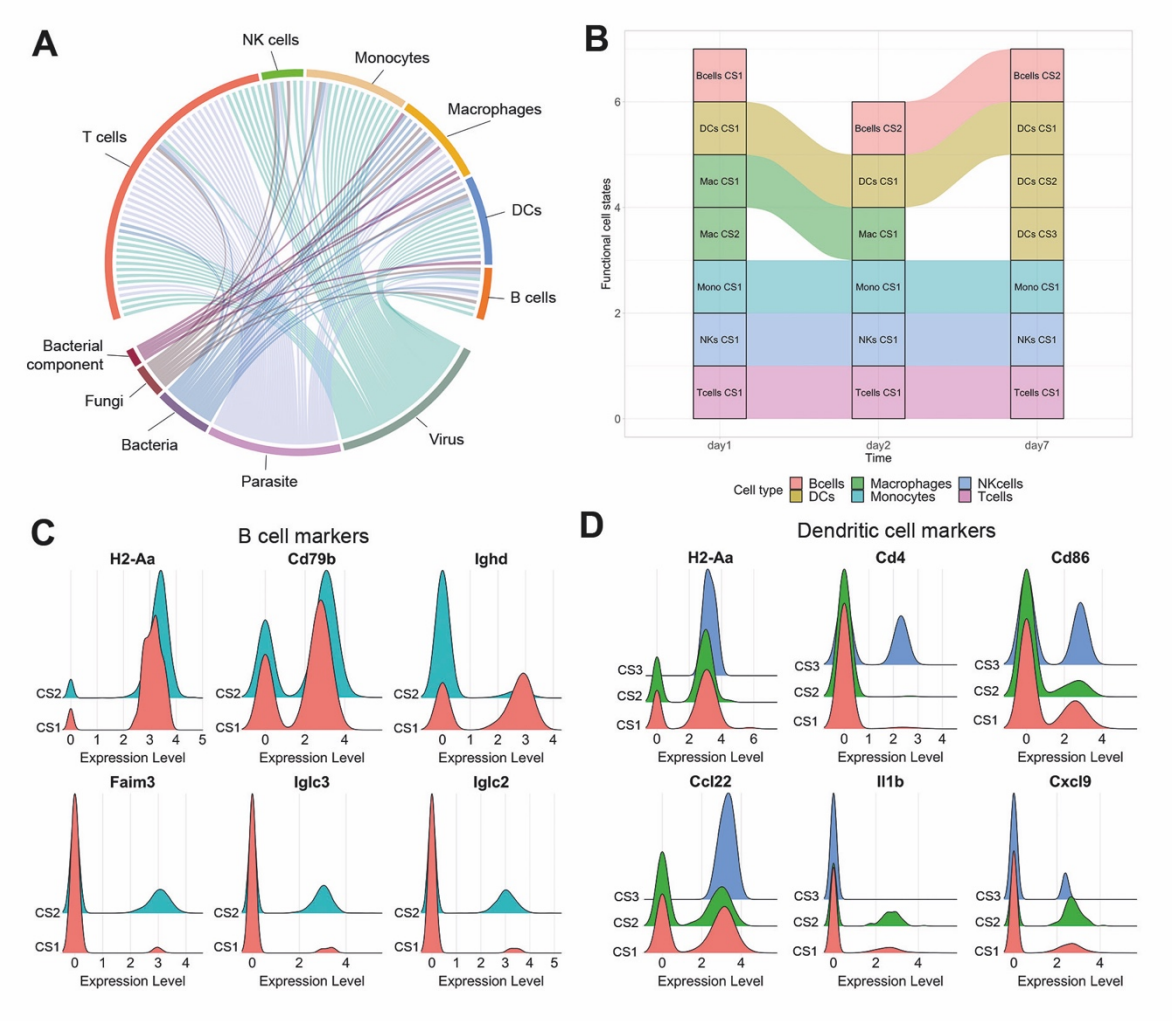

Figure 5. Metadata analysis of functional cell states.

(A) Chord diagram representing the common and unique cell states across pathogen types infections. (B) Alluvial plot of the functional states identified for the six immune cell types infected by Mycobacterium smegmatis across three time points. CS: functional Cell State, DCs: dendritic cells, Mac: macrophages, Mono: monocytes, NKs: Natural Killer cells. (C,D) Markers expression for (C) the two B cells and (D) the three DCs functional cell states shown in (B). Distributions displayed include all applicable time points.

4.2.3 Supplementary Information

Supplementary Methods

Functional partitioning algorithm

In order to reliably identify and characterize functionally relevant cell states, we developed a network-based approach combined with a recursive hierarchical clustering named FunPart. The algorithm is composed of four main parts:

(1) Cleaning and normalization of the data: The algorithm accepts any type of data (counts, UMI or normalized). In case raw data are provided, a normalization procedure will be performed using Seurat. Outliers cells will be removed using a Rosner test on the number of genes expressed in each cell. Finally, two last quality control steps are performed on the genes: (1) any gene expressed in less than 5% of the cells will not be considered and (2) genes that are too lowly expressed are removed, with a gene too lowly expressed falling below the 5% of genes expression sum distribution.

(2) Network-based set of genes identification: To identify functional sets of genes, a correlation network is constructed around all the genes. Based on the correlation scores distribution, the 2.5% of each tail are considered to be the strongest interactions and are kept for the following steps: (1) identification of cliques of transcription factors (TFs) that are positively correlated together, (2) filtering of cliques that are not unique, with a unique clique defined as a clique with less than 70% of common TFs, (3) an expression score reflecting the average expression of the clique is calculated and only the top 30% is kept, (4) in order to identify antagonistic pairs of cliques, a negative score is calculated between each pair of positive clique identified in 3., (5) if less than ten antagonistic pairs are found, all of them are used in step 6., however if more than ten are found, only the top 5% most negatives are kept, (6) the top target genes are identified for each TF of the two modules.

(3) Functional characterization of the set of genes: An enrichment analysis is then performed on the candidate pairs of modules, using manually annotated immune modules by Singhania et al. The functional enrichment is performed using the clusterProfiler R package as the following: (1) all the genes profiled in the dataset are used as the universe and genes of the module considered are used to perform the comparison, (2) a multiple test correction (Bonferroni) is performed and only the enriched annotations with an adjusted p-value less than 5% are kept, (3) enriched categories mapped to only one gene of the set are not considered, (4) a score consisting in the sum of all the resulting gene

ratio for the module is computed, (5) the two negatively connected modules need to be both enriched to be considered, (6) each pair of negatively connected modules is ranked according to the computed score. The top one enriched set, consisting of two gene modules, is then used for the hierarchical clustering.

(4) Recursive unsupervised hierarchical clustering: In order to investigate each level of resolution, a recursive binary splitting is used (unsupervised hierarchical clustering). For each level, a bi-clustering is performed by building a heatmap using the cells of the corresponding level as well as the identified genes of the two gene modules. The general workflow is the following: (1) at each level, a hierarchical tree is constructed using the single cell expression data, the best set of genes and the Pearson correlation measure using the complete aggregation approach, (2) the first level of the cells dendrogram is used to perform the binary cutting with k = 2, (3) the two distinct groups of cells identified will then be splitted separately as explained in steps 1 to 2. The algorithm stops once the groups of cells are homogeneous and no more functional gene modules are found.

FunPart deciphers functional diversity by identifying and using set of gene modules to pinpoint and characterize functional cell states. Each gene module identified is composed of TFs, forming a clique of positively co-expressed edges only, and their direct neighbor genes for which they have a strong positive interaction. Furthermore, these genes modules can be classified as intermediate modules or terminal modules. A genes module is intermediate if the group of cells identified is further splitted whereas a genes module is terminal if the group of cells identified is not further splitted (corresponds to a functional cell state and leaf in the hierarchical tree). Indeed, an intermediate gene module characterize a group of functional cell states whereas a terminal gene module characterizes a specific functional cell state.

The module attribution to a group of functional cell states or one functional cell state is performed for each binary splitting. Indeed, each binary splitting is performed using two gene modules, with each of them belonging to one of the two groups resulting from the split, according to FunPart rationale. Thus, the module attribution is performed based on the average number of cells expressing the TFs of the clique in the module. Each gene module is then assigned to the group (branch 0 or 1) in which it is expressed the most and classified as characterizing this group. This step allows the assignment of intermediate gene modules to group of functional cell states and terminal gene modules to specific functional cell states.

Validations and comparison with the state-of-the-art

The functional relevance of the predicted subpopulations by FunPart and Seurat was assessed as follow: for each dataset, a ROC test, using FindAllMarkers function from Seurat R package, has been applied to each predicted cluster; genes with an AUC greater or equal to 0.7 were considered as good candidates to classify the group of cells; genes were submitted to an enrichment analysis using annotated immune modules, a Benjamini-Hochberg correction and a p-adjusted value less than 5%. We then defined four classes to assess the functional relevance of the predictions based on each dataset:

- "True homogeneous": dataset for which one method do not identify subpopulations and the other one identifies some from which more than 50% are non-functional;
- "False homogeneous": dataset for which one method do not identify subpopulations but the other one identifies some from which more than 50% are functionally relevant;
- "True heterogeneous": dataset for which more than 50% of the cell states identified are functionally relevant;
- "False heterogeneous": dataset for which less than 50% of the cell states identified are functionally relevant.

The four non splitted datasets by both methods were discarded from this analysis. We computed a precision score such as precision = True / (True + False).

Characterization of functional cell states

The feature extractions were done using the R version of Boruta's algorithm, a wrapper built around the random forest classification algorithm, for each functional cell state. Boruta was used with default parameters and the following predictors and response vector:

- Predictors: matrix with features in columns and cells in rows. The features used consisted of the collected markers for the broad cell type of the functional cell state.
- Response vector: vector with two classes (binary classification), with class 1 for the cell state under consideration and class 0 for all the other cell states (background).

For each functional cell state, we kept markers classified as an important feature and then computed a fold change (FC) such as:

$$FC = \frac{mean(x_m^{cs})}{mean(x_m^{b})}$$

With m: marker, cs: functional cell state, b: background, x: gene expression.

A positive FC represents an overexpression of the marker in the functional cell states whereas a negative one represents a down-expression of the marker.
In order to compile markers profile for each functional cell states we identified, we computed cell expression ratios for each functional cell state and each extracted feature of the immune cell types. The ratios were computed for each functional cell states such as:

$$R_{cs}^{m} = \frac{\sum x_i^m}{n_{cs}}$$

With R the ratio, m the marker, cs the functional cell state, x the binary expression (0 or 1, with 1 = expressed) of the marker m in the cell i and n the total number of cells.

Metadata analysis

Data were integrated using the standard Seurat pipeline. Due to the high disparity between the number of cells, the integrations were performed in three steps with the biggest datasets (>1000 cells) being integrated together and then, integrated with the medium ones (>100 and <1000 cells) to finally be integrated with the smallest ones. The UMAP is computed, for each cell type, on the integrated data using Seurat and the functional set of genes characterizing the functional cell states identified using the functional splitting algorithm.

Real time PCR

The primers used corresponded to the genes Rpl19 (5'-GAC CAA GGA AGC ACG AAA GC-3' and 5'-CAG GCC GCT ATG TAC AGA CA-3'), Zfp597 (5'-ATC GGA TGA GCA GAG ACC AC-3' and 5'-TGA ACA ACG GGT GCA GCA AT-3'), Stat1 (5' -TCT GAA TAT TTC CCT CCT GGG- 3' and 5' -CGG AAA AGC AAG CGT AAT CT- 3') and Actb (5′-GAC GAT GCT CCC CGG GCT GTA TTC-3′ and 5′-TCT CTT GCT CTG GGC CTC GTC ACC-3′).

Supplementary figures

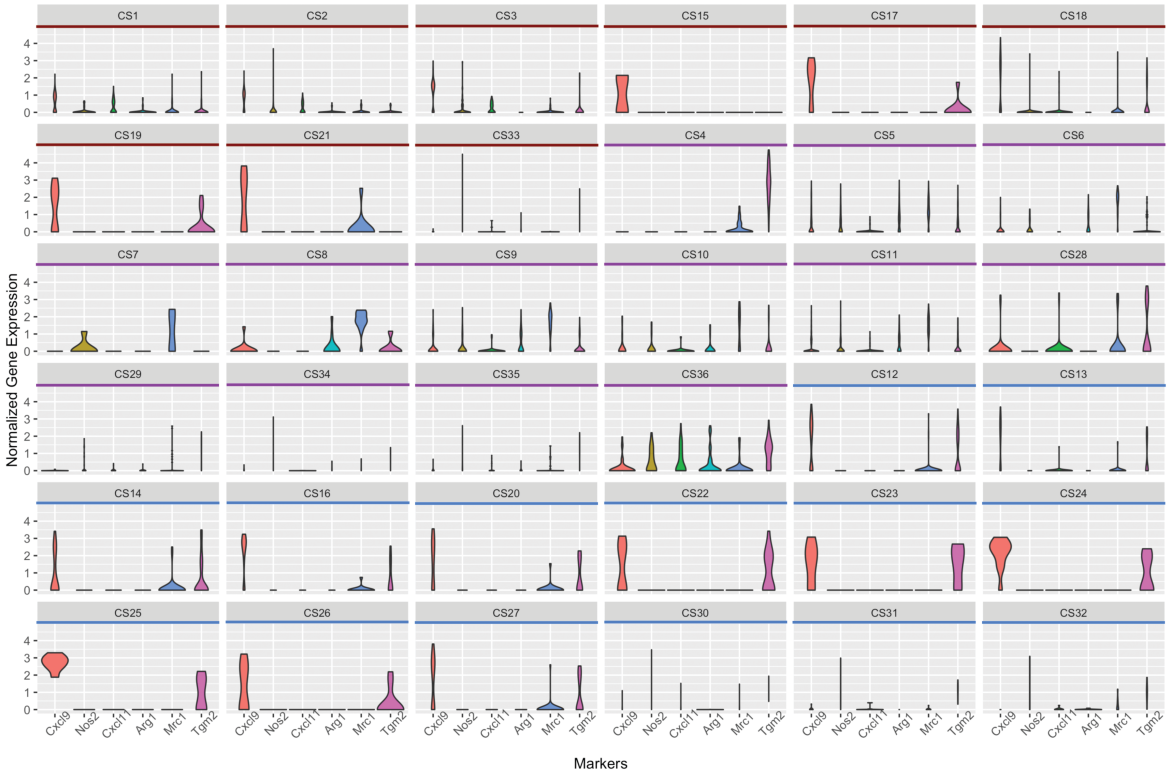

Fig. S1. M1-like and M2-like markers used to classify macrophages functional states. Representation of the M1-like (Cclx9, Nos2, Cxcl11) and M2-like (Arg1, Mrc1, Tgm2) markers distribution used to classify macrophages functional cell states (CS) as M1-like, M2-like and intermediate. CS underlined in red are classified as M1-like, CS underlined in purple are classified as M2-like and CS underlined in blue are classified as intermediate states.

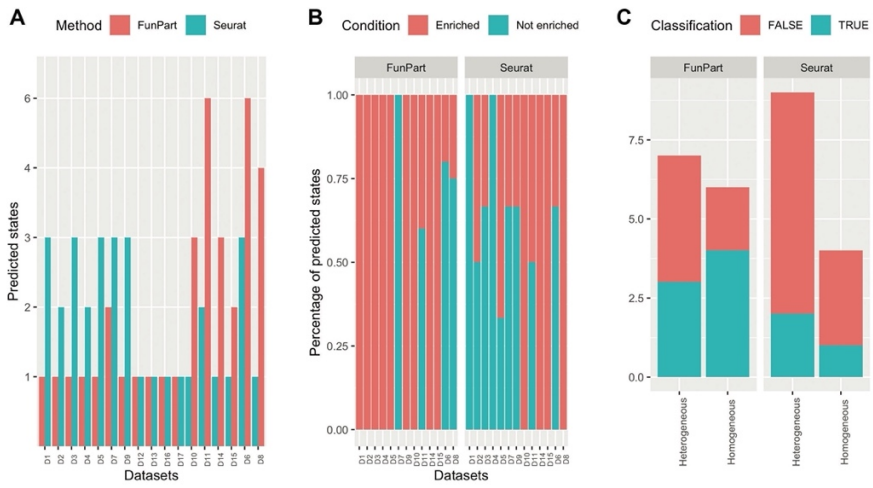

Fig. S2. FunPart validation and comparison to state-of-the-art.
(A) Predicted states by FunPart and Seurat for the 17 macrophages datasets. (B) Ratio of enriched and non-enriched predicted subpopulations for the 14 datasets for which FunPart and Seurat were not in agreement. Datasets D12, D13, D16 and D17 have been excluded from this analysis. (C) Assessment of the accuracy of both methods in distinguishing functional homogeneous datasets and identify functionally relevant subpopulations (True Heterogeneous). The computation of the different classifications is described in the Methods section.

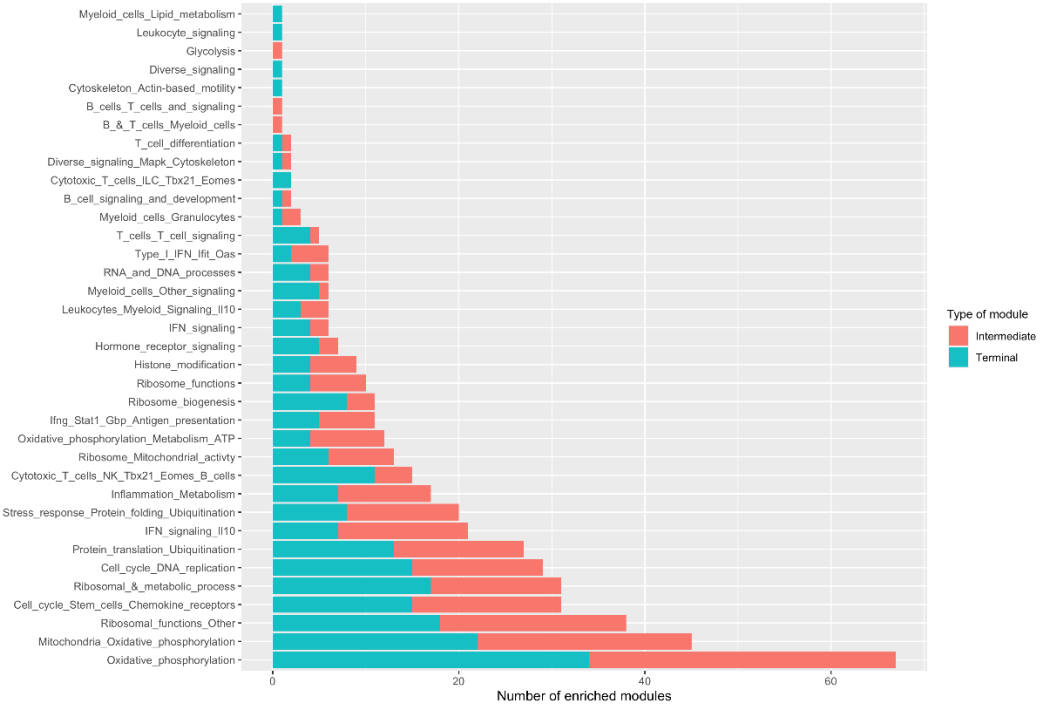

Fig. S3. Enrichment of the terminal and intermediate gene modules identified for T cells.

FunPart identified 132 terminal and 102 intermediate gene modules across the 30 T cells datasets analyzed that were enriched in diverse immune processes. Most of the modules are enriched in processes involved in broad processes such as oxidative phosphorylation, stress response and inflammation metabolism whereas fewer are enriched in more specific ones such as type I IFN and cytotoxic T cells processes.

