Prediction of Transcription Factor Activity in Single Cells

Forudsigelse af Transkriptionsfaktoraktivitet i Enkelte Celler

Master's Thesis - Computational Biomedicine
Written by

Toke Meyer

June 2025

Supervisor:

Associate Prof. Jesper Grud Skat Madsen

Co-supervisor:

Postdoc Andreas Fønss Møller

Number of Characters ≈ 117.622

SDU
Syddansk Universitet

MADLAB

Preface

I would like to thank Jesper Grud Skat Madsen, Associate Professor at the Department of Biochemistry and Molecular Biology (BMB) and principal investigator in the MadLab Group at SDU, for the opportunity to write both my bachelor's thesis, ISA, and now my master's thesis under his supervision. I have thoroughly enjoyed learning and broadening my understanding in the field of bioinformatics under his supervision. The insights I've gained will undoubtedly serve as a strong foundation for further developing and refining my bioinformatic abilities in the future. I would also like to thank Postdoc Andreas Fønss Møller for his supervision and valuable insights into mathematics and machine learning throughout my master's programme. Additionally, I would like to thank the rest of the MadLab Group and my office colleagues for our numerous engaging discussions related to the project, as well as our informal discussions and conversations. Finally, I would like to extend a special thank you to my mother and the rest of my family for their unwavering support throughout my master's project.

Table of Contents

Indholdsfortegnelse

Abstract (271 words)

During the development of a human fetus from a zygote, proteins known as transcription factors play a critical role in regulating gene expression and, consequently, the differentiation of embryonic stem cells into diverse cell types [1]. Given that transcription factors are vital for both regulating gene expression and establishing, as well as maintaining, the identities of cells, there is a natural interest in understanding the intricate interplay between transcription factors and genes [2]. This understanding may provide deeper insights into how transcriptional regulation can be disrupted or contribute to the development of disease [2].

While experimental approaches can achieve this, they are often expensive and time-consuming [3]. As a result, gene regulatory network inference tools like IMAGE can be advantageous for predicting transcription factor-gene interactions [4]. IMAGE estimates motif activities at both enhancer and gene levels, and it also predicts transcription factor target genes [4]. The enhancer motif activities generated by IMAGE were validated and compared to those of chromVAR, another computational motif activity prediction tool, using a dataset of single-cell human embryonic stem cells transduced with transcription factors to direct their differentiation [4], [5], [6].

The motif activities estimated by IMAGE appeared unstable and could potentially be unreliable, as they did not reflect the expected gradual increase in overexpressed transcription factor enrichment levels. Furthermore, when the target gene predictions of IMAGE were compared to a novel enhancer-gene prediction tool called scE2G, it appeared that many assumed to be positive enhancer-gene links were not captured by IMAGE [7]. This suggests that the current method of defining target genes in IMAGE is not suitable for modelling the complexity of gene regulatory networks.

Resume (250 Ord)

Under udviklingen af et menneskeligt foster fra en zygot, spiller proteiner kendt som transkriptionsfaktorer en afgørende rolle i reguleringen af genekspression og dermed differentieringen af embryonale stamceller til en lang række celletyper [1]. Da transkriptionsfaktorer er essentielle både for reguleringen af genekspression og for etableringen samt opretholdelsen af cellers identitet, er der naturligt stor interesse i at forstå det komplekse samspil mellem transkriptionsfaktorer og gener [2]. En sådan forståelse kan give dybere indsigt i, hvordan den transkriptionelle regulering kan forstyrres eller bidrage til udviklingen af sygdomme [2].

Eksperimentelle tilgange kan afdække disse sammenhænge, men de er ofte dyre og tidskrævende [3]. Derfor kan værktøjer til inferens af genregulatoriske netværk, såsom IMAGE, være fordelagtige til at forudsige interaktioner mellem transkriptionsfaktorer og gener [4]. IMAGE estimerer motivaktiviteter både på enhancer- og genniveau og forudsiger desuden transkriptionsfaktorers målgener [4]. De enhancer-motivaktiviteter, som IMAGE genererede, blev valideret og sammenlignet med chromVAR, et andet komputationel værktøj til forudsigelse af motivaktivitet, ved hjælp af et datasæt bestående af single-cell humane embryonale stamceller, der var transduceret med transkriptionsfaktorer for at skubbe differentieringen i en bestemt retning [4], [5], [6].

Motivaktiviteterne estimeret af IMAGE fremstod ustabile og kan potentielt være upålidelige, eftersom de ikke afspejlede den forventede gradvise stigning i berigelsesniveauet af overudtrykte transkriptionsfaktorer. Desuden viste en sammenligning af IMAGEs målgenforudsigelser med et nyt enhancer-gen-forudsigelsesværktøj, scE2G, at mange formodede positive enhancer-gen-koblinger ikke blev fanget af IMAGE. Dette indikerer, at den nuværende metode til at definere målgener i IMAGE ikke er tilstrækkelig til at modellere kompleksiteten i genregulatoriske netværke.

Aim of Study

This study aims to validate a gene regulatory inference tool known as IMAGE (Integrated Analysis of Motif Activity and Gene Expression Changes of Transcription Factors) [4]. IMAGE estimates enhancer and gene motif activities and predicts target enhancers and genes at a bulk level [4]. At the time of IMAGE's release in 2018, high-quality single-cell transcription factor overexpression data were not available for validating the motif activities estimated by IMAGE. However, a dataset published by Joung et al. 2023 has provided such high-quality data [6]. In addition to validating IMAGE with the overexpression data, it is also desirable to compare the performance of IMAGE to chromVAR and scE2G, which respectively estimate transcription factor motif activities and target genes of regulatory elements [4], [5], [7].

Introduction

Trajectory of Stem Cell Potential in Early Development

At human fertilisation, when a single totipotent stem cell, known as a zygote, is formed, it undergoes mitotic division during preimplantation embryogenesis to create an early blastocyst, as shown in Figure 1A [8]. The blastocyst consists of an outer cell wall of trophoblast cells, which can develop into the placenta, and an inner cell mass (ICM) of pluripotent stem cells, which can develop into a human fetus. During the development of the blastocyst, the totipotent stem cells' differentiation potential gradually decreases as they progress towards a state of pluripotency [8], [9]. Totipotent stem cells are characterised by their unbiased differentiation potential, which means they can turn into any cell type, including extraembryonic tissues such as trophoblasts. In contrast, pluripotent stem cells can differentiate into somatic cells of the three germ layers (ectoderm, mesoderm and endoderm) and germline cells, as seen in Figure 1A [8]. The pluripotent stem cells of the ICM will continue to differentiate and mature into multipotent cells. These cells only have a limited differentiation potential and can differentiate into a limited number of cell types. An example of this is the development of multipotential hematopoietic stem cells into lymphoid progenitor-like cells, which can differentiate into T, B, and NK cells, as illustrated in Figure 1A [9], [10].

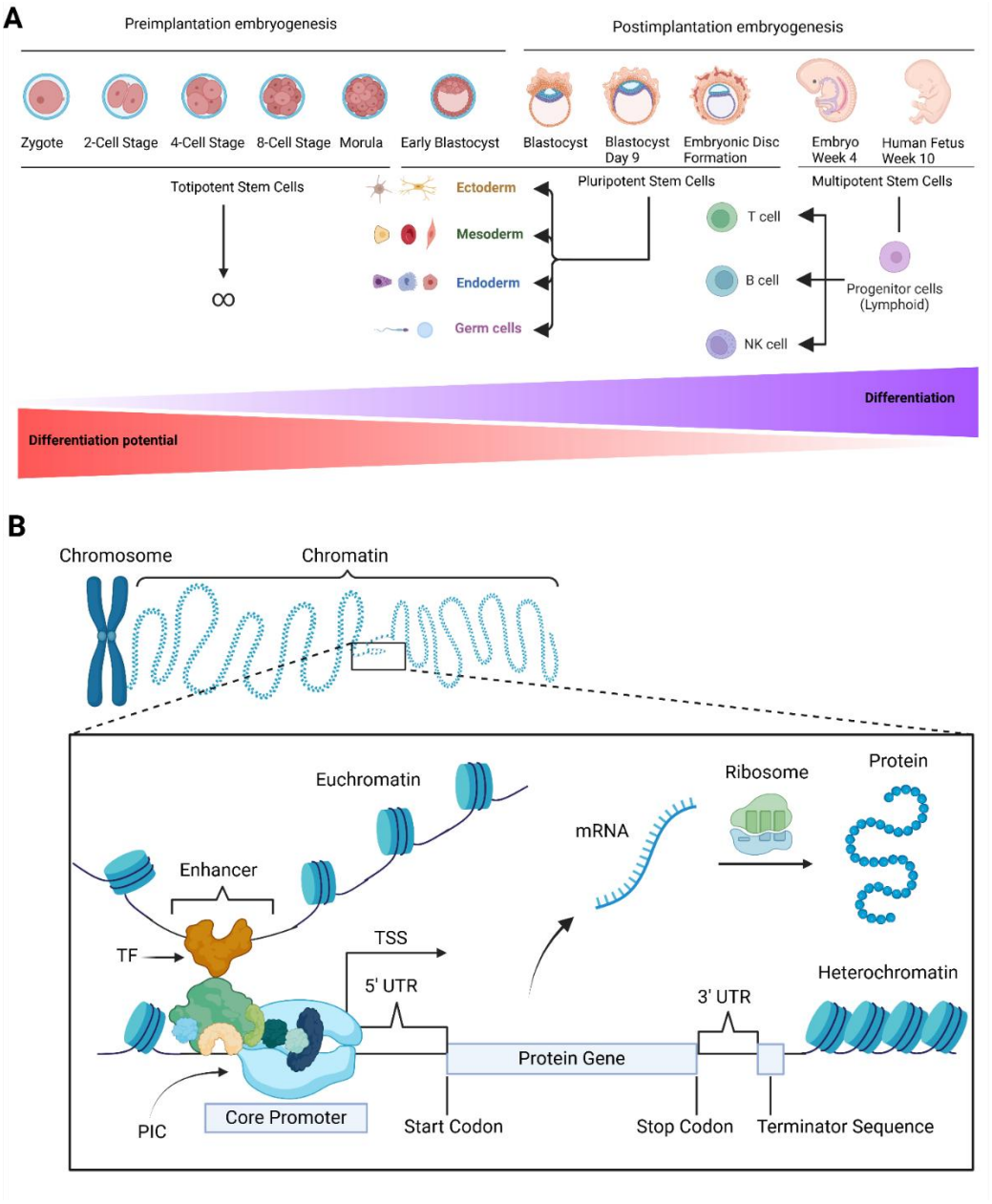


Figure 1. Cartoon schematic of embryonic development and eukaryotic gene expression. *(A) A cartoon schematic illustrating embryonic development and differentiation potency during pre-implantation and post-implantation stages of the fetus. The stem cells from the zygote until the morula stage can differentiate into any cell type (totipotency), as indicated by the infinity symbol. From the early blastocyst stage to the formation of the embryonic disc, stem cells can differentiate into somatic cells of the three germ layers and germline cells, demonstrating their multipotency. From week 4 and onwards, stem cells exhibit limited differentiation potential* ***[8], [9], [10].*** *(B) A simplified cartoon schematic depicting the eukaryotic regulation of gene expression. General TFs assemble on the Core Promoter, recruit RNA polymerase II, and form the PIC. Transcription of the gene begins at the TSS and terminates at a terminator sequence. The transcription is regulated by other TFs that bind to DNA elements such as enhancers. For a protein-encoding gene, the mRNA is translated into a protein via the ribosome. Only genes that are accessible due to open chromatin, euchromatin, can be transcribed, in contrast to genes in closed chromatin, heterochromatin* ***[1], [11], [12].*** *Abbreviations: Pre-initiation complex (PIC), five prime untranslated region (5' UTR), three prime untranslated region (3' UTR), transcription factor (TF), transcriptional start site (TSS) and messenger RNA (mRNA). Made in Biorender.com.*

The Central Role of the Epigenome in Cellular Identity

The progressive decrease in the differentiation potential of stem cells as they differentiate into more heterogeneous cell types, states, and identities is ultimately reflected and shaped by the cells' unique epigenome, transcriptome, and proteome [8], [13], [14]. A cell type can be defined, from a historical perspective, as a cell's observable features, such as its shape and biological function, both in vivo and in vitro [13]. In contrast, a cell's state refers to the dynamic and responsive changes in its epigenome, transcriptome, and proteome in response to environmental stimuli or the absence thereof, whereas a cell's identity can be considered as the cell's unique combination of these molecular profiles [8], [13]. The transcriptome and proteome are, respectively, the complete sets of RNA transcripts and proteins that a cell, tissue, or organism can express. In contrast, the epigenome encompasses concepts such as DNA accessibility and histone modification [13]. DNA that is considered accessible is referred to as euchromatin, while inaccessible DNA is designated as heterochromatin, as illustrated in Figure 1B [11].

Histones are protein octamers, consisting of two H2A, H2B, H3, and H4 subunits, that play a vital role in packaging DNA. A single DNA and histone complex is referred to as a nucleosome, whereas entire histone and DNA complexes are known as chromatin [11], [15]. The DNA and Histones form these DNA-protein complexes due to their opposite charges. DNA is negatively charged due to its phosphate backbone, whereas histones are positively charged as a result of the abundant lysine and arginine residues in histones, where two-thirds of these specific residues are concentrated in the histone tails [16]. In addition to wrapping DNA, histones regulate the chromatin structure and define euchromatin and heterochromatin regions through post-translational modifications (PTMs) of the histones. There are various PTMs, but the most common ones are methylation, phosphorylation, and acetylation [17]. Acetylation and phosphorylation of the histone lysine tails are generally associated with euchromatin because they neutralise the positive charge of the tail and thereby disrupt the electrostatic interactions between histones and DNA [18]. In contrast, methylation does not alter the electrostatic interaction between histone and DNA, but can facilitate the binding of chromatin factors, also referred to as readers [18], [19]. An example is the trimethylated (me3) lysine residue number 9 (K9) on the H3 histone subunit (H3K9me3) [18]. The methylation of this subunit allows for the binding of the reader protein HP1 to the methylated lysine, whereafter two HP1 readers can dimerise with each other and bring the nucleosomes closer and create a more compact chromatin structure or heterochromatin, which is associated with

silenced or non-actively transcribed genes [17], [18]. However, histone methylations do not only provide binding platforms for chromatin factors; they can also disrupt binding interactions. This is the case for H3K4me3, which prevents the binding of the Nucleosome Remodelling and Deacetylase (NuRD) protein complex, which regulates gene expression by acting as a transcriptional repressor [18].

Overall, chromatinized DNA is folded into stable structures of heterochromatin and euchromatin, which can be remodelled through PTMs. As a result, PTMs and chromatin architecture play crucial roles in regulating DNA-templated processes, such as transcription. The unique conformation of the epigenome and the consequent patterns in gene expression (transcriptome) and protein production (proteome) are central to progressively shaping a cell and defining its type, state, and identity during its differentiation from a totipotent stem cell.

Transcription Factors as Architects of Cellular Identity

Although chromatin structure is critical for regulating the transcription of DNA to messenger RNA (mRNA) and the differentiation process from a totipotent stem cell to a unique cell type, state and identity, the mechanisms of gene regulation and differentiation are far more intricate and encompass complexity beyond simply the chromatin structure. For instance, transcription factors (TFs) are proteins that are documented to play a key role in regulating transcription and differentiation; thereby, cellular identity can also be considered a property of TF activity [14]. These regulatory proteins bind to specific regions in the DNA known as cis-regulatory elements (CREs). Specifically, TFs bind to 6–12 bp-long degenerate DNA sequences within CREs, known as motifs, using their DNA binding domain (DBD) [1]. Degenerate DNA sequences mean that a strict sequence match at every nucleotide position in the DNA is not necessary for the TF to bind. CREs can be localised either near the transcriptional start site (TSS) or at greater distances away from the TSS [1]. CREs in the immediate vicinity of the TSS and encompassing the TSS are referred to as core promoters, as illustrated in Figure 1B [20]. Core promoters facilitate the recruitment of general TFs (GTFs) and RNA polymerase II to assemble a protein complex known as the pre-initiation complex (PIC), as seen in Figure 1B [1], [20]. TFs can be considered general if they are expressed in every cell type, or they can be regarded as cell-specific if they are only expressed in specific cells and are thus unique to those cell types [1].

The formation of the PIC is considered the rate-limiting step in transcriptional activation, and not only does it serve the purpose of forming the first few phosphodiester bonds at the active site of the

RNA Polymerase II, to ensure proper initialisation of an mRNA elongation form; but the PIC is ultimately the evolutionary result of a need to stabilise the transcriptional system and allows for more functional flexibility and control of the transcriptional process [12]. This greater control and flexibility are achieved through other CREs, such as enhancers and silencers, which respectively increase, reduce, or even inhibit the transcriptional output entirely. These can be found either proximal to or at greater distances away from the TSS [1]. Enhancers and silencers likewise serve as a platform, where typically cell-specific TFs bind to and, through a process known as looping, can interact with the PIC and regulate the gene expression [1], [21]. In the looping process, a protein called cohesion loads onto DNA and performs loop extrusion, in which DNA is pulled from both ends towards and through cohesion, bringing CREs, such as enhancers or silencers, closer to the core promoter and PIC [1], [21]. It is through a TF's effector domains that it regulates gene expression through several mechanisms, as seen in Figure 2.

It can engage in direct protein-protein interaction with the PIC and either enhance or inhibit transcription, or the TF can cooperatively bind with other TFs (cofactors) through its effector domain and interact with the PIC together with these cofactors and regulate gene expression [22]. Furthermore, it can regulate gene expression by recruiting writer enzymes such as histone acetyltransferases, deacetylases, methyltransferases, and demethylases to remodel chromatin structure through histone PTMs, including acetylation and methylation, thus shaping heterochromatin and euchromatin. Histone acetyltransferases and methyltransferases, respectively, add acetylation and methylation to histone lysine residues, while deacetylases and demethylases remove them [18], [22]. Some TFs, known as pioneer TFs, are even capable of binding to DNA in heterochromatin structures, thereby facilitating a cascade of changes. These changes include recruiting chromatin remodelling enzymes to trigger nucleosome repositioning, making the DNA more accessible and allowing the binding of other TFs to the now accessible DNA [1]. However, the binding of a pioneer TF in itself is insufficient to induce nucleosome repositioning, and recruitment of other proteins is necessary for this process [1]. Finally, in addition to facilitating protein-protein interactions and chromatin remodelling to regulate the transcriptional process, the effector domains of a TF may also regulate the activity of the TF through direct PTMs on the TF itself or by binding with molecules other than proteins known as ligands, to turn the TF on or off [22].

As evident, TFs play a critical role in the regulation of gene expression and thereby heavily contributes to shaping a cells type, state and identity, through various mechanisms ranging from direct protein-protein interactions with the PIC, with or without cofactors, remodelling the

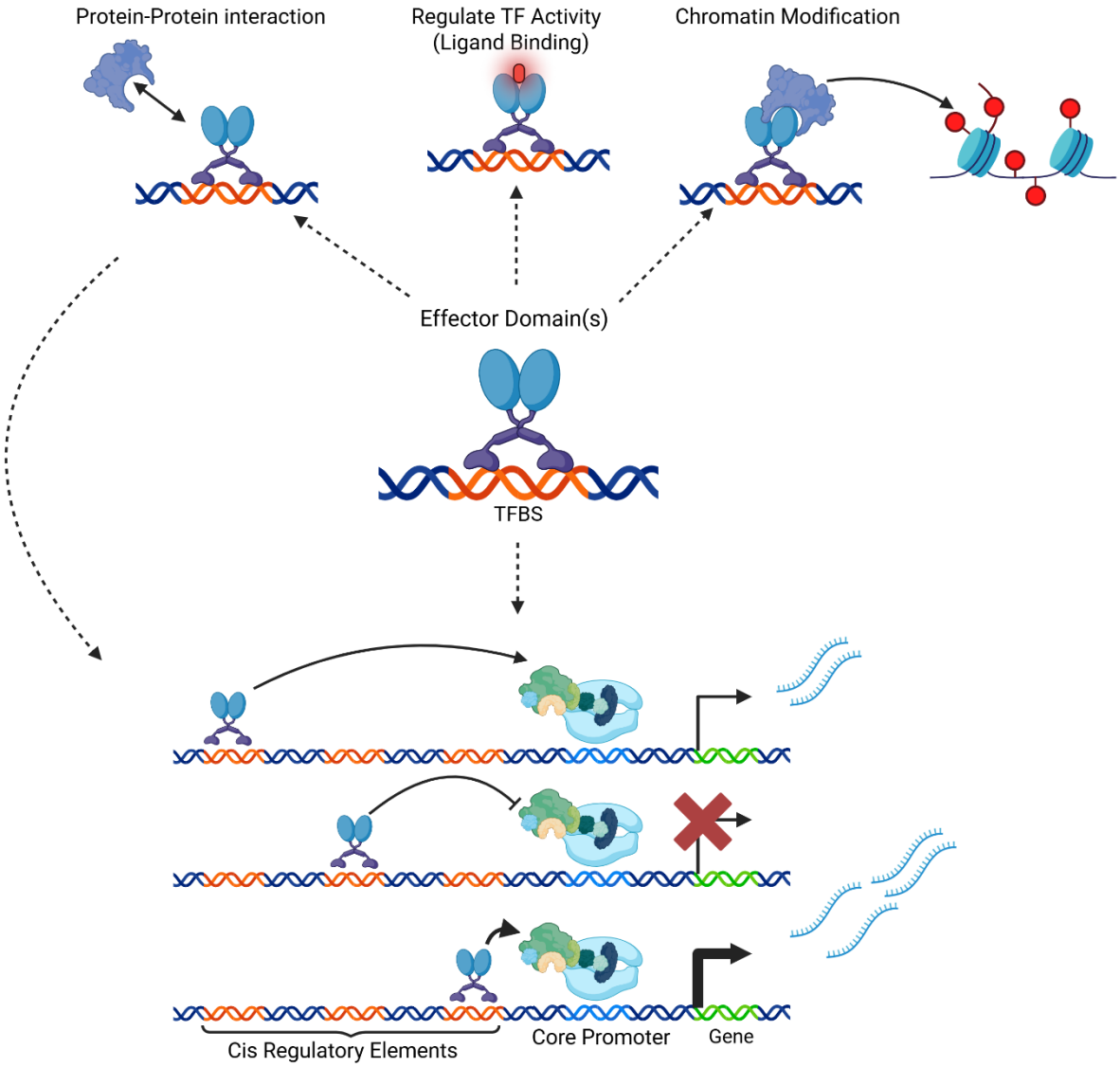


Figure 2. Cartoon illustration of TF effector domain(s) functionalities. *TFs bind to their TFBS and exert various regulatory effects through their effector domains. TFs can interact with the PIC assembled at the Core Promoter directly or from a distance. Some cis-regulatory elements that the TFs bind to may cause the TF to inhibit transcription rather than promote it. In addition to directly interacting with the PIC, the TF effector domains may also facilitate the binding of cofactors, which can then interact directly with the PIC. Through protein-protein interactions, TFs may bind to other proteins that are capable of modifying the chromatin landscape. Finally, TF effector domains may function as regulators of the TFs' activity through binding of ligands [1], [22]. Made in Biorender.com.*
Abbreviations: Transcription Factor (TF), Transcription Factor Binding Site (TFBS).

chromatin structure and regulating the activity of the TF itself, either through ligand binding or PTMs through their effector domains.

Inferring Gene Regulatory Networks: A Computational Perspective

The intricate regulation of target genes by TFs and their respective target CREs is called gene regulatory networks (GRNs). Naturally, it is of interest to study GRNs to gain insight into and understand how cellular identity is established, maintained, and disrupted in disease [2]. GRNs can

be inferred using experimental omics data, such as epigenomic, transcriptomic, and proteomic measurements [2]. An example of this would be to use epigenomic and transcriptomic assays such as chromatin immunoprecipitation followed by sequencing (ChIP-seq), and RNA-seq data, in combination with perturbation studies to investigate the interplay between genes, enhancers and TFs. RNA-seq is a technique used to quantify the entire transcriptome of a given species for each active gene [23]. ChiP-seq is a technique where DNA associated with a DNA-binding protein becomes enriched, and because of this, it can be used to identify which DNA sequences TFs bind to [24]. Utilising these approaches in combination allows for the identification of which CREs are bound by a given TF and to profile the resulting gene expression patterns. By perturbing or mutating the corresponding TF binding sites in a follow-up analysis, potential TF-gene links and interactions can be inferred from the altered gene expression. However, there are currently around 1600 DNA-binding TFs that have been identified, and to experimentally validate the binding site and regulatory significance of each TF for each gene is a time-consuming and expensive task [3], [25]. Because of this, it is of great interest to computationally infer GRNs as it would be less costly and time-consuming.

There are many different tools for computationally inferring GRNs, which leverage various techniques. Some tools are based on correlation, probabilistic models, dynamical systems, regression, and the currently very popular method, deep learning [26]. However, deep learning models are less interpretable compared to other methods, mainly because of their black-box nature, where the hidden layers perform non-linear input–output mappings [27], [28]. Therefore, it can be favourable to use computational tools that utilise other methods that do not lack interpretability to the same extent. One such tool is IMAGE (Integrated analysis of Motif Activity and Gene Expression changes of transcription factors), developed by Madsen et al. 2018, which utilises linear regression in its workflow for computing GRNs [4]. Like many other tools, IMAGE uses RNA and chromatin assays as inputs for inferring GRNs, such as RNA-seq data for quantifying the actively transcribed genes and assays for transposase-accessible chromatin using sequencing (ATAC-seq) data, which detects the chromatin landscape of cells [23], [26], [29], [30]. In short, ATAC-seq data involves isolating the nuclei of cells and exposing them to a hyperactive Tn5 transposase, which loads onto euchromatin, cuts it into fragments, and attaches sequencing adapters to the ends of the fragments used for the amplification process [30]. Using ATAC-seq data in combination with TF position weight matrices (PWMs) is a faster and more cost-effective alternative to performing ChIP-seq on all TFs, as it enables the identification and inference of TF binding sites in accessible

DNA. The PWM is a mathematical model that describes the log-likelihood of a TF binding to a DNA sequence [4]. However, the issue with ATAC-seq data is that it is very sparse, especially for single-cell ATAC-seq (scATAC-seq) data, as there are typically only 1-2 reads of a given locus for a single cell in organisms with a diploid genome [31], [32]. On the other hand, single-cell data capture heterogeneity within single-cell populations, whereas bulk data, although less sparse and noisy, mask heterogeneity [31], [32]. Noise in this case refers to detected peak signals, which are merely technical machine apparatus noise [31], [32]. Due to the trade-off between sparsity and noise versus cellular heterogeneity, a strategy could be to aggregate similar cells in single-cell data into pseudobulk clusters. By aggregating single-cell data into meaningful clusters, it may be possible to mitigate the sparsity of the data while preserving some cellular heterogeneity within the clusters [32].

The Working Principles Behind IMAGE

Motif Activity Predictions and the Issue With Multicollinearity

As such, bulk and pseudobulk RNA and chromatin assays, such as RNA- and ATAC-seq data, can be used as input for IMAGE [4]. IMAGE can be divided into two steps. Step one, where the enhancer motif activity and target enhancers are predicted, and step two, where the gene motif activity and target genes are predicted, as seen in Figure 3A [4]. The terms enhancer motif activity and gene motif activity refer to how much the TF contributes to enhancer activity and how much the TF contributes to gene expression [4]. Both steps predict motif activities through a linear regression model, as it is simple to interpret, since the regression coefficients of the TFs function serve as a proxy for their motif activities [4]. However, one of the assumptions of linear regression is that the independent variables or features are not allowed to be correlated, a phenomenon known as multicollinearity [33]. The issue with multicollinearity is that it makes the model unreliable. When two or more features are correlated, they behave in a similar manner, and linear regression analysis cannot assign accurate weights to each feature. The correlated features could potentially get a considerably large or small weight. This results in the model overfitting to the data, and as a consequence, it inflates the variance of the estimated coefficients, making the model less generalisable. This means that when the model is applied to data it has not been trained on, it performs poorly, and the standard errors of the coefficients become large [33].

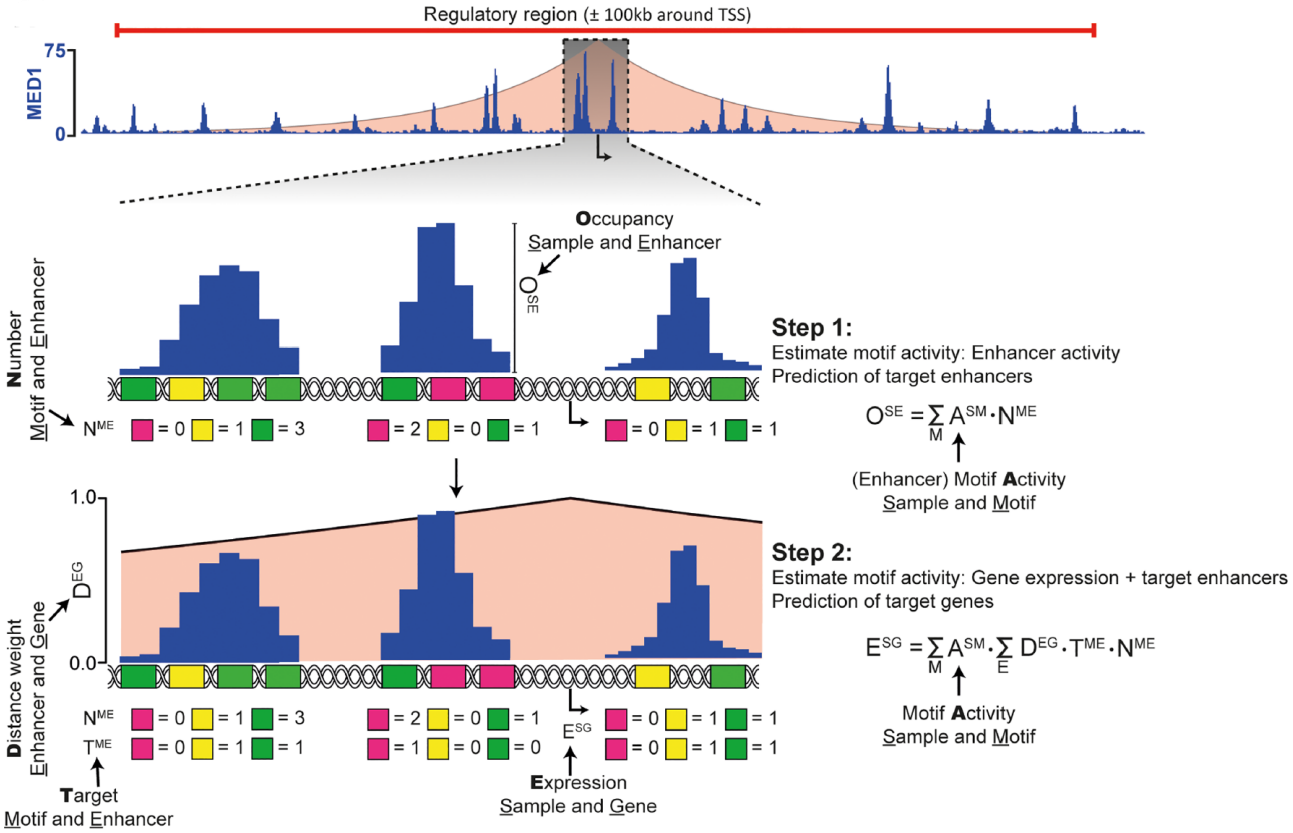


Figure 3. Schematic workflow of IMAGE. *IMAGE estimates motif activities and predicts target enhancers and genes in two steps. In step 1, IMAGE first estimates the activity of each motif in each sample using ridge regression, as shown in the equation in the figure. N^{ME} is the number of motifs of motif M in each enhancer E, while O^{SE} is the chromatin accessibility assay signal measured as the TF occupancy at the enhancer E, measured in sample S. Every transcription-factor motif M that occurs in enhancer E is summed to gain the motif activity per sample. After estimation of the enhancer motif activity, IMAGE predicts target enhancers E of motif M (T^{ME}). In step 2, the number of motifs M in the target enhancer E (T^{ME}*N^{ME}) is weighted based on the distance between gene G and enhancer E (D^{EG}) and is summed for every motif M's distance to gene G. The weight is based on a regulatory region around 100kb of the TSS. The Gene motif activity A^{SM} is identified by ridge regression and summed for each motif M to obtain the gene motif activity per sample S. Finally, target genes are predicted* ***[4]****.*

The Elastic-Net Function

IMAGE addresses the issue of multicollinearity in steps 1 and 2 for predicting motif activities using the Elastic-Net Regression function, a regularised form of linear regression [34].

$$(\hat{\beta}_0, \hat{\beta}) = \min_{(\hat{\beta}_0, \hat{\beta}) \in \mathbb{R}^{p+1}} \left[\frac{1}{2N} \cdot \sum_{i=1}^{N} (y_i - \beta_0 - x_i^T \beta)^2 + \lambda P_\alpha(\beta) \right] \quad (1)$$

The elastic-net regression in (1), outputs both a vector of coefficients for each feature (p) that takes on real numeric values ($\hat{\beta} \in \mathbb{R}^p$), and the intercept of the regression ($(\hat{\beta}_0, \hat{\beta}) \in \mathbb{R}^{p+1}$). The elastic net attempts to find the best possible coefficients and intercept, which minimise the loss function inside the squared brackets in equation (1) [33], [34]. A loss function is used to assess the model's performance by calculating the deviation in the model's predictions from the ground truth (observed

data). The first half of the loss function is the mean squared error (MSE) expressed as, $\frac{1}{2N} \cdot \sum_{i=1}^{N}(y_i - \beta_0 - x_i^T\beta)^2$, and the second half is the penalty term, $\lambda P_\alpha(\beta)$ [33], [34], [35]. The MSE measures the average of the squared differences between predicted and observed values. N is the total number of samples, y_i is the observed outcome for sample i and $x_i^T\beta$ is the predicted value for feature x in sample i [33], [34], [35].

A penalty term is added to regularise the regression and avoid inflated variance in the coefficients, making it more generalisable. It does so by shrinking the coefficients. The penalty term decides how significant the penalty should be for having large coefficients, and what kind of penalty it should be [33], [34]. Lambda (λ) in the penalty term decides the strength of the penalty and can take on values equal to or greater than 0 [36]. Choosing an appropriate value for λ is typically done via k-fold cross-validation. A grid of λ values is generated, and for each fold in the cross-validation, the elastic net function is solved for each lambda. The λ resulting in the lowest cross-validation error is selected as the final λ for the model [34], [35], [36]. The penalty function in the penalty term, as seen in equation (2), determines the type of penalty depending on the hyperparameter alpha (α) [33], [34].

$$P_\alpha(\beta) = \sum_{j=1}^{p}\left[\frac{1}{2}(1-\alpha)\beta_j^2 + \alpha\left|\beta_j\right|\right] \quad (2)$$

There are two main types of penalties, lasso ($\alpha = 1$) and ridge ($\alpha = 0$). Lasso regression punishes a large coefficient j by taking the cardinality of the coefficients and summing them, while ridge regression squares the coefficients and sums them up. Because of this, lasso functions as a form of feature selection, since it handles all identical correlated features by assigning them the value of zero, except for one of the correlated features. Ridge regression handles multicollinearity by pushing all coefficients of identical correlated features towards zero, such that the correlated features each get identical coefficients with 1/kth the size, where k is the number of correlated features [33], [34]. This is illustrated in Figure 4A, where the lasso sets correlated features to zero and the ridge pushes them towards zero, but never actually sets any of them to zero, unless the log lambda becomes infinitely large [37]. As evident from Figure 4A, the alpha hyperparameter can also take on values between 0 and 1, allowing for a mix of the two types of penalties. This results in some correlated features being set to zero, while the rest of the coefficients are pushed towards zero

[33], [34], [37]. The reason ridge does not set correlated coefficients to zero, but lasso does, is due to the mathematical nature of the two types of penalties, as illustrated in Figure 4B.

The ideal and minimal coefficients of the loss function are represented by the black dot, $\hat{\beta}$. For two coefficients, β_1 and β_2, their best possible solution in a ridge and lasso regression can be found within a constrained region, s, for every value of λ [35]. The larger λ becomes, the smaller the region s. To find the best possible solution for the loss function, it is necessary to move away from the ideal minimum, which is represented by the red contours of the MSE. At the intersection between the constrained region and the contour, the coefficients and the minimal solution to the loss function are found [35]. Lasso regression penalty for the two coefficients is in the form of $|\beta_1| + |\beta_2| \leq s$, which geometrically equates to a diamond shape, while ridge regression penalty is in the form of $\beta_1^2 + \beta_2^2 \leq s$ and geometrically equates to a circle. Because Lasso has corners at each of the axes and ridge does not, it is possible for the contour to intersect at the corner and set one of the coefficients to equal zero in Lasso regression, but they can only approach zero in Ridge regression [35].