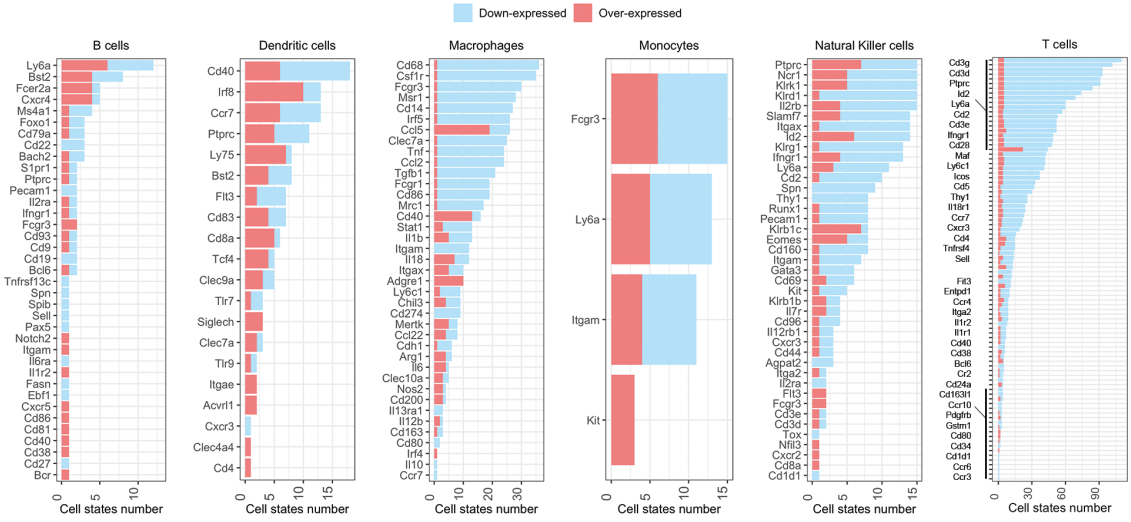

Fig. S4. Extracted features for each functional cell state.

Feature extraction was performed to identify important markers to classify the identified functional cell states. Stacked boxplots represent the frequency of each marker being found as important for the classification. Light blue parts represent markers found down-expressed in the specific functional cell considered and red parts represent over-expressed markers. We can observe that broad markers such as CD3 for T cells are more frequent than specific markers such as Tlr9 for dendritic cells, regardless of their expression level.

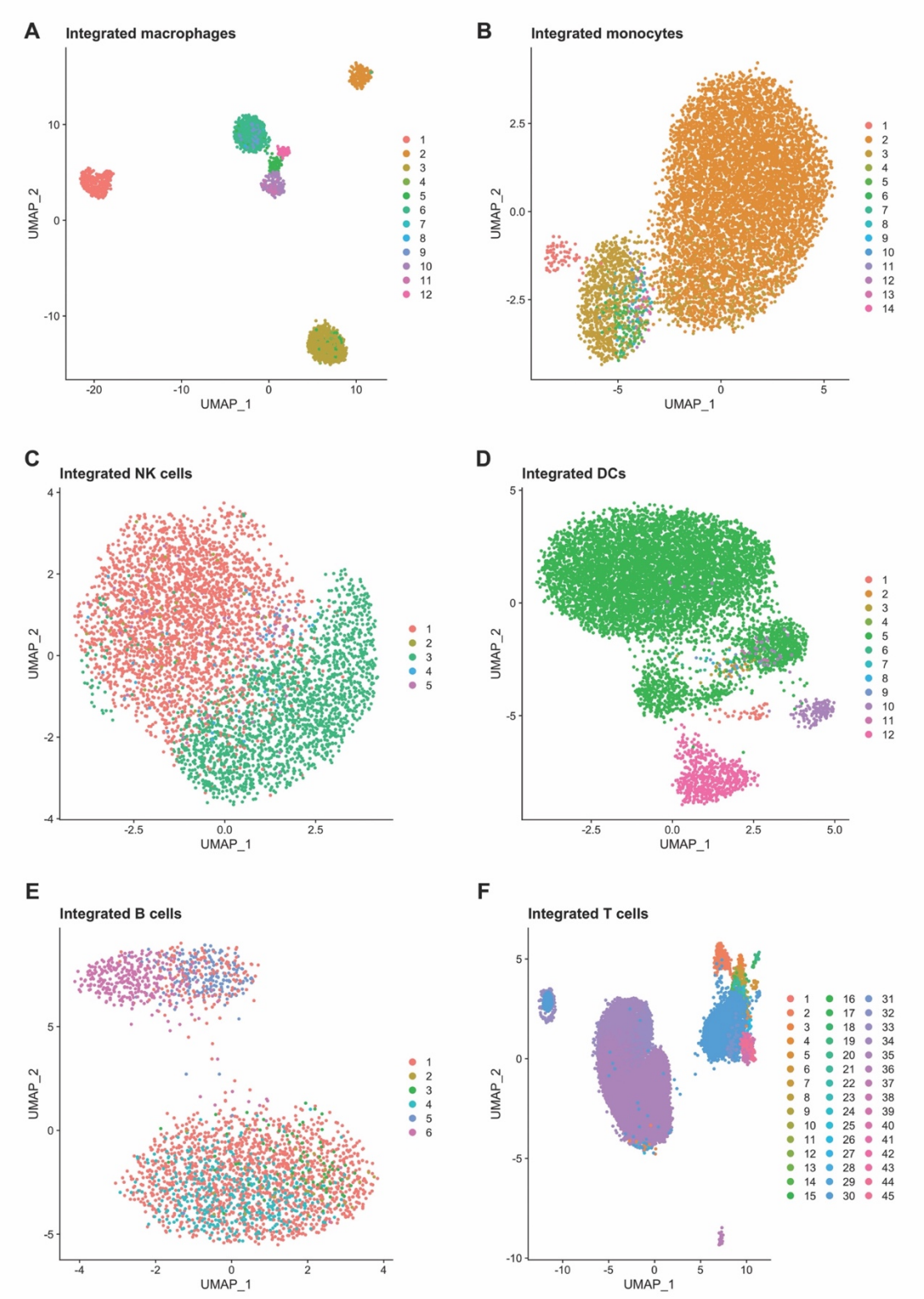

Fig. S5. UMAPs of the integrated data for the six immune cell types.

UMAP were computed for integrated data of the (A)macrophages, (B) monocytes, (C) natural killer (NK) cells, (D) Dendritic cells (DCs), (E) B cells and (F) T cells. They were built using all functionally relevant genes reported by FunPart for the non-integrated analysis for each cell type as features. Functional states identified after the integration analysis are displayed with some of them in intermediate states and not distinct. NK cells and B cells have the lowest number with 5 and 6 respectively compared to the T cells composed of 45 functional states

4.3 Deciphering impaired regulatory mechanisms in diseases

4.3.1 Preface

In this study entitled "*RNetDys: regulatory network inference to identify impaired interactions in diseases*" we present a multi-OMICS pipeline to infer comprehensive cell (sub)type and state specific GRNs and systematically identify transcriptional regulatory interactions impaired due to SNPs in diseases. RNetDys is a pipeline that aims at providing a better understanding of cell (sub)type and state specific regulatory mechanisms impaired in diseases due to SNPs. Indeed, the comprehensive view of cell (sub)type or state specific regulatory landscape impaired due to disease-related SNPs is a promising approach to have better transcriptomic regulatory mechanistic insights and guide the development of strategies for therapeutic intervention. Thus far, several strategies and methods have been developed to study the effect of SNPs and their involvement in diseases, but there is still a lack for a comprehensive view of the regulatory mechanisms that could be impaired.

In that regard, we propose RNetDys, a computational pipeline to infer comprehensive cell (sub)type or state specific GRNs and identify regulatory interactions impaired due to disease-related SNPs. We showed the better accuracy of RNetDys to infer cell (sub)types specific regulatory interactions including TF-genes and enhancer-promoters compared to state-of-the-art methods. Moreover, we applied our pipeline in five disease case studies and validated the relevance of the predicted impaired interactions using literature, GWAS and eQTL evidences. In summary, we provide a user-friendly pipeline to generate comprehensive cell (sub)type or state specific GRNs and identify transcriptional regulatory mechanisms impaired in diseases due to SNPs by leveraging the GRN information.

Contribution: I implemented the computational method, collected and processed the data, performed the benchmarking analysis, generated the cell (sub)type specific GRNs, collected the disease-related SNPs, performed the data analysis, and wrote the manuscript.

4.3.2 Manuscript

RNetDys: regulatory network inference to identify impaired interactions in diseases

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

1 Computational Biology Group, Luxembourg Centre for Systems Biomedicine (LCSB), University of Luxembourg, L-4362 Esch-sur-Alzette, Luxembourg

2 Computational Biology Group, CIC bioGUNE-BRTA (Basque Research and Technology Alliance), Derio, Spain 48160

3 Ikerbasque, Basque Foundation for Science, Bilbao, Bizkaia, 48012. Spain

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

Abstract

Gene regulation is a fundamental process largely controlled by transcription factors to activate or repress genes. The dysregulation of regulatory mechanisms due to SNPs can lead to non-physiological conditions such as disease development. However, regulatory dysregulations do not affect all cell types and subtypes equally. Therefore, having a comprehensive view of the cell (sub)type specific regulatory landscape is required to accurately decipher specific regulatory interactions impaired in diseases. Here, we present RNetDys, a pipeline that leverages multi-OMICS data to infer regulatory interactions mediated by TFs and enhancers of regulated genes for cell (sub)types or states, and to identify specific regulatory interactions impaired due to SNPs in diseases. We showed that the cell (sub)type specific GRNs inferred by RNetDys were more accurate compared to state-of-the-art methods. Moreover, we validated the ability of RNetDys to accurately identify impaired regulatory interactions due to SNPs in five disease case studies by leveraging the GRN information.

Introduction

Gene regulation is a complex and fundamental process that gives rise to highly heterogeneous gene expression signatures which define cell identity (Cooper, 2000). Indeed, transcription is largely controlled by transcription factors that bind to specific DNA loci such as promoter and enhancer regions to either express or repress gene expression (Latchman, 2011). This regulatory process is triggered in response to stimuli, and the cell (sub)type specific regulatory mechanisms are largely conferred by enhancers (Andersson *et al.*, 2014). Therefore, it plays a critical role to maintain the homeostasis, integrity and physiology of an organism (Wray *et al.*, 2003). The impairment of these regulatory interactions can lead to dysregulations that trigger pathological gene expression changes and 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 heart diseases (Hiramoto *et al.*, 2015; Akhlaghipour *et al.*, 2022). Notably, genome-wide association studies (GWAS) showed that the majority of disease-related genetic variants such as SNPs were found in enhancer regions (Claringbould and Zaugg, 2021). Thus, characterizing the gene regulatory network (GRN) describing the interactions mediated by TFs and enhancers of regulated genes is critical to understand the underlying mechanisms of gene regulation in both physiological and pathological conditions. Indeed, the characterization of the regulatory landscape impaired due to SNPs in diseases would provide better mechanistic insights and aid the development of strategies for therapeutic intervention (Uddin *et al.*, 2020).

Over the years, several GRN inference methods were developed to predict the interactions between genes using bulk transcriptomics data (Margolin *et al.*, 2006; Huynh-Thu *et al.*, 2010; Guo *et al.*, 2016). The emergence and fast development of single-cell based technologies enhanced the development of more refined computational methods to predict cell (sub)type specific genes regulatory interactions using scRNA-seq data, such as PIDC (Chan *et al.*, 2017) and SCENIC (Aibar *et al.*, 2017). However, although these methods take advantage of the high-resolution offered by scRNA-seq data, they are not designed to infer direct regulatory interactions involving enhancers. Therefore, these methods remain limited for the inference of cell (sub)type specific regulatory mechanisms, mainly driven by enhancers, that are required to provide cell (sub)type specific mechanistic insights in diseases (Andersson *et al.*, 2014; Claringbould and Zaugg, 2021). In that regard, the combination of different type of OMICS data has been shown to be a promising approach to build comprehensive GRNs by taking advantage of the high-resolution provided by single cell technologies (Zhang *et al.*, 2022). However, the applicability of such method remains limited as it requires matched data between cells, which remains poorly available (Bravo González-Blas *et al.*, 2020).

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 in diseases 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). Here we present RNetDys, a multi-OMICs pipeline combining scRNA-seq, scATAC-seq, ChIP-seq and prior-knowledge to decipher cell (sub)type specific impaired regulatory interactions due to SNPs in diseases. This pipeline exploits the GRN information, obtained from the GRN inference of RNetDys, to identify impaired regulatory mechanisms due to SNPs. In particular, RNetDys provides the binding affinity score of TFs, the sign of interactions to distinguish activation from repression and, a list of ranked TFs based on their involvement in the regulatory impairments. Notably, compared to existing strategies to study SNPs (Yu *et al.*, 2022; Nathan *et al.*, 2022), our pipeline provides a comprehensive view of the impaired regulatory landscape to provide better mechanistic insights. In addition, RNetDys does not require matched datasets hence allowing for a wider applicability. We first showed that RNetDys predicts cell (sub)type specific GRNs more accurately than existing methods. We then applied our pipeline to five diseases to study the differential cell (sub)type specific impairment and validate the relevance of the predicted impaired regulatory interactions.

Material and methods

General workflow of RNetDys

We implemented a systematic pipeline that leverages multi-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 (i) the cell (sub)type specific GRN inference, and (ii) the capture of impaired regulatory interactions due to SNPs to gain regulatory mechanistic insights for the disease condition.

Cell (sub)type specific regulatory interactions inference

The cell (sub)type specific regulatory network inference was based on a multi-OMICS approach that relied on 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 (Figure 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: correlation value, TF: transcription factor, E: enhancer, G: 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: correlation value, TF: transcription factor, E: enhancer, G: gene

Then, 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:

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

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.

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 processing

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).

Validation and comparison to state-of-the-art

We 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-gene 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. Of note, cell lines should be homogeneous and thus we assume that the performances obtained using cell line specific GS can be extrapolated for more specialized cell

(sub)populations such as cell subtypes. 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 Positive (*predicted but not found in the GS*) and FN = False Negative (*not predicted but found in the GS*).

Results

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

We implemented RNetDys, a systematic pipeline based on multi-OMICS data that systematically decipher impaired regulatory interactions due to SNPs in diseases by leveraging the information of cell (sub)type specific GRNs. RNetDys is an integrative approach relying on single cell transcriptomics and single cell chromatin accessibility from a specific cell (sub)type, 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 (Figure 1, Material and methods, Figure S1). The first part consists of the GRN inference for a healthy cell (sub)type 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. Moreover, RNetDys could be applied for any cell (sub)populations, including cell states, as it exploits the high resolution of single cell data. 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 (Figure 1).

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

RNetDys highly relies on the cell (sub)type specific regulatory landscape to identify impaired regulatory interactions due to SNPs in diseases. Therefore, we assessed the performance of RNetDys in predicting cell (sub)type specific GRNs (Figure 2). We

performed the benchmarking of both TF-gene and enhancer-promoter interactions, compared to state-of-the-art 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 (Figure 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 (Figure 2C, D). In summary, we showed the better performances of RNetDys to predict cell (sub)type specific regulatory interactions between TF-genes and enhancer-promoters. Therefore, we demonstrated that the cell (sub)type specific GRN information leveraged by our pipeline to capture impaired transcriptional regulatory mechanisms due to SNPs in diseases is accurate.

Cell (sub)type differential dysregulation in diseases

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 validated the impact of the mapped SNPs in each of the predicted impaired interactions. Across the five diseases, we were able to validate the relation SNP-target gene in 90% of our results using GWAS from ClinVar database. 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. Notably, by using the same data in PD, we were able to validate novel SNP-target genes interactions such as rs11538371, rs2072814 and rs8137714 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.*, 2008; Annese *et al.*, 2015). Furthermore, *TIMP3* has been shown to inhibit β-amyloid precursor (APP) proteolysis and hence increase β-amyloid aggregates, a major hallmark of PD dementia (Hoe *et al.*, 2007). 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 prone to be dysregulated in diseases (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 (Figure 3). Interestingly, in EPI, astrocytes and OPCs 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).

Cell (sub)type specific disease-related regulatory impairment

We finally aimed at exploiting the GRN information provided by RNetDys to further analyze the regulatory impairments of cell (sub)types (Figure 4, Figure S2-S5). We observed that in AD (Figure 4), the same enhancers were involved in every cell (sub)type specific networks with an impact on the expression of *APP* and presenilin 1 (*PSEN1*). Indeed, 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 (Figure 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 (Figure 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). Interestingly, *CREB1* was found to be the main TF regulator involved in AD and EPI in every cell (sub)types apart from astrocytes (Table 1, Figure 4, Figure S3). CREB 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) (Figure S3). In AD, *PSEN1* has been shown to be a downstream target of *CREB1*, which further supports the results obtained by our pipeline as *CREB1* was predicted to regulate *PSEN1* (Cui *et al.*, 2016) (Table 1). Moreover, *MXI1* was found to be one of the main regulators involved in impaired regulatory interactions for PD, apart from dopaminergic neurons (Table 1, Figure S2). *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). Finally, *STAT3* was overall found to be the main regulator involved in impaired interactions of T1D and T2D (Table 1, Figures S4 and S5). 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 (Mashili *et al.*, 2013), supporting its importance as a regulator of the dysregulations involved in the disease.

Discussion

The study of cell (sub)type specific regulatory interactions impaired due to SNPs in diseases is required to pave the way towards the development of novel gene-based therapies to treat diseases (Rao *et al.*, 2021). In addition, the comprehensive view of the regulatory landscape 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 get a better understanding of the disease-related dysregulations (Rao *et al.*, 2021; Bryois *et al.*, 2021). In addition, current approaches have been mainly focused on cell types, but it has been recently shown that more specialized group of cells, such as cell subtypes, are not equally involved in diseases (Nathan *et al.*, 2022; Kamath *et al.*, 2022). Here we present RNetDys, a systematic multi-OMICS pipeline to decipher cell (sub)type specific regulatory interactions impaired due to SNPs in diseases. This pipeline exploits the high-resolution of single cell to infer a comprehensive regulatory landscape used to identify impairment due to SNPs. Notably, RNetDys can be applied to more specialized cell (sub)populations such as cell states due to its design. We first ensured that the multi-OMICS approach used by RNetDys was outperforming existing methods for inferring cell (sub)type specific regulatory interactions. Notably, the main limitation of the GRN inference part of RNetDys was the use of prior-knowledge. Indeed, it strongly increases the confidence in the predicted edges but also discard the discovery of unreported ones. Nevertheless, we alleviated this limitation by using GeneHancer, the most complete prior-knowledge available to date (Fishilevich *et al.*, 2017; Oki *et al.*, 2018). We applied RNetDys to five disease cases and observed that cell (sub)type specific regulatory mechanisms were not equally impaired, suggesting their differential involvement in the studied diseases. Moreover, we validated the relevance of the impaired regulatory mechanisms and provided additional insights into the main regulators involved. In particular, the presented analysis was performed using SNPs retrieved from ClinVar, but RNetDys could be of great use to provide valuable regulatory

mechanistic insights while leveraging the GRN information 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 RNetDys could facilitate the discovery of regulatory impairments. To conclude, we foresee our pipeline 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.

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.

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 nigro-striatal 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.

Bravo González-Blas,C. *et al.* (2020) Identification of genomic enhancers through spatial integration of single-cell transcriptomics and epigenomics. *Mol Syst Biol*, **16**.

Bryois,J. *et al.* (2021) Cell-type specific cis-eQTLs in eight brain cell-types identifies novel risk genes for human brain disorders Neurology.

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.* (2008) A novel intracellular role of matrix metalloproteinase-3 during apoptosis of dopaminergic cells. *J Neurochem*, **106**, 405–415.

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

Cui,J. *et al.* (2016) Quantification of dopaminergic neuron differentiation and neurotoxicity via a genetic reporter. *Sci Rep*, **6**, 25181.

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

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.

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.