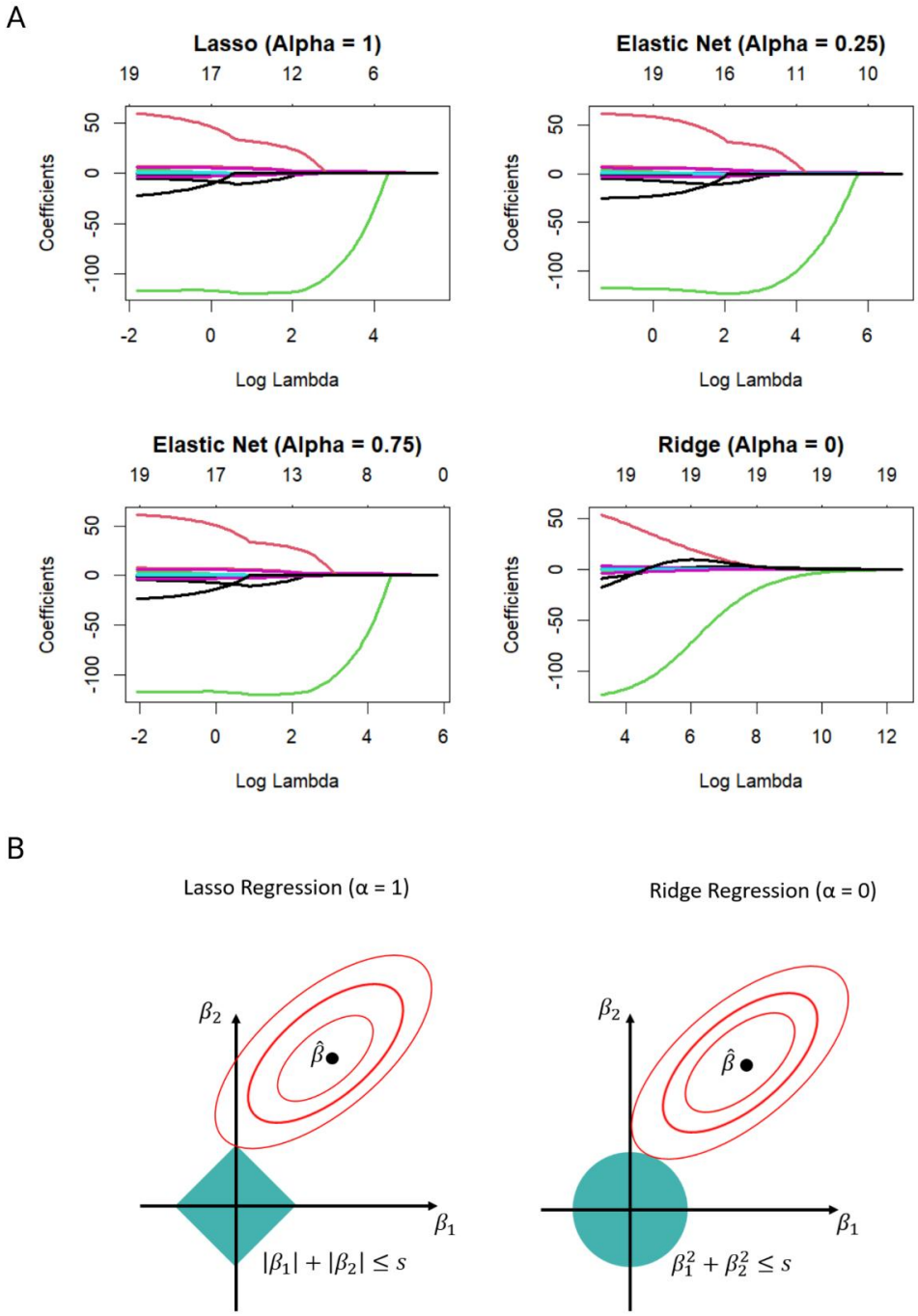


Figure 4. Visualisation of coefficient shrinkage in the elastic net function and the geometry of ridge and lasso. *(A) Visualisation of how Lasso regression (Alpha = 1) selects features by setting their coefficients to 0, resulting in a reduction in the number of variables, as illustrated by the numbers above each plot. The elastic net function enables the adjustment of the feature selection severity, as demonstrated with Alpha = 0.75 and Alpha = 0.25. Ridge regression (Alpha = 0) does not carry out feature selection; it merely shrinks the coefficients towards 0 [37]. (B) Ridge regression primarily reduces coefficients toward zero without actually reaching it, unlike lasso regression, due to the geometric characteristics of each method. The black dot ($\hat{\beta}$) signifies the optimal values for the two coefficients, β_1 and β_2. The green diamond and circle illustrate the geometrically restricted value regions (s) that the coefficients can assume under Lasso and Ridge regression penalties, respectively. The red contours represent the Mean Squared Error (MSE) and expand to encompass possible coefficient values within the given constraints. Since Lasso follows a diamond shape, its contours can touch zero precisely at the tip, a scenario not possible with the circular shape in ridge regression [35].*

Coordinate Descent

In practice, finding the optimal coefficient values that minimise the Elastic Net function would typically be done by differentiating the function with respect to all coefficients simultaneously. However, this is not possible. As seen in Figure 4B, the ridge function is a smooth and continuous function, while the Lasso is not, since it has no unique derivative at zero due to the kink [34]. Because of this, the optimal solution for equation (1) is solved through cyclical coordinate descent [34]. In cyclical coordinate descent, the coefficients are cycled through and updated one at a time, while holding the other coefficients fixed. All coefficients are initialised to 0. The current coefficient that is being updated is denoted as $\tilde{\beta}_j$, while the other fixed coefficients are denoted as $\tilde{\beta}_\ell$. The coefficients are updated one at a time using equation (3) [34].

$$\tilde{\beta}_j = \frac{S\left(\frac{1}{N}\sum_{i=1}^{N} x_{ij}\left(y_i - \tilde{y}_i^{(j)}\right),\ \lambda\alpha\right)}{\frac{1}{N}\sum_{i=1}^{N} x_{ij}^2 + \lambda(1+\alpha)} \tag{3}$$

The equation is essentially the slope of a linear regression, with the lasso and ridge penalty incorporated into it, through a soft-thresholding operator that takes care of the lasso contribution to the penalty and a shrinkage scaling factor for the ridge penalty [34], [35]. The $\tilde{y}_i^{(j)}$ is the fitted value excluding the contribution from x_j across all observations and is defined by

$$\tilde{y}_i^{(j)} = \tilde{\beta}_0 + \sum_{\ell \neq j} x_{i\ell}\tilde{\beta}_\ell \tag{4}$$

This makes $y_i - \tilde{y}_i^{(j)}$ the partial residual when fitting for $\tilde{\beta}_j$ [34]. The $S(z,\gamma)$ is the soft-thresholding operator, which decides the value of the nominator in equation (3) and is defined like so

$$S(z,\gamma) = \begin{cases} z-\gamma & if\ z > 0\ and\ \gamma < |z| \\ z+\gamma & if\ z < 0\ and\ \gamma < |z| \\ 0 & if\ |z| \leq \gamma \end{cases} \tag{5}$$

The $z = \frac{1}{N}\sum_{i=1}^{N} x_{ij}\left(y_i - \tilde{y}_i^{(j)}\right)$ and the $\gamma = \lambda\alpha$ from equation (3) [34]. Overall, when α becomes larger and resembles more lasso penalty ($\alpha = 1$) and if $\gamma = \lambda\alpha$ becomes bigger than the cardinality of the partial correlation term, $|z| = |\frac{1}{N}\sum_{i=1}^{N} x_{ij}\left(y_i - \tilde{y}_i^{(j)}\right)|$, the coefficients will be set to 0 according to equation (3) and (5). However, the smaller α becomes and resembles more ridge

penalty ($\alpha = 0$), the harder it becomes for γ to be bigger than $|z|$, and the nominator will not be set to 0, but the denominator in (3) can still scale the coefficient towards 0 the higher the value of λ becomes. This is how coordinate descent minimises the coefficients in the elastic net function piece by piece, until it reaches convergence and the coefficients are not minimised anymore [34].

The overall workflow and overview of predicting coefficients (motif activities) as described above are illustrated in the pseudocode in Table 1, inspired by the glmnet source code, Friedman et al. 2010 and Tay et al. 2023 [34], [36].

Table 1: Algorithm of coordinate descent pseudocode

Algorithm: Solving the elastic net function with coordinate descent
1) ***Input:*** *Data (X variable, Y variable),* $\alpha \in [0,1]$, $k - folds \in \mathbb{N}, k \geq 2$
2) ***Generate a grid of λ values and split the data into k-fold***
3) *For k = 1, …, m:*
(a) ***Initialisation:***
If k = 1 then $\left(\hat{\beta}_0^{(0)}(\lambda_k), \hat{\beta}^{(0)}(\lambda_k)\right) = (0,0)$,
else warm start $\left(\hat{\beta}_0^{(0)}(\lambda_k), \hat{\beta}^{(0)}(\lambda_k)\right) = \left(\hat{\beta}_0^{(0)}(\lambda_{k-1}), \hat{\beta}^{(0)}(\lambda_{k-1})\right)$
(b) ***For each λ (start at*** λ_{max} ***where*** $\left(\hat{\beta}_0^{(0)}(\lambda_k), \hat{\beta}^{(0)}(\lambda_k)\right) = (0,0)$ ***):***
- Coordinate descent (training data)
- Update coefficients one at a time
- Repeat until convergence
- Predict on held-out fold data
- Compute and store MSE error
4) ***Average errors across folds for each λ***
5) ***Extract the minimum lambda resulting in the lowest error, intercept and coefficients.***

Target Enhancers and Target Genes

After IMAGE has estimated enhancer motif activities in the `IMAGE::calculateEnhancerMotifActivity` function, it predicts target enhancers in the `IMAGE::findTargetEnhancers` function. The motif activities are used to calculate a predicted enhancer signal with the function in step 1 in Figure 3A. The predicted signal is used to calculate an

error between the observed enhancer signal j in sample i and the predicted signal of enhancer j in sample i, as seen in equation (6).

$$FullModelError_{ij} = \left(Observed\ signal_{ij} - predicted\ signal_{ij}\right)^2 \quad (6)$$

A similar error can be calculated, but for a reduced model. This is achieved through a leave-one-out-based analysis, where each motif is excluded from the error calculation (-m) to determine its contribution to the overall error, as shown in (7).

$$ReducedModelError_{ij(-m)} = \left(Observed\ signal_{ij} - predicted\ signal_{ij(-m)}\right)^2 \quad (7)$$

Based on the two error calculations, a motif contribution score can be calculated to see which motif contributes the most to the enhancer activity.

$$Contribution\ score_{jm} = \frac{\sum_{j=1}^{p} ReducedModelError_{ij(-m)} - FullModelError_{ij}}{Average\left(FullModelError_{ij}\right)} \quad (8)$$

By dividing by the average error of the full model, the motifs' contributions are normalised and become comparable across motifs, reflecting each motif's relative importance. A positive score would mean that by removing the motif and increasing the error, the motif is important for the model. Therefore, if a motif has a raw motif count above zero and a motif contribution score to the enhancer activity above zero, it is predicted to be a target enhancer [4].

For the downstream analysis of IMAGE in step 2, where the gene motif activity is calculated, the target enhancers need to be weighted according to their proximity to each gene to assess their regulatory potential, which is done in the function `IMAGE::MotifWeightMatrix` [4]. In IMAGE, this is achieved through a regulatory potential function, as shown in (9).

$$Potential(d) = \frac{e^{-\left(0.5+4\cdot\left(\frac{d}{100000}\right)\right)} - e^{-(0.5+4)}}{\max\left(e^{-\left(0.5+4\cdot\left(\frac{d}{100000}\right)\right)} - e^{-(0.5+4)}\right)} \quad (9)$$

The function illustrates a nonlinear exponential decay between the enhancers and the TSS within a 100-kilobase (kb) regulatory region. This means that within the 100 kb, the closer an enhancer is to a gene, the higher it is weighted or prioritised. The d is the absolute distance between the enhancer peak centre and the TSS. The distance is normalised or scaled between 0 and 1 by the maximum possible distance that can be obtained. The function subtracts $e^{-(0.5+4)}$ to remove the effects of

these, since the constant 4 was added to adopt more flexible shapes, while 0.5 was included to better fit ChIA-PET and Hi-C data [38].

After the motifs have been weighted with (9), the gene motif activity is estimated in step 2 similarly with the elastic net function solved through coordinate descent in the function `IMAGE::calculateMotifActivity`. Afterwards, pairwise T-test comparisons are performed on the experimental conditions to determine significant differences between motif activities. The minimum p-value is extracted from this statistical test, and used to calculate a composite weighted p-like-value as seen in equation (10). It consists of the maximum gene expression for each TF and their corresponding motif m (MaxE), maximum motif activity across conditions (MaxA), the minimal p-value from the t-tests between conditions (MinP) and the minimal false discovery rate (MinFDR) from gene expression for the given TF and its motif m.

$$WeightedPValue_m = \frac{rank(-MaxE_m)}{n} \cdot \frac{rank(-MaxA_m)}{n} \cdot \frac{rank(MinP_m)}{n} \cdot \frac{rank(MinFDR_m)}{n} \quad (10)$$

To reverse their behaviour, the MaxE and MaxA have a negative sign, such that lower scores indicate greater significance, just like a p-value. The number of motifs n normalises the ranks. TFs and their corresponding motifs are filtered, retaining only significant TFs for downstream analysis. TFs with a weighted p-value ≤ 0.001 and a minimal p-value from the t-test ≤ 0.005 are kept [4].

After this, target genes can be predicted similarly to those in (6), (7), and (8) using a leave-one-out analysis for each motif, but for gene expression instead of enhancer activity, in the function `IMAGE::findTargetGenes`. The new motif-gene contribution scores similar to (8) for the gene expression are used to calculate a P-value-like score (11) [4].

$$Score_{gm} = \frac{rank(-w_{gm})}{N} \cdot \frac{rank(-c_{gm})}{N} \quad (11)$$

The c_{gm} and w_{gm} are respectively the gene-motif contribution score and the motifs weighted based on their distance to the TSS of the gene. They are normalised with the number of genes N, such that they rank from 1/N to 1. This score is used to predict target genes. If a gene scores below 0.005 and is expressed above a raw count of 10 in CPM, it is defined as a target gene [4]. Finally, to flag the significant TFs in terms of their predicted causal significance, they are flagged as either being of rank 0, 1 or 2, where 2 is the highest and zero is the lowest. The entire workflow overview for IMAGE is presented in Table 2.

Table 2: Overview of IMAGE workflow

1) *IMAGE::countMotifs*

- *Checks compatibility between the genome hg38 and ATAC-seq sequences*
- *Identifies and counts motif matches between the ATAC-seq sequences and the PWM*

2) *IMAGE::processMotifs*

- *Normalises motifs*

3) *IMAGE::NormalizeEnhancers*

- *Correct raw peak counts for GC bias and library size (FQ-FQ normalisation)*
- *Differential accessibility test*
- *Keep significant peaks.*
- *CPM normalisation*
- *Standardisation*

4) *IMAGE::calculateEnhancerMotifActivity*

- *Estimate enhancer motif activity with ridge regression (Table 1)*

5) *IMAGE::findTargetEnhancers*

- *Calculate the predicted enhancer signal*
- *Calculate full model error (equation 6)*
- *Calculate reduced model error (equation 7)*
- *Calculate enhancer motif contribution score (equation 8)*
- *Target enhancer if the enhancer motif contribution score and raw motif count >0*

6) *IMAGE::analyzeRNA*

- *Correct gene expression for GC bias and library size (FQ-FQ normalisation)*
- *CPM normalises the gene expression data*
- *Filters out lowly expressed genes (below raw count of 10 in CPM)*
- *Differential gene expression analysis*

7) *IMAGE::MotifWeightMatrix*

- *Calculate regulatory potential (equation 9)*
- *Weight TF motifs in target enhancers based on regulatory potential*

8) *IMAGE::calculateMotifActivity*

- *Estimate gene motif activity with ridge regression (Table 1)*
- *Pairwise T-test for significant motif activity differences*
- *Calculate weighted p-value (equation 10)*
- *Filter TF motifs based on* weighted p-value ≤ 0.001 and t-test minimal p-value ≤ 0.005

9) *IMAGE::findTargetGenes*

- *Calculate the predicted gene expression*
- *Calculate full model error, reduced model error and gene motif contribution score (equations 6, 7, 8)*
- *Calculate P-value-like score (equation 11)*
- *Target gene if P-value like score < 0.005 and expressed > raw count of 10*

Multiple Roads to Understanding Regulation

ChromVAR and the Inference of Motif Activity From Sparse Accessibility Data

Many tools have been created to understand the activity of TFs and their role in GRNs [26], [29]. IMAGE is one such tool that combines the prediction of TF activity with the inference of target genes [4]. However, some tools are more modular and focus on predicting the activity of TFs and linking CREs like enhancers to their target genes, such as ChromVAR and scE2G [5], [7].

ChromVAR predicts motif activity from single-cell sparse chromatin-accessibility assays, such as ATAC-seq data, by computing a raw accessibility deviation score [5]. It does so by first computing the expected number of fragments per cell as seen in equation (12).

$$E = \frac{\sum_{i=1} x_{ij}}{\sum_{j=1}\sum_{i=1} x_{ij}} \cdot \sum_{j=1} x_{ij} \tag{12}$$

The x_{ij} represents the number of fragments from cell i in peak j. The numerator sums all counts for peak j across cells, indicating the overall accessibility of peak j. Dividing by the total number of fragments in the dataset in the denominator explains what fraction of the counts belong to peak j. Finally, by multiplying with the summed peaks j for a given cell i, the expression indicates the expected number of fragment counts of peak j in cell i, if cell i distributed its total counts in the same proportion as the dataset [5]. With E, a raw accessibility deviation score can be calculated as shown in Equation (13).

$$Y = \frac{M \cdot X^T - M \cdot E^T}{M \cdot E^T} \tag{13}$$

The M is the matrix of motif matches, where m_{kj} is 1 if motif k is present in peak j, and if it is not, then it is 0. The X is the matrix of observed peaks j in cell i. By subtracting the observed peaks from the expected peaks, the raw motif accessibility score is calculated and normalised, so that no peak dominates over others [5].

However, the authors of chromVAR call attention to the fact that the implicit aggregation across peaks via the matrix multiplication can amplify technical biases between cells due to PCR amplification, such as GC content [5]. The bias arises from a stronger base pairing between guanine (G) and cytosine (C), as they form three hydrogen bonds versus two hydrogen bonds between adenine (A) and thymine (T). Because of this, DNA sequences with a high GC content can be more

challenging to denature during PCR amplification, resulting in an underrepresentation of GC-rich sequences [5], [39]. Additionally, promoters tend to have a higher average accessibility than enhancers, and they also contain more GC content on average [5]. This means that differences in accessibility can also be seen as somewhat associated or correlated with GC content [5]. Ultimately, to ensure that the accessibility is mainly associated with motif sequences and to correct for GC-rich peaks that might appear to be artificially less accessible, they correct for the GC content by calculating background peak sets [5]. They separate the GC content and chromatin accessibility by transforming the data, and in the uncorrelated transformed space, they split it into even grids of 50 along both axes [5]. Each of the bins contains peaks with similar GC content. Iteratively, they go through each peak and pick the background peaks for the given peaks with a probability function.

$$P(x \mid x \in b_i) = \frac{f(d(i-j)|0,w)}{\tilde{n}_j} \tag{14}$$

The nominator is the probability distribution function f of a normal distribution with a mean of zero and the standard deviation w set to 0.01. The $d(i-j)$ is the distance between bin j and bin i. The denominator is the number of peaks in bin j. Overall, the function finds the probability of a given peak in bin j to have another peak x in bin i as its background peak [5]. Each background peak x for the given peak is stored in a matrix B and multiplied by the matrix of motif matches M to get a background motif match matrix. This is used to calculate a background raw deviation score [5].

$$Y' = \frac{(M \cdot B) \cdot X^T - (M \cdot B) \cdot E^T}{M \cdot E^T} \tag{15}$$

Since this is an iterative process for each peak, the mean of the background raw deviation scores is calculated and subtracted from the raw deviation scores to obtain a bias-corrected raw deviation score. Finally, this is divided by the standard deviation of Y', to achieve a motif activity z-score of the TFs [5].

$$Motif\ Activity = \frac{Y - mean(Y')}{s.d(Y')} \tag{16}$$

Single- Cell Enhancer-to-Gene (scE2G): A Novel Tool for Linking Enhancers to Genes

In contrast to ChromVAR, the novel tool scE2G does not focus on predicting motif activity but rather the interactions between enhancers and genes [7]. ScE2G predicts enhancer-gene links on

scRNA- and scATAC-seq data. Broadly speaking, a list of candidate element-gene pairs is defined as any accessible element and promoter within 5 Mb of each other, and the genes should be active [7]. It utilises logistic regression to predict enhancer and gene interactions, as shown below.

$$P(X) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n)}} \tag{17}$$

There are six input features in total ($\beta_1 x_1 : \beta_6 x_6$) that scE2G uses for its predictions: Ubiquitous expression of promoter, the number of candidate elements between enhancer and gene, number of transcription start sites (TSS) between enhancer and gene, number of candidate elements within 5 kb of the enhancer, normalized ATAC signal at promoter and finally the ARC-E2G score for multiome data or the ABC score if only chromatin accessibility assays such as ATAC-seq data is provided [7], [40]. The ubiquitous expression of the promoter is a binary classification of whether the promoter is ubiquitously active across cell types. This would indicate whether a gene is more of a housekeeper gene or cell-specific. In general, these features are measures of genomic and regulatory complexity between the element-gene pairs and reflect how likely an enhancer is to regulate a given gene compared to other regulatory elements [7].

The ARC-E2G (Activity, Responsiveness, and Contact - Enhancer 2 Gene) score is a composite metric that combines the strength of the ABC (Activity by Contact) score and the Kendall correlation [7], [40]. The ABS score measures chromatin accessibility and 3D contact between elements and the gene. It was calculated on a per-gene basis [7].

$$ABC_{score} = \frac{A_E \cdot C_{EG}}{\sum_{All\ elements\ E\ within\ 5\ Mb\ of\ G} A_E \cdot C_{EG}} \tag{18}$$

The A_E is the activity for element E, measured as the normalized ATAC signal and the $C_{E,G}$ is the contact between element E and gene G [40]. The 3D contact is estimated using a power law decay function, which, in its simplest form, is defined in (19).

$$C_{EG} \approx \frac{1}{distance} \tag{19}$$

It describes a function that resembles an exponential increase, the closer the regulatory element is to the gene, but plateaus at a certain point [7].

The Kendal correlation is a measure of rank correlation and captures the relationship between two pairs of observations, such as enhancer-gene pairs [7]. If the pair is concordant, the rank for both elements agree, and they are discordant if the ranks disagree. In other words, suppose there are two variables, X (enhancer) and Y (gene), and there are two pairs of observations, (X_1, Y_1) and (X_2, Y_2). The pair is concordant if $X_2 > X_1$ and $Y_2 > Y_1$. They are discordant if e.g. $X_2 > X_1$ and $Y_2 < Y_1$. Noticeably, due to the binarization of enhancers, there can be a lot of ties ($X_2 = X_1$), and the Kendal correlation that accounts for this is given by

$$T_b = \frac{n_c - n_d}{\sqrt{(n_0 - n_1) \cdot (n_0 - n_2)}} \tag{20}$$

The n_c is the number of concordant pairs, and n_d is the number of discordant pairs. The $n_0 = \frac{n(n-1)}{2}$, where n is the number of cells, while n_1 is the number of ties within the first variable and n_2 is the number of ties within the second variable [7].

Sheth et al. 2024 noticed that the Kendall correlation and its weight were sensitive to sequencing depth, whereas the ABC score was not. They integrated the two metrics into the ARC-E2G score by considering their linear relationship to improve robustness across sequencing depths [7]. The logistic regression model was trained on a gold-standard CRISPR perturbation dataset in K562 cells, utilising chromosome-wise cross-validation to determine the weights for each feature and to validate the enhancer-gene links predicted by scE2G [7]. It was also trained on all chromosomes to be applied to new cell types. For predicting enhancer-gene links in new cells, a list of all enhancer-gene links is created, where the pre-trained logistic regression model on all chromosomes is applied. The raw scores from the logistic regression are quantile-normalised to match the K562 model scores. This is done to adjust for sequencing depth differences [7]. Finally, lowly expressed genes are filtered out by setting their probability to 0. If binarisation of the enhancer predictions is desired, a threshold of 0.164 of the scE2G score is applied, corresponding to a 70% recall score [7].

Modulating Cell Fate: Transcription Factor-Driven Differentiation of hESCs

It is of interest to validate IMAGE and compare its performance against chromVAR and scE2G, regarding predicting enhancer motif activities and linking enhancers to genes, respectively. For this purpose, the dataset from Joung et al. 2023 was used [6]. In short, they produced a multiplexed

overexpression of regulatory factors (MORF) library of plasmids that contained barcodes, 196 unique TF open reading frames (ORFs) and two ORFs encoding control proteins GFP and mCherry [6]. The plasmids were packed into lentiviruses and transduced into H1 human embryonic stem cells (hESCs) at low multiplicity of infection (MOI), which is the ratio of the number of transducing lentiviral particles to the number of cells [6]. This is done to infect each cell with only one TF coding plasmid. After infection, the hESCs are differentiated in STEMdiff APEL media for 4 to 7 days, followed by harvesting and performing simultaneous high-throughput ATAC and RNA expression with sequencing (SHARE-seq) to acquire RNA- and ATAC-seq data as seen in Figure 5A [6], [41]. This dataset is used to validate IMAGE under the assumption that by overexpressing a single TF in a single cell, the TF should become active. Therefore, its motif activity should likewise increase, and potentially direct stem cell differentiation towards a particular cell type, state, and identity, as seen in Figure 5B [42].

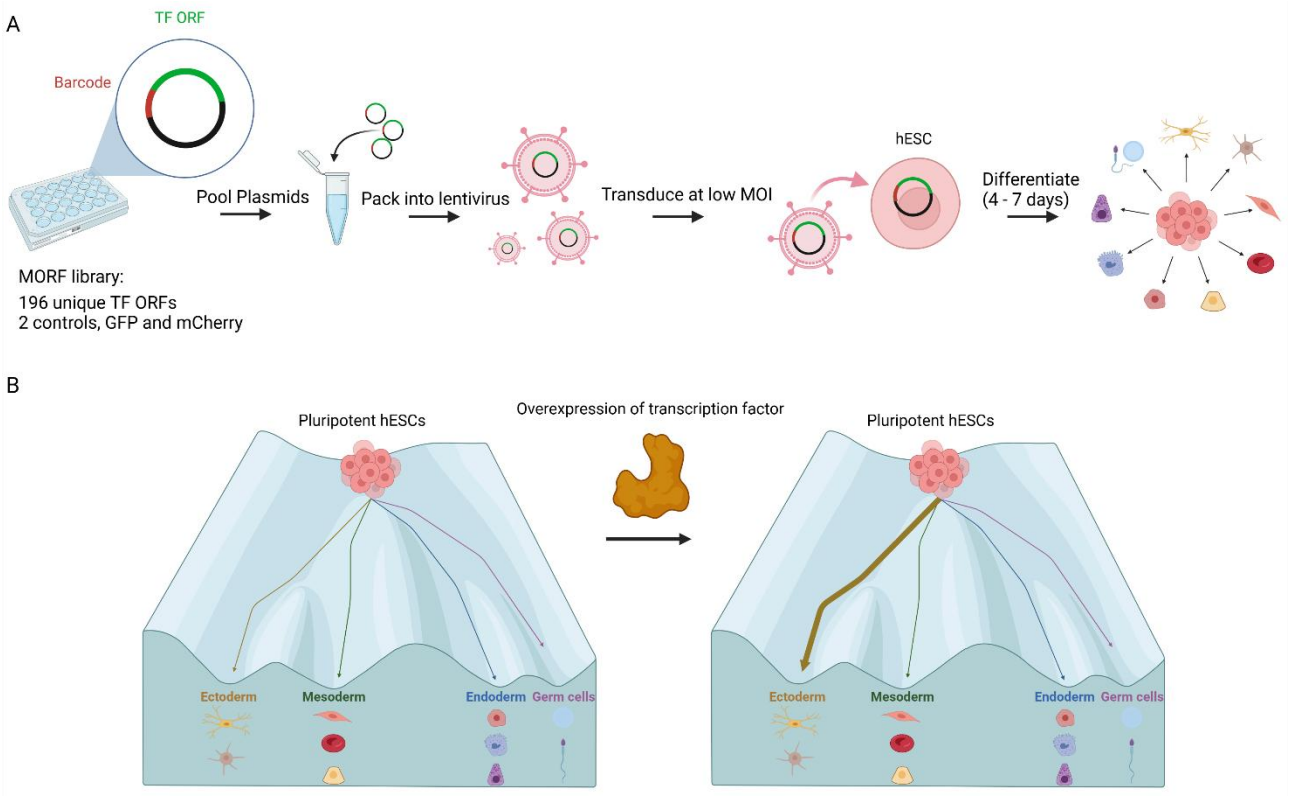


Figure 5. Cartoon illustration of the experimental setup for the overexpression of transcription factors (TFs) in single-cell SHARE-seq data and the differentiation trajectory influence of TF overexpression. *(A) A multiplexed overexpression of regulatory factors (MORF) library is constructed that contains plasmids encoding 196 unique TF open reading frames (ORFs) along with a distinctive barcode to identify which cell has overexpressed which TF. Additionally, two control proteins, namely GFP and mCherry, are included. After the plasmids have been pooled and packaged into lentiviruses, human embryonic stem cells (hESCs) are transduced with the plasmid and permitted to differentiate for 4 and 7 days. In total, 69.085 cells were harvested [42]. (B) A cartoon illustration of Waddington's epigenetic landscape, where hESCs may naturally differentiate into certain cell types. An overexpression of a TF can lead to a push towards a certain cell type fate [6]. Made in Biorender.com.*

Methods

Timepoint Analysis

Joung et al. 2023 used Seurat's workflow to perform joint chromatin accessibility and gene expression (scATAC- and scRNA-seq) multimodal analysis to show both modalities in the same space [6], [43], [44]. This resulted in SHARE-seq data that could be loaded as a Seurat object and visualised in a weighted nearest neighbour UMAP (WNN UMAP) plot, to gain an overview of the cell clusters in the data for time points day 4 and 7 [6], [43], [44]. After visualisation of the cell clusters, the data were split into their respective time points for day 4 and day 7. This was done to assign a custom label indicating their time point. The datasets are merged, and the number of cells for each respective time point in each cluster is visualised in a barplot. The Pearson correlation was calculated for the differentially expressed genes between each respective cell cluster for days 4 and 7 to determine if the data from the two timepoints could be merged for downstream analysis. To calculate differentially expressed genes, non-control cells were aggregated using `Seurat::AggregateExpression` based on their Seurat clusters and batch, and differentially expressed genes were calculated using `Seurat::FindAllMarkers`, with the default non-parametric Wilcoxon rank sum test [44]. The default is used because Butler et al. 2018 found it to be fast and generally yields good results, as indicated in a benchmarking study [44], [45]. The non-parametric Wilcoxon rank sum test identifies differentially expressed genes between two groups of cells.

$$U = T_U - \frac{n_U(n_U + 1)}{2} \tag{21}$$

It tests if the two groups come from the same population based on a feature. This could be cell type A vs cell type B. For more than two cell types, it is a one vs the rest. The null hypothesis stipulates that both cell types come from the same population, while the alternative hypothesis states that they are not from the same population. To test the null hypothesis regarding gene x, the observations (cells) of the two cell types are combined into a single vector of total sample size of N ($n_a + n_b = N$). Each observation is ranked from 1 to N in ascending order, based on their expression of gene x. When each observation is ranked based on their expression of gene x, they are split out into their original groups. Then the sum of the ranks T_U is calculated for the cell group of interest, and the U statistic is calculated. The U statistic is used to determine the p-value, and if it is smaller than a chosen critical value, typically 0.05, then the null hypothesis can be rejected. This process is done for each gene for each cell type, and p-values are adjusted for multiple testing [44], [46], [47].

As mentioned earlier, the Pearson correlation coefficient was calculated for the differentially expressed genes in the respective cell types between days 4 and 7.

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x}) \cdot (y_i - \bar{y})}{\sqrt{(\sum_{i=1}^{n}(x_i - \bar{x})^2) \cdot (\sum_{i=1}^{n}(y_i - \bar{y})^2)}} \tag{22}$$

The Pearson correlation coefficient in (22) is a dimensionless index that describes the strength of the linear association between the independent variable x and the dependent variable y. The nominator of the Pearson correlation coefficient in (22) is the residual of x multiplied by the residual of y for each pair n. If x and y deviate in the same direction with respect to their means, then the association is positive; if they deviate in the opposite direction, the association is negative. The denominator adjusts the scales of the variables to have equal units [48], [49]. From the Pearson correlation coefficient, the coefficient of determination can be calculated as shown in (23). It measures how well the variation in x explains the variation in y. For a simple linear regression model with one predictor, it is expressed as [48]:

$$R^2 = r^2 \tag{23}$$

This metric was used in the scatterplot example to illustrate the correlation between cell cluster 1 from day 4 to day 7.

Comparison of GFP and mCherry Control Cells

The gene expression of control and overexpressed cells was aggregated, log-normalised and correlated to see if the data for the GFP and mCherry control cells could be merged. To investigate the relative enrichment of control cells in each cluster compared to their overall representation across all clusters, the following equation was used:

$$control\_enrichment_i = \log_2\left(\frac{\frac{C_i}{O_i}}{\frac{C_{total}}{O_{total}}} + 1\right) \tag{24}$$

C_i is the total number of control cells in cluster i, and O_i is the number of cells that have been overexpressed with a TF in cluster i. In a similar fashion, C_{total} and O_{total} is the total number of control and overexpressed cells. Log_2 is used to transform the relative enrichment into a symmetric and interpretable scale and 1 is added to avoid the log of 0. The scale is symmetric and interpretable, since a value of 1 means 2 times enrichment and a value of -1 is a 2 times depletion.

Similarity Between Overexpressed and Control Cells

To assess how similar the cells that have been overexpressed (OE) with TFs are to the control cells, the OE and control cells were separately aggregated based on their Seurat clusters and batches. Their differentially expressed genes (DEGs) were identified with the Wilcoxon rank sum test, and the Pearson correlation coefficient was calculated for each cluster between the control and OE cells based on their DEGs. The number of shared DEGs between the clusters was also identified and plotted.

To further investigate which of the OE cells resemble stem cells the most, module scores were calculated with `Seurat::AddModuleScore` and `UCell::AddModuleScore_UCell` for selected hESC marker genes [44], [50], [51]. Module scores are essentially a measure of the relative expression of a gene set in each cell. Seurat's module score is calculated by binning all genes based on their average expression levels across the whole dataset. Then, for each gene x in the provided gene set, several control genes (default 100) are randomly selected from the same expression bin as gene x. Then, for each cluster on a cell level, the average expression level of the genes in the input gene set is calculated. This average expression is subtracted with the average expression of the control genes [50]. However, as the authors of UCell point out, by binning the expression of the genes across all cells, the results are inconsistent, as the results for the same cell can change depending on the composition of the other cell type clusters in the dataset [51]. The UCells module score avoids this by calculating a U-like statistic like in (21).

$$U_j = \sum_{i=1}^{n} r'_{ij} - \frac{n(n+1)}{2} \tag{25}$$

However, unlike in (21), this statistic is not used for significance testing between two populations. For each gene r in cell j, the gene expression is ranked. Then, for the provided gene set n, their ranks are summed together and subtracted by the best possible sum of ranks the gene set could have. Thus, the U-like statistic measures how much worse the actual ranks of the gene set are compared to the best possible case. This U-like statistic is normalised between 0 and 1 [51].

Cell Type Identification

Non-control cells were aggregated based on their Seurat clusters and their batch. Differentially expressed genes were identified as aforementioned. The top 20 differentially expressed genes for

each cell cluster were identified and plotted to visualise the gene expression. Additionally, `enrichR::DEenrichRPlot` was used, which integrates differential expression testing with gene set enrichment analysis. It performs differential gene expression with the Wilcoxon rank sum test. Overall, it identifies the same top 20 genes (specified in max.genes) and submits them to the Enrichr server to perform pathway enrichment based on a selected gene set library [52]. This analysis utilised two databases: the GO Biological Process 2023 and the Human Gene Atlas [52]. The GO Biological Process 2023 is a database with curated annotations of genes in regards to biological pathways or functions, while the Human Gene Atlas is a database that provides curated annotation of the gene expression profiles in regards to tissue and cell type [52], [53], [54]. This was done for all of the cell clusters. In two cases, the web server Panglaodb was used in addition to Enrichr for identifying cell types [55]. Panglaodb is also a comprehensive database of curated gene markers associated with specific cell types [55].

Transcription Factor Enrichment and Expression Likelihood

To analyse how enriched overexpressed TFs were in each cell type, the following equation was used.

$$Enrichment_{tf,cluster} = log2\left(\frac{\frac{N_{tf,cluster}}{N_{crtl,cluster}}}{\frac{N_{tf,total}}{N_{tf,total}}} + 1\right) \quad (26)$$

It is a TF fold-enrichment score, which measures how enriched a TF is in a specific cluster relative to control cells and the overall abundance of the TF. $N_{tf,cluster}$ is the number of cells induced with a given TF in the current cluster, relative to the number of cells in the same cluster induced with control plasmids (GFP and mCherry), $N_{crtl,cluster}$. This cluster enrichment is then ratioed with the number of cells induced with the given TF in all clusters, $N_{tf,total}$, relative to the number of cells induced with control proteins in all clusters, $N_{tf,total}$.

In contrast, Joung et al. 2023 used the percentage of cells with the indicated TF ORF for each cluster as a measure of TF enrichment.

$$TF\ transduction\ percentage_{tf,cluster} = \frac{N_{tf,cluster}}{N_{cluster}} \cdot 100 \quad (27)$$

Here, the number of cells transduced with a given TF in the current cluster is $N_{tf,cluster}$ and the denominator is the total number of cells in the current cluster of interest [6]. To see if the TF

enrichment from (26) was comparatively similar to (27) and would be appropriate to use instead for downstream analysis, the Pearson correlation coefficient of the enrichment for each TF was calculated between (26) and (27).

The likelihood of an overexpressed TF to induce its own expression was also assessed. Cells were split into a control and a non-control group. For each TF in each group, the number of cells that expressed the given TF above zero were counted and ratioed with the total number of cells as seen in (28).

$$Likelihood_{group} = log2\left(\frac{N_{expr}^{group}}{N_{total}^{group}} + 1\right) \quad (28)$$

Exploration of GC Content in Data

The fraction of GC content across peaks in non-control cells was investigated by calculating the fraction of G or C bases in each sequence of the hg38 genome with `Biostrings::letterFrequency` [56]. This analysis was followed by examining the effect of the unadjusted GC content on differential feature analysis [57]. An 8 vs. 8 mock null differentially accessible test was performed on non-control neuron-like cells [57]. The idea of an 8 vs. 8 mock null differentially accessible test is to split a single cell type into two artificial groups (control and non-control) with 8 sample replicates for each group. There should be no biological difference in the differential accessibility between the two groups as they are of the same cell type. Because of this, the magnitude of change (log2fc) should for all peaks be centred around zero, and any deviation from this when held up against the GC content indicates that GC content biases the analysis, which can lead to false biological interpretations [57].

This was done by aggregating the single-cell data of non-control neuron-like cells based on Seurat clusters and batches. This resulted in 16 replicates, which were split into artificial test groups with 8 sample replicates per group. This data is processed with the standard edgeR workflow of normalising counts, estimating dispersion and fitting a negative binomial distribution model on the data, followed by a differentially accessible test [58], [59]. The log2FC values were extracted and compared to the GC content.

To adjust for GC content and sequencing depth, two rounds of full-quantile (FQ-FQ) normalisation are applied, respectively, for within each sample and between samples [57], [60]. For within-sample FQ normalisation, genes or peaks are stratified into equally-sized bins based on GC-content. Then,

for each sample and each bin, the feature counts are ranked in ascending order, and the median of ranks across bins is calculated. The original values are substituted with these median values [57], [60], [61]. For between-sample FQ normalisation, sample values are ranked, and the median of ranks across samples is calculated and replaces the original values [57], [60], [61]. After this, an 8 vs. 8 mock null differentially accessible test is performed again to see the effect of the FQ-FQ normalisation.

Enhancer Motif Activity Validation

ChromVAR was run on both scATAC-seq data and the same pseudobulked data for all peaks for which IMAGE was run on. IMAGE was run on different pseudobulk combinations of scATAC-seq data to explore feature selection. ScATAC-seq data was pseudobulked to 4 replicates per cell type based on batch and Seurat cluster labels [4], [5], [44]. The three feature selections were based on all peaks, peaks detected in at least 1% of cells in a given cluster and differentially accessible peaks with the aforementioned standard edgeR workflow [58], [59]. For each feature selection, the following functions of IMAGE were run: `IMAGE::countMotifs`, `IMAGE::processMotifs`, `IMAGE::normalizeEnhancers` and `IMAGE::calculateEnhancerMotifActivity`. `IMAGE::countMotifs` checks for compatibility between ATAC-seq data and the hg38 genome, and finds matches between the ATAC-seq sequences and the curated motif PWM, 'human_pwms_v2'[5]. This PWM database comprises 870 non-redundant TFs and was selected to minimise false positives by excluding motifs that are too generic or nonspecific to any single TF, thereby facilitating easier interpretation of the results [5]. The `IMAGE::processMotifs` and `IMAGE::normalizeEnhancers` normalise the motif count matrix and ATAC-seq sequences, respectively. For all three ATAC-seq feature selection methods, GC content is adjusted for. Finally, `IMAGE::calculateEnhancerMotifActivity` is run to estimate enhancer motif activities by solving for the elastic net function. The motif activities for IMAGE and chromVAR were z-score transformed to get a relative measure of activity for each motif j for comparing the activity across all cell types i.

$$Z_{ij} = \frac{X_{ij} - \mu_i}{\sigma_i} \tag{29}$$

The X_{ij} is the motif activity j in cell type i, while μ_j and σ_j are the mean activities and standard deviation of motif j across all cells.

To validate the motif activities, different log2 thresholds were chosen to filter TFs based on their log2 ratio enrichment, as described in (26), to identify enriched and control TFs. The fold thresholds for the enriched TFs were restricted to natural numbers from 2 to 9, while the control TFs were set to 1. Additionally, the control TFs were filtered such that only TFs with a normalised expression below 0,2 were kept to establish as proper a control group as possible. For both IMAGE and chromVAR, and selected cell clusters, the median TF activities were calculated for enriched and control TFs. Additionally, the selected cell clusters' median of median TF activities were calculated for the enriched and control TFs for both IMAGE and chromVAR to compare their predictions.

It was also experimented with not running GC content correction and only running within-sample FQ normalisation for enhancer motif activities in peaks detected in at least 1% of cells in a given cluster, as well as differentially accessible peaks. Validation was performed as described above.

Preparation of RNA Data

The pseudobulk RNA-seq data for IMAGE was annotated (chr/start/end/strand) using BioMart for only canonical transcripts with the highest transcript count per gene [62]. Only genes from the standard chromosomes (1–22, X, Y) were kept. Additionally, genes were filtered based on their ranges for downstream compatibility. The gene expression data was analysed with `IMAGE::analyzeRNA`, which adjusts for GC content, normalises the gene expression data, and filters out lowly expressed genes. The filtering was based on a threshold corresponding to a logCPM of 10 counts, so genes in at least two samples had logCPM values that met or exceeded this threshold [63]. Additionally, the maximum gene expression should likewise be above the threshold. Finally, a differential gene expression analysis is performed in accordance with an edgeR pipeline [4], [58], [59]. After processing the gene expression data with IMAGE, target enhancers of the TFs were predicted using `IMAGE::findTargetEnhancers`, which utilised equations (6), (7), and (8) as mentioned earlier. This analysis was followed by the rest of the IMAGE pipeline, consisting of `IMAGE::MotifWeightMatrix`, `IMAGE::calculateMotifActivity` and `IMAGE::findTargetGenes`. The gene motif activity was validated similarly to the enhancer motif activity.

Characterisation of TFs

The top and bottom 5% of raw enhancer and gene motif activities were investigated to analyse extreme regulators. TFs were classified as either a positive effector if their minimum motif activity

was above zero across cell types, or a negative effector if their maximum motif activity was below zero across cell types. If none of these applied, it was considered a dual-functional effector TF. The sign consistency between enhancer and gene motif activities was assessed to analyse shifts in TF activities across the two contexts. Finally, the gene motif activities of TFs were correlated with their own gene expression and the expression of predicted target genes to assess transcriptional concordance between TF activity and gene expression.

Processing of the IMAGE Enhancer Gene Links

A per-gene dataset of transcriptional regulatory relationships predicted by IMAGE was generated. This was achieved by first subsetting the enhancers and genes based on their predicted causal TFs. All enhancers within a 100kb region of genes were identified, and their regulatory potential was calculated according to (9). Looping through each causal TF and its target genes and enhancers, each gene-enhancer-TF combination was annotated with TF identity, motif contribution to gene regulation and the motif contribution to enhancer regulation. The results were compiled into a gene-indexed list.

Evaluation of Enhancer-Gene Links

To run scE2G, scRNA- and scATAC-seq data were prepared according to the instructions of Sheth et al. 2024 and analysed with scE2G [7]. IMAGE and scE2G enhancer gene links were binarised. For IMAGE, if the enhancer and gene were predicted to be a target, it would be set to 1; otherwise, 0 [4]. For scE2G, binarisation was done based on a score threshold determined to be 0,164 [7]. To make the comparison between IMAGE and scE2G fair, only enhancers within 100kb of TSS in scE2G predictions were used. Additionally, because scE2G predicts enhancer-gene links outside of canonical peak annotations in the ATAC-seq data, a peak overlap of 40% or above of scE2G enhancer peaks covered by IMAGE was found between the predictions and used for downstream analysis [4], [7].

To visualise the enhancer-gene links, a Seurat chromatin assay was constructed, and gene annotation for hg38 was attached to it, to enable genomic visualisation. The locus plots were visualised with Seurat::CoveragePlot, and the gene TSS was used as the anchor for the links [44]. The locus from the TSS was extended 100000 bp upstream and downstream to capture all links, and the quantile-normalised scE2G score was used as the confidence score in the locus plot [7]. For IMAGE, the regulatory potential was used as the confidence score [4].

A confusion matrix was constructed by looping through each gene and finding overlaps between IMAGE and scE2G peak ranges. Through these overlaps, the binarisation from sce2g was extracted and mapped back to IMAGE, and the results were aggregated across genes.

Finally, a precision recall plot was constructed by extracting the binarisation from scE2G and mapping it to IMAGE. This was input into Precrec to generate the plot [64]. Precision is the ratio of true positives (TP) over the number of true positives with the number of false positives (FP). Overall, it is the proportion of the number of predicted positives that are actually true positives (TP) and is defined as follows [65].

$$Precision = \frac{TP}{TP + FP} \tag{30}$$

Conversely, Recall is the ratio of true positives over false negatives (FN) added together with true positives. Essentially, it is the proportion of real positive cases that are correctly predicted as positive [65].

$$Recall = \frac{TP}{TP + FN} \tag{31}$$

The area under the precision-recall curve (AUPRC) can be extracted as a summarisation measure of general performance, rather than picking a specific decision point for deciding what counts as a positive [66].

ZCA Whitening

To experiment with decorrelation of the enhancer motif activities from ridge regression, zero-phase component analysis (ZCA) whitening was performed on the normalised motif counts. Whitening is a procedure for decorrelating a set of variables, and whitening seeks to find a matrix W that completely decorrelates a data matrix X and is defined as shown in (32) [67].

$$Z = XW^T \tag{32}$$

As shown in equation (32), ZCA whitening is a linear transformation. Since its purpose is to decorrelate the variables, the covariance of the whitened data, Z, should be the identity matrix. Covariance measures the relationship between variables in a dataset, essentially describing how they move together. If the covariance matrix only has 1 on the diagonal and 0 elsewhere, it is the identity matrix, and the variables are completely uncorrelated. The matrix W, which completely

decorrelates the variables, is known to be the inverse square root of the covariance matrix of X, denoted as $\Sigma^{-1/2}$ [67]. W is transposed to match the inner dimensions for matrix multiplication. Therefore, W = $\Sigma^{-1/2}$. A version of this is ZCA-cor, which ensures that the whitened data is as maximally correlated with the original variables of X as possible [67]. This whitened data is used in ridge regression, and the coefficients therefore represent coefficients in the transformed space, not in the original space. To get back to the original space it is worth considering what the linear regression for Z looks like in (33).

$$y = Z\beta_Z + \beta_0 + \varepsilon \tag{33}$$

The Z from (32) can be inserted into (33) to get

$$y = (XW^T)\beta_Z + \beta_0 + \varepsilon \tag{34}$$

This can be slightly rewritten by pulling X out and β_Z in, since both are multiplied

$$y = X(W^T\beta_Z) + \beta_0 + \varepsilon \tag{35}$$

By comparing this (35) with a normal linear regression in (36)

$$y = X\beta + \beta_0 + \varepsilon \tag{36}$$

It can be seen that the coefficient in the transformed space multiplied by the transposed whitening matrix satisfies $X\beta = X(W^T\beta_Z)$ and the original coefficients appears to be found by

$$\beta = W^T\beta_Z \tag{37}$$

The unwhitened coefficients of enhancer motif activities was validated as mention earlier.

Results and Discussion

Single-Cell Data Integration of Time Points

As previously stated, single-cell data from hESCs have been overexpressed with a single TF selected from a library of 196 unique TFs. This was achieved by transducing the cells with plasmids that contain the ORFs of the TFS, as illustrated in Figure 5A. Some hESCs in the data function as a control group, since they have been overexpressed with either GFP or mCherry. The cells were differentiated for 4 and 7 days and are visualised in Figure 6A.

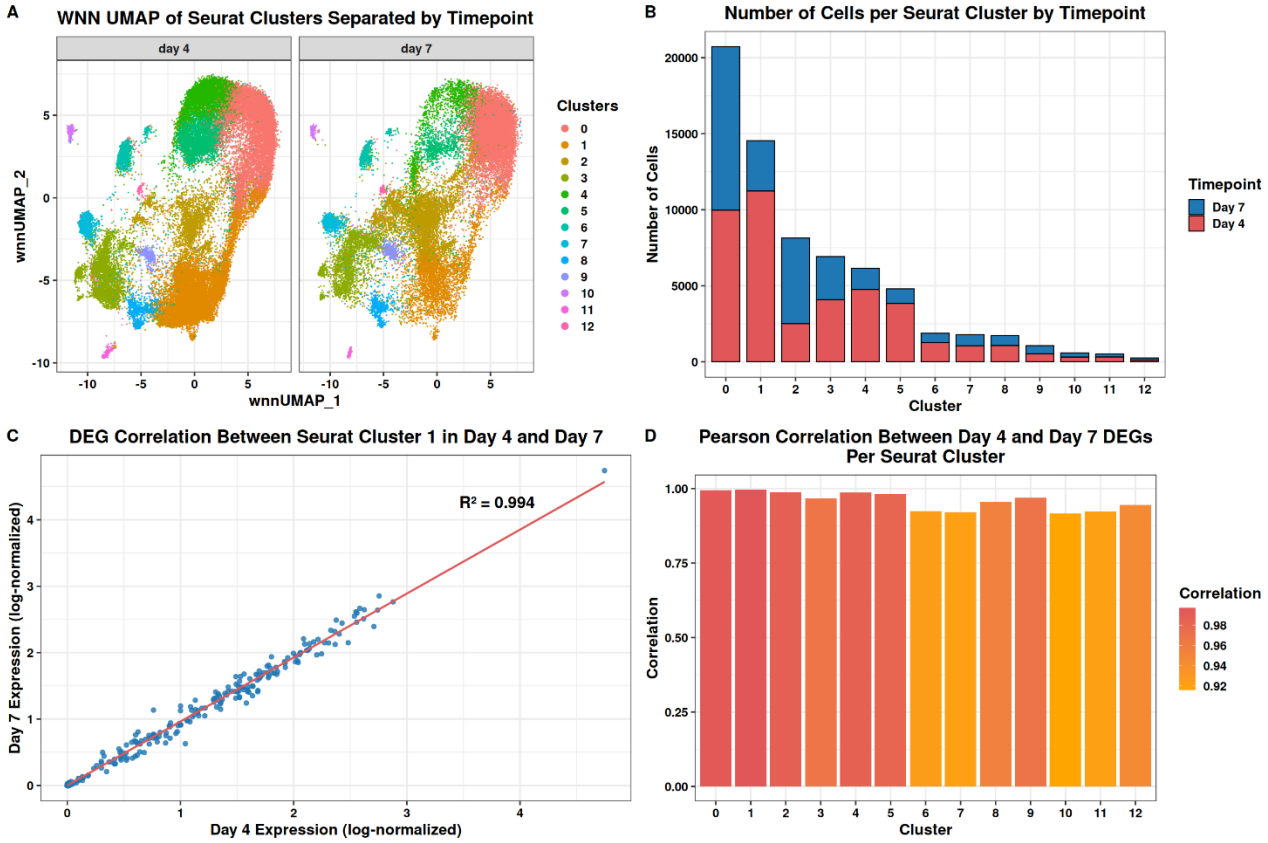


Figure 6. Timepoint integration analysis. *(A) Joint chromatin accessibility and gene expression analysis were performed by Joung et al. 2023 and the single-cell clusters for day 4 and day 7 have been visualised in a weighted nearest neighbour uniform manifold approximation and projection (WNN UMAP) plot to show both modalities in the same space [6]. There are 69085 single-cells in total. The x-axis and y-axis are the first and second unitless coordinates of the WNN UMAP. The clusters are colour-coded as shown on the right-hand side. Each dot is a cell. (B) Barplot of the number of cells per cluster for day 4 and day 7, where the x-axis is the cluster, and the y-axis is the number of cells. Day 4 is visualised in red, and day 7 is visualised in blue. (C) Scatterplot illustrating the correlation between shared genes of differentially expressed genes (DEGs) in cells from cluster 1 on Day 4 and Day 7. The x-axis represents gene expression on Day 4, log-normalised, while the y-axis represents gene expression on Day 7, log-normalised. R^2 is the coefficient of determination, which measures how well the variation in x explains the variation in y. It ranges from 0 to 1, with 1 being the optimal value. Each blue dot represents a gene. (D) Barplot of the Pearson correlation of shared genes of DEGs between day 4 and day 7 for each cell cluster. The x-axis is the cluster number, while the y-axis is the Pearson correlation value. The legend illustrates the colour coding of the Pearson correlation, where the highest correlation is represented in red and the lowest correlation is represented in orange. Pearson correlation values range from -1 to 1, where 1 indicates a positive directed correlation and -1 indicates an inverse correlation. A correlation of 0 indicates no correlation.*

It can be seen that the hESCs from day 4 and day 7 differentiated into 13 cell types and are depicted in the UMAP as 13 clusters. There appear to be more cells (dots) for day 4 in the UMAP, which can be confirmed by Figure 6B, which shows the number of cells per cluster for both day 4 and day 7. In total, there are 41040 cells for day 4 and 28045 cells for day 7. Additionally, it can be seen that there are more cells in total in cell cluster 0, and that number gradually decreases towards cell cluster 12. Since the cells for day 7 are more likely to be more differentiated, it would have been preferable to only work with those cells, but due to potential data sparsity, especially for cluster 6, 7, 8, 9, 10, 11 and 12, it was of interest to see if the data from the two timepoints could be merged.

For this purpose, the control and non-control cells were separated in the data, and the Pearson correlation was calculated for the differentially expressed genes (DEGs) for each respective cluster between day 4 and day 7 in Figures 6C and 6D for the non-control cells. The scatterplot in Figure 6C illustrates the correlation between the DEGs of cluster 1 between day 4 and day 7, with a high coefficient of determination at 0.994, which indicates that 99% of the variation in the data from day 4 can explain the variation in the data from day 7. This means that the variables correlate well, and the data from the two time points can be assumed to be highly similar for cluster 1. The same thing can be observed for the rest of the cluster, as seen in Figure 6D, with a minimum Pearson correlation as high as 0.92. This was expected, as it is well documented that stem cells are first fully differentiated around the 21-day mark [68], [69], [70]. From this, the rest of the clusters from the two time points are also assumed to be highly similar, and the data are merged for downstream analysis.

Integration and Cluster-Level Enrichment of Control Cells

Similarly, it was of interest to examine whether the control cells could be merged, since the GFP and mCherry data are very sparse, especially compared to the cells that have been overexpressed (OE) with a TF, as seen in Figure 7A. Because of this, the correlation between the gene expression of the control cells that had been transduced with GFP and mCherry was calculated, as illustrated in Figure 7B. It can be seen that there is a near-perfect correlation between the two types of control cells' gene expression, as indicated by the coefficient of determination at 0.992, which means that about 99% of the variation in the dependent variable, mCherry gene expression, can be explained by the independent variable, GFP gene expression. Because of this, it was assumed that the data could be safely merged for the two types of control cells.

Additionally, the enrichment of control cells compared to overexpressed cells in each cluster relative to their representation across all clusters was assessed by calculating the log2 fold enrichment using equation (24) in the methodology section and visualised in Figure 7C.

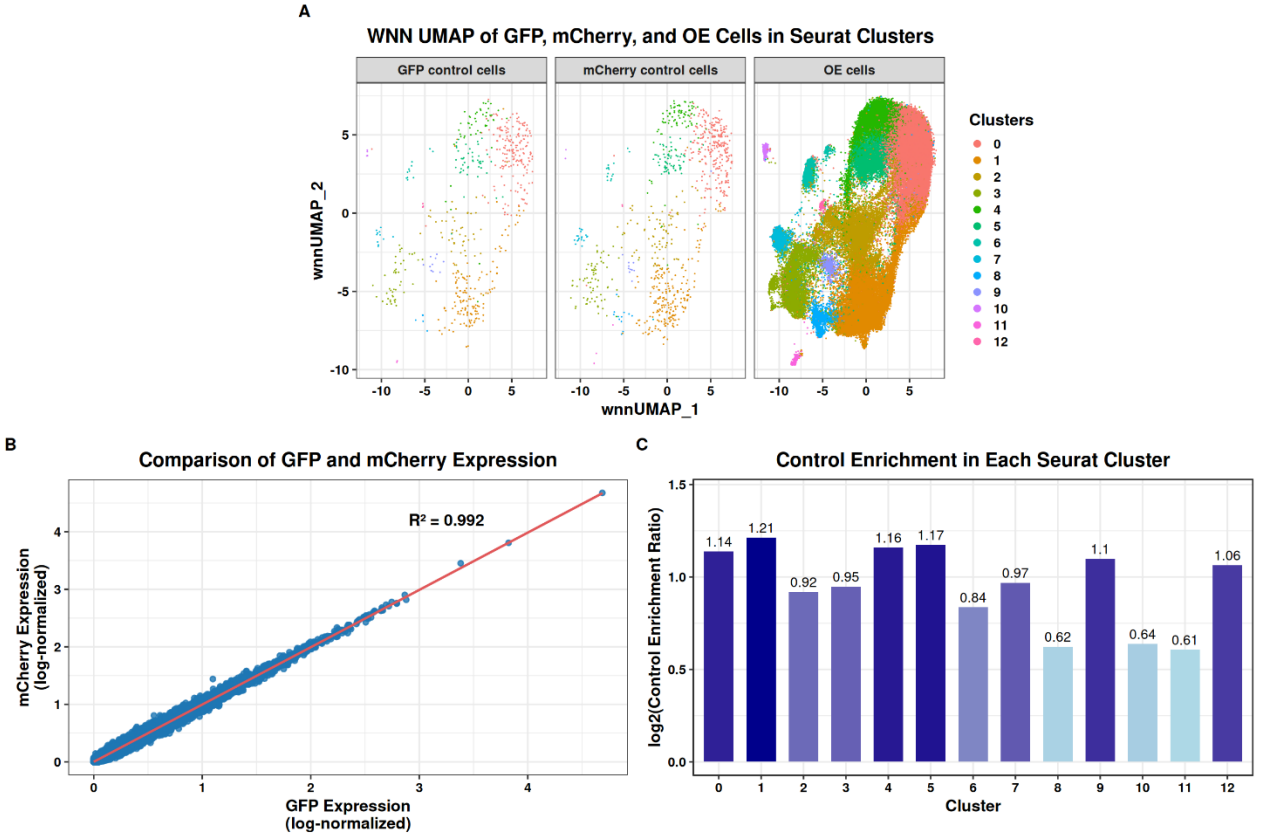


Figure 7. Integration and enrichment of control cell analysis. *(A) Joint chromatin accessibility and gene expression analysis were performed by Joung et al. 2023. The cells (dots) in each cluster that have been overexpressed (OE) with a transcription factor (TF) have been visualised in a weighted nearest neighbour uniform manifold approximation and projection (WNN UMAP) plot to show the cells with regard to both modalities in the same space [6]. Additionally, control cells (dots) that have been overexpressed with GFP and mCherry are likewise shown in a WNN UMAP. The x-axis and y-axis are the first and second unitless coordinates of the WNN UMAP, and the cell clusters are colour coded as seen in the legend on the right-hand side. (B) Scatterplot of the correlation between the gene expression of GFP and mCherry control cells, log-normalised. The x-axis is the gene expression of GFP control cells, and the y-axis is the gene expression of mCherry control cells. Each blue dot represents a gene. The R^2 is the coefficient of determination, and it measures how well the variation in x explains the variation in y. It ranges from 0 to 1, with 1 being the optimal value. (C) Barplot of the enrichment of control cells in each cluster as calculated in equation 24. The x-axis is the cluster number, and the y-axis is the log2-scaled control enrichment ratio, which is also shown on top of each barplot. A value of 1 means 2 times enrichment, and a value of -1 is a 2 times depletion.*

From Figure 7C, it can be observed that some clusters appear to be more enriched in control cells than the other clusters. A log2 ratio value of 1 indicates a 2-fold enrichment of control cells in a given cluster compared to the global enrichment of control cells. A value of zero means that there is no enrichment, and a value of -1 would indicate a two-fold depletion of the control cells. Since some of the clusters, especially clusters 0, 1, 4, 5, 9, and 12, appear to be more enriched for control cells, this raises the question of whether any of these clusters are more hESC-like than the rest of the clusters, which are not as enriched in control cells.

Evaluating Stemness Across Differentiated Cells

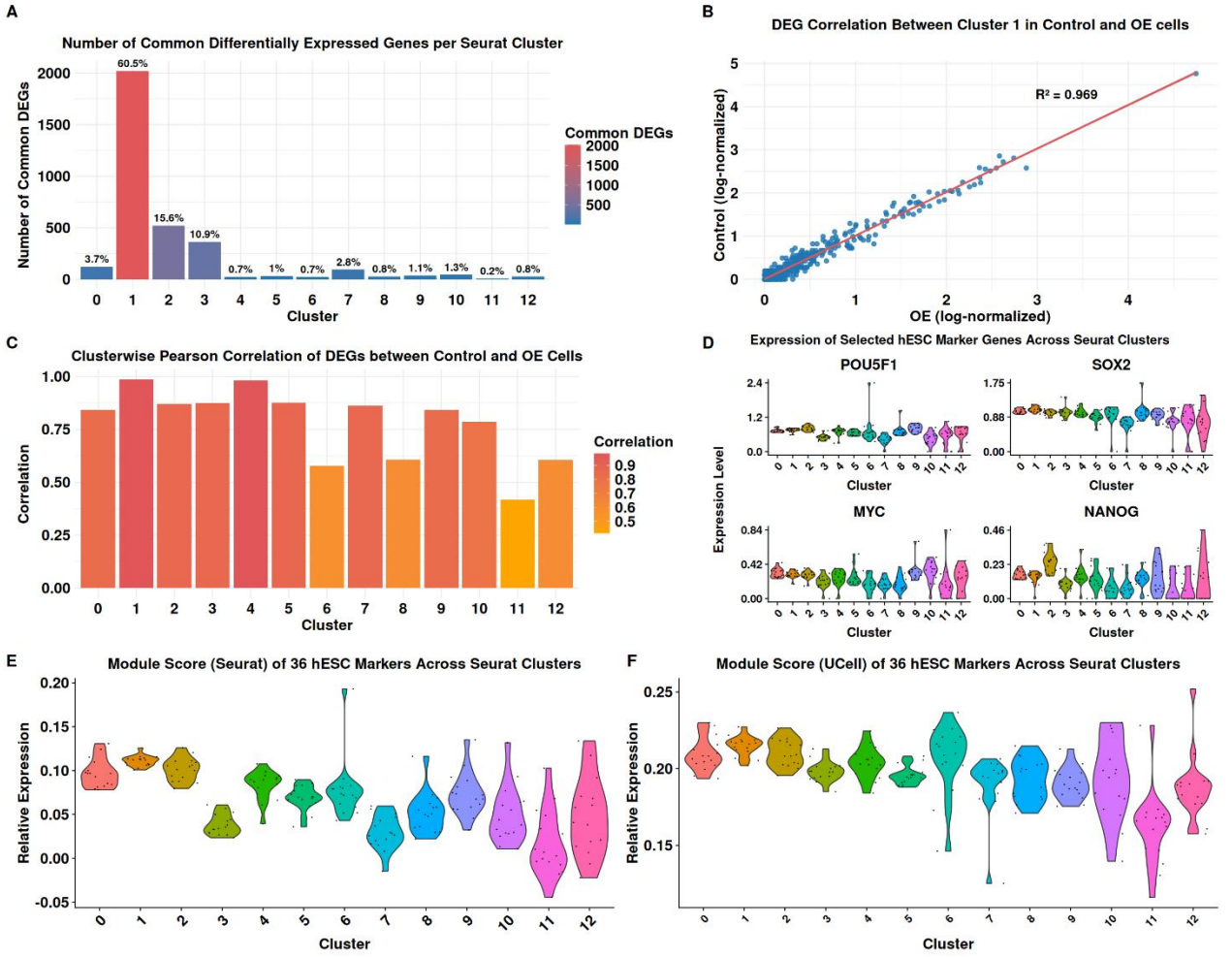


Figure 8. Cell cluster stemness analysis. *(A) Barplot of the number of differentially expressed genes (DEGs) found to be shared between overexpressed (OE) cells and control cells for each cluster. The percentage of DEGs that each cluster accounts for out of all the identified common DEGs across clusters is shown on top of each bar. The x-axis is the cluster number, and the y-axis is the number of DEGs found in common between OE and control cells. The number of common DEGs is colour-coded as shown in the legend to the right of the barplot. (B) Scatterplot of the correlation of shared genes of DEGs between cluster 1 in control and OE cells. Each blue dot is a gene. The x-axis is the gene expression of OE cells, log-normalised, and the y-axis is the gene expression of control cells, log-normalised. The coefficient of determination (R^2) measures how well the variation in x explains the variation in y, and it ranges from 0 to 1, where 1 is the best fit. (C) Barplot of the Pearson correlation of shared genes of DEGs between OE and control cells per cluster. The y-axis is the Pearson correlation, and the x-axis is the cluster number. The bars are colour-coded based on the Pearson correlation as shown in the legend on the right-hand side of the plot. (D) Violin plots of the gene expression of human embryonic stem cell (hESC) marker genes POU5F1, SOX2, MYC and NANOG in each cluster [71], [72]. The y-axis is the gene expression log-normalised for each respective gene, and the x-axis is the cell clusters. Each black dot represents a pseudobulk of the single cells based on the Seurat clusters and the cell batch number. There are 16 batches in total. (E) Violin plot of Seurat module scores of 36 hESC marker genes across cell clusters. The y-axis is the relative expression of each marker gene as described in the method section, and the cell clusters are shown on the x-axis. A higher module score or relative expression means that the set of genes is more highly expressed in the particular cell compared to the randomly selected control genes, as described in the methodology section. Each black dot represents a pseudobulk of cells based on the Seurat cluster and total batch number (16). (F) Violin plot of UCell module scores of 36 hESC marker genes. The module score or relative expression is shown on the y-axis and is described in the method section, while the x-axis shows the cell cluster number. A lower module score indicates that the normalised ranks of the genes are worse compared to the best possible rank the entire gene set could have. Each black dot is a pseudobulk of the single cells based on the Seurat clusters and the total batch number (16) [71], [72], [73], [74], [75].*

To assess the similarity between control and non-control cell clusters, the DEGs shared between each respective cluster and the control types were evaluated. In Figure 8A, it can be seen that cell cluster 1 of non-control cells has more DEGs in common with cluster 1 of the control cells. In fact, it accounts for roughly 60% of all the common DEGs found across clusters. Based on this, a correlation was calculated between the common DEGs in Figure 8B, and it can be observed that they also correlate highly. This suggests that cluster 1 may be more stem cell-like than the other clusters. However, the rest of the first couple of clusters also correlate fairly highly, despite not necessarily sharing that many DEGs in common between control and non-control cells, as seen in Figure 8A and C. Additionally, based on what was observed in Figure 7A and C, cluster 0 was likewise highly enriched in control cells, which might point towards cluster 0 also being more stem cell-like. To further analyse stemness, the gene expression of hESC marker genes, such as the Yamanaka factors, MYC, SOX2, POU5F1, and the hESC marker NANOG, can be evaluated in Figure 8D [71], [72]. It can be observed that all clusters express the marker genes to nearly the same degree, which means that, based on these specific marker genes, stemness among the clusters cannot be determined. Because of this, module scores were calculated for 36 H1 hESC marker genes to assess the relative expression of these genes across cell clusters as seen in Figure 8E and F [71], [72], [73], [74], [75]. From the Seurat and UCells module scores, it is evident that clusters 0, 1, and 2 appear more stem cell-like compared to the other clusters based on the 36 stem cell marker genes and will therefore be annotated as hESC 1, 2, and 3 for downstream analysis.

Identification of Cell Types Through Key Marker Genes

To investigate the cell types of the other clusters, the top 20 DEGs were examined, and an example of this is shown in Figure 9.