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.
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.
Guo,S. *et al.* (2016) Gene regulatory network inference using PLS-based methods. *BMC Bioinformatics*, **17**, 545.
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.
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.
Hoe,H.-S. *et al.* (2007) The metalloprotease inhibitor TIMP-3 regulates amyloid precursor protein and apolipoprotein E receptor proteolysis. *J Neurosci*, **27**, 10895–10905.
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.
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.
Lestón Pinilla,L. *et al.* (2021) Hypoxia Signaling in Parkinson's Disease: There Is Use in Asking "What HIF?" *Biology*, **10**, 723.
Malpartida,A.B. *et al.* (2021) Mitochondrial Dysfunction and Mitophagy in Parkinson's Disease: From Mechanism to Therapy. *Trends Biochem Sci*, **46**, 329–343.
Margolin,A.A. *et al.* (2006) ARACNE: An Algorithm for the Reconstruction of Gene Regulatory Networks in a Mammalian Cellular Context. *BMC Bioinformatics*, **7**, S7.
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*.
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**.
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).
Uddin,F. *et al.* (2020) CRISPR Gene Therapy: Applications, Limitations, and Implications for the Future. *Front Oncol*, **10**, 1387.
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*.
Zhang,L. *et al.* (2022) DIRECT-NET: An efficient method to discover cis-regulatory elements and construct regulatory networks from single-cell multiomics data. *Sci. Adv.*, **8**, eabl7393.
Zhu,X. *et al.* (2012) Decreased CREB levels suppress epilepsy. *Neurobiol Dis*, **45**, 253–263.

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.

Funding

C.B. is supported by funding from the Luxembourg National Research Fund (FNR) within PARK-QC DTU (PRIDE17/12244779/PARK-QC). M.R. is supported by Fonds National de la Recherche Luxembourg (C17/BM/11662681).

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).

Authors contribution

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

Competing Interests

The authors declare no competing interests.

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

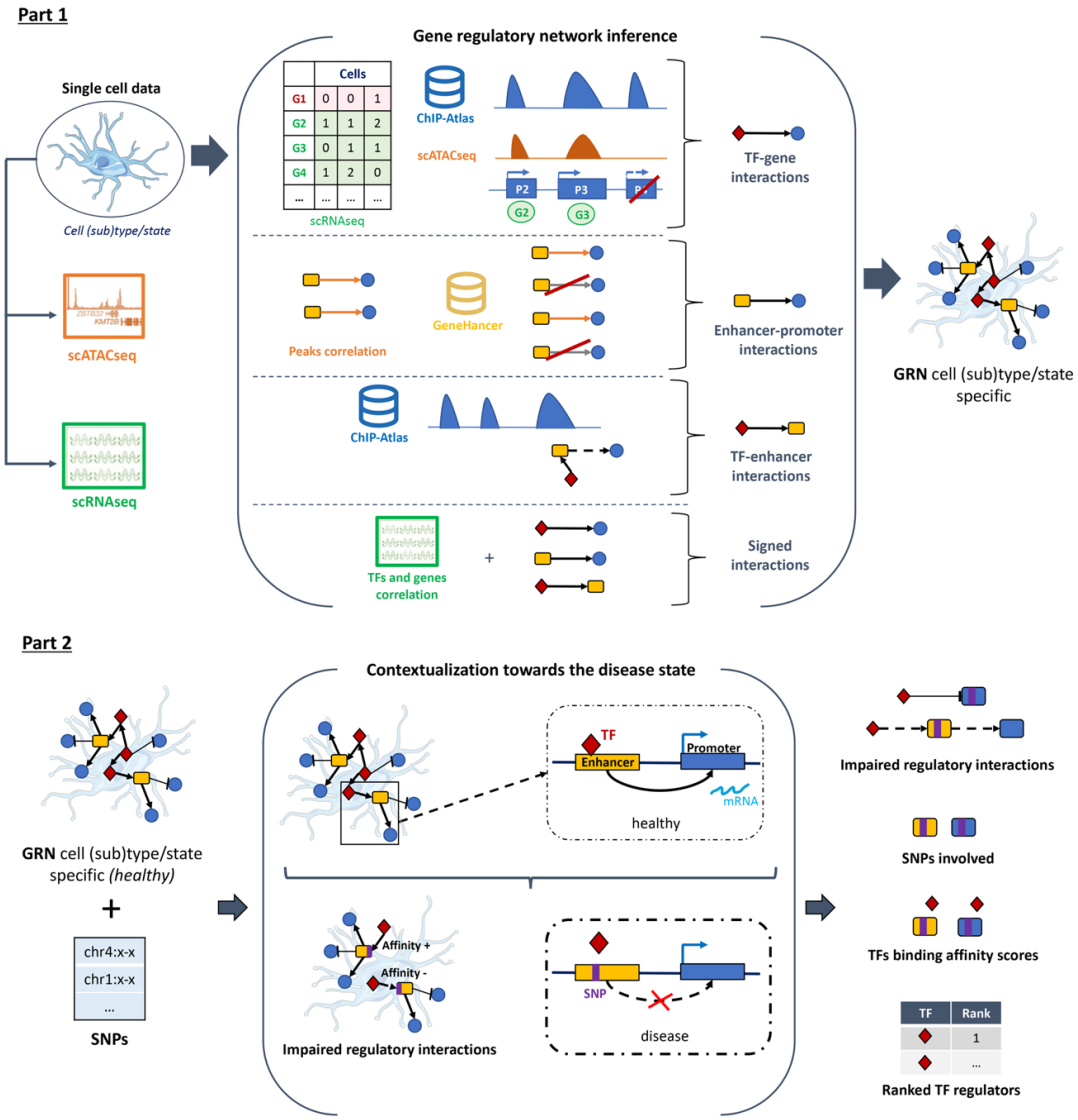

Figure 1. General workflow of RNetDys to decipher regulatory dysregulation in diseases.

RNetDys is composed of two main parts including (1) the 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. The first part provides the cell (sub)type or state specific GRN describing the regulatory interactions mediated by TFs and enhancers of regulated genes. The second part provides the list of candidate impaired regulatory interactions in the cell (sub)type, the SNPs that were mapped to these interactions, the TFs for which the binding affinity is impaired, and the regulatory TFs ranked based on their importance in the impairments.

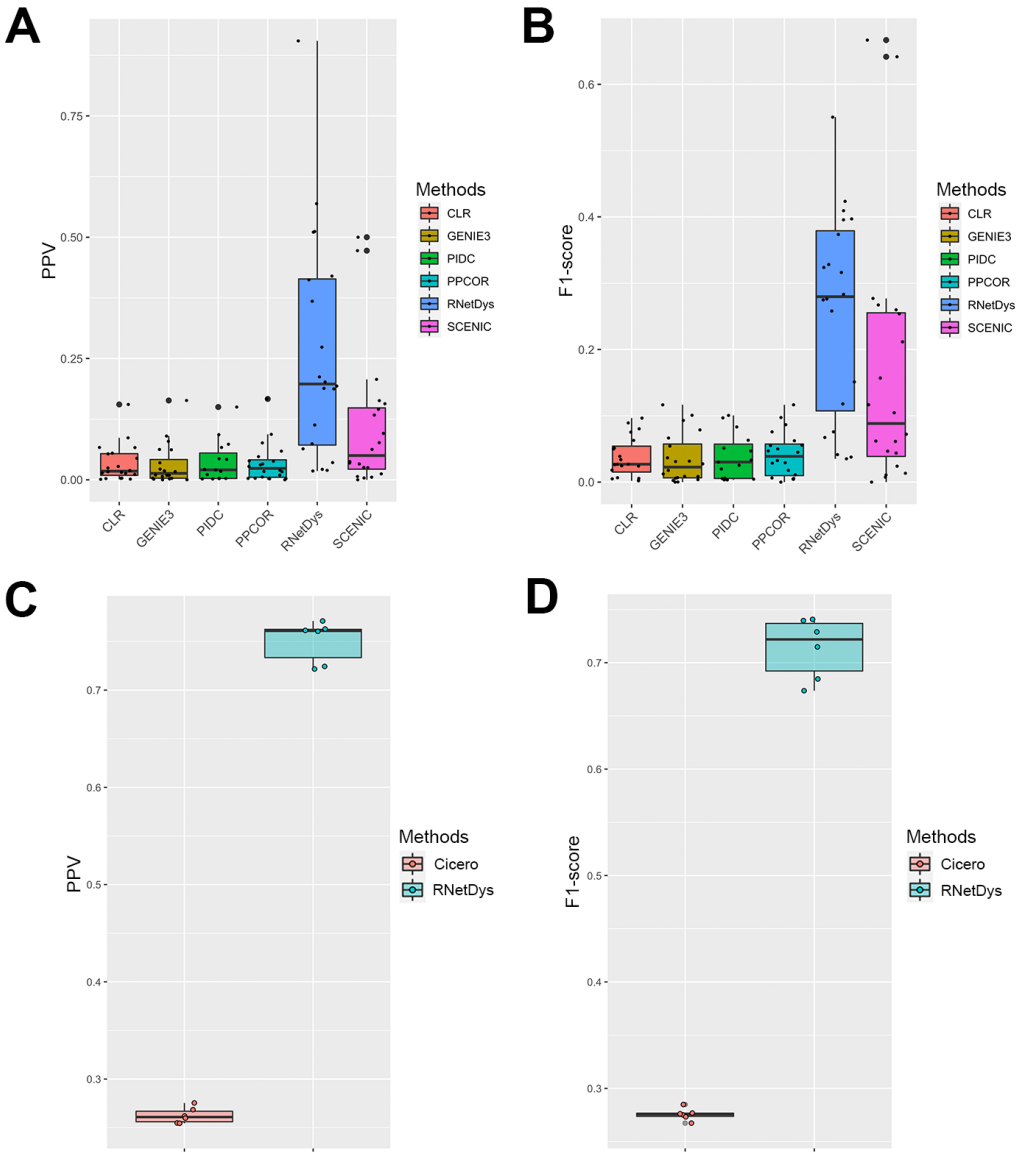

Figure 2. Performances of RNetDys and comparison to other methods.
(A, B) TF-gene regulatory interactions performances assessed using (A) the PPV and (B) the F1-score metrics. Performances were assessed for RNetDys and state-of-the-art methods on 20 datasets from six human cell lines. (C, D) Enhancer-promoter 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.

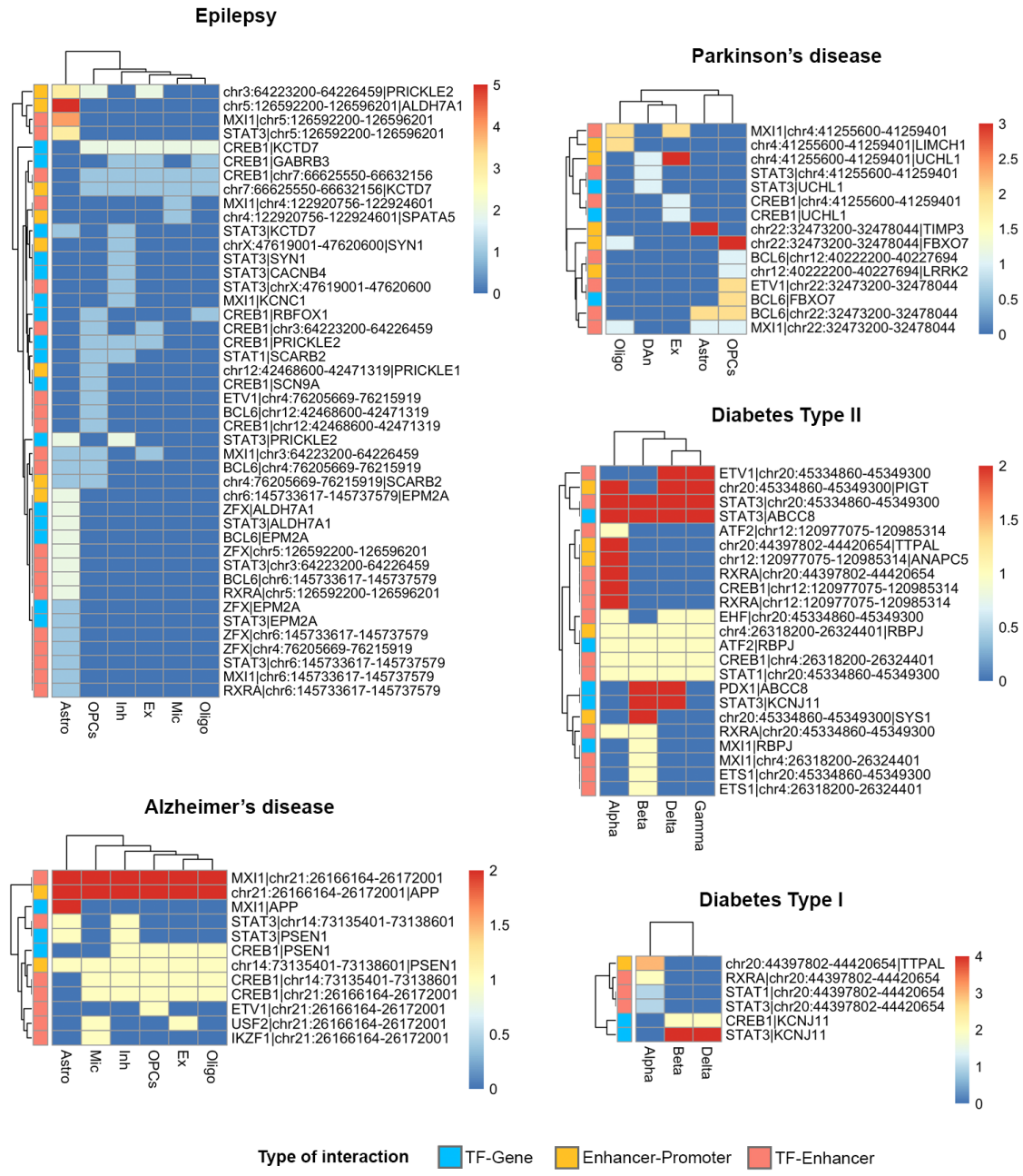

Figure 3. Cell (sub)type differential regulatory impairment in diseases.

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.

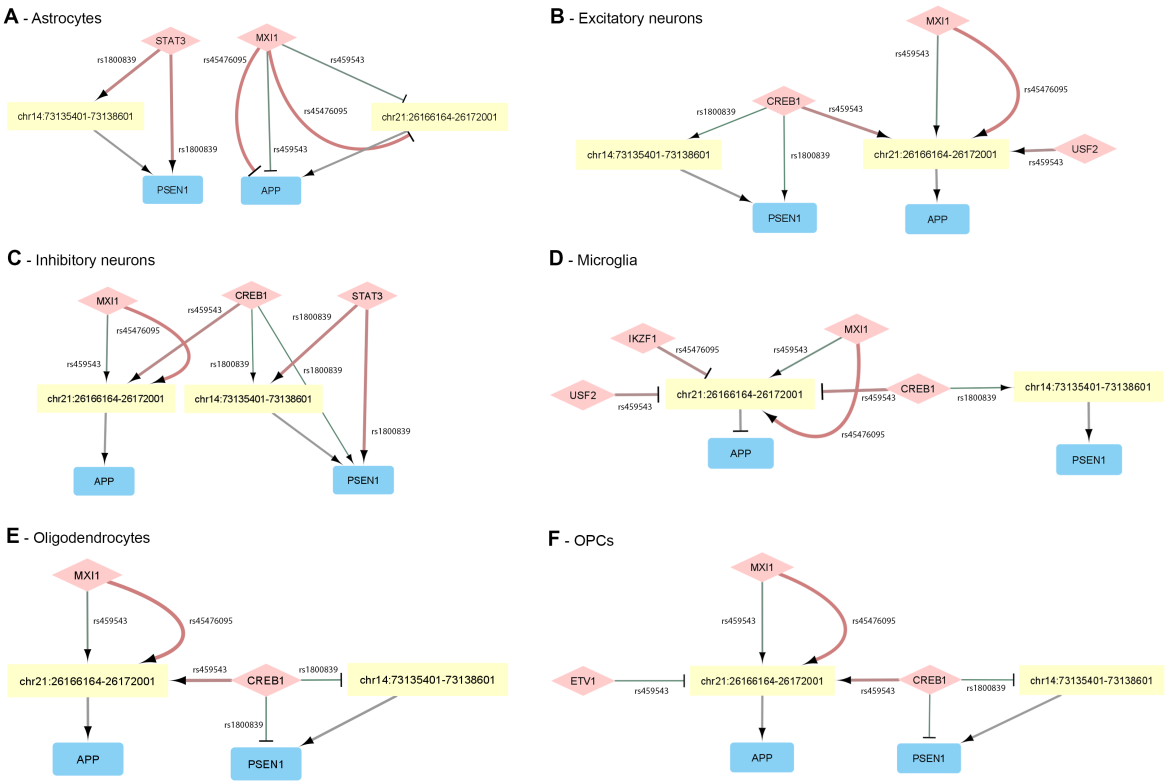

Figure 4. Cell (sub)type specific regulatory impairment in AD.
Network visualization of impaired regulatory interactions for (A) astrocytes, (B) excitatory neurons, (C) inhibitory neurons, (D) microglia, (E) oligodendrocytes and (F) OPCs. TFs are represented as diamond in light red, enhancers as yellow 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.

4.3.3 Supplementary Information

Supplementary Methods

Supplementary References

Supplementary Figures:

Supplementary Tables:

Supplementary Methods

RNetDys workflow

Cell (sub)type 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 was performed on the scRNA-seq and scATAC-seq in which any rows (gene or peaks) or columns (cells) having a sum of zero were removed from further analyses. Then, the following steps were computed to infer the cell (sub)type 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 considered 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 were 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 was considered to be accessible if its gene was expressed (conserved at least in 50% of the cells) and at least one ATAC peak was 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 allowed us to predict the cell (sub)type specific TF-gene 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 was accessible. Then, we splitted the scATAC peaks matrix such that one matrix contained 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 were computed for each pairwise correlation. Then, a Benjamini-Hochberg multiple test correction was applied 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 found 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 was intersected with all GeneHancer (Fishilevich *et al.*, 2017) reported connections.

(3) TF-Enhancers interactions. First, enhancers present in the Enhancer-Promoter network were selected. They are then intersected with the ChIP-seq data, using BEDTools and -F 1.0. Therefore, if 100% of the ChIP-seq TF peak felt inside an enhancer region, then this TF was a regulator of 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 were most likely activation whereas negative ones were most likely repression. The signs for enhancer-promoter interactions were determined by computing the sum of correlation values for the TFs binding to the enhancer regulating the specific gene. Notably, the correlation values were corresponding to the TF-gene relationship. The correlation score to determine the sign was computed such as (Figure S1) such as:

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

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

Finally, signs for TF-Enhancers were computed by summing, for each TF binding of the enhancer, the TF-genes relationship correlation values for each gene 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, the regulatory network was contextualized towards the disease condition based on a list of SNPs. First, promoter regions coordinate for which a TF binding site has been identified were retrieved from the ChIP-seq data used by RNetDys. Then, the SNPs were mapped to these regions and enhancer regions of the GRN using BEDTools, under the condition that the SNP mapped exactly inside one of the regions (parameter -F 1). This step allowed for 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 was performed on the candidate impaired binding sites. 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 candidate impaired binding region. Then, we refined the impaired regulatory interactions by selecting the ones having at least one TF

binding site significantly impacted. 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 involved in the impairments due to SNPs, and hence were 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 bedtool. 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 (aspecific) 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 (Figure S6).

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 annotations were used from their original studies. Similarly, the scATAC-seq datasets were processed using Signac (Stuart *et al.*, 2020) and annotations were kept from their respective studies. The gene expression and peaks matrices for each cell (sub)type were extracted for each tissue as follow:

- Pancreas: We performed the peak calling with Signac using 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. 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 for two different brain regions (Table S3). The 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). 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 performed a comparison with 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, no 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-score. Notably, cell line specific GS were used to assess the performances for inferring cell (sub)type specific GRNs. Indeed, cell lines are well studied and hence data is available to compile GS with confidence. In addition, we assume that the performances obtained using cell line specific GS can be extrapolated for more specialized cell (sub)populations such as subtypes due to their homogeneity.