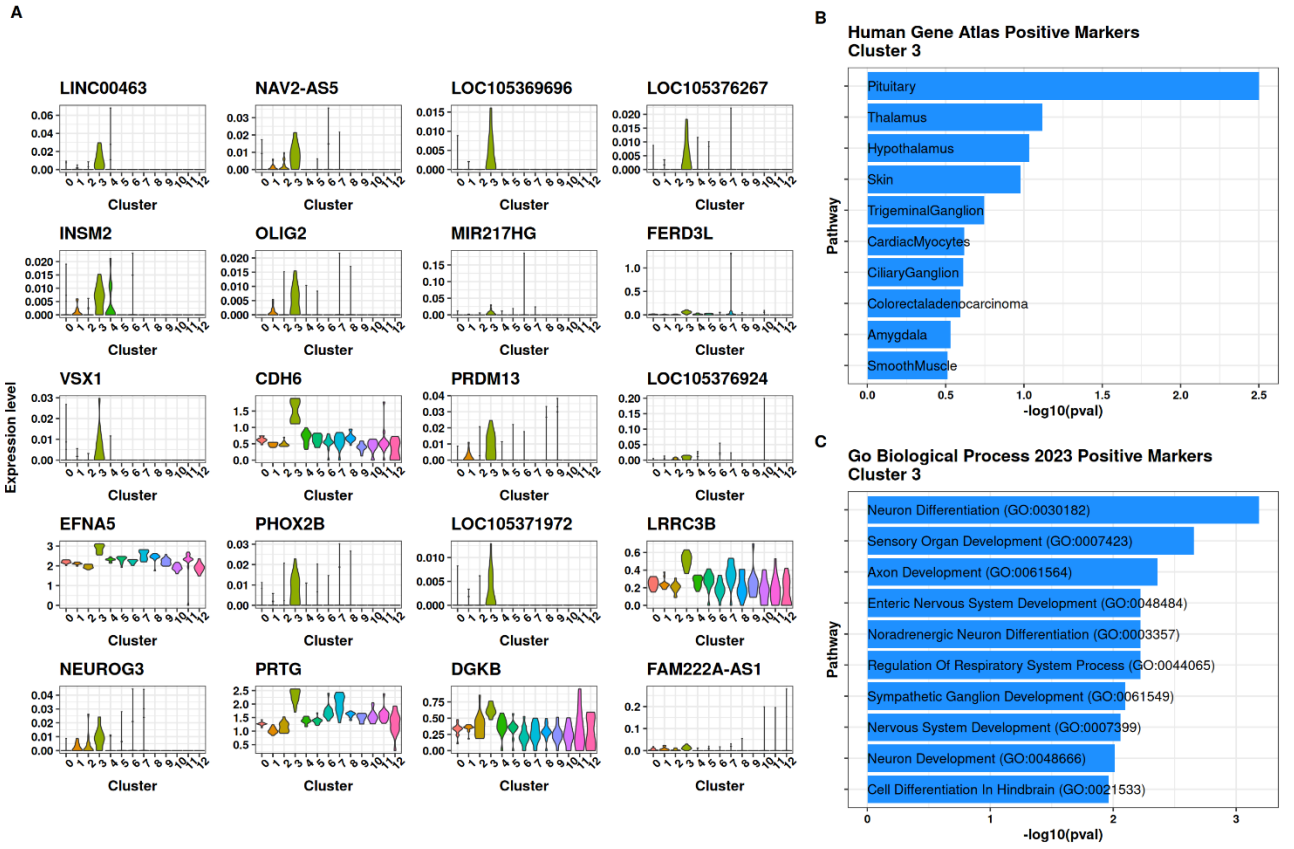


Figure 9. Top 20 marker genes and gene set enrichment analysis of cluster 3. *(A) Violin plots of the top 20 positive marker genes of cluster 3, where the log-normalised gene expression is shown on the y-axis and the cell cluster in which they are expressed on the x-axis. The names of the marker genes are shown on top of each violin plot. (B) Barplot of enriched pathways/terms in the gene set enrichment analysis using the human gene atlas database. On the y-axis, the cell type associated with the top 20 differentially expressed positive marker genes is shown according to the human gene atlas database, while the x-axis is the -log10 transformed p-value, indicating the significance of the differentially expressed positive marker genes associated with the term. (C) Barplot of enriched pathways or biological processes associated with the top 20 differentially expressed positive marker genes according to the GO Biological Process 2023 database. The enriched pathway/term is shown on the y-axis, while the x-axis shows the -log10 scaled p-value.*

The top 20 DEGs for cluster 3 are shown in Figure 9A. The y-axis shows the normalised gene expression for each gene, and the x-axis shows each cluster. The marker gene that immediately indicated this cluster might be neuron cells was NEUROG3, as it has been documented to be involved in neurogenesis as a TF [76]. However, it has also been documented to play a role in the differentiation of pancreatic progenitor cells into pancreatic islet cells [77]. Nevertheless, there are other genes that likewise point towards cluster 3 being neuron cells, such as OLIG2, PHOX2B and VSX1 [78], [79], [80], [81]. OLIG2 is a key TF that differentiates oligodendrocyte precursor cells into immature oligodendrocytes. An oligodendrocyte is a cell type that resides in the central nervous system and produces myelin and wraps it around neuronal axons [81]. PHOX2B and VSX1 are respectively associated with the differentiation of progenitor cells into noradrenergic and interneurons [78], [79], [80]. This is further supported by gene set enrichment analysis as shown in

Figure 9B and C. From Figure 9B the top 10 cell types, which are mostly associated with these 20 marker genes according to the human gene atlas database, are shown on the y-axis, while the x-axis is the –log10(p-value), where larger values indicate higher statistical significance of enrichment [52], [54]. From 9C, the top 10 biological processes according to the GO biological process 2023 database, in which the top 20 differentially expressed genes are associated with, are shown on the y-axis and the statistical significance in enrichment is likewise shown on the x-axis [52], [53]. As evident from Figure 9B and C, the highest enriched cell types and biological processes are linked to neuron cells, such as pituitary, thalamus, hypothalamus, neuron differentiation, noradrenergic neuron differentiation, neuron development, etc. However, it is important to note that the cells from cluster 3 are most likely neuron-like cells, because none of the cell clusters appear fully differentiated based on Figure 8, and the fact that most cells are only near fully differentiated around the 21-day mark, as mentioned earlier. Because of this, all cell types are annotated as like cells. This analysis was conducted for each cluster, as shown in the Supplementary Figures 1-14. Additionally, because all of the clusters represent cell-like types and are not fully differentiated, some clusters proved more difficult to annotate due to their chimeric nature. Examples of this are clusters 5, 7 and 8. The marker genes and gene set enrichment analysis did not always point towards a definitive cell type. In these cases, the database Panglaodb was used, and other markers associated with specific cell types were investigated, as shown in the supplementary material for clusters 7 and 8 [55], [82], [83], [84], [85], [86], [87], [88], [89], [90], [91].

Selected marker genes for each cluster are represented in Figure 10A, to compare the cell clusters based on the marker genes.

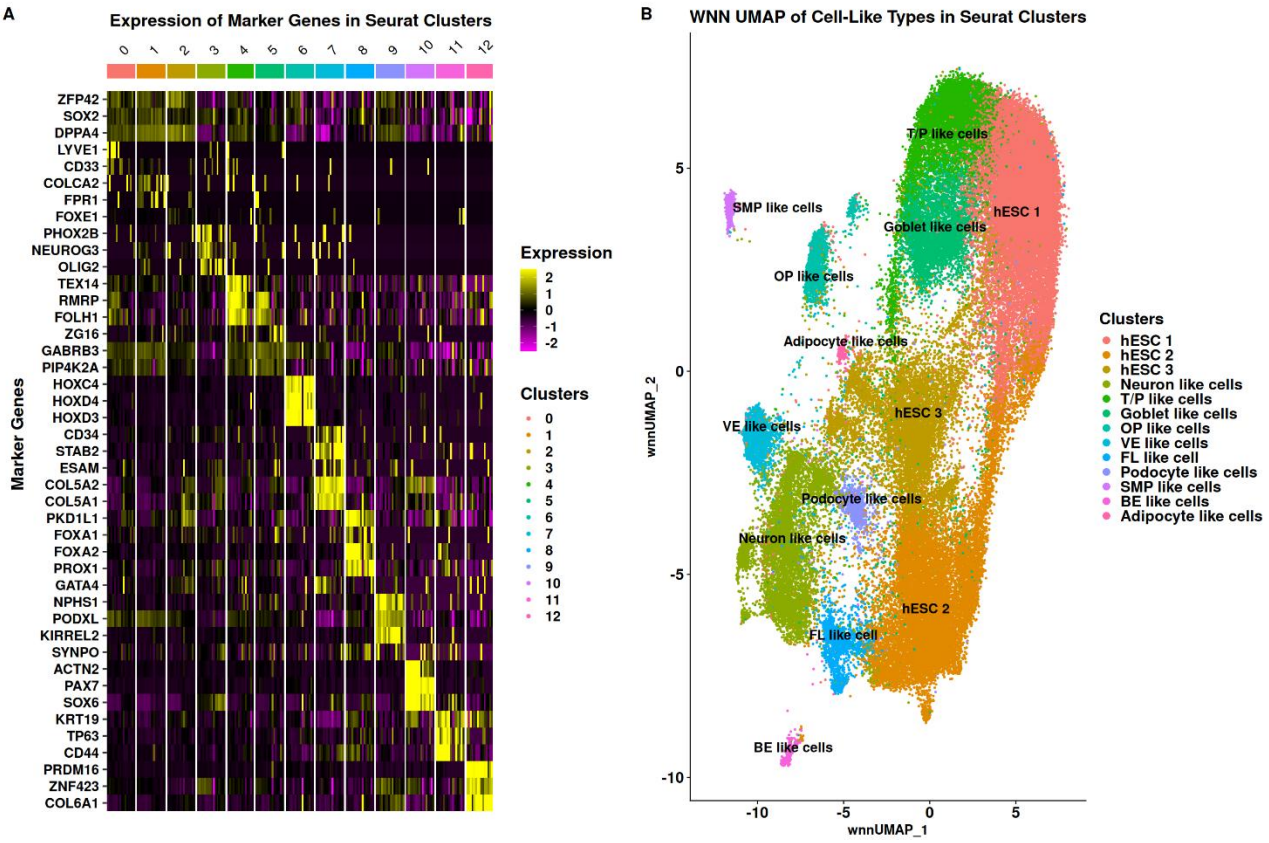


Figure 10. Selected marker gene expression comparison between Seurat clusters and cell type annotation. *(A) The heatmap illustrates selected marker genes for each cluster. The y-axis displays the chosen marker genes for every Seurat cluster, while the x-axis represents the Seurat clusters, colour-coded and numerically annotated to align with the legend on the right-hand side of the heatmap. Each square in the heatmap corresponds to a pseudobulk derived from the Seurat clusters and batch number, encompassing a total of 16 batches. Each pseudobulk is colour-coded according to the gene expression level within it. Gene expression has been log-normalised, with the highest expression represented by a yellow colour and the lowest expression shown in purple, as detailed in the second legend on the right-hand side of the heatmap. (B) Joint chromatin accessibility and gene expression analyses were conducted by Joung et al. 2023. Each dot represents a single cell visualised in a weighted nearest neighbour uniform manifold approximation and projection (WNN UMAP) plot, illustrating the cells in relation to both modalities within the same space (Joung et al. 2023). Each cell is colour-coded according to the annotated cell types displayed on the right-hand side of the WNN UMAP plot. The x-axis and y-axis are the unitless first and second coordinates of the WNN UMAP.*

From Figure 10A, it can be seen that three selected marker genes for hESCs, ZFP42, SOX2 and DPPA4 are highly expressed in cluster 0, 1 and 2, but cluster 4 and 9 likewise seem to express these more so than the other clusters [71], [72], [73], [74], [75]. Furthermore, cluster 5 also appears to express marker genes such as TEX14, RMRP, and FOLH1, which characterise cluster 4 and are associated with testis/prostate-like cells [92], [93], [94]. Additionally, cluster 4 also appears to express marker genes for cluster 5, such as ZG16, GABRB3, and PIP4K2A, which could indicate that clusters 4 and 5 are more similar to one another compared to other cell clusters. ZG16 is a component in the mucus that goblet cells secrete, while GABRB3 and PIP4K2A respectively encode for a receptor in the brain and an enzyme which has been documented to be involved in energy metabolism [95], [96], [97]. Cluster 5 was annotated as goblet-like cells, since Panglaodb

annotated it as such based on the top 20 DEGs [55]. Overall, clusters 0, 1, 2, 4, 5 and 9 appear to be more similar based on the gene expression of these marker genes, while a more transcriptionally distinct profile can be observed in the remaining clusters regarding the selected marker genes. With that in mind, the final annotations for cell types across the 13 clusters are illustrated in the UMAP in Figure 10B, with the legend aligned to the cluster order: cluster 0 corresponds to hESC 1, cluster 2 to hESC 2, and so forth.

TF Enrichment Patterns from Overexpression in Single Cells

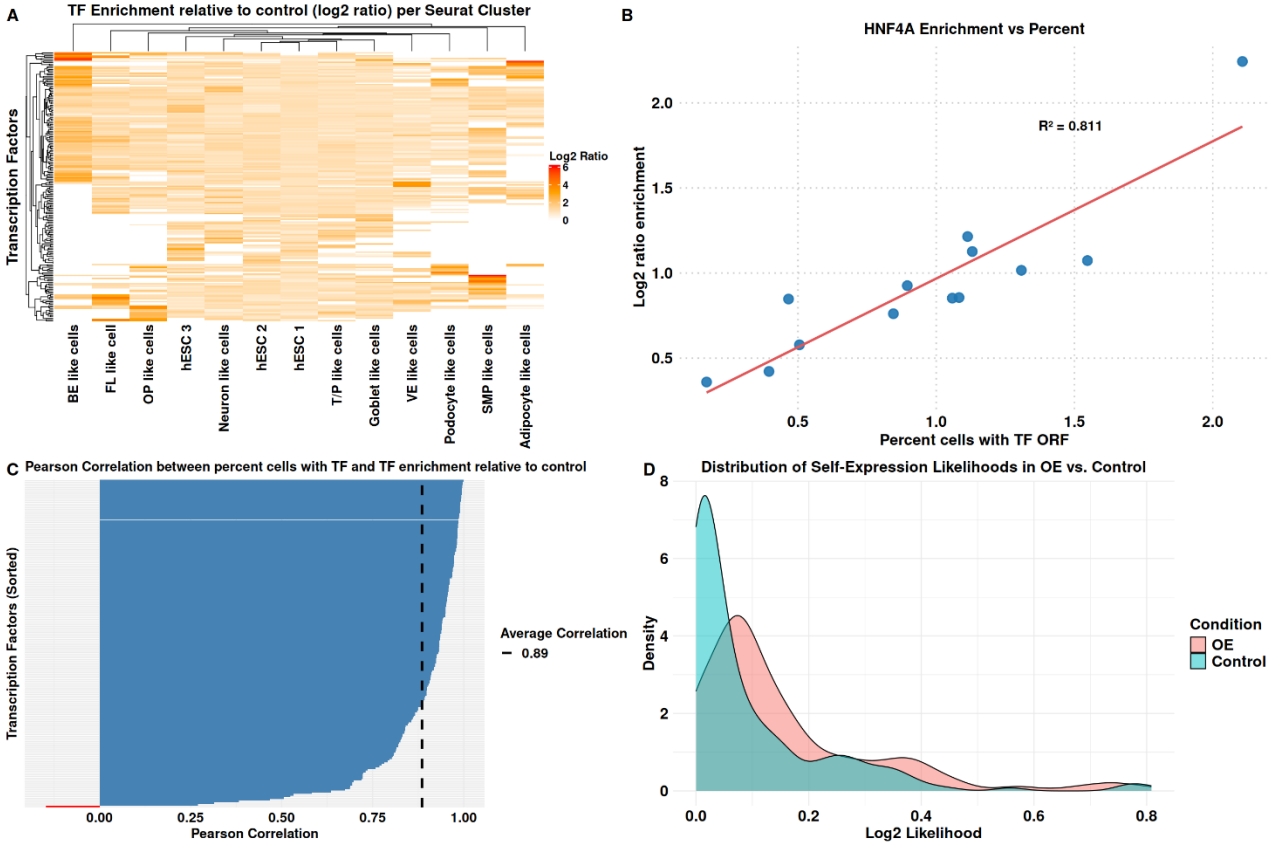


Figure 11. Transcription factor (TF) enrichment vs. percent of cells expressing the TF analysis. *(A) Heatmap of TF enrichment for each cell type. The cell types are shown on the x-axis, while the TFs (196) are shown on the y-axis. The TF enrichment is colour-coded according to the log2 ratio value, as indicated in the legend on the right-hand side of the heatmap. An enrichment value of 1 indicates a 2-fold enrichment, while a value of -1 indicates a 2-fold decrease, and the TF enrichment is calculated according to equation (26). (B) Scatterplot of the correlation between the log2 ratio enrichment value and the percentage of cells transduced with the TF open reading frame (ORF) value for the HNF4A TF. Each dot is a cell type. The x-axis is the percentage of cells with the TF ORF as calculated in equation (27), while the y-axis is the log2 ratio enrichment of the TF. The coefficient of determination (R^2) indicates how well the variation in x explains the variation in y, and the values range from 0 to 1, where 1 is the best fit. (C) Barplot of the Pearson correlation for each TF (196) between the log2 ratio enrichment and percentage of cells with TF ORF. The y-axis shows the TFs sorted based on the highest Pearson correlation on the x-axis. Pearson correlation ranges from a value of -1 to 1, where a value of 1 indicates perfect positive correlation and a value of -1 is a perfect inverse correlation. A value of 0 means no correlation. The Average correlation is indicated on the right-hand side and shown with dashed lines on the barplot. (D) Density plot showing TF self-expression likelihoods in overexpressed (OE) versus control cells, calculated using equation (28). The x-axis represents log2 likelihood of TF self-induction, and the y-axis depicts the distribution of each TF value. A value of 1 indicates a 2-fold likelihood of self-induction. Distributions for OE and control cells are colour-coded in the legend on the right.*

The cells were, as aforementioned, overexpressed with TF encoding plasmids, which resulted in the annotated cell types as seen in Figure 10B. The TFs enriched in each cell type can be examined and visualised as seen in Figure 11A. The TF enrichment score was calculated using the equation (26). It can be seen that certain TFs appear to be cell-type specific, as indicated by the high log2 ratio scores, coloured in red. Furthermore, it can be observed that the first six cell types corresponding to clusters one through six: hESC 1-3, neuron-like cells, testis/prostate-like cells and goblet-like cells, do not seem to be highly enriched for any TFs in particular. This might be interpreted as another indicator of the fact that these cell types are not nearly as differentiated, which was also observed in Figure 8C. Additionally, the lack of enrichment in these cell types also reflects that they have a higher enrichment of control cells, which was also observed in Figure 7C. This is because equation (26) utilises the ratio between overexpressed cells and control cells within each cluster in relation to the global enrichment of the TFs. It was for this reason that this metric was preferred and utilised for the downstream validation and comparison analysis of IMAGE, as it provides this extra bit of information, in contrast to the percentage of cells with TF ORF, which the authors of the data use in equation (27). To illustrate that the TF enrichment metric reflects similar biological patterns to the percentage of cells with TF ORF, the correlation between the two metrics was calculated and is shown in Figure 11B and C. Figure 11B displays an arbitrary example of a TF (HNF4A) and the correlation between the two metrics. It can be seen that the coefficient of determination is quite high at 0.8. If the correlation is performed for each of the 196 TFs, it can be observed that the average correlation for all TFs is 0.89, as shown in Figure 11C. This suggests that the TF enrichment metric used in Figure 11A seems appropriate to use instead, as it reflects similar patterns and arguably also offers a more nuanced enrichment representation.

In addition to establishing TF enrichment for the validation and comparison of IMAGE, it is essential to assess whether the overexpression of a TF results in the endogenous self-expression of the given TF. This was calculated using equation (28), and the distribution of self-expression likelihoods of TFs in overexpressed cells (OE) versus control cells is shown in Figure 11D. From the density plot, it can be seen that there is a slightly higher likelihood of inducing an endogenous self-expression of a TF when cells are overexpressed with that TF compared to control cells. This could suggest a potential positive feedback mechanism for certain TFs. However, the control cells also appear to express TFs to a certain degree, which could be attributed to technical background noise, or perhaps the TFs might simply be natively expressed in the human embryonic control stem cells. The latter would make sense, since the non-control cells are still quite similar to the control

cells as established, and because of this, there is potentially not much difference in their endogenous gene expression.

Comparative Analysis of Predicted Motif Activities

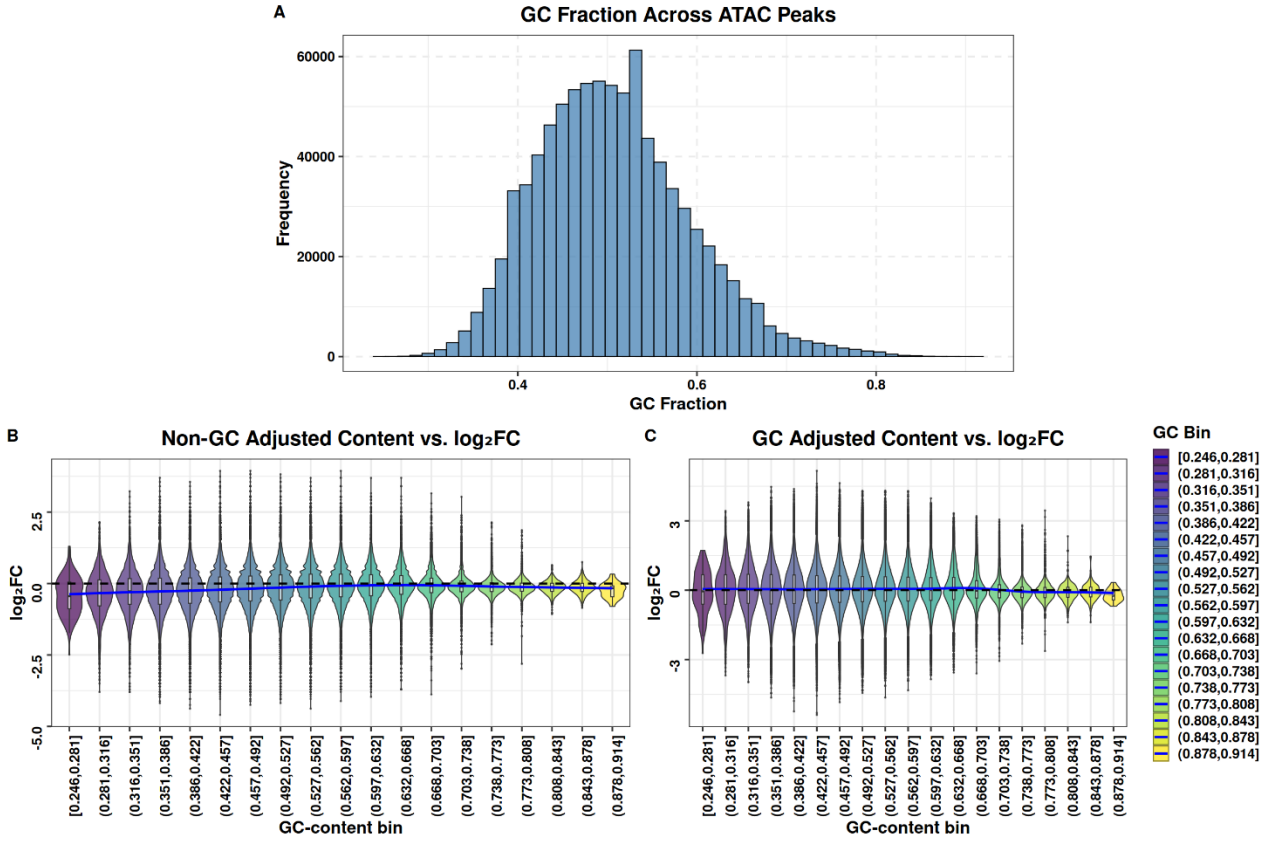


Figure 12. Examination of GC content and effect. *(A) Histogram illustrating the fraction of GC content across ATAC-seq peaks. The x-axis represents the fraction of GC content found in ATAC-seq peaks, while the y-axis indicates the frequency of this GC fraction. (B) Non-GC adjusted violin plot showcasing a mock null differential accessibility test. On the x-axis, the GC content fraction is organised into equal-sized bins, each increasing by 0.035, with violin plots coloured according to the GC bin, as detailed in the legend on the right. The log2 fold change (FC) is displayed on the y-axis, reflecting the magnitude of change, while the blue trend line summarises the behaviour of log2FC across the GC content bins in the mock null differential accessibility test. (C) GC adjusted violin plot of a mock null differential accessibility test. The y-axis displays the log2FC, and the x-axis is binned into equal-sized increments of 0.035, representing the GC fraction. The violin plots are colour-coded based on their corresponding GC content bins, as illustrated in the GC bin legend on the right. The blue trend line denotes the behaviour of log2FC across these GC content bins.*

After laying the groundwork for TF enrichment and TF self-expression likelihood for downstream motif activity validation, the enhancer motif activities of IMAGE and chromVAR were predicted. However, it was rather difficult to get IMAGE to assign enhancer motif activities that would result in any meaningful clustering of the cell types, as seen in Supplementary Figures 16 and 17, which likewise did not perform well in the validation. Because of this, biases in the data were explored, and unadjusted GC content seemed to substantially affect the performance of IMAGE. In Figure 12A, it can be seen that the vast majority of peaks have a medium amount of GC content, which

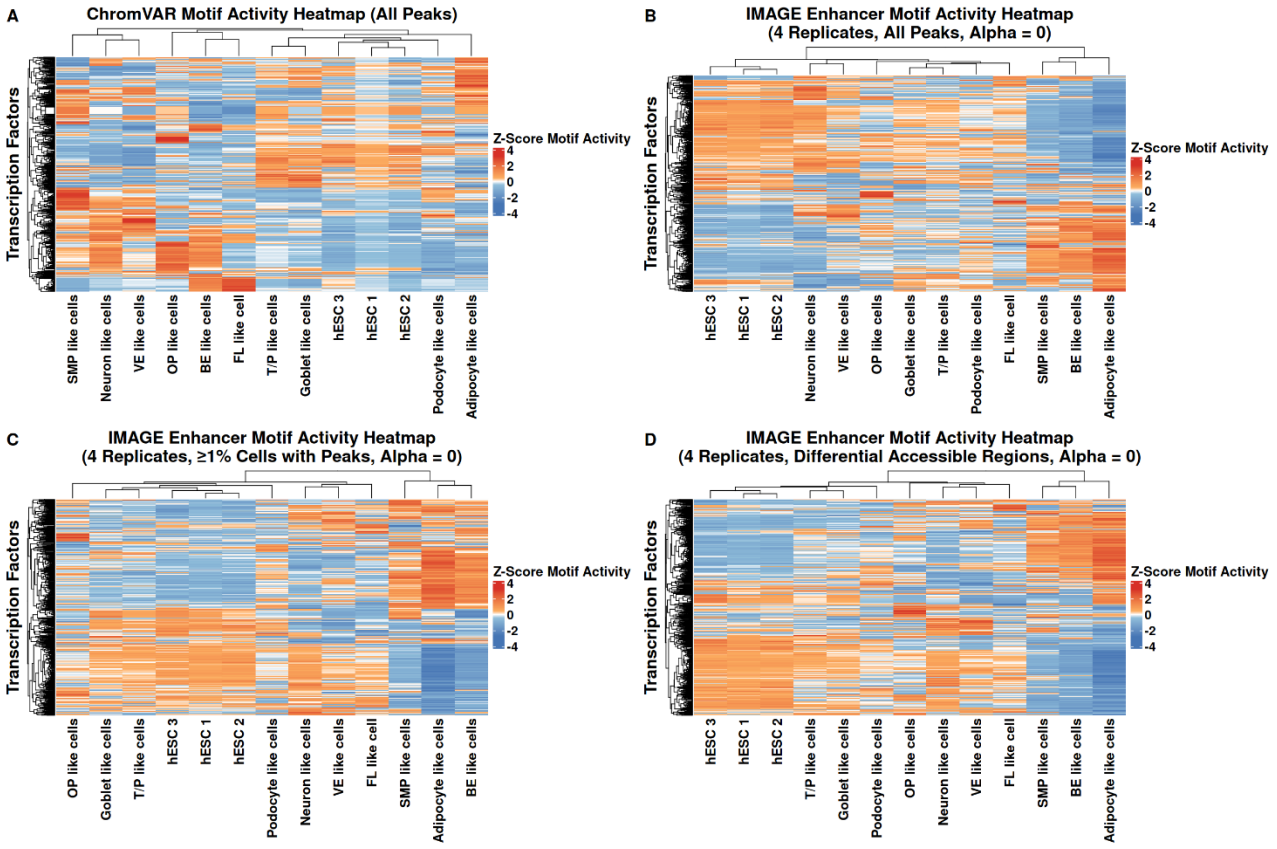


Figure 13. Motif activity heatmap analysis of IMAGE vs. chromVAR. *(A) This heatmap displays the predicted motif activities from chromVAR for all ATAC-seq peaks. The x-axis represents each cell type, while the y-axis represents the transcription factors (TFs) (870). Each TF in each cell type is colour-coded according to the z-scored motif activity, as indicated on the right side of the heatmap. A z-score for motif activity indicates whether the activity is above or below the average for that specific motif, which is defined as a value of 0. Values above 0 indicate greater than average activity, while values below 0 indicate lesser activity in comparison across cell types for that specific TF motif. The dendrogram above the heatmap illustrates how cell types are clustered based on their motif activities, whereas the dendrogram on the left illustrates the clustering of TFs based on their motif activities across different cell types. (B) Heatmap of IMAGE predicted enhancer motif activities for all ATAC-seq peaks, pseudobulked to four replicates per cell type and estimated using ridge regression (alpha = 0). The x-axis represents the cell types, while the y-axis displays each transcription factor (TF) (870). For each TF, the enhancer motif activity is colour-coded according to the z-scored enhancer motif activity of IMAGE, as indicated on the right-hand side of the heatmap. The z-score signifies whether the motif's activity is above or below its average activity across cell types, with the average activity defined as a value of zero. The dendrograms on the rows cluster the TFs based on their activity across cell types, and the dendrogram atop each column clusters the cell types according to their motif activity for each TF. The same observations apply to (C) and (D), but the distinction lies in the feature selection of peaks; (C) illustrates the motif activity for only peaks detected in one per cent or more of cells, while (D) showcases the enhancer motif activity for peaks that are differentially accessible.*

may be worth adjusting for, as illustrated in Figures 12B and 12C, as it can, at the very least, affect differential accessibility tests.

Each ATAC-seq sequence is binned based on its GC content, and a mock null differential accessibility test was performed, as described in the methodology section and Berge et al. 2022 [57]. Because this type of test is conducted on a single cell type with an artificial test and control group, the magnitude of change (log2fc) should be centred around zero across bins, as there should be no change between conditions. However, this is not the case, as seen in Figure 12B, which

indicates that the source of variation in this analysis is the GC content. When the GC content is adjusted for, the log2fc values centre around zero as intended, as shown in Figure 12C. After adjusting the GC content in the ATAC-seq peaks, IMAGE enhancer motif activities were predicted, yielding the following results.

ChromVAR predicted enhancer motif activities on a single cell level and averaged to bulk level, while IMAGE predicted enhancer motif activities on a bulk level for all peaks as seen in Figure 13A and B. A motif database of 870 non-redundant TF motifs was used for the analysis [5]. The cell types are represented on the x-axis, and the TF motifs are plotted on the y-axis. The motif activities are represented as z-scores to facilitate comparison across cell types. A motif activity z-score of 0 suggests average motif activity when comparing across cells. Conversely, a z-score above or below 0 indicates that the motif for a specific TF is more or less active relative to different cell types. For chromVAR, it is evident that the cell types hESC 1-3, testis/prostate-like cells, goblet-like cells, and podocyte-like cells are clustered together based on their motif activities. This clustering is desired, as it was these clusters that seemed most similar based on the selected marker genes in Figure 10A. Evidently, the clustering of the cell types is also reflected in their motif activities.

On the other hand, IMAGE does not produce the same type of clustering as seen in Figure 13B. Although the hESCs 1-3 are clustered together, and the testis/prostate, goblet, and podocyte-like cells are clustered together, these cell types should preferably all cluster together as seen in 13A. This might indicate that IMAGE has difficulty assigning accurate motif activities when there are too many accessible sites during ridge regression, which could act as noise. Because of this, two feature selection methods for the peaks were explored. The first method involved only retaining peaks detected in at least 1% of cells within a given cluster, while the second method selected peaks that were differentially accessible. As seen in Figures 13C and 13D, both feature selection methods result in IMAGE assigning motif activities that reflect the preferred clustering of the aforementioned cell types. The issue with the peaks detected in at least 1% of the cells is that these peaks may not differ between clusters and may potentially include ubiquitous or non-informative peaks that simply act as noise for downstream analysis. The specific threshold of 1% was chosen since it yielded 177.472 unique peaks, and a higher threshold would result in too sparse data. On the other hand, differentially accessible peaks would include regulatory elements that are functionally distinguishable between cell types and, as a result, might be more suitable for TF activity validation and GRN inference in general. Additionally, this method yielded 216.612 unique peaks, and as a result, differentially accessible peaks were used in the subsequent analyses. Interestingly,

chromVAR is not capable of clustering cell types properly based on motif activities when run on the same bulk data as IMAGE, as shown in Supplementary Figure 15. This suggests that the raw accessibility deviation of chromvar from equation (13) might work better for sparse data than for bulk data, potentially because it is easier to detect the presence of raw accessibility in sparse data due to its binary-like nature. In contrast, bulk data may obscure raw accessibility because of continuous read counts, as opposed to sparse, binary-like data.

Despite both chromVAR and IMAGE now assigning motif activities that reflect the preferred cell-type clustering, the two enhancer motif activity prediction tools do not necessarily agree with each other, as seen in Figure 14. The Pearson correlation between the z-score enhancer motif activities for each cluster is calculated. If IMAGE and chromVAR were in complete agreement, there would be a red diagonal line in the correlation matrix with a value of 1, which is not the case. Naturally, this difference in enhancer motif activities stems from the fact that IMAGE and chromVAR predict activities differently, as chromVAR utilises raw accessibility deviations, as shown in equation (13), whereas IMAGE employs ridge regression, as described in equation (1). To determine which method more accurately reflects TF activity, validation of the two tools was performed.

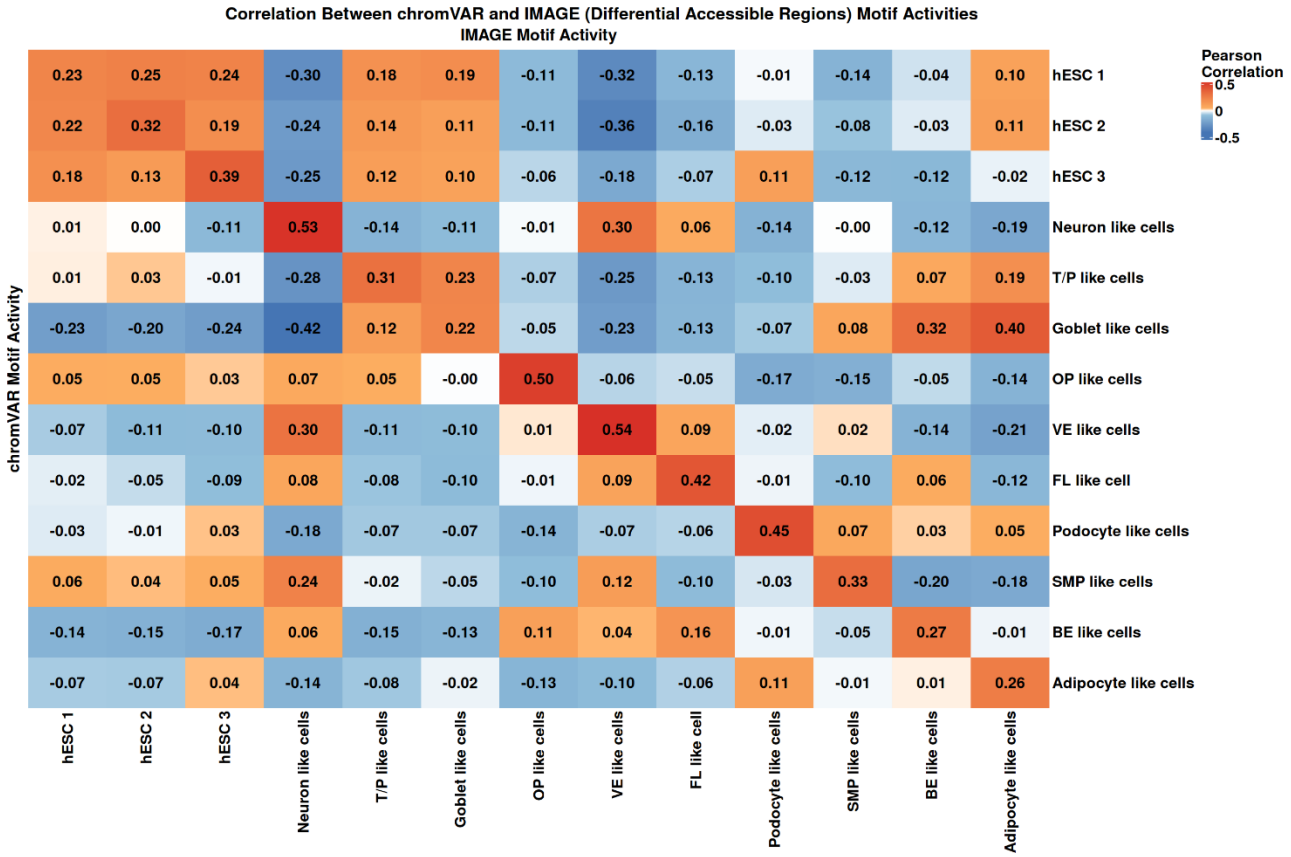


Figure 14. Heatmap of Pearson correlation comparison between IMAGE and chromVAR cell types based on enhancer motif predictions. *The y-axis and x-axis display each cell type for chromVAR and IMAGE, respectively. The values within the heatmap represent the Pearson correlation for each comparison between the cell types of IMAGE and chromVAR, colour-coded according to the Pearson correlation value as indicated in the legend on the right-hand side of the plot. A Pearson correlation value of -1 denotes an inverse correlation between the two comparisons, a value of 0 signifies no correlation, and a value of 1 indicates a perfect correlation.*

To validate the enhancer motif activities, an enriched and a control group were made. The enriched group consisted of TFs from Figure 11A, which had a log2 ratio value of 1 and above (2-fold enrichment) in each cluster, and the control group consisted of TFs with a log2 ratio value of 0 (1-fold enrichment) for each cell type cluster. Additionally, despite cells not having had a TF overexpressed in them, they can still exhibit an endogenous expression of potentially active TFs, as previously observed in Figure 11D. Therefore, to establish a stringent control set of truly inactive TFs, the control group TFs were filtered to only include TFs with a low normalised expression equal to or below 0.2. Then, the enhancer motif activities for the TFs in each group for selected cell types were extracted, and the median enhancer motif activity was calculated and visualised in Figure 15A. The enhancer motif activity in the enriched group should be higher than that in the control group in each of the selected cell types. It can be seen that IMAGE predicts a better median activity compared to the control group for podocytes and vascular endothelial (VE) like cells,

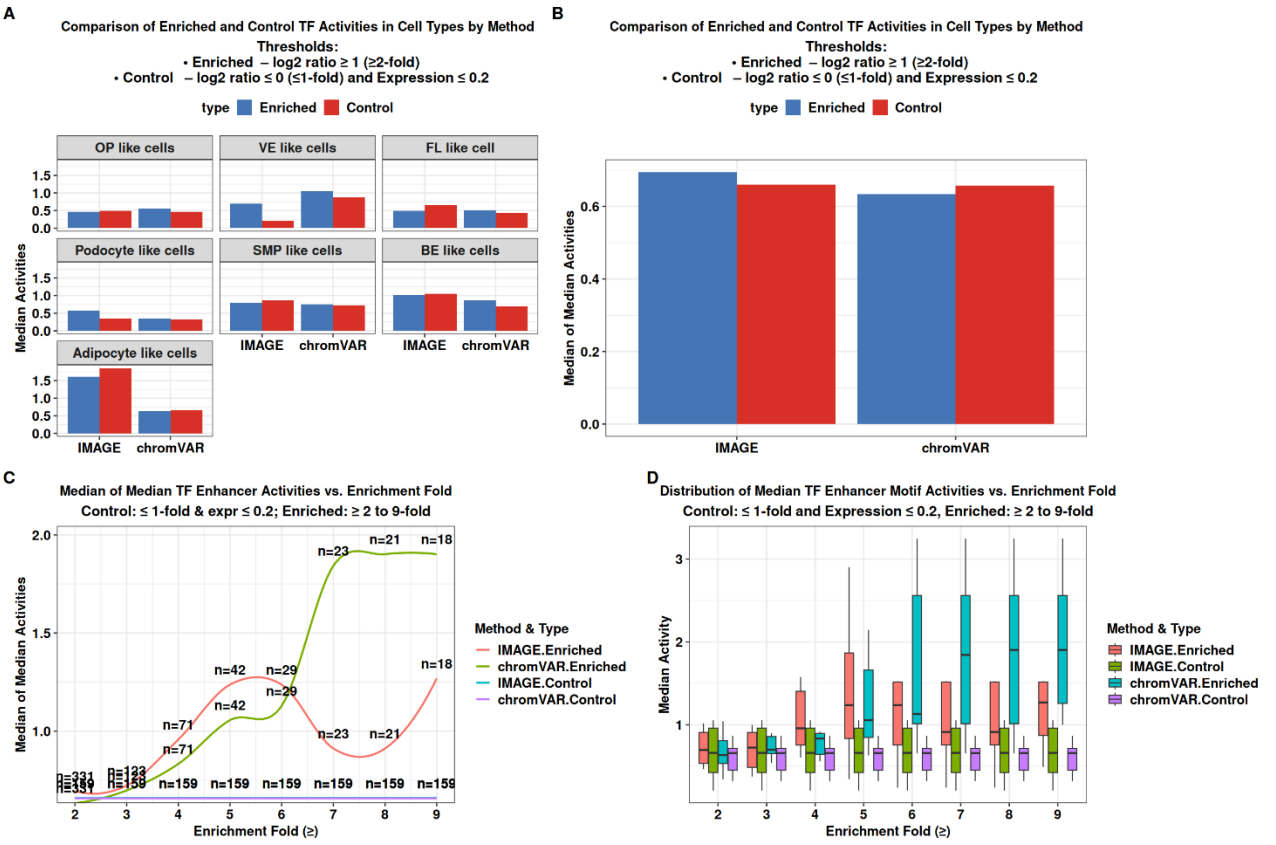


Figure 15. Validation and comparison of IMAGE and chromVAR enhancer motif activities. *(A) Barplot of median z-score enhancer motif activities for selected cell types. Enriched Transcription factors (TFs), defined by having a log2 ratio enrichment of 1 or above fold (≥2-fold), were identified, and their enhancer motif activities were extracted, and the median activity was calculated for each selected cell type. Control transcription factors (TFs) are defined as having a fold enrichment of 0 or below and having a normalised expression equal to or below 0,2. The y-axis shows the median z-score enhancer motif activity values, and the x-axis shows whether the median activity is calculated for IMAGE or chromVAR. The enriched and control TF bars are colour coded according to the legend on top of the barplot. (B) Barplot of median of median enhancer motif activities. The x-axis shows whether the activities were calculated by IMAGE or chromVAR, while the y-axis shows the median of median z-score enhancer motif activities of the cell types from 15A. In other words, the median was calculated for all the median activities of enriched and control TFs across the selected cell types predicted by IMAGE and chromVAR, respectively. Enriched and control TFs are defined as mentioned above. (C) A line plot of the median of median z-score enhancer motif activities across different enrichment thresholds. The x-axis shows an increasing fold enrichment, while the y-axis indicates the median of median z-score enhancer motif activity values. Control TFs are defined as mentioned above, while the enriched TFs are defined by their enrichment thresholds as indicated on the x-axis. The number of TFs (n) included in the analysis for each enrichment threshold is shown on the line plot. Each line is colour-coded according to whether it represents control or enriched TFs, and whether the activities were calculated by IMAGE or chromVAR, as indicated in the legend on the right-hand side of the plot. (D) Boxplot of median z-score enhancer motif activities. The x-axis shows the fold enrichment thresholds, and the y-axis shows the median z-score enhancer motif activities. On the boxplots, the interquartile range is shown as the square, containing the median of the median activities, indicated as a black line across the boxplot, while the first and third quartile marks the beginning and end of the box itself. The minimum and maximum values are found at the lower and upper tail of the boxplots respectively. The enriched TFs are defined by the fold enrichment thresholds indicated on the x-axis and control TFs are defined as above mentioned. The boxplots are colour coded according to whether the TFs are enriched or control and if the activities were predicted by IMAGE or chromVAR.*

whereas chromVAR predicts a better median activity for the other cell types. To get a better visualisation of which method predicts a better median activity across these selected cell types, the median of median activities was calculated and visualised in Figure 15B.