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 (SNVs) from ClinVar (Landrum *et al.*, 2018) and extracted SNPs such as SNVs 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 disease-related in ClinVar, and we performed a systematic extraction using regex in R 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

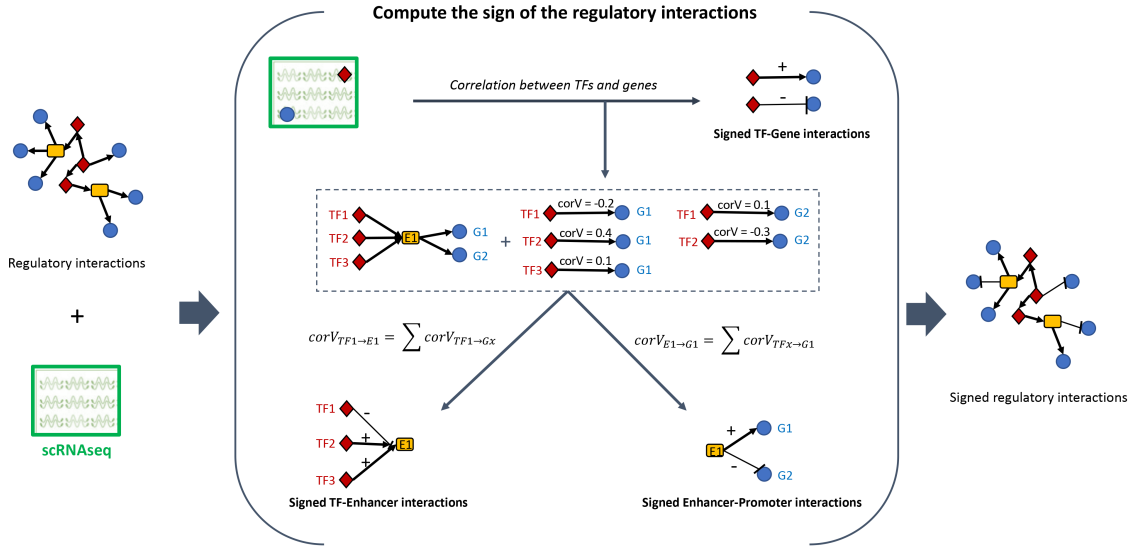

Figure S1. Strategy to compute the sign of the regulatory interactions.
The scRNA-seq dataset is used to compute the correlation between the TFs and genes of the GRN. TF-gene interactions are directly signed using the correlation values. Enhancer-promoter interactions are signed by summing the correlation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

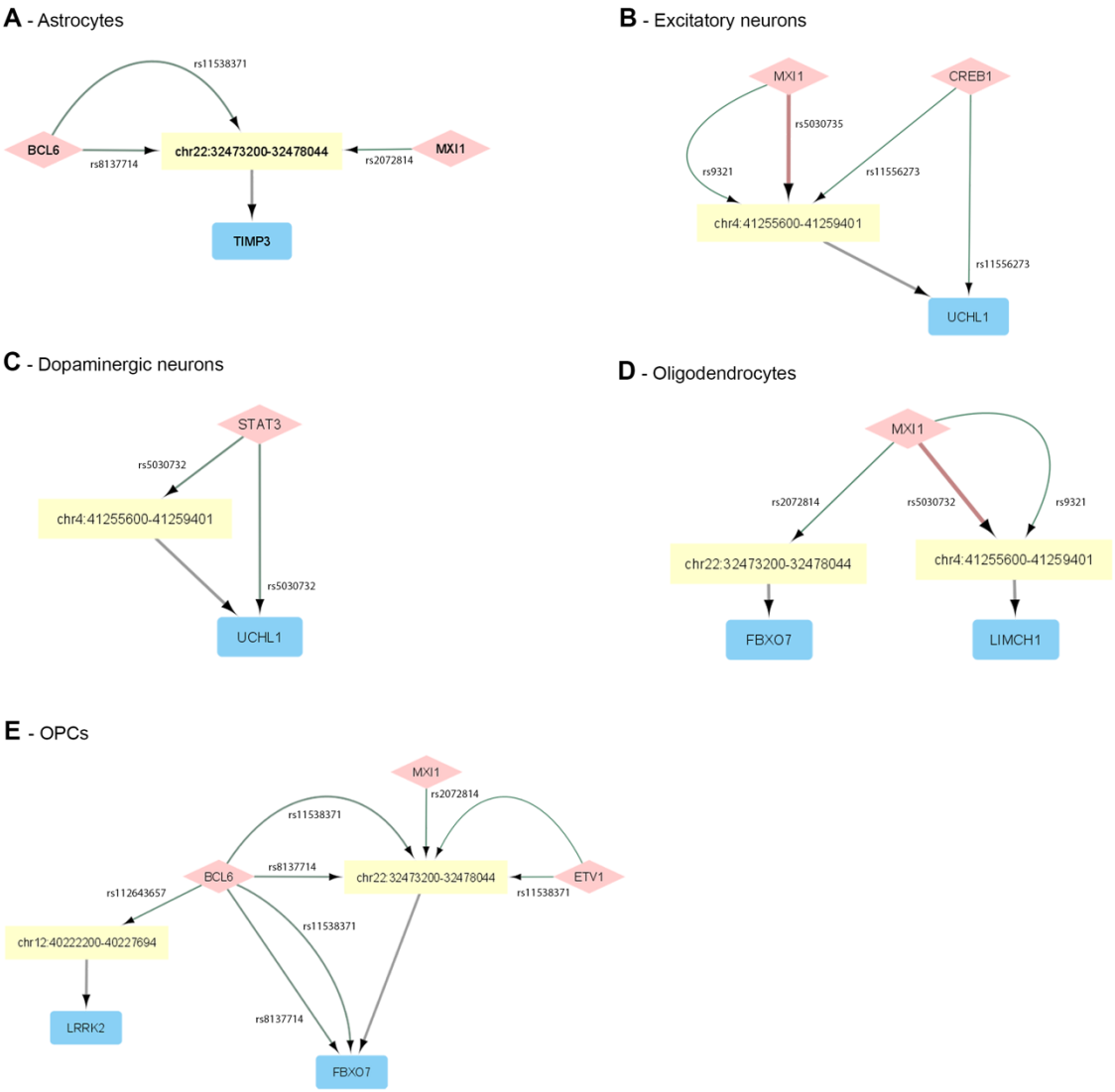

Figure S2. Cell (sub)type specific regulatory impairment in PD.
Network visualization of impaired regulatory interactions for (A) astrocytes, (B) excitatory neurons, (C) dopaminergic neurons, (D) oligodendrocytes and (E) OPCs. TFs are represented as diamond in light red, enhancers as yellow rectangles 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.

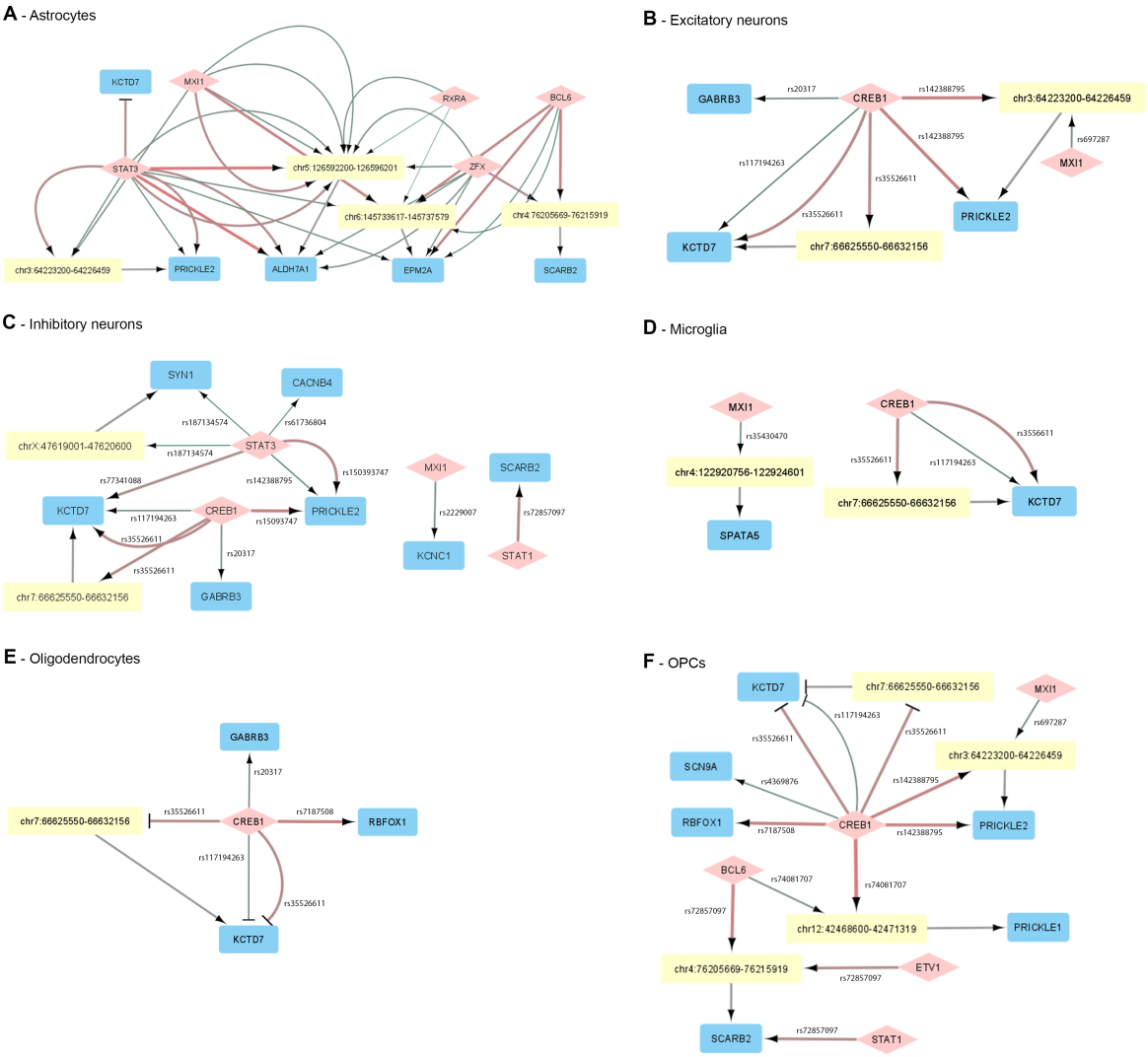

Figure S3. Cell (sub)type specific regulatory impairment in EPI.
Network visualization of impaired regulatory interactions for (A) astrocytes, (B) excitatory neurons, (C) inhibitory neurons, (D) microglia, (E) oligodendrocytes and (F) OPCs. TFs are represented as diamond in light red, enhancers as yellow 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. Notably, the labels for edges are not displayed in (A) Astrocytes due to the high number of interactions, but each edge mediated by a TF represents an impairment due to a specific SNP.

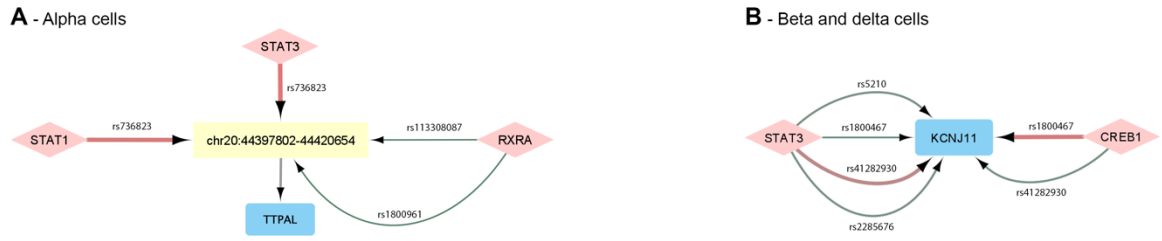

Figure S4. Cell type specific impairment in T1D.
Network visualization of impaired regulatory interactions for (A) alpha cells and (B) beta and delta cells. TFs are represented as diamond in light red, enhancers as yellow rectangles and genes in blue rectangles. Arrows represent activations. The weight 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.

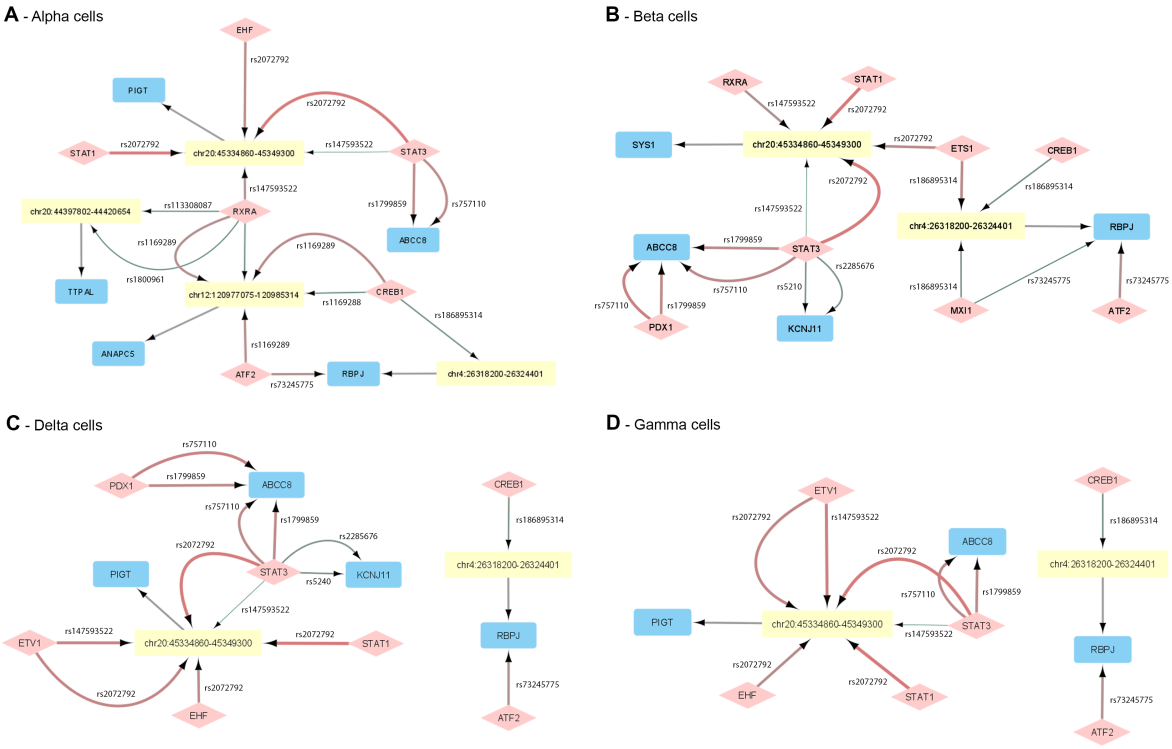

Figure S5. Cell type specific impairment in T2D.
Network visualization of impaired regulatory interactions for (A) alpha cells, (B) beta cells, (C) delta cells and (D) gamma cells. TFs are represented as diamond in light red, enhancers as yellow rectangles and genes in blue rectangles. Arrows represent 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.

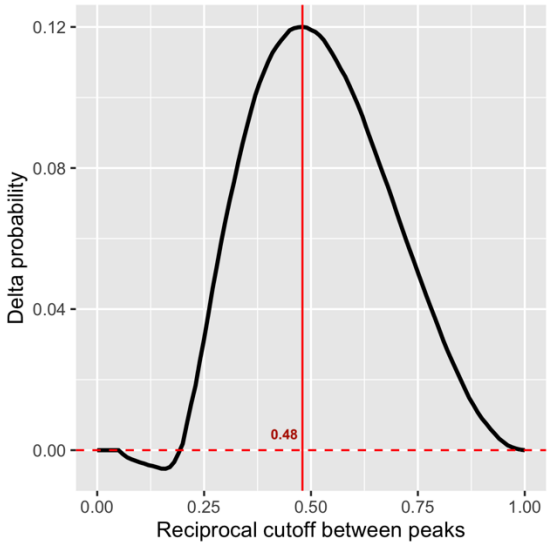

Figure S6. Threshold selection to define accessibility of promoter regions.
Delta probability between true positives and false positives. The peak of the distribution, equal to 0.48, corresponds to the highest probability to capture a true accessible promoter region in the cell (sub)type.

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		x
STAT3	UCHL1	rs5030732	DAn	x	28253266, 25370916, 22839974	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		x
STAT3	PSEN1	rs1800839	Astro	x	28821390, 11389157	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		x
CREB1	PSEN1	rs1800839	Ex	x	28821390, 11389157	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		
CREB1, STAT3	PSEN1	rs1800839	Inh	x	28821390, 11389157	
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		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			
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	No data
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

Multicellular organisms are composed of highly heterogeneous and functionally specialized cells organized into different layers of complexity such as tissue or organs (Morris *et al.*, 2019; Arendt *et al.*, 2016; Regev *et al.*, 2017). Cells display specific expression patterns governed by complex regulatory mechanisms that turn off and on transcriptomic programs (Wray *et al.*, 2003). Internal and external stimuli trigger cellular responses that lead cells to a change of activity or state can enhance physiological processes but also pathological ones in case of dysregulations (Miller-Jensen *et al.*, 2007; Carson and Ribeiro, 1993; Bartsch and Wulff, 2015). The increasing prevalence of single-cell OMICS data contributed to the discovery of new or rare cell (sub)populations (e.g., subtypes, states) and to a better understanding of disease heterogeneity and complexity (Strzelecka *et al.*, 2018). Indeed, single cell technologies led to the generation of organism-wide atlases at an unprecedented resolution and allowed the implementation of computational approaches with more detailed models to dissect the heterogeneity at the cellular and molecular levels (The Tabula Muris Consortium *et al.*, 2018; The Tabula Sapiens Consortium and Quake, 2021; Efremova and Teichmann, 2020). The characterization of cells escaping the healthy cellular landscape to go towards pathological states and the identification of candidate molecules to prevent or treat diseases are part of the challenges addressed by computational systems biology approaches (Morris, 2019; Moreau and Tranchevent, 2012). Over the past few years, several computational methods were developed to analyze single-cell data and provide new biological insights to pave the way towards new therapeutic and personalized medicine approaches (Stuart *et al.*, 2019; Oulas *et al.*, 2019). Whereas the development of such methods contributed to the advance of the field, they present several limitations that need to be addressed (Lähnemann *et al.*, 2020; Morris, 2019). This thesis presents three computational approaches to study different aspects of disease modelling to overcome existing limitations and contribute to solving open challenges in systems biology. In particular, the approaches developed focus on the characterization of cell identity, the identification of functional cell states and candidate genes for cell state conversion, and the inference of comprehensive GRNs to guide the identification of impaired regulatory mechanisms in diseases.

5.1 Revising the characterization of cell identity

For years, cells were classified based on different features such as their anatomical location or morphology which has been shown to be limited and inaccurate. Indeed, the emergence

of single-cell based data provided an unprecedented resolution of cell features that uncovered the molecular and cellular complexity of biological systems and refuted the previous classification system (Morris, 2019). The accurate and extensive characterization of the cellular identity landscape for different organisms would be valuable to better understand physiological processes but also detect cells displaying non-physiological patterns (Altschuler and Wu, 2010; Ikeda *et al.*, 2018). However, the characterization of cell identity and underlying genes defining it remains a central challenge. Indeed, the fact that cell identity is acquired during the developmental process and shaped by the niche to perform specific functions makes the identification of identity genes a non-trivial task (Morris, 2019). The deciphering of identity genes highly relies on the biological context in which cells are characterized, accordingly with their hierarchical classification as cell type, subtype or phenotype. In addition, it has been shown that gene expression levels are involved in different functional outcomes for the same cell (sub)type (Huang, Yang, George W Ye, *et al.*, 2021; Shats *et al.*, 2017). In that regard, current computational methods present several limitations when identifying identity genes. Indeed, they rely on the comparison of gene expression profiles of a target cell population with other cell populations in given tissues that are usually incomplete and composed of mixed cell types, subtypes and/or phenotypes. Moreover, they categorize the gene expression as expressed or non-expressed which discard any intermediate level of expression which could lead to different functions and hence be part of the cell identity. The combination of these two limitations highly hinders the accurate characterization of cell identity. To address these limitations, we developed HCellig, a computational method relying on the hierarchical organization of cell identity (cell types, subtypes and phenotypes) and accounting for intermediate levels of gene expression to accurately capture identity genes (section 4.1).

5.1.1 Scope and utility

HCellig has been implemented as a general method to capture identity genes of any cell (sub)population, including cell types, subtypes and phenotypes, in physiological and pathological conditions. A priori, no annotation is required to use the method and identify identity genes of an unknown target cell population. Indeed, to characterize the cell identity of a cell type, the cell type background will be used to determine its identity genes. However, for more precise levels of resolution such as cell subtype and phenotype, it is required to gather additional information to select the most relevant background to use HCellig. Therefore, in case of an unknown target cell subpopulation or phenotype, one strategy could

first consist into the capture of identity genes by comparing with the cell type background. These identity genes would help determining, based on expert knowledge and/or with the help of well-defined cell type markers (X. Zhang *et al.*, 2019), to which cell type belongs the target cell subpopulation or phenotype. Then, to get a more refined characterization of its identity, one could use this information to select the correct cell subtype background to use for deciphering identity genes. Finally, to characterize the cell identify of an unknown cell phenotype, one can use a similar strategy to first find to which cell subtype it belongs. Then, the right cell phenotype background can be selected to define the cell identity at its most refined level of resolution and obtain the list of identity genes for the target cell phenotype. To illustrate the discussed strategy, let us assume that we obtained an unknown cell (sub)population after clustering our scRNA-seq data. We would first identity its identity genes compared to the cell type layer using HCellig. Then, based on expert knowledge and well-defined markers we would determine that our unknown cell (sub)population belongs to neurons. We can then select the neuron subtype background to refine the identity genes of our cell (sub)population. Using the same strategy, we can either observe a mix of neuron subtypes, in which case we can consider that our cell (sub)population correspond to neurons or determine that our unknown cell subpopulation are dopaminergic neurons. Finally, we can select the dopaminergic neurons phenotype background to obtain the identity genes of the specific phenotype (e.g., dopaminergic neurons from midbrain).

The accurate identification of identity genes is required to characterize cell (sub)populations of interest and is of great use to perform downstream analyses such as the discovery of regulatory modules of cell identity or the characterization of core biological processes. Indeed, the use of identity genes to perform downstream analyses could be seen as a feature selection step to capture the most relevant and informative genes defining the target cell (sub)population. These identity genes could then be used to guide the identification of key TFs regulators of cell identity (Almeida *et al.*, 2021) to guide cell conversion or get a better understanding of identity destabilization or disruption in diseases (Ikeda *et al.*, 2018; Brumbaugh *et al.*, 2019; Jung *et al.*, 2021). For instance, HCellig could be used to capture identity genes of a target cell (sub)population in healthy and disease condition. The comparison between the two sets of identity genes would allow the identification of the ones that might be lost in the disease condition. Moreover, the generation of identity cores, by building the GRN around the identity genes, could provide a more comprehensive view of the cell identity destabilization. Indeed, it would allow the identification of key regulator of

identity that might be involved in the dysregulations. Notably, the GRN could be built based on the transcriptomic data used with HCellig by using state-of-the-art approaches (Aibar *et al.*, 2017). In addition, the identity core could be used to identify functional modules of genes, based on functional enrichment analyses (Wu *et al.*, 2021), and provide additional information into the functional impairment specific to the characterized cell (sub)populations in diseases.

5.1.2 Strengths

HCellig relies on a hierarchical cell identity model to capture identity genes of any cell (sub)population, accordingly with their hierarchical classification as cell type, subtype and phenotype. For instance, HCellig would determine genes characterizing neurons by comparing them to any cell type of the organism. Then, it would identify genes defining dopaminergic neurons by comparing them to all subtype neurons of the organism and finally it would capture identity genes of dopaminergic neurons specific to substantia nigra by comparing them to all different locations in which this subtype can be found. We assume that this hierarchical approach better reflects the biological reality than tissue-wise comparisons. Indeed, whereas tissue-based comparison were reasonable approaches to identify cell (sub)type markers used by experimentalists for cell extraction or sorting (X. Zhang *et al.*, 2019), this characterization of cell identity remains limited. The hierarchical cell identity model used by our computational method considers the hierarchical classification of cells organism-wide and hence allow for a more accurate characterization of their identity. Moreover, HCellig uses a discretization of gene expression strategy implemented in RefBool (Jung *et al.*, 2017), that was adapted for single cell data to quantify genes into three levels of expression including low, medium and high. The advantage of these three levels is that, compared to traditional methods categorizing genes as expressed or not (Stuart *et al.*, 2019), it allows for the distinction between medium and high expression which has been shown to be important for the functional outcomes (Huang, Yang, George W Ye, *et al.*, 2021; Shats *et al.*, 2017). In that regard, this less stringent categorization of gene expression into three levels leads to a more accurate capture of identity genes and hence a better characterization of cell identity for which the functional features are critical (Morris, 2019). In summary, the two main advances implemented in HCellig address limitations of current methods to provide an accurate characterization of cellular identity.

We pre-compiled a large-scale repository of backgrounds at each hierarchical layer including cell type, subtype and phenotype for mouse and human that can be used to characterize the cell identity of any know or unknown cell (sub)population. Notably, HCellig is a user-friendly R package that requires few parameters and computational resources to capture the identity genes and their expression level for a query cell (sub)population. Our method could be used to extend the current knowledge (X. Zhang *et al.*, 2019) of the cellular landscape by allowing the accurate characterization of cell (sub)populations in physiological and non-physiological conditions. In that regard, we generated high-resolution cell identity atlases for mouse and human that can complete the current knowledge available for these cellular landscapes (Regev *et al.*, 2017; The Tabula Muris Consortium *et al.*, 2018; The Tabula Sapiens Consortium and Quake, 2021; Morris, 2019). Indeed, we observed, as expected, that markers described in literature for specific cell populations were captured by HCellig. However, we also highlighted a high number of unreported and unknown identity genes, especially the ones expressed at a medium level, and for the phenotypes. Therefore, we expect our atlases to highlight the importance of medium identity genes that could lead to different functional outcomes, physiological or pathological, in case their expression level would be perturbed.

5.1.3 Limitations

The study performed in "*Quantification of gene level to characterize hierarchical cell identity*" (section 4.1) has several advantages but some limitations remain. First, the backgrounds for mouse were not as extensive as the ones built for human. Indeed, due to the lack of data, low sequencing depth and limited annotations, very few cell subtype backgrounds were generated, and no cell phenotype backgrounds were compiled. In addition, the compilation of large backgrounds, especially the cell type layer, requires a lot of computational resources. Therefore, it would be required to generate new backgrounds or extend the current ones by using a High-Performance Computing structure to meet the computational resources requirements. Nevertheless, we mitigated this limitation by providing a repository containing several pre-compiled backgrounds for each hierarchical layer. However, with the growing availability of organism-wide scRNA-seq data, it could be interesting to extend the current cell type backgrounds or increase the list of available subtype and phenotype ones. HCellig is adapted for UMI-based single cell data only, due to the lack of state-of-the-art normalization approaches for non-UMI data. Indeed, it is well accepted that the normalization and batch effect correction approaches differ between UMI

and non-UMI data (Lytal *et al.*, 2020; Chen *et al.*, 2019). However, compared to scTransform which is widely used for UMI data (Hafemeister and Satija, 2019), no consensus has been found for non-UMI data. Finally, this study provides two main novelties with the hierarchical model and the three levels of expression, but no experimental validations were performed. Indeed, it would be interesting to perform functional assays to support the impact of expression level changes for the medium identity genes on the cell (sub)population functional outcome (section 5.5).

5.2 Identifying functional cell states and immunomodulators

Cellular identity is defined by a set of genes characterizing cells specific features such as their functions. The functional specialization of cells is acquired during the development and is further shaped by external signals. In response to stimuli, the same cell (sub)type can exhibit different phenotypes, corresponding to different functional cell states that are characterized by specific molecular features (Trapnell, 2015). Whereas stimuli are part of physiological processes in place to maintain the integrity and homeostasis of the organism, they can also trigger dysregulations that can lead to pathological states (Rué and Martinez Arias, 2015; Lutshumba *et al.*, 2021). Computational biology models aim at leveraging the discovery of cell states to have a better understanding of the underlying heterogeneity in physiological and pathological conditions. In addition, these models aim at identifying potential candidate genes that could be used for cell states conversion, and for instance revert a disease state towards a healthy one (Wei *et al.*, 2022). Despite recent efforts to decipher cellular states, there is room for improvement to accurately decipher them while identifying the key functional genes and functional processes that characterize them. To tackle this challenge, we developed FunPart, a computational method that decipher functional cell states, capture the key genes and their related functional processes to characterize them (section 4.2). FunPart was applied to the mouse immune system, widely studied over the years (Chaplin, 2010; P. Fang *et al.*, 2018; Iwasaki and Akashi, 2007), as the identification of their functional states and transcriptional characterization would be pivotal for the development of therapy strategies relying on immunomodulators. Notably, FunPart could be applied to any type of cells to identify functional cell states and provide candidate modulators for cellular conversion in different disease conditions.

5.2.1 Scope and utility

FunPart has been developed as a general method to decipher functional cell states and systematically capture the key genes that characterize them, in physiological and pathological conditions. It allows the dissection of the functional heterogeneity by accounting for subtle differences to identify groups of cells, named functional cell states, that share similar transcriptomic profiles and functions. In addition, the method deciphers functional module of genes and the key transcription factors characterizing these states. In the study (section 4.2), we applied FunPart to decipher the functional heterogeneity of immune cells across different types of infection but, it can be widely applied to any type of cells in both pathological and physiological conditions.

Functional heterogeneity is a fundamental property of biological systems that needs to be dissected and characterized to gather biological insights in physiological and pathological conditions (Gough *et al.*, 2017). Indeed, functional heterogeneity have been shown to play a critical role in homeostasis and maintenance of tissue integrity (Krieger and Simons, 2015). Moreover, functional heterogeneity have been shown to play a crucial role in non-physiological conditions but, functional cell states identification and characterization remains elusive (Clarke *et al.*, 2021; Chan *et al.*, 2022). Therefore, the study of functional heterogeneity is pivotal to have a better understanding of both physiological and pathological conditions and pave the way towards the development of novel therapies. FunPart can be applied to discover novel functional states and their key genes. Depending on the research question, the uncovered cell states can then be further analyzed to understand their functional specialization or implication in diseases (Li and Boussiotis, 2011; Clarke *et al.*, 2021). Notably, the transcription factors provided by FunPart are candidate modulators for cell state conversion, but the genes related to these TFs and functional enrichment, provided by the method, can be used to get a better understanding of the functional specialization. Nevertheless, FunPart provides the most relevant functional gene modules according to the designed criteria including the strength of negative correlation between the modules and their functional enrichment. Therefore, to get a more extensive view of the functional specialization of the cell state, it would be required to perform enrichment analyses of other set of genes. Notably, these set of genes could be identified using the same strategy implemented in FunPart, by building a correlation network using the gene expression information of the cell states and identifying modules of genes.

5.2.2 Strengths

This study has several strengths regarding the implemented computational method and the *Catalogus Immune Muris* resource generated. First of all, FunPart systematically detect functional cell states in a semi-supervised manner, by relying on the data and the provided functional annotations that can be specific to a subset of BPs, as we did in our study by focusing on immune processes (Singhania *et al.*, 2019). Indeed, our method relies on the combination of a feature selection strategy based on the concept of functional gene modules, and a recursive clustering approach which we showed is more accurate than the state-of-the-art approach. In addition, FunPart provides for each identified cell states the set of genes, including TFs, that characterize them as well as the biological processes in which these genes are involved to provide insights into the functional differences between the functional cell states identified. Moreover, the set of TFs identified as being characteristic and driver of the functional cell state are usually small, which can be seen as a prioritization of candidates to use for cellular conversion between functional states, as we demonstrated with *Zfp597* in our study. Finally, FunPart is a user-friendly R package that can be widely used by the scientific community to decipher and characterize new functional states in physiological and pathological conditions. This study generated a *Catalogus Immune Muris,* a large-scale catalogue of immune functional cell states, identified in different types of infections, that report all functional modules (TFs and co-expressed genes). Therefore, it contains a molecular characterization of these immune functional states that can be leveraged to design novel immunomodulatory strategies (Iqbal Yatoo *et al.*, 2021). In that regard, we found *Zfp597* to be an immunomodulator of macrophages infected by *Salmonella* and showed that the knockout of this TF was reverting the macrophages towards a pro-inflammatory state. Thus, *Zfp597* could be used to modulate the response of macrophages infected by *Salmonella* to switch their states towards pro-inflammatory or anti-inflammatory functions.

5.2.3 Limitations

The main limitation of FunPart is that the method highly relies on the functional annotations provided for the clustering approach to decipher the functional states. Indeed, to limit an over-clustering of the algorithm, it is recommended to remove broad BPs categories such as the default ones from the Gene Ontology annotations (Ashburner *et al.*, 2000). Specific or more specialized BPs categories should be provided to FunPart to ensure the functional relevance of the identified cell states. Moreover, the presented strategy considers modules composed of TFs that are potentially difficult to target or not preferred approaches for

therapeutic uses, as they might give rise to mutagenesis or unexpected off-target effects (Ben-David and Benvenisty, 2011; Yamanaka, 2020). Indeed, the development of immunomodulatory therapies is typically based on the use of drugs or chemical compounds to alter cellular functions, which has been shown safer than TFs perturbations for cellular conversion (Kumar and Mali, 2020). However, chemical compounds or molecules to specifically target the candidate TFs identified by FunPart could be identified to tackle this limitation by making use of existing approaches (Zheng, 2021).

5.3 Gene regulatory network to decipher impaired regulatory mechanisms

Transcriptomics based GRN inference methods are a promising approach to study dysregulation in diseases, but they partially model the regulatory machinery. Nevertheless, due to data availability limitations, these methods are commonly used as single cell transcriptomics data is widely available compared to other types of OMICS data (e.g., scATAC-seq) (Lee *et al.*, 2020; Chen *et al.*, 2019). Indeed, the exploitation of OMICS data to characterize regulatory mechanisms of heterogeneous cell (sub)populations still remains a challenge, mainly due to the lack of single cell sequencing techniques or data (Bravo González-Blas *et al.*, 2020). Moreover, it has been shown that the majority of SNPs related to diseases lie in intronic regions, especially enhancers, for which the regulatory mechanisms remain unresolved (Claringbould and Zaugg, 2021; Boix *et al.*, 2021; Nasser *et al.*, 2021). Therefore, it is required to have a comprehensive GRN describing the underlying regulatory mechanisms mediated by TFs and enhancers of regulated genes to translate SNPs risk-variants into mechanistic insights. Indeed, the exploitation of a comprehensive regulatory landscape would help to have better mechanistic insights to understand diseases conditions, and it would guide the dissection of cell (sub)type specific impairment. In that regard, we propose RNetDys, a computational pipeline relying on multi-OMICs data to infer comprehensive cell (sub)type and state specific GRNs and identify candidate regulations impacted by leveraging the GRN information.

5.3.1 Scope and utility

RNetDys consists of a systematic pipeline that leverages multi-OMICS data to build comprehensive cell (sub)type and state specific GRNs and identify regulatory interactions that can be impaired in diseases due to SNPs. Our pipeline gives additional information to better understand the regulatory dysregulations by leveraging the GRN information. Indeed, it provides a comprehensive view of the regulatory interactions mediated by TFs and

enhancers of regulated genes for a specific cell (sub)population or state of interest. In addition, it identifies impaired regulatory mechanisms due to SNPs in diseases, provides information about impaired TFs binding sites, the type of regulatory mechanisms impacted (activation and repression), and identifies the main TF regulators involved in the impairment. RNetDys can be applied to study any disease of interest, under the condition that healthy scRNA-seq and scATAC-seq are available to build the GRNs, and that SNPs of interest can be provided to the pipeline. Notably, scRNA-seq and scATAC-seq does not need to come from the same cell measurement (unmatched data), but they need to belong to the same cell (sub)type or state. In that regard, the confidence in the annotations is crucial to ensure the accuracy of the predicted GRN. In case the scRNA-seq or scATAC-seq data is not annotated, or if the degree of confidence in the annotations is low, one strategy consists of integrating or mapping the two types of data (Stuart *et al.*, 2020). This approach allows to either annotate the dataset and extract the cell (sub)populations of interest, or to ensure that the cells share similar profiles and validate that they most likely belong to the same cell (sub)type or state.

One central challenge in genomics is to find out how genetic variations such as SNPs can lead to complex diseases (Shastry, 2007; Degtyareva *et al.*, 2021). The development of NGS technologies strengthened the development of functional genomics to better identify SNPs and their involvement in gene expression dysregulations (Cano-Gamez and Trynka, 2020). RNetDys can be used to complete the current knowledge provided by GWAS and eQTL studies (Coetzee *et al.*, 2016; Cano-Gamez and Trynka, 2020) by providing a comprehensive view of the regulatory impairments due to SNPs. Indeed, the pipeline provides valuable insights including the impaired binding affinity score of TFs, the impaired regulatory mechanisms mediated by these TFs, the SNPs that could impair these regulatory mechanisms and the main regulator TFs that are involved. In particular, for the research project presented in section 4.3, few SNPs were analyzed as they were retrieved from a prior-knowledge database (Landrum *et al.*, 2018) for validation purposes. The use of genotyping data from patients having a specific disease would allow for a larger-scale analysis of the potential impact of SNPs on cell (sub)types or state specific regulatory mechanisms and provide a better understanding of their differential impairment in the disease.

5.3.2 Strengths

RNetDys is a comprehensive computational pipeline that first infers the regulatory landscape of a specific cell (sub)type or state and then systematically identify impaired interactions in

disease conditions due to SNPs. The combination of scRNA-seq, scATAC-seq and prior-knowledge, including ChIP-seq TF binding evidences (Oki *et al.*, 2018) and GeneHancer database (Fishilevich *et al.*, 2017), allowed us to build comprehensive GRNs describing regulatory relationships mediated by TFs and enhancers of regulated genes. In addition, we showed that the use of multi-OMICS increased the overall accuracy to predict regulatory interactions compared to existing methods. Moreover, RNetDys can be used to infer the specific GRN of any cell type, subtype or state, under the condition that both scRNA-seq and scATAC-seq data are available. Therefore, it provides a valuable strategy to describe regulatory mechanisms more accurately in physiological and pathological conditions. Moreover, the network contextualization towards the disease condition only requires a list of SNPs related to the disease of interest to identify candidate impaired regulatory mechanisms specific to the cell (sub)type studied. Indeed, the pipeline provides valuable information to guide our understanding into the cell (sub)type specific transcriptional mechanisms impaired including the list of candidate impaired interactions, the TFs binding affinity scores and the TF regulators involved in the impairments. In summary, compared to existing strategies, RNetDys provides a systematic approach, that takes advantage of the single cell to guide the study of regulatory mechanisms specific to cell (sub)types, up to cell states, and provide insights into their differential impairment at the regulatory level in disease conditions.