In Figure 15B, IMAGE shows a higher separation between the enriched and control groups in the median of median enhancer motif activities compared to chromVAR, which suggests a greater sensitivity to biologically relevant TF enrichment. However, this is only for a 2-fold enrichment threshold and above. If the threshold is gradually increased from a 2-to-9-fold enrichment, it can be observed that IMAGE performs better than chromVAR up to a 6-fold enrichment, as shown in Figure 15C. However, after the 6-fold threshold, chromVAR performs vastly better, and the enhancer motif activities of IMAGE seem to become unstable or inconsistent. This might be due to how highly correlated variables get pushed towards zero in ridge regression [34]. Therefore, this might possibly reduce and misrepresent the actual influence that the TFs exert on the enhancer activity. In contrast, chromVAR utilises the raw accessibility deviation as a proxy for motif activity and may be a more unbiased and stable method for predicting enhancer motif activities [5].

To assess the robustness of the median of median activities in Figure 15C, the median activity distributions are visualised in Figure 15D. It can be observed that for a 2-to-5 enrichment threshold, the median activities of the cell types are relatively compact for IMAGE and chromVAR, compared to a 6-to-9-fold enrichment threshold, where the median activities are more spread out, which is especially the case for chromVAR. This variability could be a result of a too small sample size, since only 29 to 18 TFs are included in the analysis at those thresholds. This might indicate that the results at lower thresholds are more robust and reliable, in which case IMAGE seem to predict motif activities that reflect the enrichment better. However, it is still noteworthy that at higher enrichment thresholds, chromVAR performs noticeably better by assigning higher activities to those TFs, despite the small sample size, which IMAGE ideally should also be able to do, but potentially can't due to the nature of ridge regression.

A possible way to circumvent the multicollinearity and thereby the distribution of 1/k equally sized coefficients for highly correlated variables in ridge regression could potentially be to decorrelate the variables by transforming the data. This can be achieved with ZCA whitening, which transforms the data and decorrelates the variables completely, while retaining as much of the transformed data as possible, closely approximating the original data. ZCA whitening does not reduce the dimensionality of the data [67]. Then, ridge regression can be run on the transformed data in the whitened space, and the resulting coefficients can be reverse-transformed to obtain completely decorrelated motif activities. An example of this can be seen in Supplementary Figure 18, where it results in a much more stable validation compared to what was observed in Figure 15C. However, the clustering of cell types does not yield the desired result, as shown in Supplementary Figure 18.

This could be because the enhancer motif activity is technically over-normalised, as discussed below.

It should be noted that the results in Figures 13, 14, 15, and onward regarding IMAGE are a product of technically double-normalising the ATAC-seq and RNA-seq data. The EDASeq package used to normalise for GC content utilises both the `EDASeq::withinLaneNormalization` and `EDASeq::betweenLaneNormalization` functions [60]. The `EDASeq::withinLaneNormalization` normalises for GC content, while the `EDASeq::betweenLaneNormalization` normalises for sequencing depth [60]. However, the ATAC-seq sequences are also normalised for sequencing depth by CPM normalisation, which effectively means that the data is normalised twice for sequencing depth. Double normalisation could be thought to result in distorted and misleading biological interpretations, as normalisation changes the scale of the data and therefore the original context of the values is lost due to over-normalisation and data transformation. However, it was only by double normalising the data that the desired clustering of the cell types in Figure 14 for IMAGE was observed. Even when only `EDASeq::withinLaneNormalisation` is performed for differentially accessible regions, the preferred clustering of the cell types is not reflected in the motif activities as seen in Supplementary Figure 19. Additionally, the author of the EDASeq package notes that the pseudocounts generated after running both `EDASeq::withinLaneNormalisation` and `EDASeq::betweenLaneNormalization` still show typical features of a negative binomial distribution [98]. A negative binomial distribution for raw count data is expected, which could potentially imply that CPM normalising the data after `EDASeq::betweenLaneNormalization` might not have been as detrimental as initially thought. Especially considering that the desired clustering of the cell types was reflected in the motif activities as observed in Figure 13C and D.

Interestingly enough, when only `EDASeq::withinLaneNormalisation` is run together with ZCA whitening, the exact desired clustering of the cell types is not observed; however, the validation appears to be the best observed thus far, as it follows a curve resembling a sigmoid function of the TF enrichment in Supplementary Figure 20. It's possible that the information on the expected clustering of cell types based on Figure 10A may be insufficient and not fully capture the true similarity of cell types based on gene expression, as it's only based on around 3-5 marker genes for each cell type. Potentially, by including more marker genes in the analysis, a different pattern of similarity may have been observed for Figure 10A, which hypothetically could resemble the one seen in Supplementary Figure 20. In this case, IMAGE would not only appear to perform better or

on par with chromVAR but also reflect a clustering of cell types based on motif activities at a bulk level compared to the single-cell level. Furthermore, it may be worth updating the model for predicting enhancer motif activities by considering whether the TF is expressed, as it would not make sense for a TF to be active if it is not expressed. This could be done by min-max scaling the gene expression and multiplying it by the normalised motif counts.

Finally, after having validated the enhancer motif activity, the gene motif activity was calculated and validated in a similar fashion. In Figure 16, it is evident that the desired clustering of cell types is also reflected in the predicted gene motif activities. Based on the observation that hESC 1-3, testis/prostate, goblet and podocyte-like cells are also clustered together, the predicted gene motif activities are validated and compared to the predicted enhancer motif activities by IMAGE, as seen in Figure 17A and B.

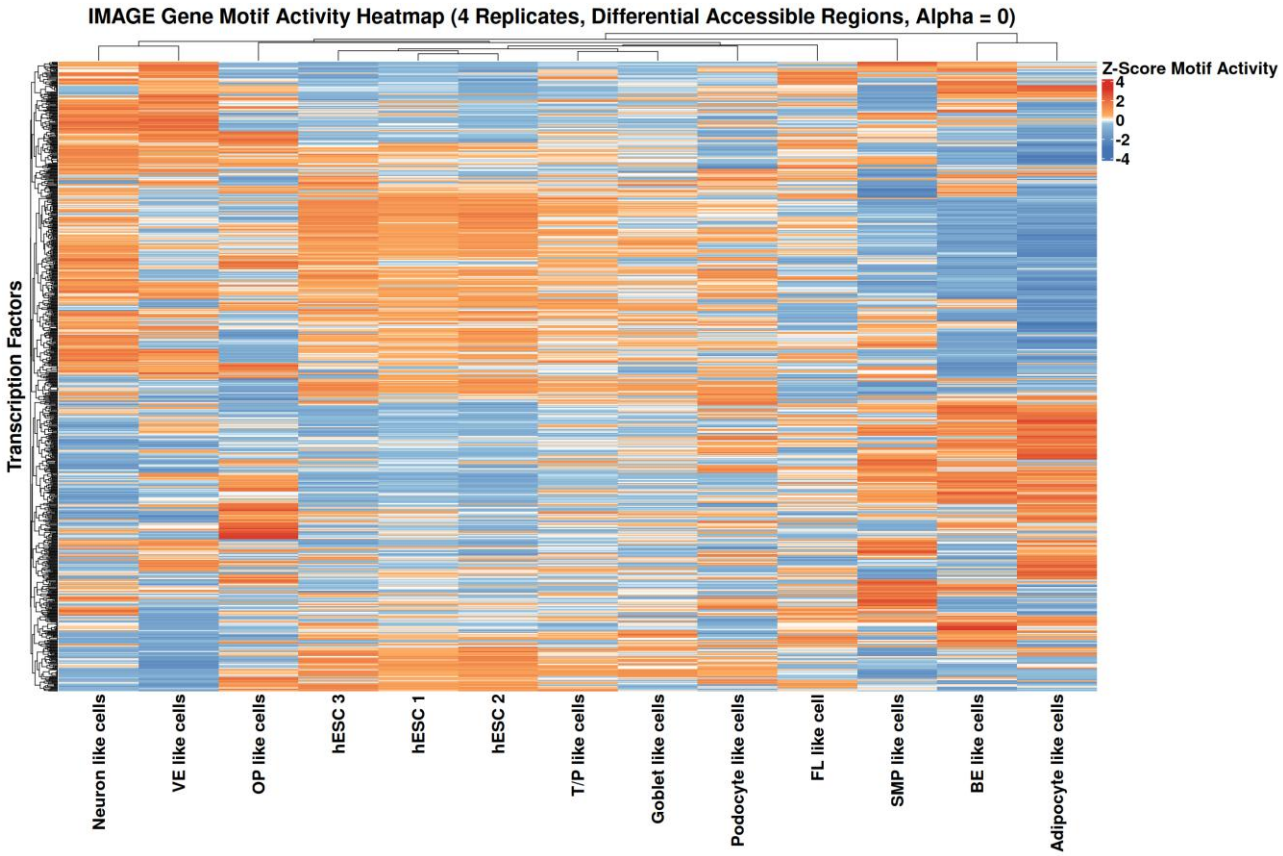


Figure 16. Heatmap of IMAGE estimated gene motif activities. *This heatmap illustrates the predicted gene motif activities of IMAGE for all ATAC-seq peaks, pseudobulked to four replicates per cell type and estimated via ridge regression (alpha = 0). The x-axis denotes the various cell types, while the y-axis lists each transcription factor (TF) (870). The gene motif activity for each TF is colour-coded based on the z-scored motif activity of IMAGE, as shown on the right side of the heatmap. The z-score indicates whether a motif's activity is above or below its average across cell types, with an average activity set at zero. The row dendrograms group the TFs according to their activity across cell types, while the column dendrogram organises the cell types based on their motif activity for each TF.*

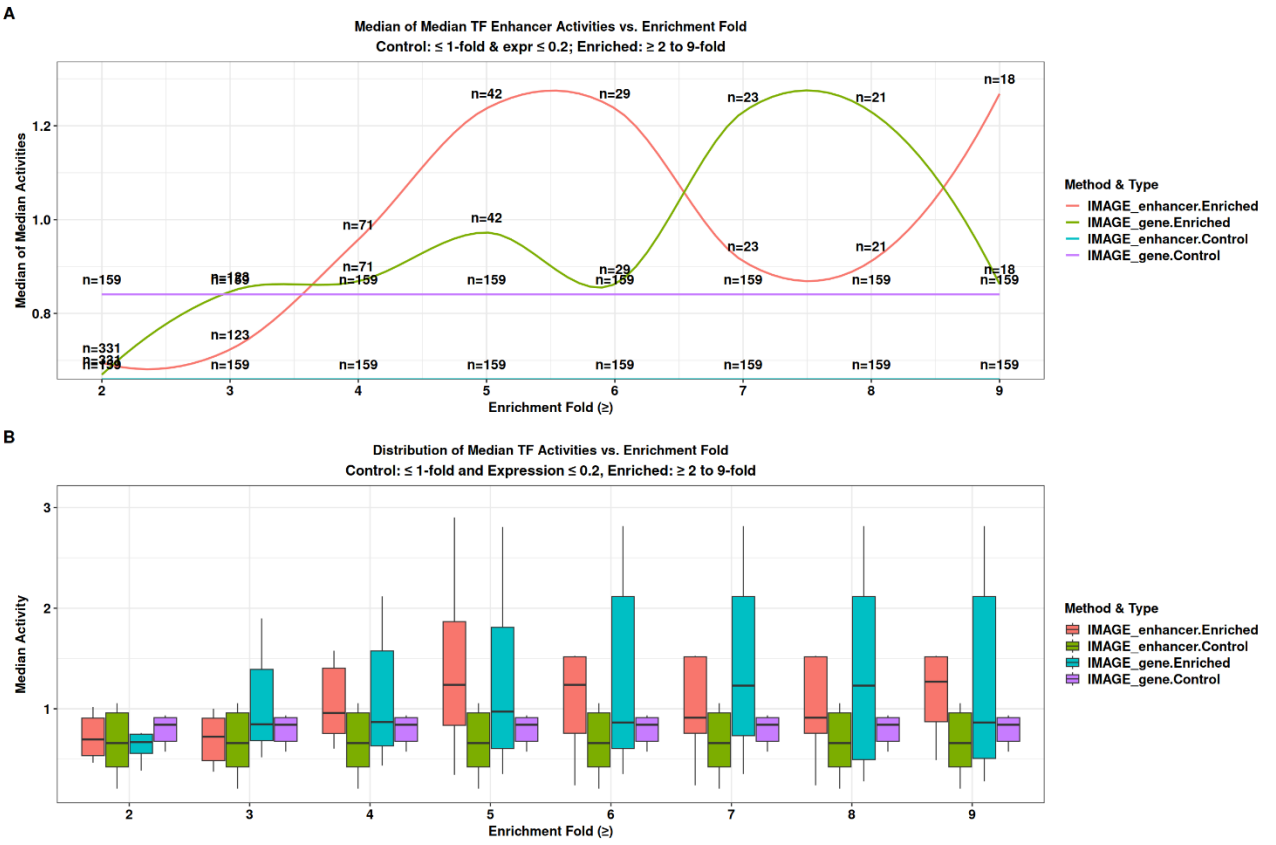


Figure 17. Validation and comparison of IMAGE enhancer and gene motif activities. *(A) A line plot illustrating the median of median z-score enhancer motif activities across varying enrichment thresholds. The x-axis represents increasing fold enrichment, while the y-axis displays the median of median z-score enhancer motif activity values. Control transcription factors (TFs) are defined as having a fold enrichment of 0 or below and having a normalised expression equal to or below 0,2, while enriched TFs are identified by their corresponding enrichment thresholds on the x-axis. The number of TFs (n) analysed for each enrichment threshold is depicted on the line plot. Each line is colour-coded based on whether it signifies control or enriched TFs, as well as whether it reflects the enhancer or gene motif activities predicted by IMAGE, as detailed in the legend on the right side of the plot. (D) Boxplot displaying median z-score enhancer motif activities. The x-axis indicates the fold enrichment thresholds, while the y-axis presents the median z-score enhancer motif activities. In the boxplots, the interquartile range is illustrated as a square, encapsulating the median of the median activities, which is marked by a black line across the boxplot. The first and third quartile values represent the beginning and endpoints of the box itself, while the minimum and maximum values are located at the lower and upper tails of the boxplots, respectively. The enriched TFs are defined by the fold enrichment thresholds shown on the x-axis, whereas control TFs are defined as previously noted. The boxplots are colour-coded to indicate whether the TFs are enriched or control, and if the activities are for enhancer or gene motifs predicted by IMAGE.*

The same thresholds are applied here as in Figures 15C and D. As shown in Figure 17A, not only are the control TFs assigned a much higher activity, but the gene motif activities of the enriched TFs at the different thresholds also appear to be significantly more unstable compared to the enhancer motif activity. This may once again be due to multicollinearity and coefficient shrinkage [34]. Additionally, in Figure 17B, it can also be seen that the median activities are less robust and more spread out for higher thresholds. This could once again be due to the small sample size of TFs at these thresholds. In general, these results suggest that the model for estimating gene motif activities may not be as effective as the model for predicting enhancer motif activities. However, it could also

be due to the double normalisation that the gene motif activities appear unstable and therefore may reflect inaccurate biological interpretations.

Regulatory Behaviour of TFs: From Enhancer Activity to Gene Expression Impact

To characterise the TFs in regard to their enhancer and gene motif activities, the raw activities were investigated. The reason for examining the raw activities instead of z-scores, is because the raw motif activities retain biologically meaningful magnitude and directionality of the TFs, as opposed to z-scores which would distort the magnitude and directionality. However, motif activity z-scores are suitable for validation, as they enable the comparison of relative TF activities across cell types and between tools, as they are standardised to the same scale. The distribution of the raw enhancer and gene motif activities is illustrated in Figures 18A and 18B, respectively.

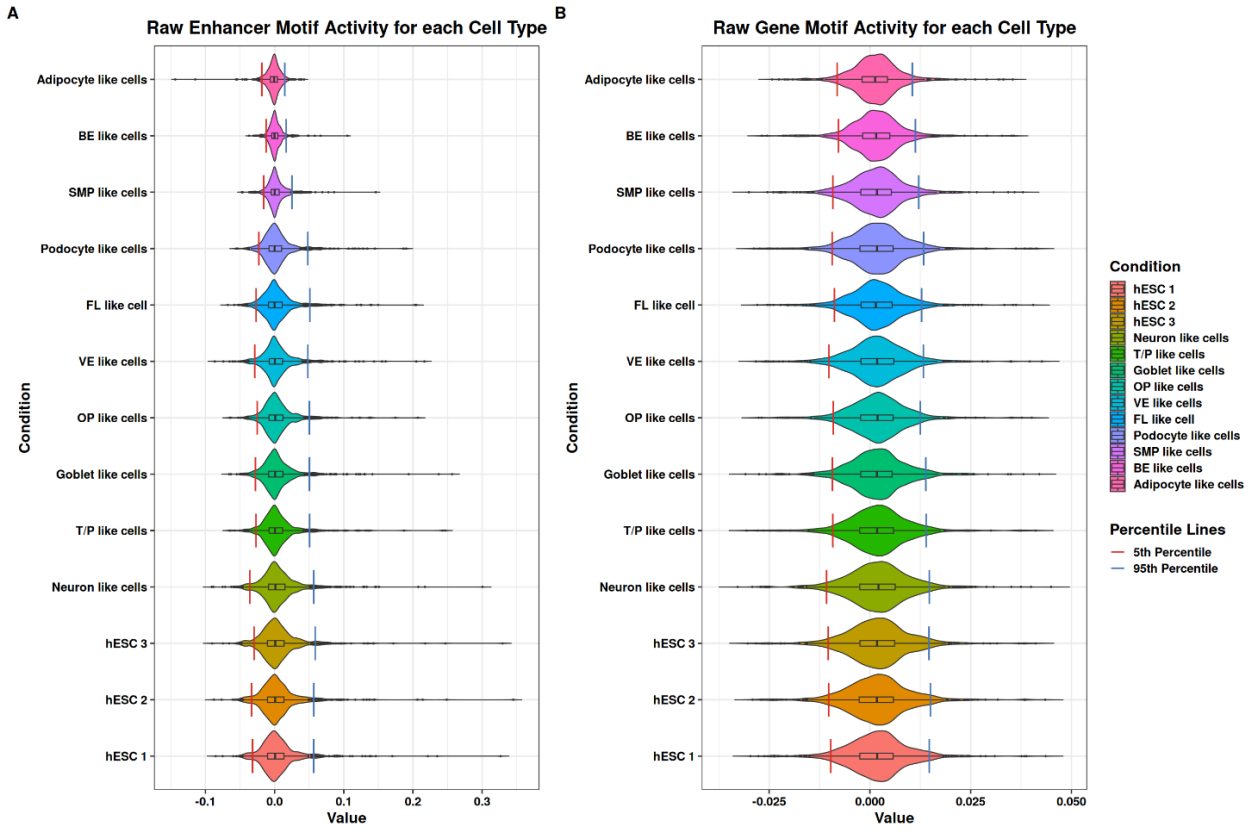


Figure 18. Raw enhancer and gene motif activities by IMAGE. *(A) Violin plot of raw enhancer motif activities, with values represented on the x-axis and cell type or condition shown on the y-axis. Each violin plot contains a boxplot, where the box illustrates the interquartile range, which includes the first quartile, the median, and the third quartile. The upper and lower tails of each boxplot indicate the maximum and minimum values, respectively. The bottom and top 5% of the motif activity values are marked by red and blue lines at the 5th and 95th percentiles for each violin plot. Each violin plot is colour-coded according to its corresponding cell type, as indicated in the legend on the right-hand side of the plot. (B) Violin plot of raw gene motif activities, where the same aforementioned descriptions apply.*

From Figures 18A and B, it can be observed that the values seem quite small, which could indicate highly correlated variables across all cell types. Since all coefficients are closely pushed towards zero, there is a higher likelihood that a TF will fall on either side of zero by coincidence and thus fluctuate between signs. Therefore, to only characterise TFs that are more robust and may reflect better biological interpretability, only the top and bottom 5% of TFs were characterised as illustrated in Figure 18, with the 5th and 95th percentiles indicated with a red and blue cutoff, respectively. Interestingly, the coefficients in Figure 18A appear to increase slightly from top to bottom across the cell types. This appears to be a similar trend to that observed in Figure 6B, where there is a higher cell count in cluster 0 (hESC 1), and it gradually decreases in cell number towards cluster 12 (Adipocyte-like cells). This skewness might arise from the inherently sparse nature of ATAC-seq data combined with a small cell count for the later clusters [31], [32]. The sparsity could be argued to increase the variance in the data, due to a lower signal-to-noise ratio, which ridge regression ultimately has to accommodate for by shrinking the coefficients to make it more generalisable [34], [99]. Notably, the values in 18A are larger in magnitude compared to those in 18B. This might indicate that, in the gene expression data, the variables are still highly correlated; however, they are not as sparse as the ATAC-seq data and therefore do not display the same skewness.

The sign of the top and bottom 5% of TFs is considered across all cell types to characterise them, as seen in Figure 19. It can be observed for TFs in regard to enhancer motif activities, that there is a greater number of TFs that have a dual function as opposed to TFs only being a positive or negative effector. In contrast, there are only two TFs that have a dual function in regard to the gene motif activity, while the rest of the TFs are either positive or negative effectors. An effector, in this case, refers to whether the TF contributes positively or negatively to enhancer activity and gene expression. In other words, it describes how the presence of TF motifs correlates with accessibility and gene expression data, and may generally function as either a repressor or an activator according to the model's predictions.

The contrast observed in Figures 19A and B suggests that the TFs are quite heterogeneous in their contributions to enhancer activity across different cell types. However, they appear to be more consistent in their roles in gene expression. Thus, even though the TFs may or may not facilitate the CRE being active in the presence of the TF motif across cell types, the majority of TFs ultimately serve the same functional role in gene expression regulation across these cell types. This seems quite contrary to current research beliefs, which emphasises that the gene regulatory behaviour of

TFs as activators or repressors can change depending on several factors, such as PTMs, coregulators, and the CRE they bind to, since their functions have been documented to be position-dependent relative to the TSS [100], [101], [102], [103]. Some TF effector domains have even recently been experimentally documented to exhibit bifunctional effects, both in the sense that the same region can activate and repress transcription, but also independent regions within the same effector domain can activate or repress gene expression [104]. Additionally, it has also recently been experimentally examined that multiple effector domains, within a single TF, can in combination exert an activating and repressing effect on a target gene at the same time [105]. However, there are also TFs that are, as of date, purely categorised as activators or repressors, such as PU.1 and ZEB2, respectively [103]. It may be the case that only such TFs are captured in the top and bottom 5% of gene motif activities. Overall, the figure implies that regulatory complexity is more visible at the enhancer level for these TFs. Naturally, this complexity cannot be inferred by looking at the individual cell types in isolation, as seen in Figure 19C and D, but needs to be contextualised with respect to between-cell type comparisons, as illustrated in Figure 19E and F. Evidently, the analysis in Figure 19E and F depends on which cell types are being compared; in this case, it is simply the incrementally added comparison between the cell types in this dataset to illustrate the point.

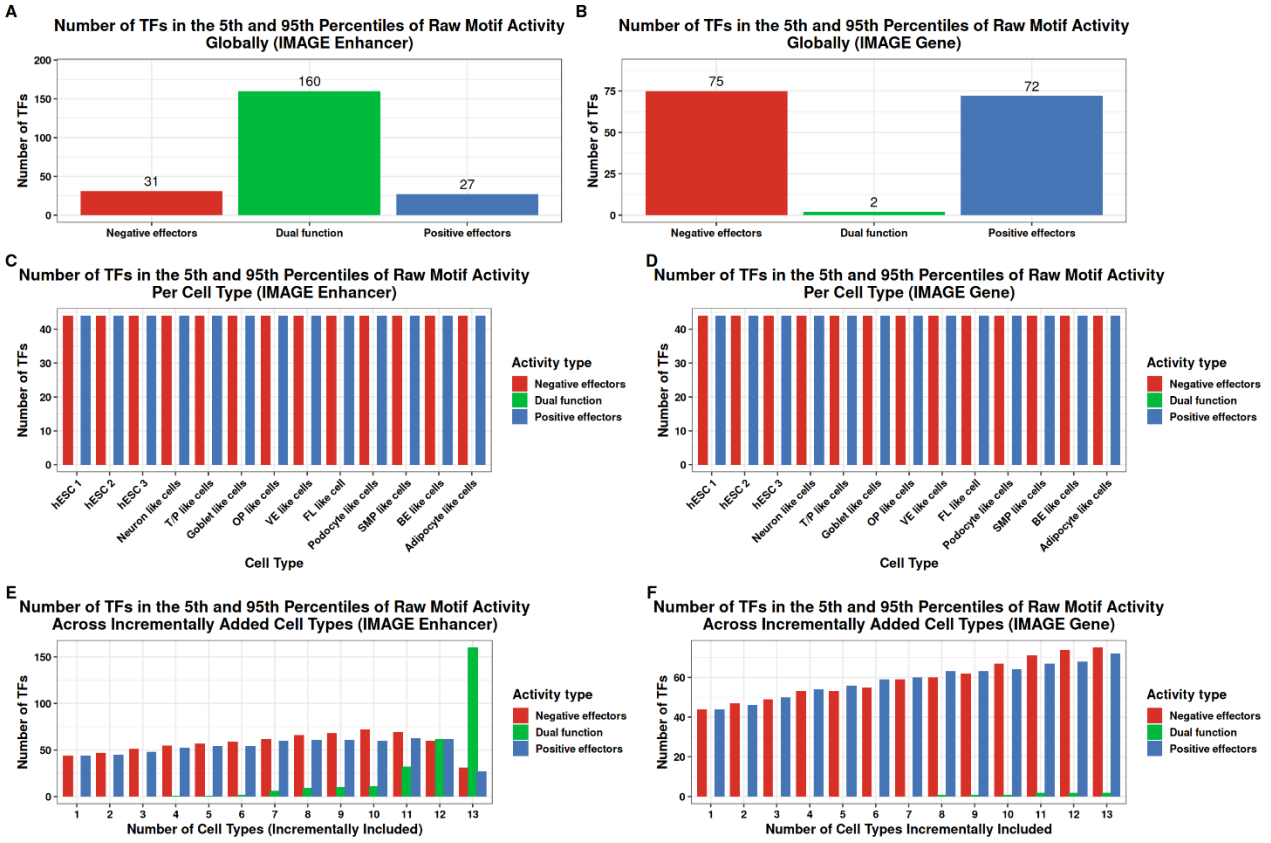


Figure 19. Characterisation of transcription factors (TFs) found the 5th and 95th percentiles of raw enhancer and gene motif activity predicted by IMAGE*. (A) A barplot of the number of TFs characterised as a negative effector, positive effector, or dual functional effector can be seen on the y-axis, with the x-axis representing the respective categories. The number of TFs in each category is displayed above each bar. The characterisation is based on the 5th and 95th percentile raw enhancer motif activity sign across all cell types (globally). If a TF only have a raw enhancer motif activity that is negative, it is characterised as a negative effector, and if the TF only have a positive raw enhancer motif activity across cell types, it is a positive effector. Otherwise, it is a dual-functional effector. (B) A similar barplot where the prior descriptions hold true, but for the gene motif activity and across all cell types (globally). (C) A barplot that shows the number of TFs in the 5th and 95th percentiles of raw enhancer motif activities per cell type. On the y-axis, the number of TFs is depicted, and the cell types are shown on the x-axis. Each bar is colour coded based on whether it has been categorised as a negative, positive or dual functional effector, as indicated on the legend on the right-hand side of the plot. The characterisation of a TF as a negative, positive or dual functional effector is the same as described above, but only within each cell type. (D) A similar barplot to (C), but for the raw gene motif activities. (E) Barplot of the number of TFs in the 5th and 95th percentiles of the raw enhancer motif activities, but across incrementally added cell types. The y-axis represents the number of TFs, and the x-axis depicts the total number of cell types included in the characterisation, which increases incrementally by adding a cell type one at a time. The bars are colour-coded based on whether the TFs are characterised as positive, negative, or dual functional effector, as indicated by the legend. The characterisation is once again based on the sign of the TFs' raw enhancer motif activities across the incrementally added cell types. (F) A barplot similar to (E) but for gene motif activities.*

Furthermore, it can be examined if the signs of the TFs are consistent between the enhancer and gene motif activities. To examine this, the intersection between TFs for both activities was identified, and their signs were assessed, as shown in Figure 20A. It can be observed that it is seemingly TFs that remain as a positive effector between the two motif activities that dominate the analysis, and that there are very few TFs which shift from being a positive effector concerning the enhancer motif activity to being a negative effector in the gene motif activity. A consistency in the sign as a positive effector indicates that when the motif of the TF is present, then the chromatin is

accessible and the gene is expressed. On the contrary, if the sign is consistently negative, then the chromatin is generally inaccessible in the presence of the TF motif, and the gene is not expressed. A consistent negative sign might indicate that the TF might recruit writer or eraser enzymes, such as histone methyltransferases and demethylases, which can alter the chromatin landscape into heterochromatin [18], [106]. Additionally, it may also indicate a coregulator that does not bind to the CRE but instead binds to another TF and represses or inhibits gene expression, thereby acting as a corepressor. In Figure 20B, an example of this could be KLF12, which has been documented to act as a co-repressor that interacts with p53 and downregulates target genes of p53, such as p21 and PUMA [107]. However, KLF12 cannot be classified as a co-repressor by definition, since it does not lack a DBD [107]. The tumour suppressor p53 is activated in response to various stress conditions, including DNA damage and tumorigenesis [108], [109]. In general, p53 is a key regulator of cell cycle, apoptosis, senescence, DNA repair, cell differentiation and angiogenesis (formation of new blood vessels) [109]. It can activate genes, such as p21 and PUMA, which respectively lead to cell cycle arrest and trigger apoptosis [108]. KLF12 enhances proliferation by blocking the interaction between p53 and the co-activator p300. This inhibition prevents the acetylation of p53 by p300, allowing for its ubiquitination and degradation, which consequently reduces the expression of p21 and PUMA [107]. Based on this, it could be speculated that KLF12 serves the purpose of suppressing cell cycle arrest and promoting the proliferation of the hESCs in this dataset.

In contrast, a consistent positive sign might indicate that the TF can also recruit writers in a similar fashion and remodel the chromatin into euchromatin. Naturally, it could also simply be that the chromatin is already accessible, allowing the TF to bind and positively regulate gene expression as an activator. An example of this in Figure 20B is NRF1, which is documented to be important for primarily resistance to oxidative stress, but also adipogenesis, maintenance of normal neuronal cell function, osteoblast differentiation, and embryonic development, since a global disruption of NRF1 leads to embryonic lethality in mice, likely due to an increase in intracellular reactive oxygen species [110]. From this, it is evident that NRF1 is an important multifunctional protein involved in various biological processes, and it promotes transcription of downstream genes by forming a heterodimer complex with a protein called sMAF [110]. Given its documented importance in embryonic development and differentiation, it would make sense to observe it as a positive effector across all cell types in Figure 20B. Examples of NRF1 target genes related to osteoblast

differentiation are Osterix and DSPP, while some other target genes related to antioxidant defence are GCLC, GCLM, GST, HO-1, NQO1, xCT and MT1/2 [110].