5.3.3 Limitations

RNetDys has two main limitations related to the GRN inference part. First, only reported information in the prior-knowledge used can be predicted. Indeed, whereas the use of experimental-based evidences increases the confidence into the regulatory mechanisms predicted, it does not allow the prediction of novel interactions that have never been reported. Nevertheless, we used the most extensive knowledge up-to-date to mitigate this limitation by using all ChIP-seq TF binding evidence from ChIP-Atlas (Oki *et al.*, 2018) and enhancer-promoter connections reported in GeneHancer (Fishilevich *et al.*, 2017). Notably, GeneHancer is a prior-knowledge database reporting enhancer for human only, and hence the current implementation of RNetDys is limited to human studies. However, it could be further extended to account for mouse by using other enhancer prior-knowledge resources such as EnhancerDB (Kang *et al.*, 2019). In addition, these resources are regularly updated and hence the prior-knowledge used by RNetDys could be expanded to alleviate this limitation. Second, we assume that an enhancer is active if it is accessible, at least one TF is

expressed and binding to its region, and it regulates at least one promoter of an expressed gene. However, we have no evidence that the enhancer is actually active, as we did not use methylation and/or acetylation marks, that are still poorly available at the single cell level (Clark *et al.*, 2016). In addition, for the contextualization towards the disease state, RNetDys relies on prior-knowledge TF motifs used by Perfectos-ape (E. Vorontsov *et al.*, 2015) and hence only the TFs for which this information is available can be predicted as involved in the impairment of regulatory interactions. However, similarly as before, the prior-knowledge used could be regularly updated to mitigate this limitation. Notably, the creation of potential binding sites for TFs due to SNPs (Degtyareva *et al.*, 2021) is not considered by RNetDys that exclusively rely on binding sites reported in prior-knowledge ChIP-seq data (Oki *et al.*, 2018).

5.4 Relationship between the computational methods implemented

5.4.1 Cell identity

Multicellular organisms are composed of highly heterogeneous cells displaying specific expression patterns that define their identity (Morris, 2019). The identification of subtle differences between group of cells, such as cell states, as well as the characterization of their identity is an ongoing challenge that has given rise to different strategies and points of view to attempt solving it (Trapnell, 2015; Morris, 2019). In this thesis, two computational methods – HCellig (section 4.1) and FunPart (section 4.2) – were developed to help resolving the identification and characterization of cell identity in physiological and pathological conditions. HCellig is a general approach that captures identity genes and their expression level for any cell type, subtype or phenotype provided as an input. FunPart does not require the group of cells to be provided as an input to identity the different functional group of cells. Both methods capture genes characterizing the cell (sub)populations, with HCellig providing a more exhausting list than FunPart that rather focuses on a small set of TFs and genes to prioritize potential candidates for cell states conversion. Notably, the identity genes captured by HCellig could be prioritized and a strategy could be implemented to help guiding cellular conversion protocols. Whereas HCellig is a general approach to capture identity genes for any cell type, subtype and phenotype, FunPart is a more specialized method to identity functional cell states, corresponding to the cell phenotypes for HCellig. In addition, FunPart also captures set of genes which characterize the functional cell states identified, that could be used to modulate these states, as demonstrated with *Zfp597* found to be an unreported immunomodulator of macrophages infected by *Salmonella*.

5.4.2 Disease modelling

This thesis focused on the development of computational methods for disease modelling to unravel cell identity, functional cell states and transcriptional regulatory mechanisms in physiological and pathological conditions. In particular, this thesis addressed different challenges of computational systems biology with the implementation of three computational strategies. Each method focuses on different aspects, ranging from the cellular identity to the regulatory mechanisms, to aid our understanding of systems complexity by characterizing cell identity, dissecting functional heterogeneity and modelling transcriptional regulatory mechanisms. These methods and related findings aim at providing a better understanding of physiological and pathological processes to pave the way towards the development of novel therapeutic strategies such as disease treatment. They could be used in combination to identify functional cell states, characterize their identity and decipher the regulatory mechanisms to study a specific disease from different angles. Indeed, used in combination they would allow to decipher heterogeneous cells, that might be specific to the disease, characterize them to identify candidates for cellular conversion, and study the regulatory mechanisms that could be impaired to validate or expand the candidates for therapeutic approaches (section 5.5.2).

5.5 Outlook

Several perspectives of optimization and extension for the research projects presented in this thesis could be performed in the future. First, the optimization and further development of the computational methods could be done to address the limitations previously mentioned. Then, an extension or combination of these methods could be implemented to create a general and widely applicable workflow for disease modelling including the guidance to design cell conversion protocols, to revert disease phenotypes, and to systematically identify target genes or molecules to pave the way towards new therapeutic strategies.

5.5.1 Address the limitations and gather experimental validations support

Some of the aforementioned limitations could be overcome by addressing the technical limitations and performing experimental validations as a proof-of-concept or additional support for the findings. In addition, the methods implemented could be further extended to increase their accuracy and scope of applicability.

Overcoming technical limitations and extending the methods

First of all, as previously mentioned, HCellig is limited to UMI data due to the lack of state-of-the art normalization for non-UMI data (Lytal *et al.*, 2020; Tran *et al.*, 2020; Vallejos *et al.*, 2017). It would be needed to further extend the approach for non-UMI data (e.g. Smart-seq2 technology) to cover all single cell transcriptomics sequencing techniques, and hence extend the applicability of the method for any type of scRNA-seq datasets. However, it would require a well-accepted approach to normalize and account for batch-effect correction on non-UMI data. Once such state-of-the-art method will be available, HCellig could account for UMI and non-UMI data by selecting the right normalization strategy depending on the type of data provided as an input. In that regard, extensive pre-compiled backgrounds datasets using non-UMI datasets could be generated. In addition, these data could be used to extend the high-resolution atlases already generated. Moreover, the Tabula Muris atlas used was highly limited to generate cell subtype backgrounds and no phenotype ones could be produced (The Tabula Muris Consortium *et al.*, 2018). It would be interesting to extend the current background by collecting individual study datasets that usually provides a deeper sequencing depth and hence a higher resolution of the cell groups that can be identified.

Then, it could be of interest to generate a *Catalogus Immune Sapiens* using FunPart to build an atlas of human immune functional states and potential immunomodulators, as compiled for mouse with the *Catalogus Immune Muris*. Indeed, with the large availability of immune scRNA-seq data in disease conditions (Ner-Gaon et al., 2017), the use of FunPart on a compendium of human datasets would be highly valuable to decipher unknown functional states, extend the current knowledge of immunomodulators and pave the way towards the development of new therapeutic strategies. This would allow the extension of the infection and disease panel (Kuhn *et al.*, 2019; Elsland and Neefjes, 2018; Jochems *et al.*, 2018) but also aid the identification of human immunomodulators, as the ones identified for mouse models might not translate to human setups (Mestas and Hughes, 2004). Moreover, as previously stated, candidate modulators identified by FunPart are TFs, which are not preferred approaches in therapeutics setups as they might give rise to unexpected off-target effects or tumorigenic (Ben-David and Benvenisty, 2011; Yamanaka, 2020). Indeed, the development of immunotherapy strategies is usually based on drugs or chemical compounds to indirectly perturb the specific TF(s) for cellular conversion (Kumar and Mali, 2020). It would hence be interesting to implement or integrate an approach that would identify

candidate compounds to perturb the candidate immunomodulator TFs identified by FunPart and, hence benefit the drug discovery field (Moustaqil et al., 2020).

Finally, RNetDys currently relies on GeneHancer (Fishilevich *et al.*, 2017), a prior-knowledge database specific to human, for the cell (sub)type and state specific GRN inference. It could be further extended for mouse by collecting enhancer-promoter prior-information from the available extensive databases (Gao and Qian, 2019). However, as the regulatory interaction inference is highly relying on the prior-knowledge used, it would be necessary to ensure the quality of the collected information and perform a benchmarking to verify the accuracy of the predicted GRNs. In addition, the integration of different regulatory layers has been proposed as a promising strategy to gather better mechanistic insights in physiological and pathological conditions (Hu *et al.*, 2020). Currently, RNetDys uses a multi-OMICS approach involving single cell transcriptomics and single cell chromatin accessibility data. However, it could be further extended to make use of enhancer activity marks, such as H3K4me1 and H3K27ac (Kimura, 2013) at the single-cell level. Indeed, currently RNetDys assumes that if an enhancer is accessible, regulating genes expressed in the cell (sub)type/state and TFs are binding to its region, then this enhancer is considered to be active. Whereas this assumption is reasonable with the lack of activity marks, the integration of the histone modification would allow for a more accurate GRN. In the future, more single-cell histone modification datasets should be generated (Bartosovic *et al.*, 2021). Therefore, it will be of interest to integrate this layer of information while keeping advantage of the high-resolution provided by single cell technologies, already leveraged by the method.

Performing experimental validations

Experimental validations are usually required to support the computational model implemented and validate or provide support for the *in-silico* predictions generated. Such experiments were already performed in the "*A Catalogus Immune Muris of the mouse immune responses to diverse pathogens*" study (section 4.2) but it would be of great interest to have experimental support for the study entitled *"Quantification of gene level to characterize hierarchical cell identity"* (section 4.1). It would be of high interest to perform experimental validations for some identity genes identified in the high-resolution identity atlases, especially the ones expressed at a medium level. Indeed, by modifying the level of expression of such genes, we would expect to observe a different functional outcome as previously reported in some studies (Huang, Yang, George W Ye, *et al.*, 2021; Shats *et al.*,

2017). The type of experiments to perform could include the perturbation of a medium identity TF, by inducing its over-expression or knocking it down, to then observe the differential functional outcome of the specific cell type, subtype or phenotype tested. For instance, medium identity TFs identified for mouse dopaminergic neurons could be perturbed as aforementioned and we could quantify the levels of dopamine release. Moreover, immunostaining and/or electrophysiology analysis (Cui *et al.*, 2016; Farassat *et al.*, 2019; Mahajani *et al.*, 2019) could be performed to study the differential functional outcome to support the impairment or destabilization of their identity.

5.5.2 Combine the developed methods in one framework

The three computational methods presented in this thesis could be expanded and used in combination due to their close relationship. In the future, and with the described perspectives, they could contribute to the development of novel cell-based and gene-based therapeutic strategies.

HCellig and RNetDys could be used in combination to implement a computational framework aiming at guiding cellular conversion protocols for cell-based therapies (Vasan *et al.*, 2021; Grath and Dai, 2019). The accurate characterization of cell identity combined with the GRN approach provided by RNetDys would lead to the identification of master regulator TFs (MRTFs) defining cellular identity. Indeed, the captured MRTFs would be promising candidates to manipulate the cell fate (F. Fang *et al.*, 2018; Jung *et al.*, 2021) and hence guide more efficient cell conversion protocols to pave the way towards novel cell-based therapies. HCellig could be first used to capture identity genes of the target cell (sub)population at each hierarchical layer. For instance, assuming the target cells are midbrain dopaminergic neurons, HCellig could be applied to first capture identity genes of the neurons (cell type layer), then dopaminergic neurons (cell subtype layer) and then dopaminergic neurons of the midbrain (cell phenotype layers). Then, the regulatory network around the captured identity genes at each hierarchical layer could be generated using the GRN inference part of RNetDys, hence resulting in three identity networks. The identification of MRTFs and their prioritization by importance, which could be determined using the graph properties (e.g., outdegree), at each hierarchical identity layer could then be used to guide the generation of the target cells of interest. Finally, the described workflow could be automatized towards the implementation a computational pipeline combining the cell identity characterization (HCellig) and the capture of candidate MRTFs using the GRN

inference part of RNetDys. Moreover, FunPart could be added to the strategy described earlier. Indeed, HCellig requires group of cells to characterize their identity and hence is not a strategy to identify cell (sub)types or states. Whereas existing clustering approaches were shown to be promising for deciphering cell (sub)types (Kiselev, Tallulah S Andrews, *et al.*, 2019), they are limited for the capture of functional cell states. Therefore, the use of FunPart for the identification of more subtle differences including functional cell states would be of great interest. Notably, FunPart identifies functional genes characterizing these states and hence, it would be expected to observe commonalities with the identity genes that would be captured by HCellig.

The three computational tools could be combined to implement a workflow aiming at providing a better understanding of non-physiological conditions and guide the development of new disease treatments. Indeed, similarly as previously described, FunPart could be first used to dissect the functional heterogeneity using scRNA-seq datasets from patients and healthy controls. It would allow the identification of functional cell states and key genes characterizing them. Then, HCellig could be applied to capture the identity genes of each functional cell state. For each functional cell state, a functional enrichment of the identity genes could be performed to identify all the BPs and pathways characterizing them. Notably, FunPart already provides the main BP characterizing each functional cell state, but this step would allow to extend the list of specific functions each state might perform. The comparison between the functions of healthy and disease cell states would allow for a better understanding of functional dysregulations. In addition, this comparison would guide the identification of cell states that might be specific to the disease and thus be of particular interest. Finally, the GRN inference part of RNetDys could be used to build the functional cell states specific GRNs in both healthy and disease conditions. Notably, the scATAC-seq data would need to be mapped (Stuart *et al.*, 2020) to the cell (sub)type and states characterized by HCellig with the scRNA-seq data to extract the required datasets for the GRN inference using RNetDys. Genes dysregulated or involved in the disease as well as their main regulators could then be identified. Notably, DEG analyses mapped on the GRNs would guide the identification of main regulators, that could be identity genes, involved in the dysregulations. This approach would provide a better understanding of the heterogeneity of cell (sub)populations in the disease and the regulatory mechanisms involved. Therefore, it would be a valuable strategy to guide the identification of genes related to the pathological conditions and help the development of therapies for disease treatments. Notably, another

approach for the workflow could be to only focus on scRNA-seq from healthy controls, use FunPart and HCellig as previously described, and then apply RNetDys to build the cell (sub)types and states specific GRNs to identify impaired regulatory mechanisms due to disease-related SNPs. In that regard, disease-related SNPs could be collected from GWAS or eQTL studies specific to then identify cell (sub)type and state specific regulatory mechanisms impaired. SNPs could also be obtained from genotyping data of patients having the specific disease. In addition, if scRNA-seq data from these patients is available, additional analyses could be performed. Indeed, by having SNPs and gene expression data from the same patients, it would be possible to further refine the impaired regulatory mechanisms predicted by RNetDys by mapping DEGs. Whereas not all genes predicted to be impaired by RNetDys are expected to be significantly DEGs, the focus on impaired regulatory mechanisms involving DEGs could guide the identification of genes, regulators and SNPs most likely involved in the disease and its dysregulations.

6 Conclusion

The characterization of cells escaping the physiological landscape, the understanding of pathological mechanisms and the identification of candidate targets are critical to pave the way towards new therapeutic strategies for disease treatment. Indeed, the accurate characterization of cell identity and the capture of key TFs for cell state conversion holds great promises to revert disease states towards healthy ones. In addition, the study of the GRN is required to have a better understanding of the dysregulated regulatory mechanisms and guide the identification of candidate targets for disease treatment. To date, several computational methods focusing on cell identity and GRN inference have been implemented, but they have several limitations that hinder their accuracy and contribution to disease understanding. The aims of this thesis were to develop more accurate, comprehensive and systematic computational methods that address the main limitations of existing approaches, as well as extending the current knowledge in the field. In summary, this thesis provided the following contributions:

- **Implementation of a method to characterize cell identity**: HCellig is a hierarchical cell identity-based computational method that quantifies genes into three levels of expression to accurately capture identity genes for any cell type, subtype and phenotype. Compared to existing methods, it leverages the hierarchical classification of cells to not mix different layer of complexity and account for an intermediate level of expression, shown to lead to different functional outcomes. HCellig is a user-friendly R package available at: https://github.com/BarlierC/HCellig, and all pre-compiled backgrounds for mouse and human at the cell type, subtype and phenotype levels are publicly available at: https://gitlab.com/C.Barlier/HCellig_backgrounds.

- **Generation of two high-resolution identity atlases**: Using HCellig, we generated high-resolution identity atlases that reports identity genes and their level of expression for all described cell types, subtypes and phenotypes in mouse and human. The atlases for both organisms are available at: https://gitlab.com/C.Barlier/HCI.

- **Development a method to decipher functional states and the key genes characterizing them:** FunPart is a computational method to decipher functional cell states in physiological and pathological conditions. In addition, it captures the key genes characterizing these states and provides insights for their function. FunPart is an R package available at: https://github.com/BarlierC/FunPart.git.

- **Compilation of a *Catalogus Immune Muris***: This mouse atlas reports large-scale immune functional cell states for different type of infections as well as the key functional genes characterizing them. This resource contains potential candidate immunomodulators and could be exploited to aid the development of immunotherapy strategies. Notably, its potential was demonstrated with *Zfp597,* a functionally relevant gene of a macrophage cell state for which its inhibition resulted in a significant decrease in surviving bacteria. The Catalogus Immune Muris is available in Table S3 of the published paper: https://www.nature.com/articles/s41419-021-04075-y. A shiny app is also available at: https://gitlab.com/C.Barlier/immunofunmap.git.

- **Comprehensive approach to infer cell (sub)types and states specific GRN:** RNetDys is a multi-OMICS pipeline, relying on single cell data and prior-knowledge to first infer cell (sub)type or state specific regulatory interactions mediated by TFs and enhancers of regulated genes. It requires as an input scRNA-seq and scATAC-seq datasets of a specific cell (sub)type or state of interest to infer the GRN. RNetDys is a user-friendly pipeline available at: https://github.com/BarlierC/RNetDys, with its first part corresponding to the GRN inference.

- **Systematic identification of cell (sub)type candidate regulatory interactions impaired due to SNPs in diseases:** Based on the healthy GRN for a cell (sub)type or state of interest, RNetDys then systematically identifies regulatory interactions potentially impaired due to disease-related SNPs. It provides the list of regulatory interactions impaired and leverage the GRN information to provide insights into the dysregulated mechanisms. This corresponds to the second part of RNetDys pipeline available at: https://github.com/BarlierC/RNetDys.

In conclusion, the three computational methods presented in this thesis are of great value to contribute to the advance of computational disease modelling. These methods are widely applicable to characterize cell identity, dissect functional heterogeneity, identify key genes for cell state conversion, and identify impaired regulatory mechanisms in diseases. The methods and findings of this thesis highly contribute to the systems biology field with a strong potential to guide experimental strategies for disease treatment.

7 References

Abdelaal,T. *et al.* (2019) A comparison of automatic cell identification methods for single-cell RNA sequencing data. *Genome Biol*, **20**, 194.

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

Almeida,N. *et al.* (2021) Employing core regulatory circuits to define cell identity. *EMBO J*, **40**, e106785.

Altschuler,S.J. and Wu,L.F. (2010) Cellular Heterogeneity: Do Differences Make a Difference? *Cell*, **141**, 559–563.

Aly,R.M. (2020) Current state of stem cell-based therapies: an overview. *Stem Cell Investig*, **7**, 8.

Andreatta,M. *et al.* (2021) Interpretation of T cell states from single-cell transcriptomics data using reference atlases. *Nat. Commun.*, **12**, 1–19.

Andrews,T.S. and Hemberg,M. (2018) Identifying cell populations with scRNASeq. *Molecular Aspects of Medicine*, **59**, 114–122.

Antonarakis,S.E. and Beckmann,J.S. (2006) Mendelian disorders deserve more attention. *Nat Rev Genet*, **7**, 277–282.

Antony,P.M. *et al.* (2012) From Systems Biology to Systems Biomedicine. *Current Opinion in Biotechnology*, **23**, 604–608.

Arendt,D. *et al.* (2016) The origin and evolution of cell types. *Nat. Rev. Genet.*, **17**, 744–757.

Ashburner,M. *et al.* (2000) Gene Ontology: tool for the unification of biology. *Nat Genet*, **25**, 25–29.

Bahrami,S. and Drabløs,F. (2016) Gene regulation in the immediate-early response process. *Advances in Biological Regulation*, **62**, 37–49.

Barh,D. *et al.* (2020) In silico disease model: from simple networks to complex diseases. In, *Animal Biotechnology*. Elsevier, pp. 441–460.

Bartosovic,M. *et al.* (2021) Single-cell CUT&Tag profiles histone modifications and transcription factors in complex tissues. *Nat Biotechnol*, **39**, 825–835.

Bartsch,T. and Wulff,P. (2015) The hippocampus in aging and disease: From plasticity to vulnerability. *Neuroscience*, **309**, 1–16.

Belmonte-Mateos,C. and Pujades,C. (2022) From Cell States to Cell Fates: How Cell Proliferation and Neuronal Differentiation Are Coordinated During Embryonic Development. *Front. Neurosci.*, **15**, 781160.

Belokopytova,P.S. *et al.* (2020) Quantitative prediction of enhancer-promoter interactions. *Genome Res*, **30**, 72–84.

Ben-David,U. and Benvenisty,N. (2011) The tumorigenicity of human embryonic and induced pluripotent stem cells. *Nat Rev Cancer*, **11**, 268–277.

Bigas,A. and Espinosa,L. (2012) Hematopoietic stem cells: to be or Notch to be. *Blood*, **119**, 3226–3235.

Boix,C.A. *et al.* (2021) Regulatory genomic circuitry of human disease loci by integrative epigenomics. *Nature*, **590**, 300–307.

Bravo González-Blas,C. *et al.* (2020) Identification of genomic enhancers through spatial integration of single-cell transcriptomics and epigenomics. *Mol Syst Biol*, **16**.

Broekema,R.V. *et al.* (2020) A practical view of fine-mapping and gene prioritization in the post-genome-wide association era. *Open Biol.*, **10**, 190221.

Brumbaugh,J. *et al.* (2019) Reprogramming: identifying the mechanisms that safeguard cell identity. *Development*, **146**.

Bryois,J. *et al.* (2021) Cell-type specific cis-eQTLs in eight brain cell-types identifies novel risk genes for human brain disorders Neurology.

Buchberger *et al.* (2019) Cloudy with a Chance of Insights: Context Dependent Gene Regulation and Implications for Evolutionary Studies. *Genes*, **10**, 492.
Budday,S. *et al.* (2015) Physical biology of human brain development. *Front. Cell. Neurosci.*, **0**.
Butler,A. *et al.* (2018) Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nat Biotechnol*, **36**, 411–420.
Cahan,P. *et al.* (2021) Computational Stem Cell Biology: Open Questions and Guiding Principles. *Cell Stem Cell*, **28**, 20–32.
Cano-Gamez,E. and Trynka,G. (2020) From GWAS to Function: Using Functional Genomics to Identify the Mechanisms Underlying Complex Diseases. *Front. Genet.*, **11**, 424.
Carson,D.A. and Ribeiro,J.M. (1993) Apoptosis and disease. *The Lancet*, **341**, 1251–1254.
Cha,J. and Lee,I. (2020) Single-cell network biology for resolving cellular heterogeneity in human diseases. *Exp Mol Med*, **52**, 1798–1808.
Chan,L. *et al.* (2022) Neutrophil Functional Heterogeneity and Implications for Viral Infections and Treatments. *Cells*, **11**, 1322.
Chan,T.E. *et al.* (2017) Gene Regulatory Network Inference from Single-Cell Data Using Multivariate Information Measures. *Cell Systems*, **5**, 251-267.e3.
Chaplin,D.D. (2010) Overview of the immune response. *Journal of Allergy and Clinical Immunology*, **125**, S3–S23.
Chen,G. *et al.* (2019) Single-Cell RNA-Seq Technologies and Related Computational Data Analysis. *Front Genet*, **10**, 317.
Chen,R. and Snyder,M. (2012) Systems biology: personalized medicine for the future? *Curr Opin Pharmacol*, **12**, 623–628.
Chen,S. and Mar,J.C. (2018) Evaluating methods of inferring gene regulatory networks highlights their lack of performance for single cell gene expression data. *BMC Bioinformatics*, **19**, 232.
Chen,X. *et al.* (2021) Tissue-specific enhancer functional networks for associating distal regulatory regions to disease. *Cell Systems*, **12**, 353-362.e6.
Chiou,J. *et al.* (2021) Interpreting type 1 diabetes risk with genetics and single-cell epigenomics. *Nature*, **594**, 398–402.
Choi,Y.H. and Kim,J.K. (2019) Dissecting Cellular Heterogeneity Using Single-Cell RNA Sequencing. *Mol Cells*, **42**, 189–199.
Claringbould,A. and Zaugg,J.B. (2021) Enhancers in disease: molecular basis and emerging treatment strategies. *Trends in Molecular Medicine*, **27**, 1060–1073.
Clark,S.J. *et al.* (2016) Single-cell epigenomics: powerful new methods for understanding gene regulation and cell identity. *Genome Biol*, **17**, 72.
Clarke,B.E. *et al.* (2021) Regionally encoded functional heterogeneity of astrocytes in health and disease: A perspective. *Glia*, **69**, 20–27.
Cliff,J.M. *et al.* (2004) Differential Gene Expression Identifies Novel Markers of $CD4^+$ and $CD8^+$ T Cell Activation Following Stimulation by *Mycobacterium tuberculosis*. *J Immunol*, **173**, 485–493.
Codega,P. *et al.* (2014) Prospective identification and purification of quiescent adult neural stem cells from their in vivo niche. *Neuron*, **82**.
Coetzee,S.G. *et al.* (2016) Enrichment of risk SNPs in regulatory regions implicate diverse tissues in Parkinson's disease etiology. *Scientific Reports*, **6**, 30509.
Collin,C.B. *et al.* (2022) Computational Models for Clinical Applications in Personalized Medicine—Guidelines and Recommendations for Data Integration and Model Validation. *Journal of Personalized Medicine*, **12**, 166.

Cooper,G.M. (2000) Regulation of Transcription in Eukaryotes. In, *The Cell: A Molecular Approach. 2nd edition*. Sinauer Associates.
Cui,J. *et al.* (2016) Quantification of dopaminergic neuron differentiation and neurotoxicity via a genetic reporter. *Sci Rep*, **6**, 25181.
Cummings,J. *et al.* (2021) Alzheimer's disease drug development pipeline: 2021. *A&D Transl Res & Clin Interv*, **7**.
Dao,L.T.M. and Spicuglia,S. (2018) Transcriptional regulation by promoters with enhancer function. *Transcription*, **9**, 307–314.
De Luca,C. *et al.* (2020) Neurons, Glia, Extracellular Matrix and Neurovascular Unit: A Systems Biology Approach to the Complexity of Synaptic Plasticity in Health and Disease. *Int. J. Mol. Sci.*, **21**, 1539.
Degtyareva,A.O. *et al.* (2021) Regulatory SNPs: Altered Transcription Factor Binding Sites Implicated in Complex Traits and Diseases. *Int J Mol Sci*, **22**, 6454.
Delaney,C. *et al.* (2019) Combinatorial prediction of marker panels from single-cell transcriptomic data. *Mol Syst Biol*, **15**.
Dolgalev,I. and Tikhonova,A.N. (2021) Connecting the Dots: Resolving the Bone Marrow Niche Heterogeneity. *Front. Cell Dev. Biol.*, **9**, 622519.
Doostparast Torshizi,A. *et al.* (2020) Cell Type-Specific Annotation and Fine Mapping of Variants Associated With Brain Disorders. *Front. Genet.*, **11**, 575928.
Dueck,H. *et al.* (2016) Variation is function: Are single cell differences functionally important?: Testing the hypothesis that single cell variation is required for aggregate function. *BioEssays*, **38**, 172–180.
Dusonchet,J. *et al.* (2014) A Parkinson's disease gene regulatory network identifies the signaling protein RGS2 as a modulator of LRRK2 activity and neuronal toxicity. *Human Molecular Genetics*, **23**, 4887–4905.
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.
Ebrahimi,B. (2016) Chemicals as the Sole Transformers of Cell Fate. *International journal of stem cells*, **9**.
Edgar,L. *et al.* (2020) Regenerative medicine, organ bioengineering and transplantation. *Br J Surg*, **107**, 793–800.
Efremova,M. and Teichmann,S.A. (2020) Computational methods for single-cell omics across modalities. *Nat. Methods*, **17**, 14–17.
Elsland,D. and Neefjes,J. (2018) Bacterial infections and cancer. *EMBO Rep*, **19**.
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**.
Epstein,D.J. (2009) Cis-regulatory mutations in human disease. *Brief Funct Genomic Proteomic*, **8**, 310–316.
Factor,D.C. *et al.* (2020) Cell Type-Specific Intralocus Interactions Reveal Oligodendrocyte Mechanisms in MS. *Cell*, **181**, 382-395.e21.
Faith,J.J. *et al.* (2007) Large-Scale Mapping and Validation of Escherichia coli Transcriptional Regulation from a Compendium of Expression Profiles. *PLoS Biol*, **5**, e8.
Fang,F. *et al.* (2018) A distinct isoform of ZNF207 controls self-renewal and pluripotency of human embryonic stem cells. *Nat Commun*, **9**, 4384.
Fang,P. *et al.* (2018) Immune cell subset differentiation and tissue inflammation. *J Hematol Oncol*, **11**, 97.

Farassat,N. *et al.* (2019) In vivo functional diversity of midbrain dopamine neurons within identified axonal projections. *eLife*, **8**, e48408.
Fernandes,H.J.R. *et al.* (2020) Single-Cell Transcriptomics of Parkinson's Disease Human In Vitro Models Reveals Dopamine Neuron-Specific Stress Responses. *Cell Reports*, **33**, 108263.
Fishilevich,S. *et al.* (2017) GeneHancer: genome-wide integration of enhancers and target genes in GeneCards. *Database*, **2017**.
Ford Versypt,A.N. (2021) Multiscale modeling in disease. *Current Opinion in Systems Biology*, **27**, 100340.
Gabhann,F.M. *et al.* (2010) Gene Therapy from the perspective of Systems Biology. *Curr. Opin. Mol. Ther.*, **12**, 570.
Gao,T. and Qian,J. (2019) EnhancerAtlas 2.0: an updated resource with enhancer annotation in 586 tissue/cell types across nine species. *Nucleic Acids Research*, gkz980.
Gause,W.C. *et al.* (2020) Heterogeneity in the initiation, development and function of type 2 immunity. *Nat Rev Immunol*, **20**, 603–614.
Gill,N. *et al.* (2014) Computational Disease Gene Prioritization: An Appraisal. *Journal of Computational Biology*, **21**, 456–465.
Gitler,A.D. *et al.* (2017) Neurodegenerative disease: models, mechanisms, and a new hope. *Disease Models & Mechanisms*, **10**, 499–502.
Gough,A. *et al.* (2017) Biologically Relevant Heterogeneity: Metrics and Practical Insights. *SLAS Discov*, **22**, 213–237.
Grath,A. and Dai,G. (2019) Direct cell reprogramming for tissue engineering and regenerative medicine. *Journal of Biological Engineering*, **13**, 14.
Gu,Z. *et al.* (2012) Gene regulation is governed by a core network in hepatocellular carcinoma. *BMC Syst Biol*, **6**, 32.
Guo,M. *et al.* (2015) SINCERA: A Pipeline for Single-Cell RNA-Seq Profiling Analysis. *PLoS Comput. Biol.*, **11**, e1004575.
Guo,Rongqun *et al.* (2021) Single-cell map of diverse immune phenotypes in the acute myeloid leukemia microenvironment. *Biomarker Research*, **9**, 1–16.
Gyun Jee Song,K.S. (2017) Pharmacological Modulation of Functional Phenotypes of Microglia in Neurodegenerative Diseases. *Front. Aging Neurosci.*, **9**.
Hafemeister,C. and Satija,R. (2019) Normalization and variance stabilization of single-cell RNA-seq data using regularized negative binomial regression. *Genome Biol*, **20**, 296.
Hansson,O. (2021) Biomarkers for neurodegenerative diseases. *Nat Med*, **27**, 954–963.
Hariprakash,J.M. and Ferrari,F. (2019) Computational Biology Solutions to Identify Enhancers-target Gene Pairs. *Computational and Structural Biotechnology Journal*, **17**, 821–831.
Hartmann,A. *et al.* (2019) Modeling Cellular Differentiation and Reprogramming with Gene Regulatory Networks. In, Cahan,P. (ed), *Computational Stem Cell Biology*, Methods in Molecular Biology. Springer New York, New York, NY, pp. 37–51.
Hassan,M. *et al.* (2018) Computational modeling and biomarker studies of pharmacological treatment of Alzheimer's disease (Review). *Mol. Med. Rep.*, **18**, 639–655.
Hu,X. *et al.* (2020) Integration of single-cell multi-omics for gene regulatory network inference. *Computational and Structural Biotechnology Journal*, **18**, 1925–1938.
Hua,X. and Thompson,C.B. (2001) Quiescent T cells: actively maintaining inactivity. *Nat. Immunol.*, **2**, 1097–1098.
Huang,C., Yang,D., Ye,George W, *et al.* (2021) Vascular Notch Signaling in Stress Hematopoiesis. *Front. Cell Dev. Biol.*, **0**.
Huang,C., Yang,D., Ye,George W., *et al.* (2021) Vascular Notch Signaling in Stress Hematopoiesis. *Front. Cell Dev. Biol.*, **8**, 606448.

Huynh-Thu,V.A. *et al.* (2010) Inferring Regulatory Networks from Expression Data Using Tree-Based Methods. *PLoS ONE*, **5**, e12776.
Iacono,G. *et al.* (2019) Single-cell transcriptomics unveils gene regulatory network plasticity. *Genome Biol*, **20**, 110.
Ikeda,T. *et al.* (2018) Srf destabilizes cellular identity by suppressing cell-type-specific gene expression programs. *Nat. Commun.*, **9**, 1–15.
Iqbal Yatoo,Mohd. *et al.* (2021) Immunotherapies and immunomodulatory approaches in clinical trials - a mini review. *Human Vaccines & Immunotherapeutics*, 1–13.
Iwasaki,H. and Akashi,K. (2007) Myeloid Lineage Commitment from the Hematopoietic Stem Cell. *Immunity*, **26**, 726–740.
Januzaj,E. *et al.* (2004) DBDC: Density Based Distributed Clustering. In, *Advances in Database Technology - EDBT 2004*. Springer, Berlin, Heidelberg, pp. 88–105.
Jenner,A.L. *et al.* (2020) Leveraging Computational Modeling to Understand Infectious Diseases. *Curr. Pathobiol. Rep.*, **8**.
Jin,T. *et al.* (2021) scGRNom: a computational pipeline of integrative multi-omics analyses for predicting cell-type disease genes and regulatory networks. *Genome Med*, **13**, 95.
Jochems,S.P. *et al.* (2018) Inflammation induced by influenza virus impairs human innate immune control of pneumococcus. *Nat Immunol*, **19**, 1299–1308.
Jung,S. *et al.* (2021) A computer-guided design tool to increase the efficiency of cellular conversions. *Nat. Commun.*, **12**, 1–12.
Jung,S. *et al.* (2017) RefBool: a reference-based algorithm for discretizing gene expression data. *Bioinformatics*, **33**, 1953–1962.
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.
van Kampen,A.H.C. and Moerland,P.D. (2016) Taking Bioinformatics to Systems Medicine. In, Schmitz,U. and Wolkenhauer,O. (eds), *Systems Medicine*, Methods in Molecular Biology. Springer New York, New York, NY, pp. 17–41.
Kanehisa,M. *et al.* (2017) KEGG: new perspectives on genomes, pathways, diseases and drugs. *Nucleic Acids Res*, **45**, D353–D361.
Kang,R. *et al.* (2019) EnhancerDB: a resource of transcriptional regulation in the context of enhancers. *Database*, **2019**.
Kartha,V.K. *et al.* (2021) Functional Inference of Gene Regulation using Single-Cell Multi-Omics. *bioRxiv*.
Kikuchi,M. *et al.* (2019) Enhancer variants associated with Alzheimer's disease affect gene expression via chromatin looping. *BMC Med. Genomics*, **12**, 1–16.
Kim,H.J. *et al.* (2021) Defining cell identity beyond the premise of differential gene expression. *Cell Regen*, **10**, 20.
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.
Kim,S.Y. and Nair,M.G. (2019) Macrophages in wound healing: activation and plasticity. *Immunol Cell Biol*, **97**, 258–267.
Kimura,H. (2013) Histone modifications for human epigenome analysis. *J Hum Genet*, **58**, 439–445.
Kiselev,V.Y., Andrews,Tallulah S., *et al.* (2019) Challenges in unsupervised clustering of single-cell RNA-seq data. *Nat. Rev. Genet.*, **20**, 273–282.
Kiselev,V.Y., Andrews,Tallulah S, *et al.* (2019) Challenges in unsupervised clustering of single-cell RNA-seq data. *Nat. Rev. Genet.*, **20**, 273–282.

Kiselev,V.Y. *et al.* (2017) SC3: consensus clustering of single-cell RNA-seq data. *Nat Methods*, **14**, 483–486.
Kiser,A.K. and Pronovost,P.J. (2009) Management of diseases without current treatment options: something can be done. *JAMA*, **301**, 1708–1709.
Kitano,H. (2002) Computational systems biology. *Nature*, **420**, 206–210.
Klemm,S.L. *et al.* (2019) Chromatin accessibility and the regulatory epigenome. *Nat. Rev. Genet.*, **20**, 207–220.
Knip,M. *et al.* (2005) Environmental Triggers and Determinants of Type 1 Diabetes. *Diabetes*, **54**, S125–S136.
Kouli,A. *et al.* (2018) Chapter 1: Parkinson's Disease: Etiology, Neuropathology, and Pathogenesis. In, *Parkinson's Disease: Pathogenesis and Clinical Aspects*. Brisbane (AU): Codon Publications.
Koutrouli,M. *et al.* (2020) A Guide to Conquer the Biological Network Era Using Graph Theory. *Front. Bioeng. Biotechnol.*, **8**, 34.
Krieger,T. and Simons,B.D. (2015) Dynamic stem cell heterogeneity. *Development*, **142**, 1396–1406.
Kuhn,J.H. *et al.* (2019) Classify viruses — the gain is worth the pain. *Nature*, **566**, 318–320.
Kumar,A. and Mali,P. (2020) Mapping regulators of cell fate determination: Approaches and challenges. *APL Bioengineering*, **4**, 031501.
Kwon,H.S. and Koh,S.-H. (2020) Neuroinflammation in neurodegenerative disorders: the roles of microglia and astrocytes. *Transl. Neurodegener.*, **9**, 1–12.
La Manno,G. *et al.* (2021) Molecular architecture of the developing mouse brain. *Nature*, **596**, 92–96.
Lähnemann,D. *et al.* (2020) Eleven grand challenges in single-cell data science. *Genome Biol*, **21**, 31.
Lam,S. *et al.* (2020) A systems biology approach for studying neurodegenerative diseases. *Drug Discovery Today*, **25**, 1146–1159.
Landrum,M.J. *et al.* (2018) ClinVar: improving access to variant interpretations and supporting evidence. *Nucleic Acids Res*, **46**, D1062–D1067.
Lee,J. *et al.* (2020) Single-cell multiomics: technologies and data analysis methods. *Exp Mol Med*, **52**, 1428–1442.
Lee,J.T.H. and Hemberg,M. (2019) Supervised clustering for single-cell analysis. *Nat Methods*, **16**, 965–966.
Li,L. and Boussiotis,V.A. (2011) Molecular and functional heterogeneity of T regulatory cells. *Clinical Immunology*, **141**, 244–252.
Li,W. and Ding,S. (2010) Small molecules that modulate embryonic stem cell fate and somatic cell reprogramming. *Trends in Pharmacological Sciences*, **31**, 36–45.
Li,X. and Wang,C.-Y. (2021) From bulk, single-cell to spatial RNA sequencing. *Int J Oral Sci*, **13**, 36.
Liu,E. *et al.* (2019) Gene Regulatory Network Review. In, *Encyclopedia of Bioinformatics and Computational Biology*. Elsevier, pp. 155–164.
Liu,Q. *et al.* (2016) The cytokine storm of severe influenza and development of immunomodulatory therapy. *Cell Mol Immunol*, **13**, 3–10.
Liu,S.X. *et al.* (2020) Trajectory analysis quantifies transcriptional plasticity during macrophage polarization. *Sci Rep*, **10**, 12273.
Liu,Y. *et al.* (2018) CRISPR Activation Screens Systematically Identify Factors that Drive Neuronal Fate and Reprogramming. *Cell Stem Cell*, **23**.
Luecken,M.D. and Theis,F.J. (2019) Current best practices in single-cell RNA-seq analysis: a tutorial. *Mol Syst Biol*, **15**.