If, however, the TF changes from being a positive effector to a negative effector, then it might be that the TF binds to CREs, such as silencers or insulators, but when it then interacts with the PIC, it inhibits gene expression [1]. It can be observed that, e.g. CTCF go from being a positive effector to a negative effector across all but the last two cell types in Figure 20B. CTCF is a somatically expressed TF that can act as an insulator protein [111]. An example of its insulator function is seen in the H19/Igf2 locus. In this locus, the two genes are separated by a region called the imprinting control region (ICR) [111]. On the maternal allele, the ICR region is not methylated, and the CTCF TF can bind to it and insulate an enhancer from interacting with the core promoter for IGF2 and repressing it, but the H19 gene can be transcribed. However, on the paternal allele, the ICR region is methylated, and the CTCF TF cannot bind to it. Consequently, the IGF2 gene can be transcribed while the H19 gene is repressed [111]. Other target genes of CTCF as an insulator are HS4, Rxrb and α-globin [111]. Since experimental observations indicate that CTCF can mediate intra- and inter-chromosomal interactions and facilitate direct communication between promoters and regulatory elements in mouse embryonic stem cells, CTCF is regarded as a key organiser at the global genome level [111].
Interestingly enough, CTCFL (CTCF-like) is the paralog to CTCF created by a duplication event, yet it does not phenocopy CTCF's function as an insulator [112]. Instead, it is documented to regulate pluripotency and testis-specific genes in spermatogenesis, where it acts as an activator for e.g. the testis-specific gene CST [112], [113]. In contrast, this is not observed in Figure 20B, which could indicate an inaccuracy in the model's prediction of gene motif activities, especially considering that the testis/prostate-like cell square is not marked in blue, since CTCFL acts as an activator.

Finally, if the TF transitions from being a negative effector to a positive effector, this may indicate that it functions as a coregulator and thereby does not contribute to accessibility, but rather to gene expression [1]. Seemingly, the only example of this in Figure 20B is TOPORS, which functions as a co-activator [109]. TOPORS is a coactivator of p53, and the target gene of the complex was observed to be p21, MDM2 and Bax [109]. Respectively, Bax is a pro-apoptotic protein that regulates the balance between cellular life and death, while MDM2 serves as a key regulator of p53 by ubiquitinating p53 in a negative feedback loop, thereby targeting p53 for degradation [114], [115]. Overall, it appears that it is possible to identify some TF

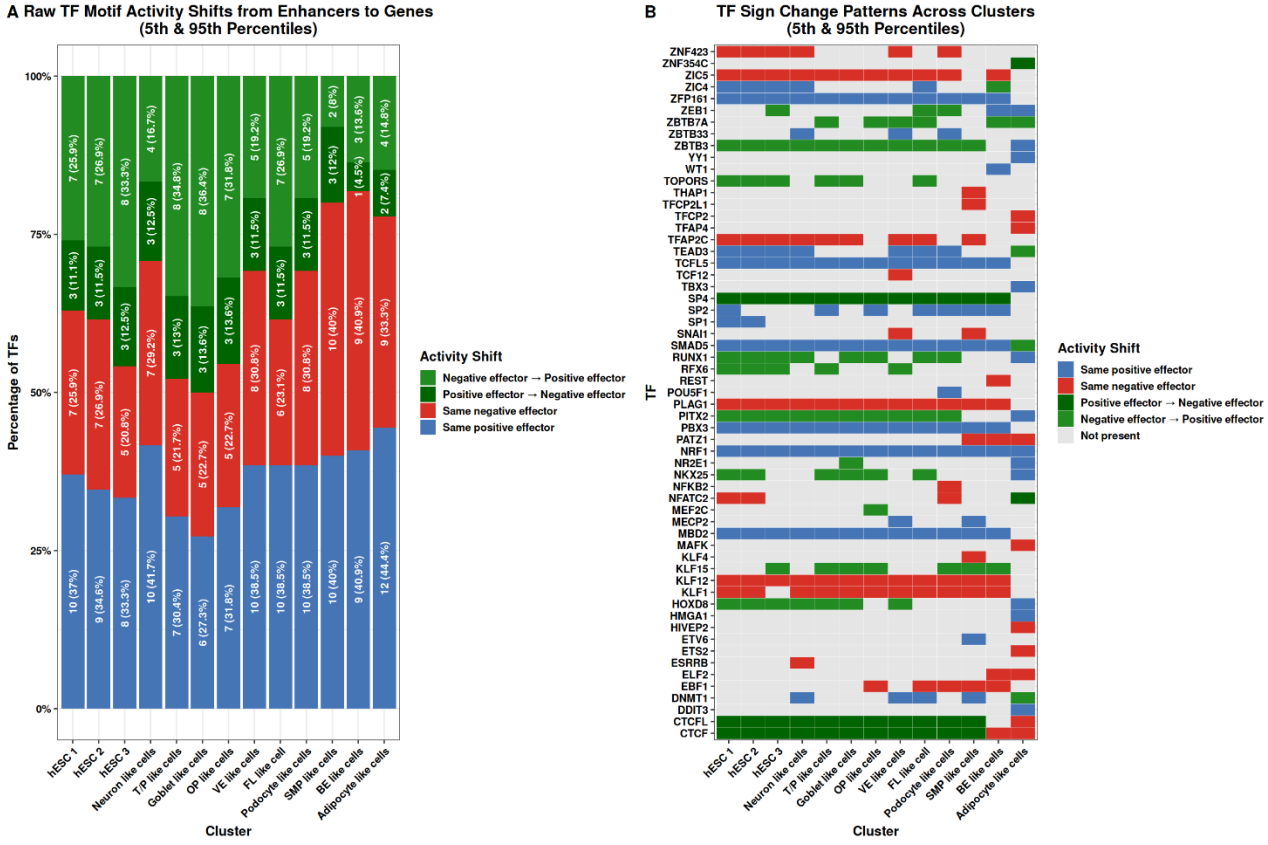


Figure 20. Transcription factor (TF) raw enhancer and gene motif activity sign shift analysis. *(A) A percentage bar graph that displays the percentage of TFs on the y-axis that either shift in sign (+ or -) from a raw enhancer to gene motif activity, or maintain either a positive or negative raw enhancer and gene motif activity. This shift is colour-coded, as indicated in the legend on the right-hand side of the plot. The x-axis represents each cell type. Percentages and absolute numbers are written in white for each of the activity shift categories inside the bars. (B) Heatmap of TF raw enhancer-to-gene motif activity shift sign patterns across cell types. The TFs are identified from the top and bottom 5% of the raw enhancer and gene motif activities. The y-axis shows the TF names, and the x-axis shows the cell type. The activity shift patterns are colour coded according to the legend on the right-hand side of the heatmap.*

where documented literature supports the observed sign changes between enhancer and gene motif activities in Figure 20, which may point towards their regulatory function.

To examine the self-regulatory role of TFs globally in the analysis, the correlation between their gene motif activity and their own expression is visualised in Figure 21. In Figure 21A, three extreme examples of this are illustrated for POUf1, RBPJ and TBX3. It can be observed that the gene motif activity of POUF1 positively correlates with its own gene expression, indicating a potential positive feedback loop and suggesting that POU5F1 may act as an activator of its own gene expression. It has also been recorded that POU5F1 can bind to an enhancer with composite sox-oct elements, which the SOX2-POU5F1 complex can recognise together to regulate their own expression and each other's [116]. In contrast, there appears to be no literature supporting the notion that TF RBPJ regulates its own expression in a negative feedback loop, thereby functioning as a

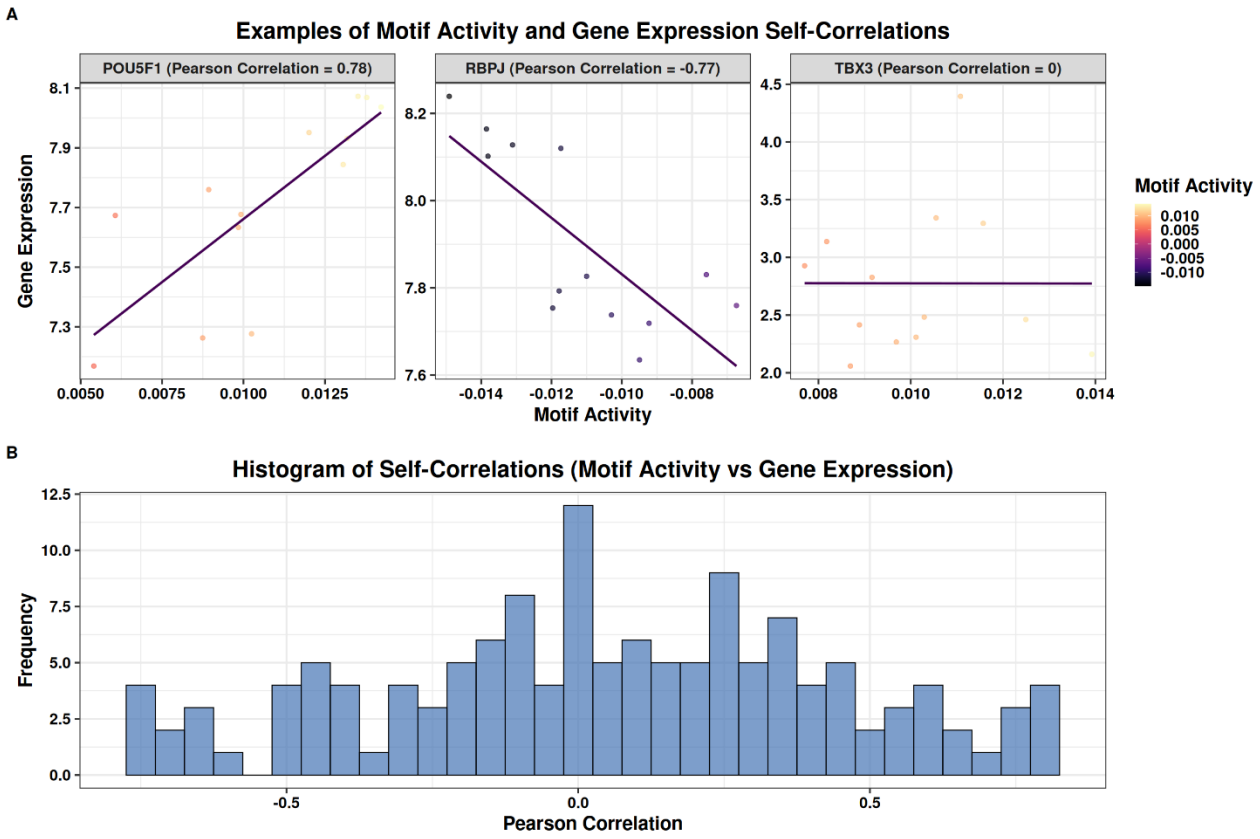


Figure 21. Transcription factor (TF) gene motif activity and self-expression correlation analysis. *(A) Three scatterplots illustrate the correlation between the gene motif activities of POU5F1, RBPJ, and TBX3 and their normalised gene expression. The gene motif activities appear on the x-axis, while the normalised gene expression is displayed on the y-axis. Additionally, the Pearson correlation values are presented above each scatterplot. These values range from -1 to 1, where 1 represents a perfect positive correlation and -1 represents a perfect negative correlation. A value of 0 indicates no correlation. Each dot in the scatterplot represents a cell type, colour-coded by its raw gene motif activity, as shown in the legend on the right-hand side of the scatterplot. (B) A histogram that depicts the Pearson correlation between TF gene motif activity and their own gene expression at a global level. The x-axis represents the Pearson correlation, while the y-axis indicates the frequency of TFs' own gene expression and their predicted gene motif activities.*

repressor. RBPJ is a key mediating TF in the short-range cell-to-cell notch signalling pathway, which is crucial for embryonic development and adult tissue homeostasis [117]. This could be explained by the model's inaccuracy.

If, however, there is neither a strong negative nor a positive correlation between the gene motif activity and the TF's own expression, it might indicate that the motif activity is not predictive of that gene's expression. Alternatively, it is also possible that the TF can regulate its own expression; however, due to post-transcriptional and post-translational regulations, such as small interfering RNA (siRNA), PTMs, or ligands, this analysis may not indicate whether the TF can induce an autoregulatory feedback loop. This does not appear to be the case for TBX3, as there seems to be no documentation indicating that it autoregulates its own expression. TBX3 has also been shown to be

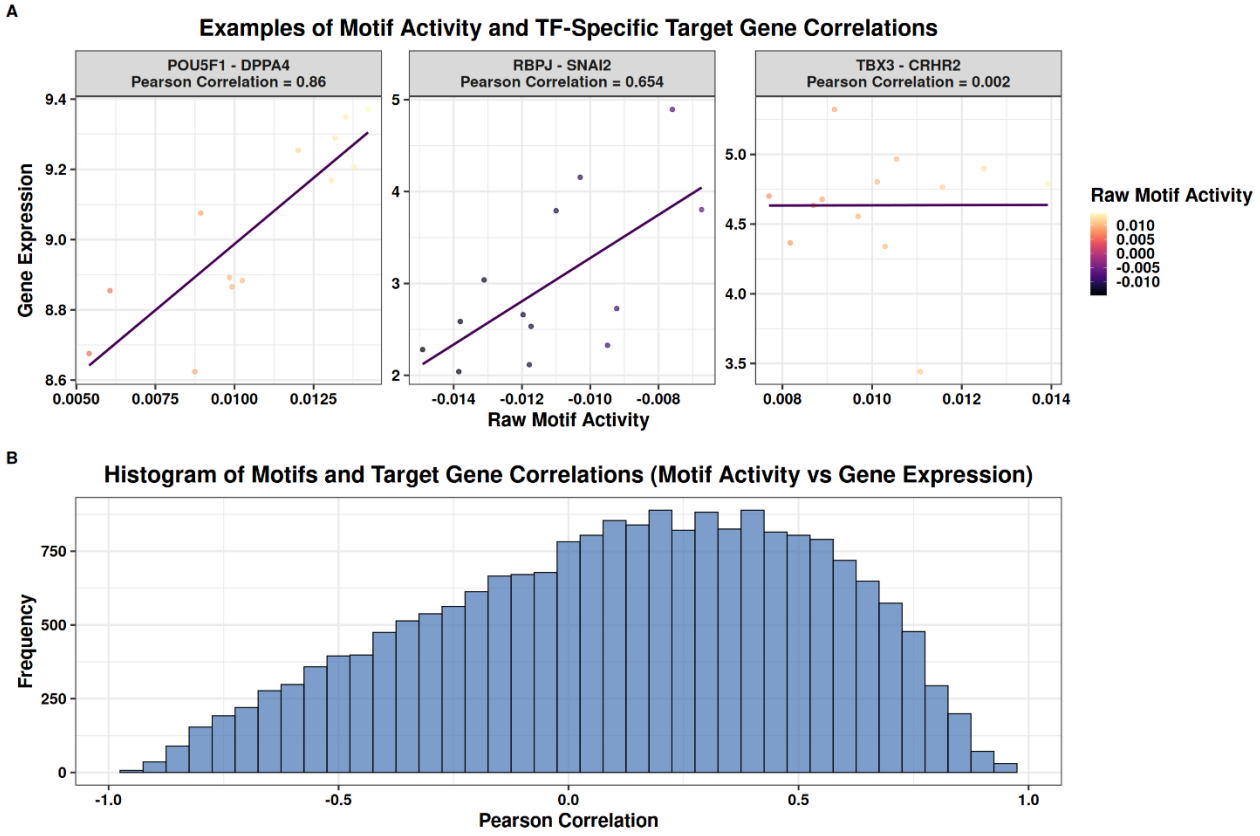


Figure 22. Transcription factor (TF) gene motif activity and predicted target gene expression correlation analysis. *(A) Three scatterplots demonstrate the relationship between the gene motif activities of POU5F1, RBPJ, and TBX3 and the normalised gene expression of their predicted target genes, DPPA4, SNAI2, and CRHR2. The gene motif activities are plotted on the x-axis, with the normalised gene expression shown on the y-axis. Above each scatterplot, the Pearson correlation values, which range from -1 to 1, are indicated. A value of 1 denotes a perfect positive correlation, while a value of -1 signifies a perfect negative correlation. A correlation value of 0 shows no correlation. Each point on the scatterplot represents a cell type, colour-coded according to its raw gene motif activity, as illustrated in the legend on the right side of the scatterplot. (B) A histogram illustrates the Pearson correlation between TF gene motif activity and the gene expression of the TFs' predicted target genes at a global level. The x-axis displays the Pearson correlation, while the y-axis shows the frequency of the TFs' own gene expression along with their predicted gene motif activities.*

crucial for maintaining stem cell pluripotency and self-renewal, as the dynamic expression of TBX3 in ESCs is associated with cell fate decisions [118]. A high TBX3 expression is associated with the maintenance of stem cell pluripotency and self-renewal, while a low expression can predispose cells toward differentiation [118]. Globally, it seems to be the case that the vast majority of TF motifs are not predictive of their expression or that they are potentially post-transcriptional or post-translational regulated, as seen in the histogram in Figure 21B. This is also true for the aforementioned TFs, KLF12, CTCF, NRF1, and TOPORS, which exhibited self-correlation ranging from -0.132 to 0.356; there is likewise no current literature that provides definitive evidence for autoregulation.

To accurately characterise TFs as activators or repressors requires evaluating the connection between gene motif activity and gene expression. After having reviewed seven examples of autoregulation, it can be analysed whether the TFs function as activators or repressors for their predicted target genes, as seen in Figure 22A. In Figure 22A, it can be observed that the predicted gene motif activity of POU5F1 is positively correlated with DPPA4 gene expression, suggesting that POU5F1 is an activator that positively regulates DPPA4, a gene essential for maintaining pluripotency in stem cells and embryogenesis [119]. This is in accordance with the documentation of DPPA4 being a well-known target gene of POU5F1 in embryonic stem cells [119]. There have likewise been documented instances of SNAI2 (also known as SLUG) to be a direct target of RBPJ in the notch signalling pathway [120], [121]. However, CRHR2 has been predicted to be a target gene of TBX3, but there are seemingly no documented instances that confirm CRHR2 as a target gene of TBX3 based on the available literature. Alternatively, if literature could confirm the target gene, a zero correlation might indicate the complexity of gene regulation, which again might involve post-translational events, such as PTMs, and ligands. Globally, the majority of TFs and their target genes seem to be more positively correlated than not in Figure 22B, which suggests that most TFs function as activators. Furthermore, upon closer data inspection, only TOPORS compared to NRF1, CTCF and KLF12 were predicted to be a causal TF and have target genes, but only MDM2 was identified, and it had a gene motif activity and gene expression correlation of 0.197, which does lean towards TOPORS being an activator, but there is seemingly no documentation for MDM2 to be a direct target gene of TOPORS.

Enhancer-Gene Link Validation

Since inaccuracies were observed regarding predicting target genes, the enhancer-gene prediction tool scE2G was used to validate IMAGE's enhancer-gene predictions under the assumption that scE2G is the ground truth. A singular example of enhancers predicted to be linked to the DPPA4 gene is shown in Figure 23A and B.

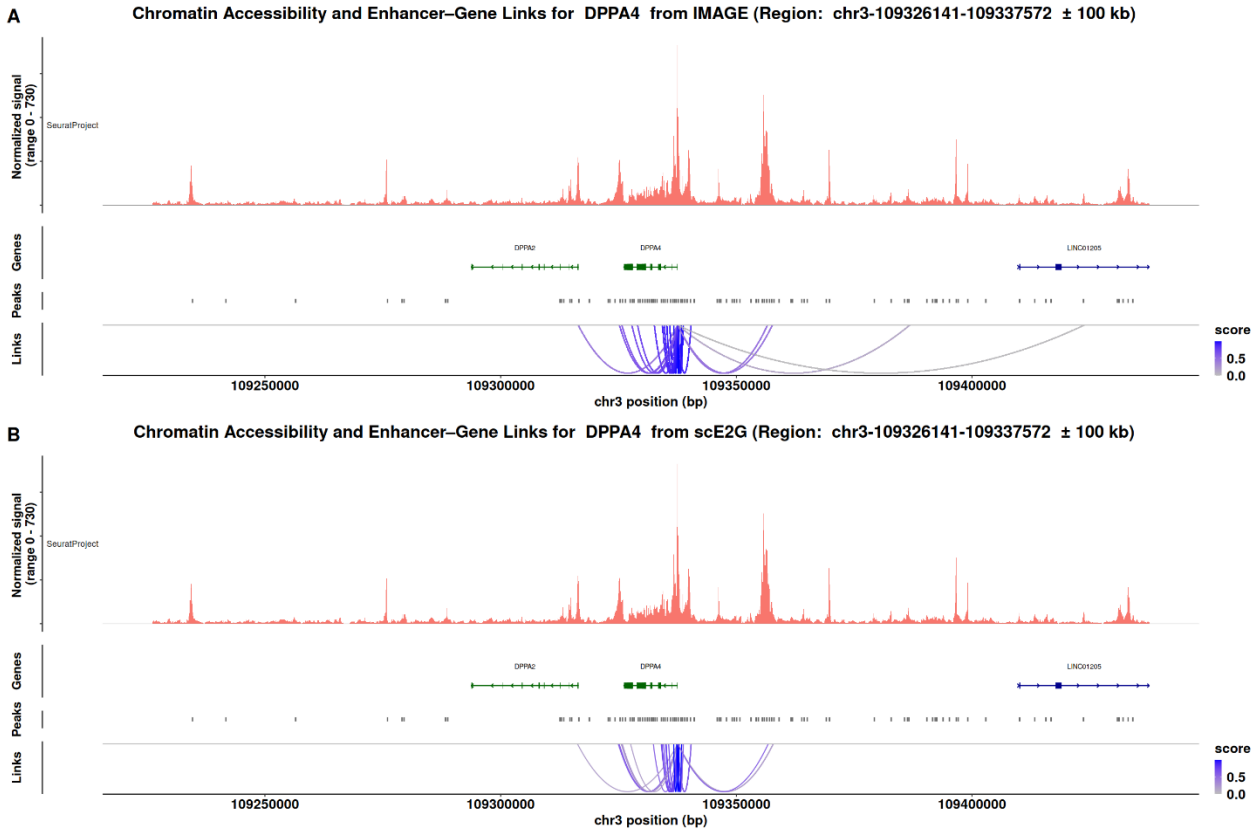


Figure 23. Enhancer-gene links predicted by IMAGE and scE2 for the DPPA4 gene. *(A) A Locus plot depicting links between the DPPA4 gene and its regulatory elements predicted by IMAGE. The links in the links track on the y-axis are coloured based on the Regulatory potential score in equation (9). The peaks in the peaks track are statistically confident peaks identified in the smoothed and library-size–corrected pseudo-bulk view of the raw ATAC signal, as seen in the normalised signal track. The DPPA4 gene is located on the third chromosome (chr3) and is positioned within a region of 109326141 to 109337572 (± 100 kb). In the gene track, exons are represented by square boxes, and introns are represented by connecting lines between the boxes. The arrowheads point toward the direction of transcription and indicate which strand the gene is located on. Arrows pointing to the right indicate a reading direction to the right of the transcription start site (TSS), and that the gene is located on the positive strand (coding strand), while an arrow pointing to the left indicates that the gene is on the negative strand (template strand) and the reading direction is to the left of the TSS. (B) A Locus plot illustrating the connections between the DPPA4 gene and its predicted regulatory elements, as identified by sce2g. The previous description for (A) also applies to this plot, with the distinction that the score used is the scE2G score rather than the regulatory potential score.*

The enhancers linked to DPPA4 predicted by IMAGE are visualised in Figure 23A. The DPPA4 gene is located on chromosome 3 within a region spanning 109326141 to 109337572 (±100 kb). The square boxes in the gene represent exons, while the lines connecting the boxes represent introns. The Arrows indicate the reading direction and the strand on which the gene is located. An arrow pointing to the right indicates that the gene (LINC01205) is located on the coding strand and that the TSS is near the leftmost exon, while an arrow pointing to the left is positioned on the template strand and that the TSS is near the rightmost exon [122]. The normalised signal track in Figure 23A is a smoothed, library-size–corrected pseudo-bulk view of the raw ATAC signal, while squares in the peaks track are the statistically confident peaks called on that signal [123]. The score

in Figure 23A represents the regulatory potential of each enhancer for the gene, as calculated in equation (9), and the link tracks indicate the predicted enhancer-gene links identified by IMAGE [4]. Some enhancers appear to be intergenic, while others are not. When comparing Figure 23A for IMAGE with the enhancer-gene links predicted by scE2G in Figure 23B, it is clear that both tools concur on seemingly the majority of links, while differing on others. Additionally, the colour scoring is different since the score for scE2G represents the scE2G score [7]. To obtain a global assessment of the concurrence between IMAGE and scE2G, a confusion matrix was constructed, as shown in Figure 24A.

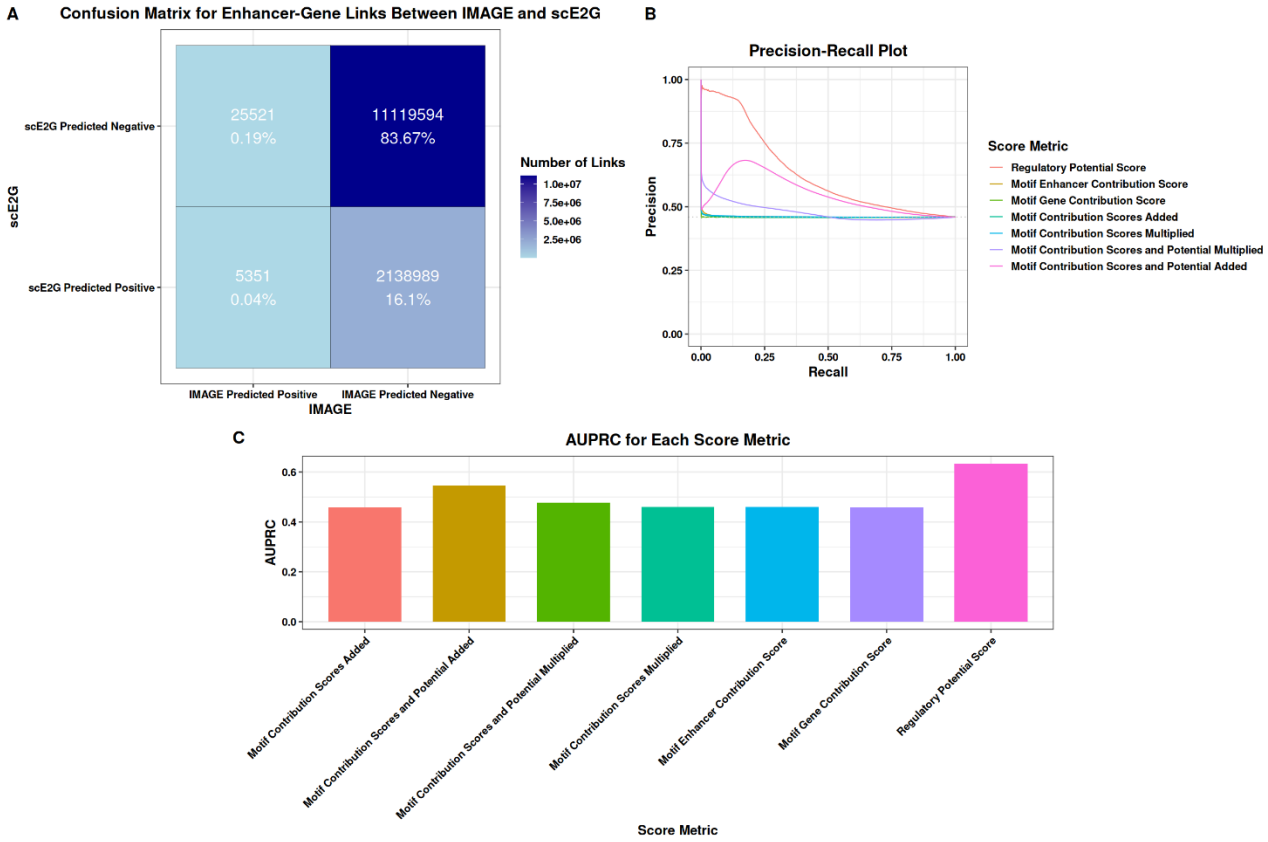


Figure 24. Validation of IMAGE enhancer-gene link predictions. *(A) Confusion matrix of enhancer-gene links between IMAGE and scE2G. The x-axis displays the positive and negative predicted enhancer gene links identified by IMAGE, and the y-axis shows the positive and negative predicted enhancer gene links determined by scE2G. The values inside the confusion matrix show both the absolute and relative number of links found in the intersection between the two prediction tool categories. The squares for each intersection are colour coded according to the number of enhancer-gene links in each intersection, as illustrated in the legend. (B) A Precision recall plot of IMAGE score metrics. The x-axis represents recall, and the y-axis represents precision. Each line represents how well the score metrics of IMAGE are capable of retrieving the binarised predictions. The metrics include the enhancer and gene motif contribution scores, as well as the regulatory potential score, as shown in Equations 8 and 9, respectively. Composite metrics of the three are likewise included, where they have been either added or multiplied together. They are colour coded according to the legend on the right-hand side of the plot. (C) Shows a barplot of the area under the precision-recall curve (AUPRC) for each of the aforementioned metrics and composite metrics. The AUPRC on the y-axis is a unitless metric, and on the x-axis, the metrics are showcased.*

From the confusion matrix, it can be seen that both IMAGE and scE2G agree that the vast majority (83%) of potential enhancer-gene links are non-significant. It does, however, seem like under the assumption that scE2G is the ground truth, IMAGE misses out on a lot of predicted positive links since it only catches 5351 (0,04%) of them, while there are still 2138989 (16%) of them that are predicted to be non-significant links but actually are under the assumption that scE2G is the ground truth. Additionally, there are also 25521 links that IMAGE actually does predict to be significant, but aren't according to scE2G. This strongly suggests that the existing method for defining target enhancers and genes in IMAGE is insufficient to accurately reflect the regulatory interactions between enhancers and genes. It was examined whether any current metrics in IMAGE, used for identifying target enhancers, genes, or their regulatory potential, can more effectively retrieve the binarised predicted enhancer-gene relationships in the Recall-Precision plot as seen in Figure 24B. Here, the enhancer and gene motif contribution scores, as calculated in equation (8), the regulatory potential in equation (9), and the composite metrics of the three are held up against each other to examine whether they are better for retrieving the predictions. Recall is the proportion of real positive cases that are correctly predicted as positive, while precision is the proportion of the number of predicted positives that are actually true positives [65]. This indicates that if the metric successfully retrieves all the positive predicted links first in the correct order, followed by the negative links, the line extends from a y-axis value of 1 to the end of the x-axis. Therefore, maximising the height of the line curve is desirable. It can be observed in Figure 24B that the regulatory potential retrieves the most predictions of scE2G's enhancer gene links. This is followed by the regulatory potential and the two contribution scores added together. This is additionally confirmed in Figure 24C, where the area under the precision-recall curves is shown. As mentioned earlier, it is desired to have the highest curve, and therefore, the most area under the curve. A potential reason why the regulatory potential is better at retrieving the predicted scE2G links might be that the contact between element E and gene G is estimated with a power law decay function, which in its simplest form resembles an exponential function that plateaus at a certain point, whereas the regulatory potential function is merely an exponential function. In that sense, the distances are similar yet distinct, and there are additional features incorporated into the scE2G model that account for genomic complexity, which IMAGE does not accommodate. Because of this, the regulatory potential score cannot fully retrieve the scE2G predictions. Furthermore, adding or multiplying the regulatory potential score by the contribution scores appears to introduce added noise to the regulatory potential score, effectively decreasing the area under the curve.

Conclusion

In conclusion, all the cells still strongly resemble stem cells and are not fully differentiated after four and seven days of being transduced with plasmids encoding TF ORFs. Based on the similarity in gene expression of selected marker genes between the cell types, a desired clustering of the cell types was anticipated, as the clustering should be reflected in the motif activities. The estimated TF motif activities by chromVAR did reflect the expected clustering of the cell types; however, it was only after identifying differentially accessible peaks for feature selection that IMAGE was able to predict motif activities that likewise reflected the expected clustering of the cell types. These peaks were used for the downstream analysis of IMAGE and for validation. During validation, it became evident that chromVAR could predict enhancer motif activities that increased gradually in line with the TF fold enrichment, unlike the current version of IMAGE, which also exhibited a gradual increase in enhancer motif activities relative to the TF fold enrichment but became unstable at higher fold enrichments. This instability might have arisen from the ATAC-seq data being technically over-normalised. Nevertheless, this was the only method that yielded the desired clustering of the cell types, such that this clustering was reflected in the motif activities. Alternatively, it could also have been due to the nature of ridge regression, which shrinks the coefficients of highly correlated variables towards zero when compared to whitened enhancer motif activities. Similarly, the gene motif activity also appeared to be visibly unstable, which could be due to the aforementioned reasons.

When characterising TFs based on changing signs (+ or -) in the top and bottom 5% of enhancer and gene motif activities, it was observed that some TFs, such as KLF12, NRF1, CTCF, and TOPORS, exhibited varying patterns in their enhancer and gene motif activities. These patterns may reflect regulatory roles such as corepressor, activator, silencer, and coactivator, which have also been documented in instances where that is true. However, to accurately characterise TFs as activators or repressors based on motif activity predictions, it may be more informative to consider the correlation between gene motif activities and the expression of target genes. For some TFs, such as POU5F1, there is documentation indicating that it both autoregulates its own expression and acts as an activator, positively regulating the expression of genes like DPPA4. However, there are also instances where IMAGE predicted target genes of TFs that the current literature seemingly does not support, such as TBX3 regulating CRHR2. Because of this, IMAGE was compared to scE2G, which was assumed to function as the ground truth. The comparison revealed that IMAGE overlooked numerous assumed-to-be-positive enhancer-gene links.

Overall, considering that IMAGE does not accurately reflect the increase in TF enhancer and gene motif activities in alignment with an increase in TF enrichment, potentially due to multicollinearity as backed up by Supplementary Figure 18, and that IMAGE fails to predict accurate target enhancer-gene links based on the assumption that scE2G is the ground truth, it appears that the current implementation of IMAGE is not capable of reliably identifying transcription factor and gene regulatory networks.

Future Perspectives

For future use of a dataset similar to the one used in this study by Joung et al 2023. It would potentially be beneficial to allow the cells to differentiate until the 21-day mark to avoid ambiguous cellular identities and to get a more distinct chromatin landscape, gene expression and motif activities as well.

As concluded, the current implementation of IMAGE does not appear to be adequate enough to reliably estimate motif activities and predict target genes, which primarily seems to stem from the prevailing issue of multicollinearity, over-normalised data and a definition of target enhancers and genes that does not properly capture the assumed to be true positively enhancer genes links when compared to scE2G. However, this does not mean that linear models cannot be used to accurately predict motif activities or predict enhancer-gene links, as demonstrated with scE2G. In the future, it would be interesting to update the current enhancer motif activity model by multiplying the enhancer motif activity with min-max normalised gene expression, such that if the gene expression of the TF is zero, then the motif activity of that sample likewise becomes zero, since it would not make sense that the motif is active if it is not even expressed. Additionally, it would be interesting to include more marker genes to see if the desired clustering of cell types changes and matches the whitened, not double-normalised, enhancer motif activity clustering better. Finally, exploring additional methods to circumvent multicollinearity would be advantageous, and the IMAGE model for predicting target genes would need to be reworked, as it appears not to capture the assumed true enhancer-gene regulatory relationships.

Bibliography

[1] F. Spitz and E. E. M. Furlong, "Transcription factors: from enhancer binding to developmental control," *Nat Rev Genet*, vol. 13, no. 9, pp. 613–626, 2012, doi: 10.1038/nrg3207.

[2] P. Badia-i-Mompel *et al.*, "Gene regulatory network inference in the era of single-cell multi-omics," *Nat Rev Genet*, vol. 24, no. 11, pp. 739–754, 2023, doi: 10.1038/s41576-023-00618-5.

[3] X. Jiang and X. Zhang, "RSNET: inferring gene regulatory networks by a redundancy silencing and network enhancement technique," *BMC Bioinformatics*, vol. 23, no. 1, p. 165, 2022, doi: 10.1186/s12859-022-04696-w.

[4] J. G. S. Madsen, A. Rauch, E. L. Van Hauwaert, S. F. Schmidt, M. Winnefeld, and S. Mandrup, "Integrated analysis of motif activity and gene expression changes of transcription factors," *Genome Res*, vol. 28, no. 2, pp. 243–255, Feb. 2018, doi: 10.1101/gr.227231.117.

[5] A. N. Schep, B. Wu, J. D. Buenrostro, and W. J. Greenleaf, "ChromVAR: Inferring transcription-factor-associated accessibility from single-cell epigenomic data," *Nat Methods*, vol. 14, no. 10, pp. 975–978, Oct. 2017, doi: 10.1038/nmeth.4401.

[6] J. Joung *et al.*, "A transcription factor atlas of directed differentiation," *Cell*, vol. 186, no. 1, pp. 209-229.e26, Jan. 2023, doi: 10.1016/j.cell.2022.11.026.

[7] M. U. Sheth *et al.*, "Mapping enhancer-gene regulatory interactions from single-cell data," Nov. 24, 2024. doi: 10.1101/2024.11.23.624931.

[8] Y. H. Che, H. Lee, and Y. J. Kim, "New insights into the epitranscriptomic control of pluripotent stem cell fate," *Exp Mol Med*, vol. 54, no. 10, pp. 1643–1651, 2022, doi: 10.1038/s12276-022-00824-x.

[9] S. Mitalipov and D. Wolf, "Totipotency, Pluripotency and Nuclear Reprogramming," in *Engineering of Stem Cells*, U. Martin, Ed., Berlin, Heidelberg: Springer Berlin Heidelberg, 2009, pp. 185–199. doi: 10.1007/10_2008_45.

[10] T. Reya, S. J. Morrison, M. F. Clarke, and I. L. Weissman, "Stem cells, cancer, and cancer stem cells," *Nature*, vol. 414, no. 6859, pp. 105–111, 2001, doi: 10.1038/35102167.

[11] O. Morrison and J. Thakur, "Molecular complexes at euchromatin, heterochromatin and centromeric chromatin," *Int J Mol Sci*, vol. 22, no. 13, Jul. 2021, doi: 10.3390/ijms22136922.

[12] S. Malik and R. G. Roeder, "Regulation of the RNA polymerase II pre-initiation complex by its associated coactivators," *Nat Rev Genet*, vol. 24, no. 11, pp. 767–782, 2023, doi: 10.1038/s41576-023-00630-9.

[13] L. Mincarelli, A. Lister, J. Lipscombe, and I. C. Macaulay, "Defining Cell Identity with Single-Cell Omics," Sep. 01, 2018, *Wiley-VCH Verlag*. doi: 10.1002/pmic.201700312.

[14] R. Stadhouders, G. J. Filion, and T. Graf, "Transcription factors and 3D genome conformation in cell-fate decisions," May 16, 2019, *Nature Publishing Group*. doi: 10.1038/s41586-019-1182-7.

[15] G. Millán-Zambrano, A. Burton, A. J. Bannister, and R. Schneider, "Histone post-translational modifications — cause and consequence of genome function," Sep. 01, 2022, *Nature Research*. doi: 10.1038/s41576-022-00468-7.

[16] Y. Lorch, R. D. Kornberg, and B. Maier-Davis, "Role of the histone tails in histone octamer transfer," *Nucleic Acids Res*, vol. 51, no. 8, pp. 3671–3678, May 2023, doi: 10.1093/nar/gkad079.

[17] G. Millán-Zambrano, A. Burton, A. J. Bannister, and R. Schneider, "Histone post-translational modifications — cause and consequence of genome function," Sep. 01, 2022, *Nature Research*. doi: 10.1038/s41576-022-00468-7.

[18] A. J. Bannister and T. Kouzarides, "Regulation of chromatin by histone modifications," Mar. 2011. doi: 10.1038/cr.2011.22.

[19] J. L. Miller and P. A. Grant, "The role of DNA methylation and histone modifications in transcriptional regulation in humans," *Subcell Biochem*, vol. 61, pp. 289–317, May 2013, doi: 10.1007/978-94-007-4525-4_13.

[20] B. Lenhard, A. Sandelin, and P. Carninci, "Metazoan promoters: Emerging characteristics and insights into transcriptional regulation," Apr. 2012. doi: 10.1038/nrg3163.

[21] J. H. Yang and A. S. Hansen, "Enhancer selectivity in space and time: from enhancer–promoter interactions to promoter activation," Jul. 01, 2024, *Nature Research*. doi: 10.1038/s41580-024-00710-6.

[22] L. F. Soto *et al.*, "Compendium of human transcription factor effector domains," Feb. 03, 2022, *Cell Press*. doi: 10.1016/j.molcel.2021.11.007.

[23] Z. , G. M. & S. M. Wang, "RNA-Seq: a revolutionary tool for transcriptomics," *Nat Rev Genet 10, 57–63 (2009).*, Jan. 2009, doi: https://doi.org/10.1038/nrg2484.

[24] P. J. Park, "ChIP-seq: Advantages and challenges of a maturing technology," Oct. 2009. doi: 10.1038/nrg2641.

[25] H. E. Pratt *et al.*, "Factorbook: An updated catalog of transcription factor motifs and candidate regulatory motif sites," *Nucleic Acids Res*, vol. 50, no. D1, pp. D141–D149, Jan. 2022, doi: 10.1093/nar/gkab1039.

[26] D. Kim, A. Tran, H. J. Kim, Y. Lin, J. Y. H. Yang, and P. Yang, "Gene regulatory network reconstruction: harnessing the power of single-cell multi-omic data," Dec. 01, 2023, *Nature Research*. doi: 10.1038/s41540-023-00312-6.

[27] C. Rudin, "Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead," May 01, 2019, *Nature Research*. doi: 10.1038/s42256-019-0048-x.

[28] Y. Lecun, Y. Bengio, and G. Hinton, "Deep learning," May 27, 2015, *Nature Publishing Group*. doi: 10.1038/nature14539.

[29] D. Hecker *et al.*, "Computational tools for inferring transcription factor activity," Dec. 01, 2023, *John Wiley and Sons Inc*. doi: 10.1002/pmic.202200462.

[30] F. C. Grandi, H. Modi, L. Kampman, and M. R. Corces, "Chromatin accessibility profiling by ATAC-seq," Jun. 01, 2022, *Nature Research*. doi: 10.1038/s41596-022-00692-9.

[31] H. Chen *et al.*, "Assessment of computational methods for the analysis of single-cell ATAC-seq data," *Genome Biol*, vol. 20, no. 1, 2019, doi: 10.1186/s13059-019-1854-5.

[32] S. Persad *et al.*, "SEACells infers transcriptional and epigenomic cellular states from single-cell genomics data," *Nat Biotechnol*, vol. 41, no. 12, pp. 1746–1757, Dec. 2023, doi: 10.1038/s41587-023-01716-9.

[33] J. Y. Le Chan *et al.*, "Mitigating the Multicollinearity Problem and Its Machine Learning Approach: A Review," Apr. 01, 2022, *MDPI*. doi: 10.3390/math10081283.

[34] J. Friedman, T. Hastie, and R. Tibshirani, "Regularization Paths for Generalized Linear Models via Coordinate Descent."

[35] G. James, D. Witten, T. Hastie, and R. Tibshirani, "An Introduction to Statistical Learning with Applications in R Second Edition," 2023.

[36] J. K. Tay, B. Narasimhan, and T. Hastie, "Elastic Net Regularization Paths for All Generalized Linear Models," *J Stat Softw*, vol. 106, 2023, doi: 10.18637/jss.v106.i01.

[37] B. Boehmke and B. M. Greenwell, *Hands-On Machine Learning with R*, 1st ed. Chapman & Hall/CRC, 2020. doi: 10.1201/9780367816377.

[38] Q. Tang *et al.*, "A comprehensive view of nuclear receptor cancer cistromes," *Cancer Res*, vol. 71, no. 22, pp. 6940–6947, Nov. 2011, doi: 10.1158/0008-5472.CAN-11-2091.

[39] Y. Benjamini and T. P. Speed, "Summarizing and correcting the GC content bias in high-throughput sequencing," *Nucleic Acids Res*, vol. 40, no. 10, May 2012, doi: 10.1093/nar/gks001.

[40] C. P. Fulco *et al.*, "Activity-by-contact model of enhancer–promoter regulation from thousands of CRISPR perturbations," Dec. 01, 2019, *Nature Research*. doi: 10.1038/s41588-019-0538-0.

[41] S. Ma *et al.*, "Chromatin Potential Identified by Shared Single-Cell Profiling of RNA and Chromatin," *Cell*, vol. 183, no. 4, pp. 1103-1116.e20, Nov. 2020, doi: 10.1016/j.cell.2020.09.056.

[42] S. Bhattacharya, Q. Zhang, and M. E. Andersen, "A deterministic map of Waddington's epigenetic landscape for cell fate specification," *BMC Syst Biol*, vol. 5, May 2011, doi: 10.1186/1752-0509-5-85.

[43] Y. Hao *et al.*, "Integrated analysis of multimodal single-cell data," *Cell*, vol. 184, no. 13, pp. 3573-3587.e29, Jun. 2021, doi: 10.1016/j.cell.2021.04.048.

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

[45] C. Soneson and M. D. Robinson, "Bias, robustness and scalability in single-cell differential expression analysis," *Nat Methods*, vol. 15, no. 4, pp. 255–261, Apr. 2018, doi: 10.1038/nmeth.4612.

[46] R. 2004 Shier, "Statistics: 2.3 The Mann-Whitney U Test."

[47] "The Corsini Encyclopedia of Psychology - 2010 - McKnight - Mann-Whitney U Test".

[48] C. Thorn. Ekstrøm and H. Sørensen, *Introduction to statistical data analysis for the life sciences /*, 2. ed. Boca Raton, Fla: CRC Press, 2015.

[49] J. L. Rodgers and ; W Alan Nicewander, "Thirteen Ways to Look at the Correlation Coefficient," 1988.

[50] I. Tirosh *et al.*, "Dissecting the multicellular ecosystem of metastatic melanoma by single-cell RNA-seq," *Science (1979)*, vol. 352, no. 6282, pp. 189–196, Apr. 2016, doi: 10.1126/science.aad0501.

[51] M. Andreatta and S. J. Carmona, "UCell: Robust and scalable single-cell gene signature scoring," *Comput Struct Biotechnol J*, vol. 19, pp. 3796–3798, Jan. 2021, doi: 10.1016/j.csbj.2021.06.043.

[52] Z. Xie *et al.*, "Gene Set Knowledge Discovery with Enrichr," *Curr Protoc*, vol. 1, no. 3, Mar. 2021, doi: 10.1002/cpz1.90.

[53] S. Carbon *et al.*, "The Gene Ontology Resource: 20 years and still GOing strong," *Nucleic Acids Res*, vol. 47, no. D1, pp. D330–D338, Jan. 2019, doi: 10.1093/nar/gky1055.

[54] S. Carbon *et al.*, "Expansion of the gene ontology knowledgebase and resources: The gene ontology consortium," *Nucleic Acids Res*, vol. 45, no. D1, pp. D331–D338, Jan. 2017, doi: 10.1093/nar/gkw1108.

[55] O. Franzén, L.-M. Gan, and J. L. M. Björkegren, "PanglaoDB: a web server for exploration of mouse and human single-cell RNA sequencing data," *Database*, vol. 2019, p. baz046, Jan. 2019, doi: 10.1093/database/baz046.

[56] "Package 'Biostrings' Title Efficient manipulation of biological strings," 2025.

[57] K. Van den Berge *et al.*, "Normalization benchmark of ATAC-seq datasets shows the importance of accounting for GC-content effects," *Cell Reports Methods*, vol. 2, no. 11, p. 100321, 2022, doi: https://doi.org/10.1016/j.crmeth.2022.100321.

[58] A. T. L. Lun, Y. Chen, and G. K. Smyth, "It's DE-licious: a recipe for differential expression analyses of RNA-seq experiments using quasi-likelihood methods in edgeR," 2015. [Online]. Available: www.bioconductor.org/install

[59] M. D. Robinson, D. J. McCarthy, and G. K. Smyth, "edgeR: A Bioconductor package for differential expression analysis of digital gene expression data," *Bioinformatics*, vol. 26, no. 1, pp. 139–140, Nov. 2009, doi: 10.1093/bioinformatics/btp616.

[60] D. Risso, K. Schwartz, G. Sherlock, and S. Dudoit, "GC-Content Normalization for RNA-Seq Data," *BMC Bioinformatics*, vol. 12, no. 1, Dec. 2011, doi: 10.1186/1471-2105-12-480.

[61] Y. Zhao, L. Wong, and W. W. Bin Goh, "How to do quantile normalization correctly for gene expression data analyses," *Sci Rep*, vol. 10, no. 1, p. 15534, 2020, doi: 10.1038/s41598-020-72664-6.

[62] D. Smedley *et al.*, "BioMart – biological queries made easy," *BMC Genomics*, vol. 10, no. 1, p. 22, 2009, doi: 10.1186/1471-2164-10-22.

[63] Y. Chen, A. T. L. Lun, and G. K. Smyth, "From reads to genes to pathways: differential expression analysis of RNA-Seq experiments using Rsubread and the edgeR quasi-likelihood pipeline," *F1000Res*, vol. 5, p. 1438, Jun. 2016, doi: 10.12688/f1000research.8987.1.

[64] T. Saito and M. Rehmsmeier, "Precrec: fast and accurate precision–recall and ROC curve calculations in R," *Bioinformatics*, vol. 33, no. 1, pp. 145–147, Jan. 2017, doi: 10.1093/bioinformatics/btw570.

[65] D. M. W. Powers and Ailab, "EVALUATION: FROM PRECISION, RECALL AND F-MEASURE TO ROC, INFORMEDNESS, MARKEDNESS & CORRELATION."

[66] K. Boyd, K. H. Eng, and C. D. Page, "Area under the Precision-Recall Curve: Point Estimates and Confidence Intervals," in *Machine Learning and Knowledge Discovery in Databases*, H. Blockeel, K. Kersting, S. Nijssen, and F. Železný, Eds., Berlin, Heidelberg: Springer Berlin Heidelberg, 2013, pp. 451–466.

[67] A. Kessy, A. Lewin, and K. Strimmer, "Optimal whitening and decorrelation," Dec. 2015, doi: 10.1080/00031305.2016.1277159.

[68] L. Ronzoni, P. Bonara, D. Rusconi, C. Frugoni, I. Libani, and M. D. Cappellini, "Erythroid differentiation and maturation from peripheral CD34+ cells in liquid culture: Cellular and molecular characterization," *Blood Cells Mol Dis*, vol. 40, no. 2, pp. 148–155, 2008, doi: https://doi.org/10.1016/j.bcmd.2007.07.006.

[69] M. E. Wechsler, B. P. Hermann, and R. Bizios, "Adult Human Mesenchymal Stem Cell Differentiation at the Cell Population and Single-Cell Levels Under Alternating Electric Current," *Tissue Eng Part C Methods*, vol. 22, no. 2, pp. 155–164, Feb. 2016, doi: 10.1089/ten.tec.2015.0324.

[70] D. Zujur *et al.*, "Stepwise strategy for generating osteoblasts from human pluripotent stem cells under fully defined xeno-free conditions with small-molecule inducers," *Regen Ther*, vol. 14, pp. 19–31, 2020, doi: https://doi.org/10.1016/j.reth.2019.12.010.

[71] R. Calloni, E. A. A. Cordero, J. A. P. Henriques, and D. Bonatto, "Reviewing and updating the major molecular markers for stem cells," *Stem Cells Dev*, vol. 22, no. 9, pp. 1455–1476, May 2013, doi: 10.1089/scd.2012.0637.

[72] X. Liu *et al.*, "Yamanaka factors critically regulate the developmental signaling network in mouse embryonic stem cells," *Cell Res*, vol. 18, no. 12, pp. 1177–1189, 2008, doi: 10.1038/cr.2008.309.

[73] N. P. Mongan, K. M. Martin, and L. J. Gudas, "The putative human stem cell marker, Rex-1 (Zfp42): Structural classification and expression in normal human epithelial and carcinoma cell cultures," *Mol Carcinog*, vol. 45, no. 12, pp. 887–900, Dec. 2006, doi: 10.1002/mc.20186.

[74] R. H. Klein, P. Y. Tung, P. Somanath, H. J. Fehling, and P. S. Knoepfler, "Genomic functions of developmental pluripotency associated factor 4 (Dppa4) in pluripotent stem cells and cancer," *Stem Cell Res*, vol. 31, pp. 83–94, Aug. 2018, doi: 10.1016/j.scr.2018.07.009.

[75] W. Zhao, X. Ji, F. Zhang, L. Li, and L. Ma, "Embryonic Stem Cell Markers," *Molecules*, vol. 17, no. 6, pp. 6196–6236, May 2012, doi: 10.3390/molecules17066196.

[76] W. Ma, R.-T. Yan, W. Mao, and S.-Z. Wang, "Neurogenin3 promotes early retinal neurogenesis," *Molecular and Cellular Neuroscience*, vol. 40, no. 2, pp. 187–198, 2009, doi: https://doi.org/10.1016/j.mcn.2008.10.006.

[77] N. A. J. Krentz *et al.*, "Phosphorylation of NEUROG3 Links Endocrine Differentiation to the Cell Cycle in Pancreatic Progenitors," *Dev Cell*, vol. 41, no. 2, pp. 129-142.e6, Apr. 2017, doi: 10.1016/j.devcel.2017.02.006.

[78] C. Francius *et al.*, "Vsx1 transiently defines an early intermediate V2 interneuron precursor compartment in the mouse developing spinal cord," *Front Mol Neurosci*, vol. 9, no. Dec2016, Dec. 2016, doi: 10.3389/fnmol.2016.00145.

[79] C. Gaultier, H. Trang, S. Dauger, and J. Gallego, "Pediatric Disorders with Autonomic Dysfunction: What Role for PHOX2B?," *Pediatr Res*, vol. 58, no. 1, pp. 1–6, 2005, doi: 10.1203/01.PDR.0000166755.29277.C4.

[80] A. Pattyn, X. Morin, H. Cremer, C. Goridis, and J.-F. Brunet, "The homeobox gene Phox2b is essential for the development of autonomic neural crest derivatives," *Nature*, vol. 399, no. 6734, pp. 366–370, 1999, doi: 10.1038/20700.

[81] K. Zhang *et al.*, "The Oligodendrocyte Transcription Factor 2 OLIG2 regulates transcriptional repression during myelinogenesis in rodents," *Nat Commun*, vol. 13, no. 1, p. 1423, 2022, doi: 10.1038/s41467-022-29068-z.

[82] L. E. Sidney, M. J. Branch, S. E. Dunphy, H. S. Dua, and A. Hopkinson, "Concise Review: Evidence for CD34 as a Common Marker for Diverse Progenitors," *Stem Cells*, vol. 32, no. 6, pp. 1380–1389, Jun. 2014, doi: 10.1002/stem.1661.

[83] G. Y. Lee, J.-H. Kim, G. T. Oh, B.-H. Lee, I. C. Kwon, and I.-S. Kim, "Molecular targeting of atherosclerotic plaques by a stabilin-2-specific peptide ligand," *Journal of Controlled Release*, vol. 155, no. 2, pp. 211–217, 2011, doi: https://doi.org/10.1016/j.jconrel.2011.07.010.

[84] M. Inoue *et al.*, "Endothelial cell-selective adhesion molecule modulates atherosclerosis through plaque angiogenesis and monocyte–endothelial interaction," *Microvasc Res*, vol. 80, no. 2, pp. 179–187, 2010, doi: https://doi.org/10.1016/j.mvr.2010.04.005.

[85] J. R. Privratsky and P. J. Newman, "PECAM-1: regulator of endothelial junctional integrity," *Cell Tissue Res*, vol. 355, no. 3, pp. 607–619, 2014, doi: 10.1007/s00441-013-1779-3.

[86] O. Bondareva *et al.*, "Single-cell profiling of vascular endothelial cells reveals progressive organ-specific vulnerabilities during obesity," *Nat Metab*, vol. 4, no. 11, pp. 1591–1610, 2022, doi: 10.1038/s42255-022-00674-x.

[87] Y. Xie *et al.*, "FGF/FGFR signaling in health and disease," *Signal Transduct Target Ther*, vol. 5, no. 1, p. 181, 2020, doi: 10.1038/s41392-020-00222-7.

[88] N. Liu, A. Wang, M. Xue, X. Zhu, Y. Liu, and M. Chen, "FOXA1 and FOXA2: the regulatory mechanisms and therapeutic implications in cancer," *Cell Death Discov*, vol. 10, no. 1, p. 172, 2024, doi: 10.1038/s41420-024-01936-1.

[89] B. Sosa-Pineda, J. T. Wigle, and G. Oliver, "Hepatocyte migration during liver development requires Prox1," *Nat Genet*, vol. 25, no. 3, pp. 254–255, 2000, doi: 10.1038/76996.

[90] B. G. Jun *et al.*, "Relation of fibroblast growth factor receptor 2 expression to hepatocellular carcinoma recurrence after liver resection," *PLoS One*, vol. 15, no. 1, Jan. 2020, doi: 10.1371/journal.pone.0227440.

[91] L. He *et al.*, "Deleting Gata4 in hepatocytes promoted the progression of NAFLD via increasing steatosis and apoptosis, and desensitizing insulin signaling," *J Nutr Biochem*, vol. 111, p. 109157, 2023, doi: https://doi.org/10.1016/j.jnutbio.2022.109157.

[92] Y. Cho, Y. Nam, H. H. Lee, and R. Chang, "Inhibition mechanism of testis-expressed gene 14 (TEX14) in cytokinetic abscission: Well-tempered metadynamics simulation studies," *J Chem Phys*, vol. 159, no. 1, p. 015102, Jul. 2023, doi: 10.1063/5.0153799.

[93] S. Vakkilainen *et al.*, "The human long non-coding RNA gene RMRP has pleiotropic effects and regulates cell-cycle progression at G2," *Sci Rep*, vol. 9, no. 1, p. 13758, 2019, doi: 10.1038/s41598-019-50334-6.

[94] K. R. Noss, S. A. Wolfe, and S. R. Grimes, "Upregulation of prostate specific membrane antigen/folate hydrolase transcription by an enhancer," *Gene*, vol. 285, no. 1, pp. 247–256, 2002, doi: https://doi.org/10.1016/S0378-1119(02)00397-9.

[95] S. Chen *et al.*, "Pharmacological inhibition of PI5P4Kα/β disrupts cell energy metabolism and selectively kills p53-null tumor cells", doi: 10.1073/pnas.2002486118/-/DCSupplemental.

[96] A. Agrud *et al.*, "Gabrb3 endothelial cell-specific knockout mice display abnormal blood flow, hypertension, and behavioral dysfunction," *Sci Rep*, vol. 12, no. 1, p. 4922, 2022, doi: 10.1038/s41598-022-08806-9.

[97] G. M. H. Birchenough, M. E. V Johansson, J. K. Gustafsson, J. H. Bergström, and G. C. Hansson, "New developments in goblet cell mucus secretion and function," *Mucosal Immunol*, vol. 8, no. 4, pp. 712–719, 2015, doi: 10.1038/mi.2015.32.

[98] D. Risso, "EDASeq: Exploratory Data Analysis and Normalization for RNA-Seq (vignette)," Apr. 15, 2025. doi: 10.18129/B9.bioc.EDASeq.

[99] N. Sadeghi *et al.*, "Analysis of the contribution of experimental bias, experimental noise, and inter-subject biological variability on the assessment of developmental trajectories in diffusion MRI studies of the brain," *Neuroimage*, vol. 109, pp. 480–492, Apr. 2015, doi: 10.1016/j.neuroimage.2014.12.084.

[100] P. Weidemüller, M. Kholmatov, E. Petsalaki, and J. B. Zaugg, "Transcription factors: Bridge between cell signaling and gene regulation," Dec. 01, 2021, *John Wiley and Sons Inc*. doi: 10.1002/pmic.202000034.

[101] L. F. Soto *et al.*, "Compendium of human transcription factor effector domains," *Mol Cell*, vol. 82, no. 3, pp. 514–526, Feb. 2022, doi: 10.1016/j.molcel.2021.11.007.

[102] R. Martinez-Corral, D. Friedrich, R. Frömel, L. Velten, J. Gunawardena, and A. H. DePace, "Emergence of activation or repression in transcriptional control under a fixed molecular context," Jun. 02, 2024. doi: 10.1101/2024.05.29.596388.

[103] S. H. Duttke *et al.*, "Position-dependent function of human sequence-specific transcription factors," *Nature*, vol. 631, no. 8022, pp. 891–898, 2024, doi: 10.1038/s41586-024-07662-z.

[104] N. DelRosso *et al.*, "Large-scale mapping and mutagenesis of human transcriptional effector domains," *Nature*, vol. 616, no. 7956, pp. 365–372, 2023, doi: 10.1038/s41586-023-05906-y.

[105] A. X. Mukund *et al.*, "High-throughput functional characterization of combinations of transcriptional activators and repressors," *Cell Syst*, vol. 14, no. 9, pp. 746-763.e5, Sep. 2023, doi: 10.1016/j.cels.2023.07.001.

[106] T. G. Gillette and J. A. Hill, "Readers, writers, and erasers: Chromatin as the whiteboard of heart disease," Mar. 27, 2015, *Lippincott Williams and Wilkins*. doi: 10.1161/CIRCRESAHA.116.303630.

[107] Y. Li *et al.*, "KLF12 promotes the proliferation of breast cancer cells by reducing the transcription of p21 in a p53-dependent and p53-independent manner," *Cell Death Dis*, vol. 14, no. 5, p. 313, 2023, doi: 10.1038/s41419-023-05824-x.

[108] J. Yao *et al.*, "GLIS2 promotes colorectal cancer through repressing enhancer activation," *Oncogenesis*, vol. 9, no. 5, p. 57, 2020, doi: 10.1038/s41389-020-0240-1.

[109] L. Lin *et al.*, "topors, a p53 and topoisomerase I-binding RING finger protein, is a coactivator of p53 in growth suppression induced by DNA damage," *Oncogene*, vol. 24, no. 21, pp. 3385–3396, 2005, doi: 10.1038/sj.onc.1208554.

[110] H. Zhang *et al.*, "Understanding the Transcription Factor NFE2L1/NRF1 from the Perspective of Hallmarks of Cancer," Jul. 01, 2024, *Multidisciplinary Digital Publishing Institute (MDPI)*. doi: 10.3390/antiox13070758.

[111] S. Kim, N.-K. Yu, and B.-K. Kaang, "CTCF as a multifunctional protein in genome regulation and gene expression," *Exp Mol Med*, vol. 47, no. 6, pp. e166–e166, 2015, doi: 10.1038/emm.2015.33.

[112] M. Nishana *et al.*, "Defining the relative and combined contribution of CTCF and CTCFL to genomic regulation," *Genome Biol*, vol. 21, no. 1, p. 108, 2020, doi: 10.1186/s13059-020-02024-0.

[113] T. Suzuki *et al.*, " Expression of a Testis-Specific Form of Gal3st1 (CST), a Gene Essential for Spermatogenesis, Is Regulated by the CTCF Paralogous Gene BORIS ," *Mol Cell Biol*, vol. 30, no. 10, pp. 2473–2484, May 2010, doi: 10.1128/mcb.01093-09.

[114] J.-C. Marine and G. Lozano, "Mdm2-mediated ubiquitylation: p53 and beyond," *Cell Death Differ*, vol. 17, no. 1, pp. 93–102, 2010, doi: 10.1038/cdd.2009.68.

[115] N. B. Bloch, T. E. Wales, M. S. Prew, H. R. Levy, J. R. Engen, and L. D. Walensky, "The conformational stability of pro-apoptotic BAX is dictated by discrete residues of the protein core," *Nat Commun*, vol. 12, no. 1, p. 4932, 2021, doi: 10.1038/s41467-021-25200-7.

[116] J.-L. Chew *et al.*, " Reciprocal Transcriptional Regulation of Pou5f1 and Sox2 via the Oct4/Sox2 Complex in Embryonic Stem Cells ," *Mol Cell Biol*, vol. 25, no. 14, pp. 6031–6046, Jul. 2005, doi: 10.1128/mcb.25.14.6031-6046.2005.

[117] D. Huynh *et al.*, "Effective in vivo binding energy landscape illustrates kinetic stability of RBPJ-DNA binding," *Nat Commun*, vol. 16, no. 1, p. 1259, 2025, doi: 10.1038/s41467-025-56515-4.

[118] R. Russell *et al.*, "A Dynamic Role of TBX3 in the Pluripotency Circuitry," *Stem Cell Reports*, vol. 5, no. 6, pp. 1155–1170, 2015, doi: https://doi.org/10.1016/j.stemcr.2015.11.003.

[119] L. Ferraris *et al.*, "Combinatorial binding of transcription factors in the pluripotency control regions of the genome," *Genome Res*, vol. 21, no. 7, pp. 1055–1064, Jul. 2011, doi: 10.1101/gr.115824.110.

[120] W. Zhou, K. M. Gross, and C. Kuperwasser, "Molecular regulation of Snai2 in development and disease," *J Cell Sci*, vol. 132, no. 23, p. jcs235127, Dec. 2019, doi: 10.1242/jcs.235127.

[121] K. Niessen, Y. X. Fu, L. Chang, P. A. Hoodless, D. McFadden, and A. Karsan, "Slug is a direct Notch target required for initiation of cardiac cushion cellularization," *Journal of Cell Biology*, vol. 182, no. 2, pp. 315–325, Jul. 2008, doi: 10.1083/jcb.200710067.

[122] G. Perez *et al.*, "The UCSC Genome Browser database: 2025 update," *Nucleic Acids Res*, vol. 53, no. D1, pp. D1243–D1249, Jan. 2025, doi: 10.1093/nar/gkae974.

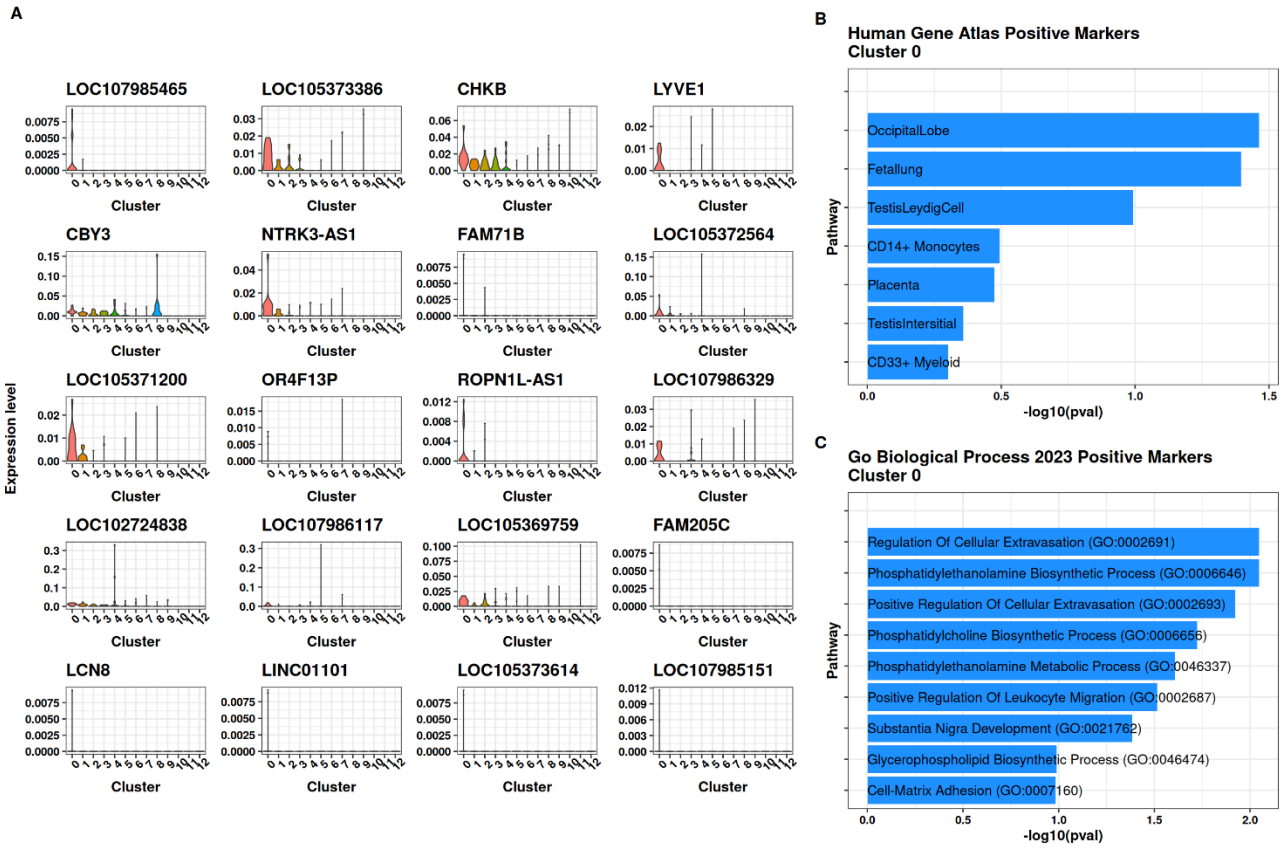
[123] T. Stuart, A. Srivastava, S. Madad, C. A. Lareau, and R. Satija, "Single-cell chromatin state analysis with Signac," *Nat Methods*, vol. 18, no. 11, pp. 1333–1341, 2021, doi: 10.1038/s41592-021-01282-5.