Lutshumba,J. *et al.* (2021) Dysregulation of Systemic Immunity in Aging and Dementia. *Front. Cell. Neurosci.*, **15**, 652111.
Lytal,N. *et al.* (2020) Normalization Methods on Single-Cell RNA-seq Data: An Empirical Survey. *Front. Genet.*, **11**, 41.
Lyu,P. *et al.* (2021) Gene regulatory networks controlling temporal patterning, neurogenesis, and cell-fate specification in mammalian retina. *Cell Reports*, **37**, 109994.
Ma,S.-X. and Lim,S.B. (2021) Single-Cell RNA Sequencing in Parkinson's Disease. *Biomedicines*, **9**, 368.
Mahajani,S. *et al.* (2019) Homogenous generation of dopaminergic neurons from multiple hiPSC lines by transient expression of transcription factors. *Cell Death Dis*, **10**, 898.
Manners,H.N. *et al.* (2016) Computational Methods for Detecting Functional Modules from Gene Regulatory Network. In, *Proceedings of the Second International Conference on Information and Communication Technology for Competitive Strategies - ICTCS '16*. ACM Press, Udaipur, India, pp. 1–6.
Marbach,D. *et al.* (2010) Revealing strengths and weaknesses of methods for gene network inference. *Proc. Natl. Acad. Sci. U.S.A.*, **107**, 6286–6291.
Mardinoglu,A. and Nielsen,J. (2016) Editorial: The Impact of Systems Medicine on Human Health and Disease. *Front. Physiol.*, **7**.
Margolin,A.A. *et al.* (2006) ARACNE: An Algorithm for the Reconstruction of Gene Regulatory Networks in a Mammalian Cellular Context. *BMC Bioinformatics*, **7**, S7.
Masuda,T. *et al.* (2020) Microglia Heterogeneity in the Single-Cell Era. *Cell Reports*, **30**, 1271–1281.
Matharu,N. and Ahituv,N. (2020) Modulating gene regulation to treat genetic disorders. *Nat Rev Drug Discov*, **19**, 757–775.
Maurano,M.T. *et al.* (2015) Large-scale identification of sequence variants influencing human transcription factor occupancy in vivo. *Nat Genet*, **47**, 1393–1401.
Mayor,R. (2019) Cell fate decisions during development. *Science*, **364**, 937–938.
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.
Mercatelli,D. *et al.* (2020) Gene regulatory network inference resources: A practical overview. *Biochimica et Biophysica Acta (BBA) - Gene Regulatory Mechanisms*, **1863**, 194430.
Mestas,J. and Hughes,C.C.W. (2004) Of Mice and Not Men: Differences between Mouse and Human Immunology. *J Immunol*, **172**, 2731–2738.
Meyer,P. and Saez-Rodriguez,J. (2021) Advances in systems biology modeling: 10 years of crowdsourcing DREAM challenges. *Cell Systems*, **12**, 636–653.
Miao,Z. *et al.* (2020) Putative cell type discovery from single-cell gene expression data. *Nat. Methods*, **17**, 621–628.
Miller-Jensen,K. *et al.* (2007) Common effector processing mediates cell-specific responses to stimuli. *Nature*, **448**, 604–608.
Miyamoto,K. *et al.* (2018) Direct In Vivo Reprogramming with Sendai Virus Vectors Improves Cardiac Function after Myocardial Infarction. *Cell Stem Cell*, **22**.
Moore,C.B. *et al.* (2010) Short hairpin RNA (shRNA): design, delivery, and assessment of gene knockdown. *Methods Mol Biol*, **629**, 141–158.
Moreau,Y. and Tranchevent,L.-C. (2012) Computational tools for prioritizing candidate genes: boosting disease gene discovery. *Nat Rev Genet*, **13**, 523–536.
Morris,S.A. *et al.* (2019) The evolving concept of cell identity in the single cell era. *Development*, **146**.
Morris,S.A. (2019) The evolving concept of cell identity in the single cell era. *Development*, **146**, dev169748.

Mortada,I. *et al.* (2021) Immunotherapies for Neurodegenerative Diseases. *Front. Neurol.*, **12**, 654739.
Mostafa,D. *et al.* (2020) Loss of β-cell identity and diabetic phenotype in mice caused by disruption of CNOT3-dependent mRNA deadenylation. *Commun Biol*, **3**, 476.
Motta,S. and Pappalardo,F. (2013) Mathematical modeling of biological systems. *Briefings in Bioinformatics*, **14**, 411–422.
Mou,T. *et al.* (2020) Reproducibility of Methods to Detect Differentially Expressed Genes from Single-Cell RNA Sequencing. *Front. Genet.*, **10**, 1331.
Moustaqil,M. *et al.* (2020) Biophysical Techniques for Target Validation and Drug Discovery in Transcription-Targeted Therapy. *Int J Mol Sci*, **21**, E2301.
Nasser,J. *et al.* (2021) Genome-wide enhancer maps link risk variants to disease genes. *Nature*, **593**, 238–243.
Nathan,A. *et al.* (2022) Single-cell eQTL models reveal dynamic T cell state dependence of disease loci. *Nature*.
Ner-Gaon,H. *et al.* (2017) JingleBells: A Repository of Immune-Related Single-Cell RNA-Sequencing Datasets. *J Immunol*, **198**, 3375–3379.
Nguyen,A. *et al.* (2018) Single Cell RNA Sequencing of Rare Immune Cell Populations. *Front. Immunol.*, **9**, 1553.
Nguyen,H. *et al.* (2021) A comprehensive survey of regulatory network inference methods using single cell RNA sequencing data. *Briefings in Bioinformatics*, **22**, bbaa190.
Nica,A.C. and Dermitzakis,E.T. (2013) Expression quantitative trait loci: present and future. *Phil. Trans. R. Soc. B*, **368**, 20120362.
Niewiadomska,G. *et al.* (2011) The cholinergic system, nerve growth factor and the cytoskeleton. *Behavioural Brain Research*, **221**, 515–526.
Nilsen,T.W. and Graveley,B.R. (2010) Expansion of the eukaryotic proteome by alternative splicing. *Nature*, **463**, 457–463.
Nimmo,R.A. *et al.* (2015) Primed and ready: understanding lineage commitment through single cell analysis. *Trends in Cell Biology*, **25**, 459–467.
Ogbeide,S. *et al.* (2022) Into the multiverse: advances in single-cell multiomic profiling. *Trends in Genetics*, S0168952522000774.
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**.
Ottman,R. *et al.* (1996) Relations of genetic and environmental factors in the etiology of epilepsy. *Ann Neurol*, **39**, 442–449.
Oulas,A. *et al.* (2019) Systems Bioinformatics: increasing precision of computational diagnostics and therapeutics through network-based approaches. *Briefings in Bioinformatics*, **20**, 806–824.
Papalexi,E. and Satija,R. (2017) Single-cell RNA sequencing to explore immune cell heterogeneity. *Nat. Rev. Immunol.*, **18**, 35–45.
Pappalardo,F. *et al.* (2016) Computational modeling of brain pathologies: the case of multiple sclerosis. *Brief. Bioinform.*, **19**, 318–324.
Pérez-Tur,J. (2006) Parkinson's disease genetics: a complex disease comes to the clinic. *The Lancet Neurology*, **5**, 896–897.
Perkel,J.M. (2021) Single-cell analysis enters the multiomics age. *Nature*, **595**, 614–616.
Pliner,H.A. *et al.* (2018) Cicero Predicts cis-Regulatory DNA Interactions from Single-Cell Chromatin Accessibility Data. *Mol Cell*, **71**, 858-871.e8.
Pliner,H.A. *et al.* (2019) Supervised classification enables rapid annotation of cell atlases. *Nat Methods*, **16**, 983–986.
Potter,S.S. (2018) Single-cell RNA sequencing for the study of development, physiology and disease. *Nat Rev Nephrol*, **14**, 479–492.

Pratapa,A. *et al.* (2020) Benchmarking algorithms for gene regulatory network inference from single-cell transcriptomic data. *Nat Methods*.
Prinz,M. and Priller,J. (2017) The role of peripheral immune cells in the CNS in steady state and disease. *Nat. Neurosci.*, **20**, 136–144.
Regev,A. *et al.* (2017) Science Forum: The Human Cell Atlas.
Reimand,J. *et al.* (2019) Pathway enrichment analysis and visualization of omics data using g:Profiler, GSEA, Cytoscape and EnrichmentMap. *Nat Protoc*, **14**, 482–517.
Rozenblatt-Rosen,O. *et al.* (2017) The Human Cell Atlas: from vision to reality. *Nature*, **550**, 451–453.
Rué,P. and Martinez Arias,A. (2015) Cell dynamics and gene expression control in tissue homeostasis and development. *Mol Syst Biol*, **11**, 792.
Sandy,A.R. and Maillard,I. (2009) Notch signaling in the hematopoietic system. *Expert Opinion on Biological Therapy*, **9**, 1383–1398.
Satija,R. *et al.* (2015) Spatial reconstruction of single-cell gene expression data. *Nat. Biotechnol.*, **33**, 495–502.
Satija,R. and Shalek,A.K. (2014) Heterogeneity in immune responses: from populations to single cells. *Trends in Immunology*, **35**, 219–229.
Schett,G. *et al.* (2021) Why remission is not enough: underlying disease mechanisms in RA that prevent cure. *Nat Rev Rheumatol*, **17**, 135–144.
Schwartz,M. *et al.* (2013) How Do Immune Cells Support and Shape the Brain in Health, Disease, and Aging? *J. Neurosci.*, **33**, 17587–17596.
Shastry,B.S. (2007) SNPs in disease gene mapping, medicinal drug development and evolution. *J Hum Genet*, **52**, 871–880.
Shats,I. *et al.* (2017) Expression level is a key determinant of E2F1-mediated cell fate. *Cell Death Differ.*, **24**, 626–637.
Sima,C. *et al.* (2009) Inference of gene regulatory networks using time-series data: a survey. *Curr. Genomics*, **10**, 416–429.
Singh,A.J. *et al.* (2018) Differential gene regulatory networks in development and disease. *Cell. Mol. Life Sci.*, **75**, 1013–1025.
Singhania,A. *et al.* (2019) Transcriptional profiling unveils type I and II interferon networks in blood and tissues across diseases. *Nat Commun*, **10**, 2887.
del Sol,A. and Jung,S. (2021) The Importance of Computational Modeling in Stem Cell Research. *Trends in Biotechnology*, **39**, 126–136.
Song,Q. *et al.* (2017) Identification of regulatory modules in genome scale transcription regulatory networks. *BMC Syst Biol*, **11**, 140.
Spitz,F. and Furlong,E.E.M. (2012) Transcription factors: from enhancer binding to developmental control. *Nat. Rev. Genet.*, **13**, 613–626.
Squair,J.W. *et al.* (2021) Confronting false discoveries in single-cell differential expression. *Nat Commun*, **12**, 5692.
Statisticat, LLC. (2021) LaplacesDemon: Complete Environment for Bayesian Inference. R package version 16.1.6.
Strzelecka,P.M. *et al.* (2018) Dissecting human disease with single-cell omics: application in model systems and in the clinic. *Dis Model Mech*, **11**, dmm036525.
Stuart,T. *et al.* (2019) Comprehensive Integration of Single-Cell Data. *Cell*, **177**, 1888-1902.e21.
Stuart,T. *et al.* (2020) Multimodal single-cell chromatin analysis with Signac Genomics.
Sun,X. *et al.* (2022) A comprehensive comparison of supervised and unsupervised methods for cell type identification in single-cell RNA-seq. *Briefings in Bioinformatics*, **23**, bbab567.

Szabo,P.A. *et al.* (2019) Single-cell transcriptomics of human T cells reveals tissue and activation signatures in health and disease. *Nat. Commun.*, **10**, 1–16.
The Tabula Muris Consortium *et al.* (2018) Single-cell transcriptomics of 20 mouse organs creates a Tabula Muris. *Nature*, **562**, 367–372.
The Tabula Sapiens Consortium *et al.* (2022) The Tabula Sapiens: A multiple-organ, single-cell transcriptomic atlas of humans. *Science*, **376**, eabl4896.
The Tabula Sapiens Consortium and Quake,S.R. (2021) The Tabula Sapiens: a multiple organ single cell transcriptomic atlas of humans Cell Biology.
Tong Ihn Lee,R.A.Y. (2013) Transcriptional Regulation and its Misregulation in Disease. *Cell*, **152**, 1237.
Tran,H.T.N. *et al.* (2020) A benchmark of batch-effect correction methods for single-cell RNA sequencing data. *Genome Biology*, **21**, 12.
Trapnell,C. (2015) Defining cell types and states with single-cell genomics. *Genome Res.*, **25**, 1491–1498.
Tusi,B.K. *et al.* (2018) Population snapshots predict early haematopoietic and erythroid hierarchies. *Nature*, **555**, 54–60.
Vallejos,C.A. *et al.* (2017) Normalizing single-cell RNA sequencing data: challenges and opportunities. *Nat Methods*, **14**, 565–571.
Vasan,L. *et al.* (2021) Direct Neuronal Reprogramming: Bridging the Gap Between Basic Science and Clinical Application. *Frontiers in Cell and Developmental Biology*, **9**.
Visscher,P.M. *et al.* (2017) 10 Years of GWAS Discovery: Biology, Function, and Translation. *The American Journal of Human Genetics*, **101**, 5–22.
Visscher,P.M. *et al.* (2021) Discovery and implications of polygenicity of common diseases. *Science*, **373**, 1468–1473.
Võsa,U. *et al.* (2021) 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.
Wang,D. and Bodovitz,S. (2010) Single cell analysis: the new frontier in 'omics'. *Trends in Biotechnology*, **28**, 281–290.
Wang,F. *et al.* (2019) SCMarker: Ab initio marker selection for single cell transcriptome profiling. *PLoS Comput. Biol.*, **15**, e1007445.
Wang,N.B. *et al.* (2020) Engineering cell fate: Applying synthetic biology to cellular reprogramming. *Current Opinion in Systems Biology*, **24**, 18–31.
Wang,Y. and Navin,N.E. (2015) Advances and applications of single-cell sequencing technologies. *Mol Cell*, **58**, 598–609.
Wang,Y.-C. *et al.* (2013) Protein post-translational modifications and regulation of pluripotency in human stem cells. *Cell Res.*, **24**, 143–160.
Ward,L.D. and Kellis,M. (2012) Interpreting noncoding genetic variation in complex traits and human disease. *Nat. Biotechnol.*, **30**, 1095–1106.
Warner,T.T. and Schapira,A.H.V. (2003) Genetic and environmental factors in the cause of Parkinson's disease. *Ann Neurol*, **53 Suppl 3**, S16-23; discussion S23-25.
Watanabe,K. *et al.* (2019) Genetic mapping of cell type specificity for complex traits. *Nat Commun*, **10**, 3222.
Watcham,S. *et al.* (2019) New insights into hematopoietic differentiation landscapes from single-cell RNA sequencing. *Blood*, **133**, 1415–1426.
Weatherall,D.J. (2000) Single gene disorders or complex traits: lessons from the thalassaemias and other monogenic diseases. *BMJ*, **321**, 1117–1120.
Wei,T. *et al.* (2022) Regeneration of β cells from cell phenotype conversion among the pancreatic endocrine cells. *Chronic Diseases and Translational Medicine*, **8**, 1–4.
Weighill,D. *et al.* (2021) Gene Targeting in Disease Networks. *Front. Genet.*, **12**, 649942.

Wild,S.L. and Tosh,D. (2021) Molecular mechanisms of transcription factor mediated cell reprogramming: conversion of liver to pancreas. *Biochemical Society Transactions*, **49**, 579–590.
Wixon,J. (2001) Pathway databases. *Comp Funct Genomics*, **2**, 391–397.
Wolf,I.R. *et al.* (2021) Three topological features of regulatory networks control life-essential and specialized subsystems. *Sci Rep*, **11**, 24209.
Wolkenhauer,O. *et al.* (2013) The road from systems biology to systems medicine. *Pediatr. Res.*, **73**, 502–507.
Wong,A.K. *et al.* (2021) Decoding disease: from genomes to networks to phenotypes. *Nat Rev Genet*, **22**, 774–790.
Wray,G.A. *et al.* (2003) The Evolution of Transcriptional Regulation in Eukaryotes. *Mol. Biol. Evol.*, **20**, 1377–1419.
Wu,T. *et al.* (2021) clusterProfiler 4.0: A universal enrichment tool for interpreting omics data. *The Innovation*, **2**, 100141.
Xia,B. *et al.* (2020) Machine learning uncovers cell identity regulator by histone code. *Nat Commun*, **11**, 2696.
Xie,J. *et al.* (2019) It is time to apply biclustering: a comprehensive review of biclustering applications in biological and biomedical data. *Brief. Bioinform.*, **20**, 1450.
Yamanaka,S. (2020) Pluripotent Stem Cell-Based Cell Therapy-Promise and Challenges. *Cell Stem Cell*, **27**, 523–531.
Yang,P. *et al.* (2021) Feature selection revisited in the single-cell era. *Genome Biol*, **22**, 321.
Yip,S.H. *et al.* (2019) Evaluation of tools for highly variable gene discovery from single-cell RNA-seq data. *Brief Bioinform*, **20**, 1583–1589.
Yu,F. *et al.* (2022) Variant to function mapping at single-cell resolution through network propagation. *Nature Biotechnology*.
Zarayeneh,N. *et al.* (2017) Integration of multi-omics data for integrative gene regulatory network inference. *Int. J. Data Min. Bioinform.*, **18**.
Zhang,A.W. *et al.* (2019) Probabilistic cell-type assignment of single-cell RNA-seq for tumor microenvironment profiling. *Nat Methods*, **16**, 1007–1015.
Zhang,J. and Zhang,S. (2013) Modular Organization of Gene Regulatory Networks. In, Dubitzky,W. *et al.* (eds), *Encyclopedia of Systems Biology*. Springer New York, New York, NY, pp. 1437–1441.
Zhang,K. *et al.* (2021) A single-cell atlas of chromatin accessibility in the human genome. *Cell*, **184**, 5985-6001.e19.
Zhang,L. *et al.* (2022) DIRECT-NET: An efficient method to discover cis-regulatory elements and construct regulatory networks from single-cell multiomics data. *Sci. Adv.*, **8**, eabl7393.
Zhang,X. *et al.* (2019) CellMarker: a manually curated resource of cell markers in human and mouse. *Nucleic Acids Research*, **47**, D721–D728.
Zhao,Mingming *et al.* (2020) The Application of Single-Cell RNA Sequencing in Studies of Autoimmune Diseases: a Comprehensive Review. *Clin. Rev. Allergy Immunol.*, **60**, 68–86.
Zheng,M. (2021) Computational guided frameworks to identify signalling perturbations for cellular transitions: application to cellular conversion, disease and regeneration.