Code Availability

Github_Link:

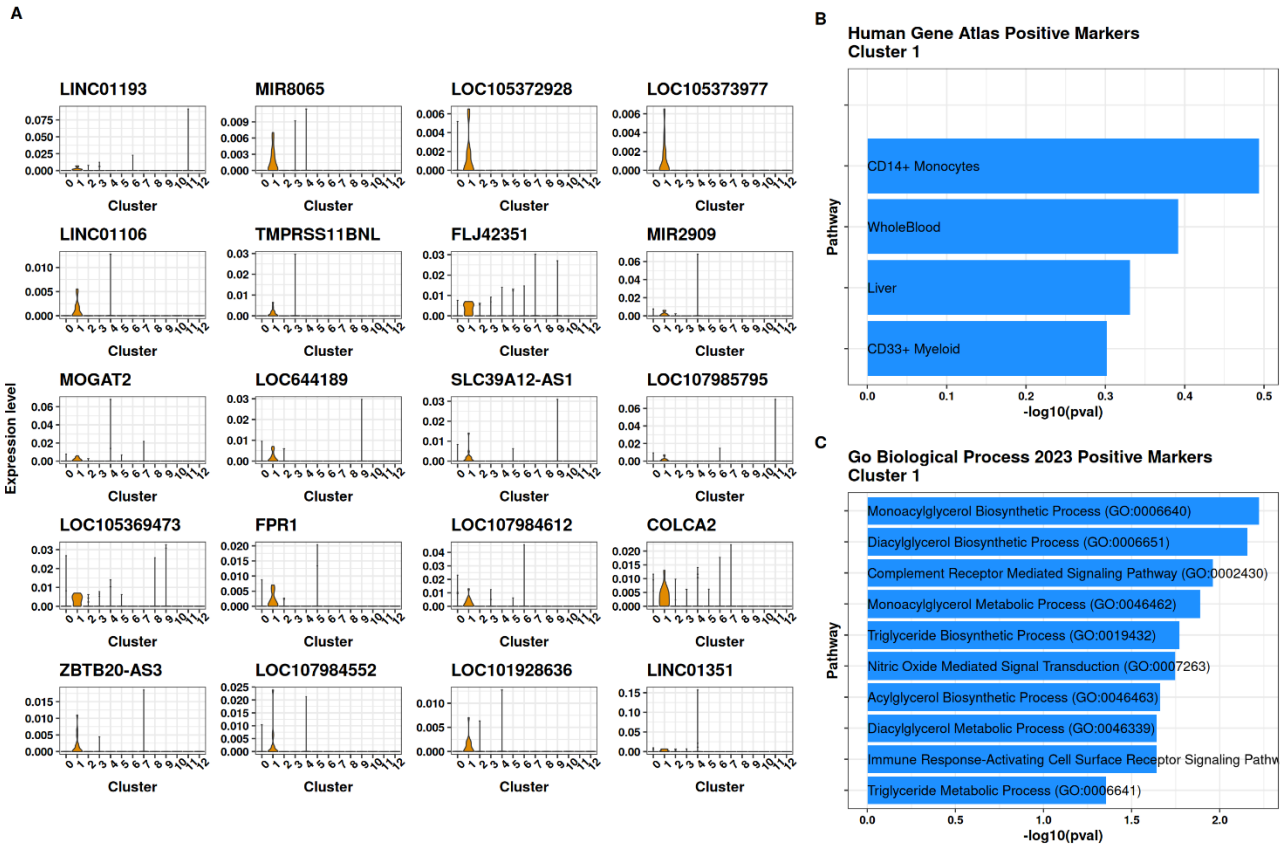
https://github.com/Tomey20/code-for-masters-thesis

Appendix

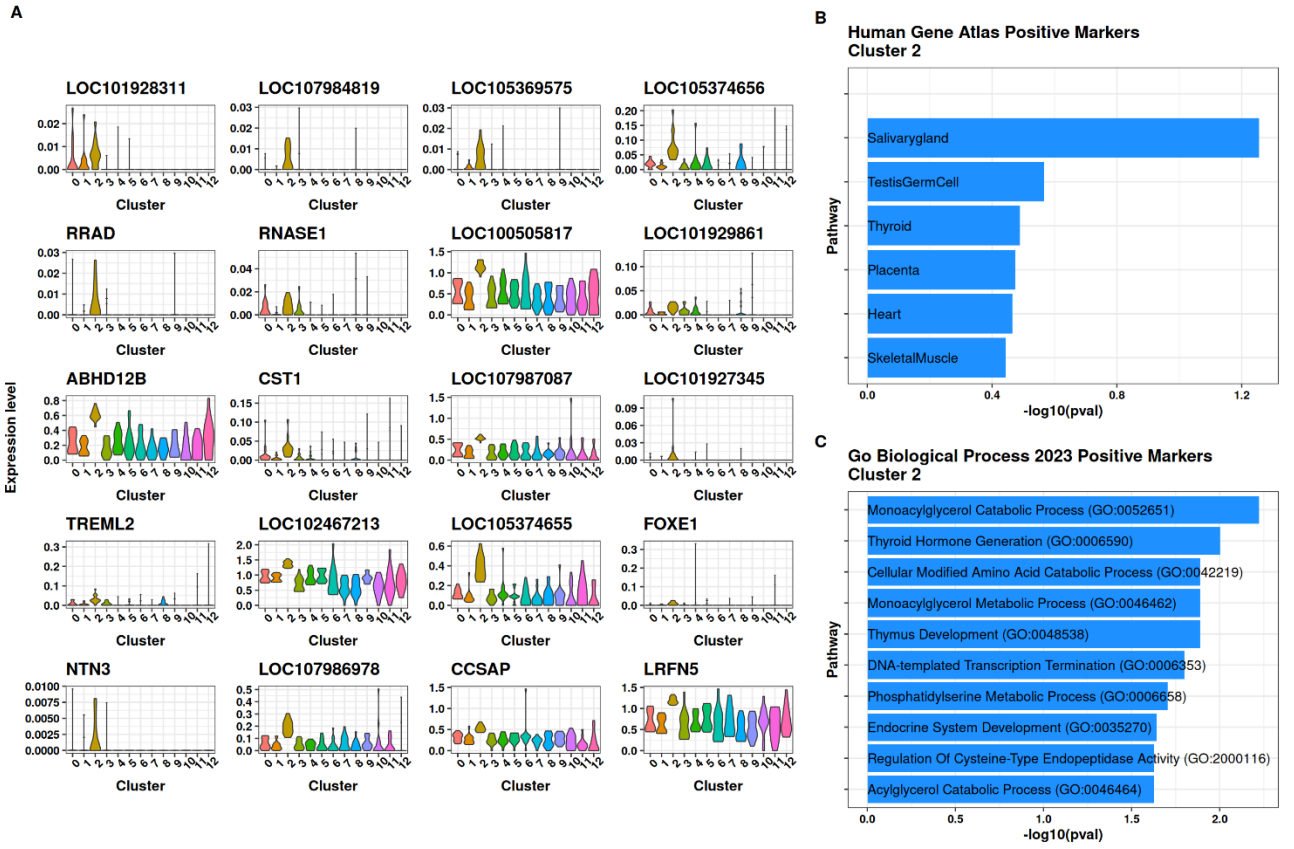


Supplementary Figure 1. Top 20 marker genes and gene set enrichment analysis of cluster 0. *(A) Violin plots of the top 20 positive marker genes of cluster 0, where the log-normalised gene expression is shown on the y-axis and the cell cluster in which they are expressed on the x-axis. The names of the marker genes are displayed above each violin plot. (B) Barplot of enriched pathways/terms in the gene set enrichment analysis using the human gene atlas database. On the y-axis, the cell type associated with the top 20 differentially expressed positive marker genes is displayed according to the Human Gene Atlas database, while the x-*

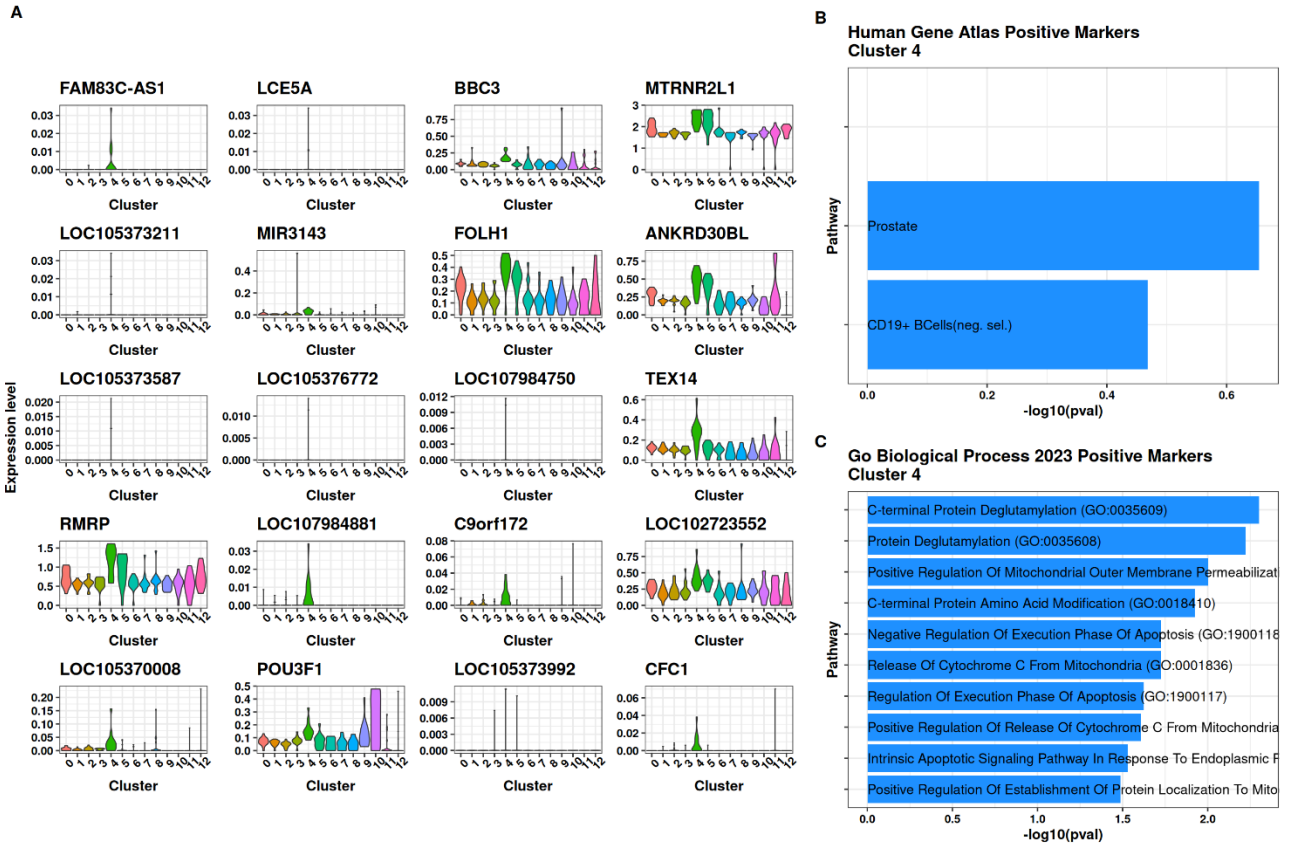
axis represents the -log10 transformed p-value, indicating the significance of the differentially expressed positive marker genes associated with the term. (C) Barplot of enriched pathways or biological processes associated with the top 20 differentially expressed positive marker genes according to the GO Biological Process 2023 database. The enriched pathway/term is shown on the y-axis, while the x-axis shows the -log10 scaled p-value.



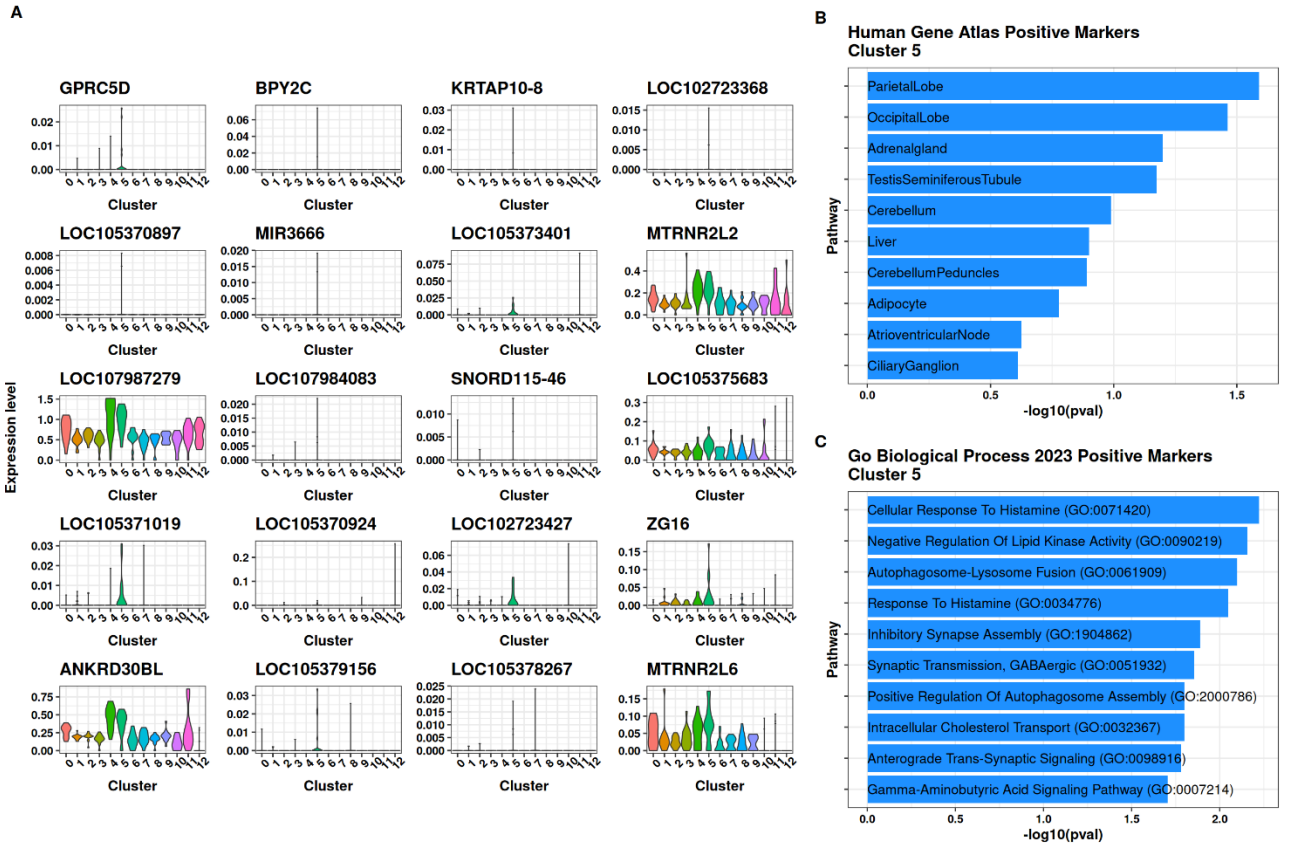
Supplementary Figure 2. Top 20 marker genes and gene set enrichment analysis of cluster 1. *(A) Violin plots of the top 20 positive marker genes of cluster 1, where the log-normalised gene expression is shown on the y-axis and the cell cluster in which they are expressed on the x-axis. The names of the marker genes are displayed above each violin plot. (B) Barplot of enriched pathways/terms for cluster 1 in the gene set enrichment analysis using the human gene atlas database. On the y-axis, the cell type associated with the top 20 differentially expressed positive marker genes is displayed according to the Human Gene Atlas database, while the x-axis represents the -log10 transformed p-value, indicating the significance of the differentially expressed positive marker genes associated with the term. (C) Barplot of enriched pathways or biological processes associated with the top 20 differentially expressed positive marker genes for cluster 1 according to the GO Biological Process 2023 database. The enriched pathway or term is shown on the y-axis, while the x-axis displays the -log10 scaled p-value.*



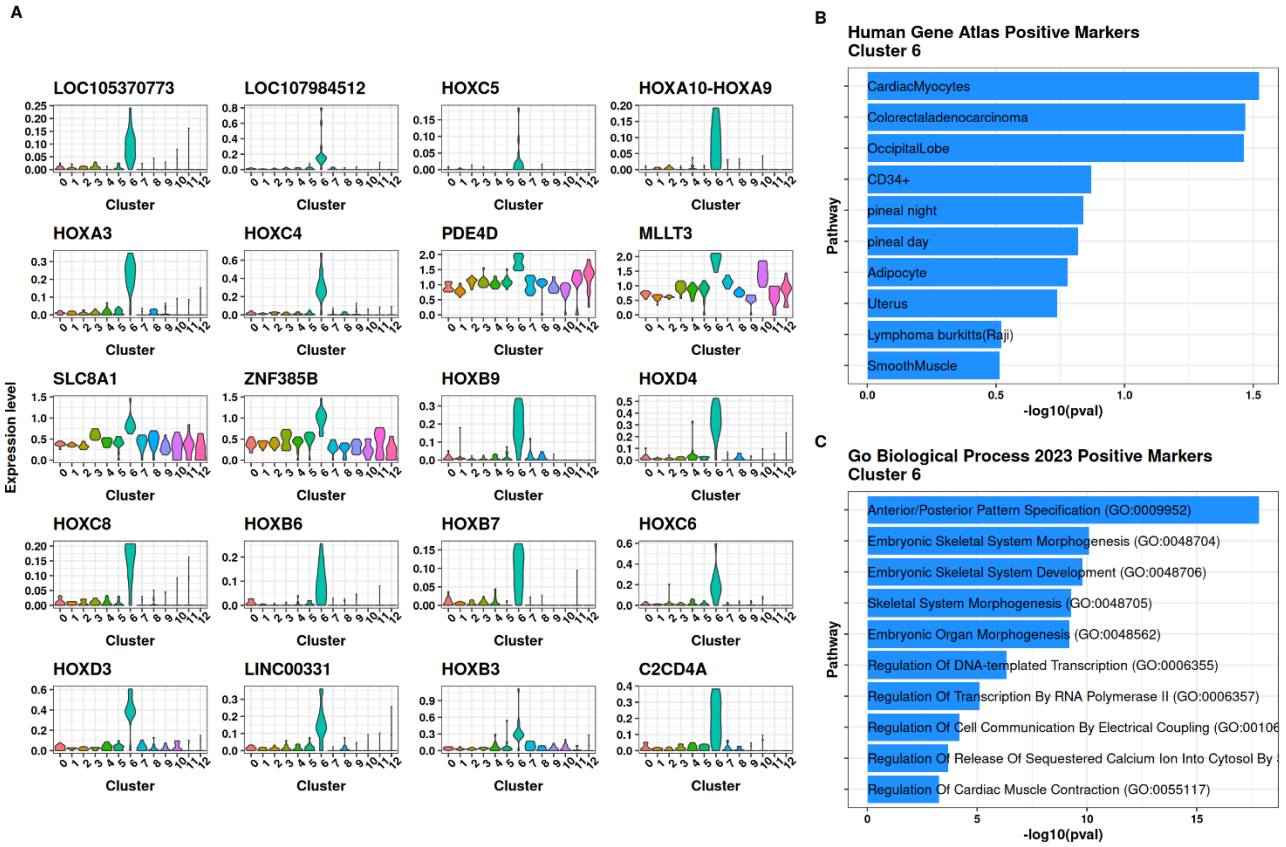
Supplementary Figure 3. Top 20 marker genes and gene set enrichment analysis of cluster 2. *(A) Violin plots illustrate the top 20 positive marker genes of cluster 2, with log-normalised gene expression plotted on the y-axis and the corresponding cell clusters on the x-axis. The marker gene names are positioned above each violin. (B) A bar plot depicting enriched pathways/terms for cluster 2, derived from gene set enrichment analysis using the Human Gene Atlas database. Here, the y-axis displays the cell types associated with the top 20 differentially expressed positive marker genes, as per the Human Gene Atlas database, while the x-axis represents the -log10 transformed p-value, indicating the significance of these genes in relation to the term. (C) Another barplot showcases enriched pathways or biological processes corresponding to the top 20 differentially expressed positive marker genes for cluster 2, based on the GO Biological Process 2023 database. The y-axis displays the enriched pathway or term, and the x-axis presents the -log10 scaled p-value.*



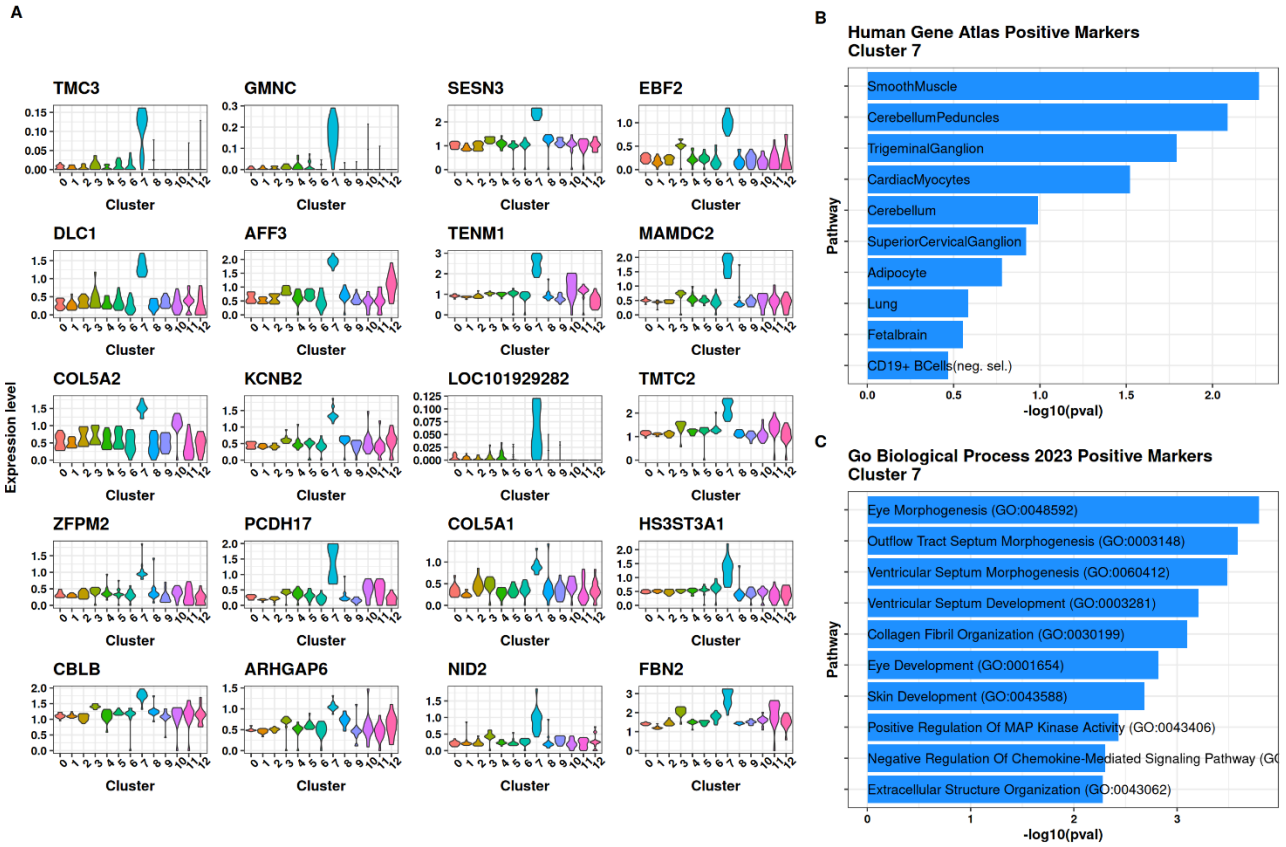
Supplementary Figure 4. Top 20 marker genes and gene set enrichment analysis of cluster 4. *(A) Violin plots show the top 20 positive marker genes of cluster 4, with log-normalised gene expression on the y-axis and corresponding cell clusters on the x-axis. The names of the marker genes are located above each violin. (B) A bar plot illustrates enriched pathways/terms for cluster 4, based on gene set enrichment analysis from the Human Gene Atlas database. In this plot, the y-axis indicates the cell types associated with the top 20 differentially expressed positive marker genes, as per the Human Gene Atlas database, while the x-axis represents the -log10 transformed p-value, highlighting the significance of these genes in relation to the term. (C) Another bar plot presents enriched pathways or biological processes associated with the top 20 differentially expressed positive marker genes for cluster 4, utilising the GO Biological Process 2023 database. The y-axis shows the enriched pathway or term, whereas the x-axis displays the -log10 scaled p-value.*



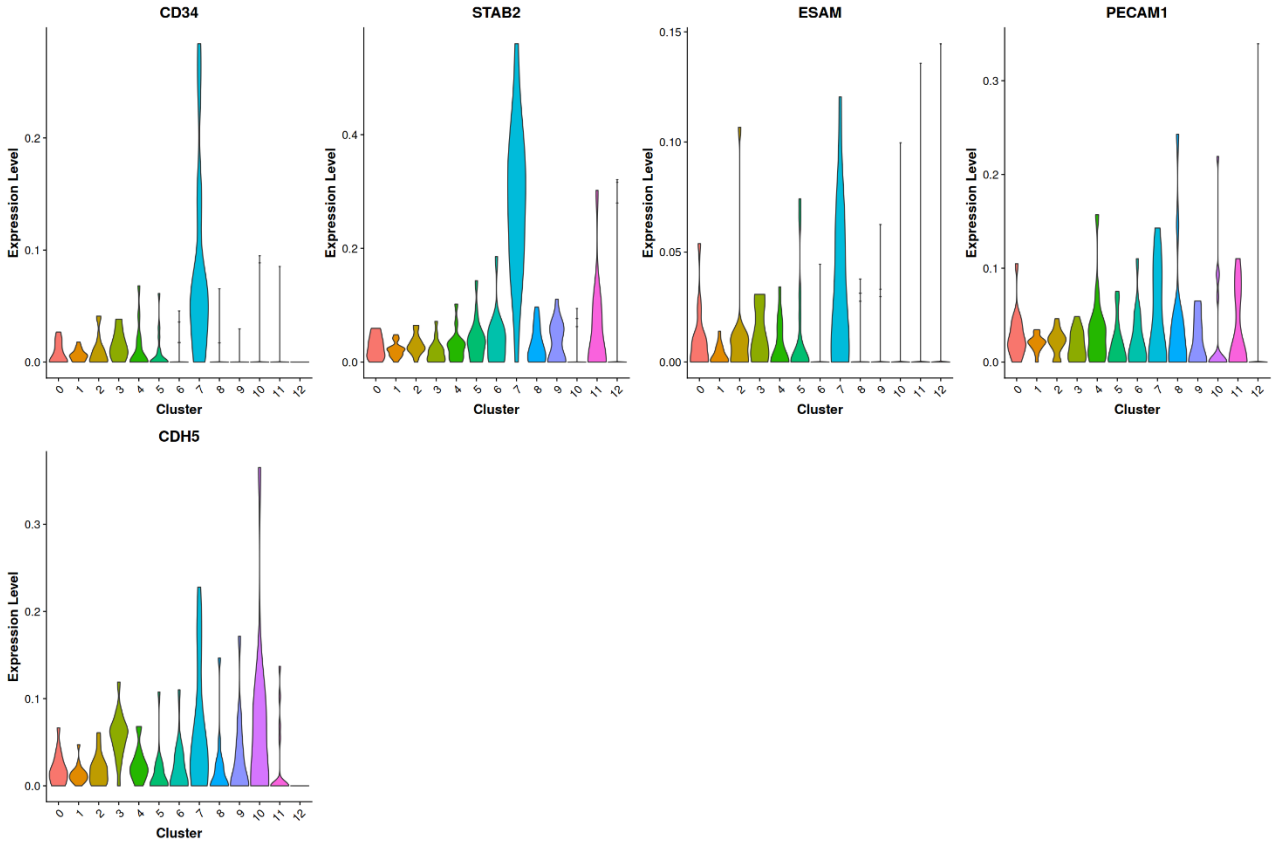
Supplementary Figure 5. Top 20 marker genes and gene set enrichment analysis of cluster 5. *(A) Violin plots display the top 20 positive marker genes from cluster 5, with log-normalised gene expression shown on the y-axis and corresponding cell clusters on the x-axis. The names of the marker genes are placed above each violin. (B) A bar plot illustrates the enriched pathways and terms for cluster 5, obtained from gene set enrichment analysis using the Human Gene Atlas database. In this plot, the y-axis represents the cell types associated with the top 20 differentially expressed positive marker genes, as determined by the Human Gene Atlas, while the x-axis displays the -log10 transformed p-value, indicating the significance of these genes in relation to the term. (C) Another bar plot presents enriched pathways or biological processes linked to the top 20 differentially expressed positive marker genes for cluster 5, based on the GO Biological Process 2023 database. Here, the y-axis indicates the enriched pathway or term, and the x-axis displays the -log10 scaled p-value.*



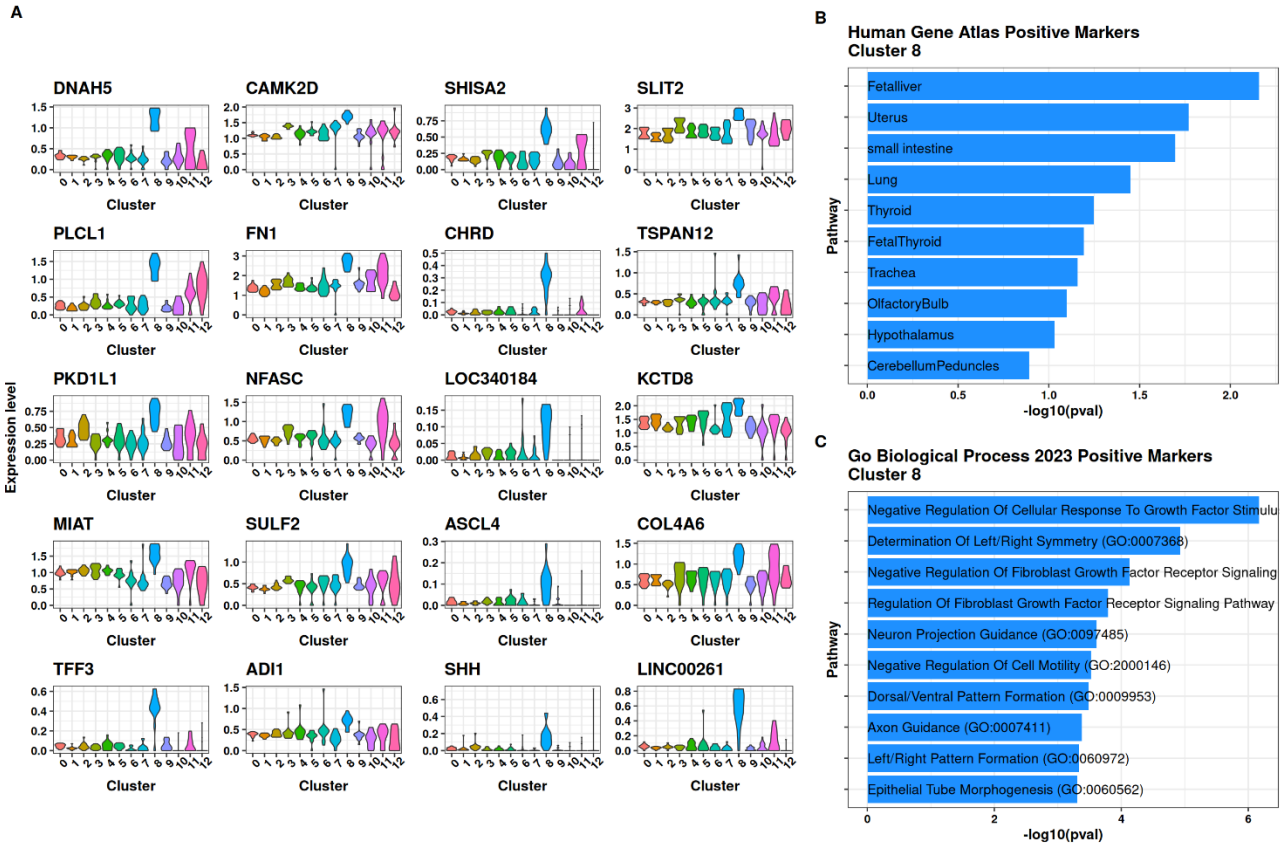
Supplementary Figure 6. Top 20 marker genes and gene set enrichment analysis of cluster 6. *(A) Violin plots demonstrate the top 20 positive marker genes of cluster 6, displaying log-normalised gene expression on the y-axis and the respective cell clusters along the x-axis. The names of the marker genes are shown above each violin. (B) A bar plot illustrates the enriched pathways/terms for cluster 6, obtained from gene set enrichment analysis using the Human Gene Atlas database. In this plot, the y-axis represents the cell types related to the top 20 differentially expressed positive marker genes according to the Human Gene Atlas, while the x-axis indicates the -log10 transformed p-value, reflecting the significance of these genes concerning the term. (C) Another bar plot highlights enriched pathways or biological processes linked to the top 20 differentially expressed positive marker genes for cluster 6, based on the GO Biological Process 2023 database. The y-axis indicates the enriched pathway or term, while the x-axis shows the -log10 scaled p-value.*



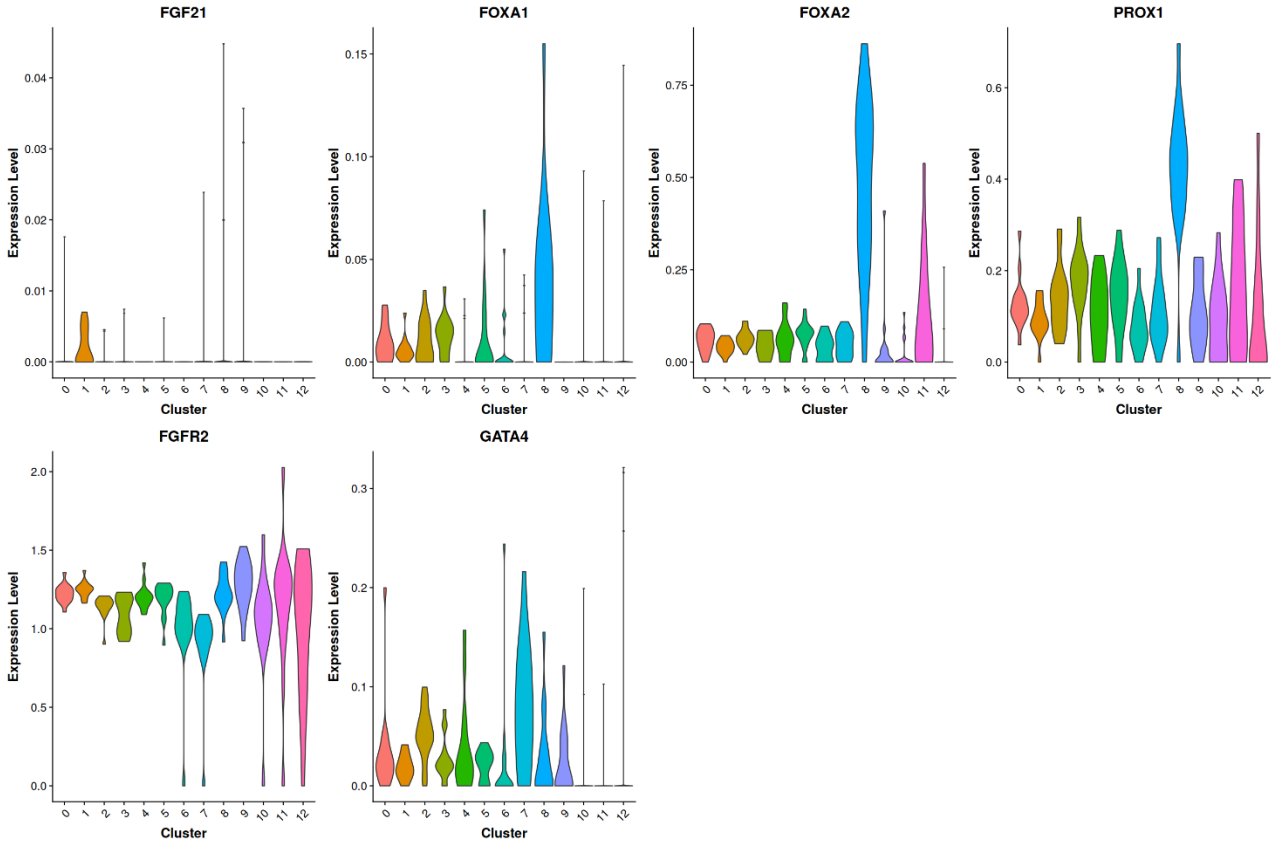
Supplementary Figure 7. Top 20 marker genes and gene set enrichment analysis of cluster 7. *(A) Violin plots display the top 20 positive marker genes for cluster 7, with the y-axis representing log-normalised gene expression and the x-axis showing the relevant cell clusters. Marker gene names are located above each violin. (B) A bar plot illustrates enriched pathways/terms related to cluster 7, obtained through gene set enrichment analysis using the Human Gene Atlas database. In this case, the y-axis represents the cell types associated with the top 20 differentially expressed positive marker genes, as determined by the Human Gene Atlas, while the x-axis displays the -log10 transformed p-value, indicating the significance of these genes in relation to the term. (C) Additionally, another bar plot highlights enriched pathways or biological processes pertaining to the top 20 differentially expressed positive marker genes for cluster 7, utilising the GO Biological Process 2023 database. The y-axis features the enriched pathway or term, whereas the x-axis displays the -log10 scaled p-value.*



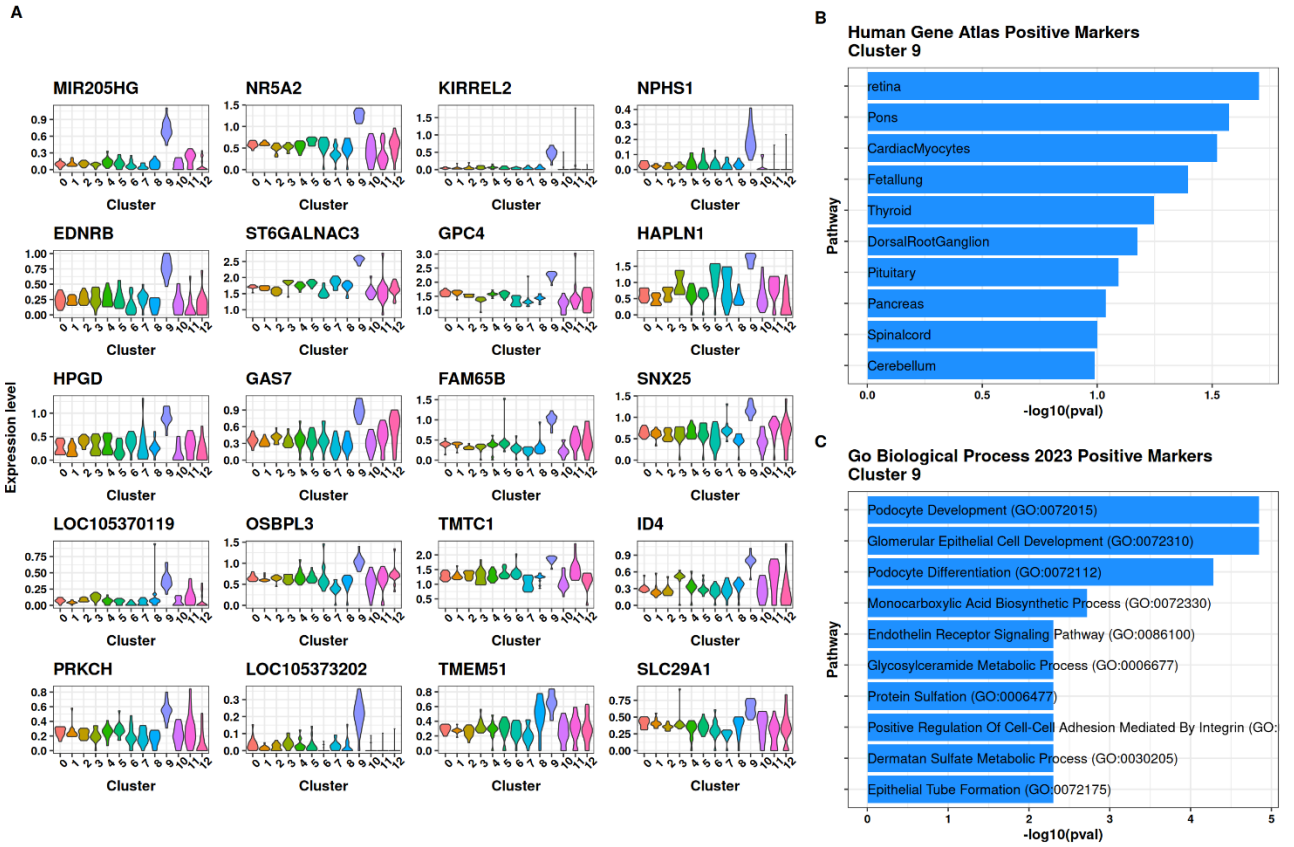
Supplementary Figure 8. Selected marker genes for cluster 7. *The figure shows violin plots for positively associated marker genes for Cluster 7, where the x-axis indicates the cluster number and the y-axis indicates the normalised gene expression [82], [83], [84], [85], [86].*



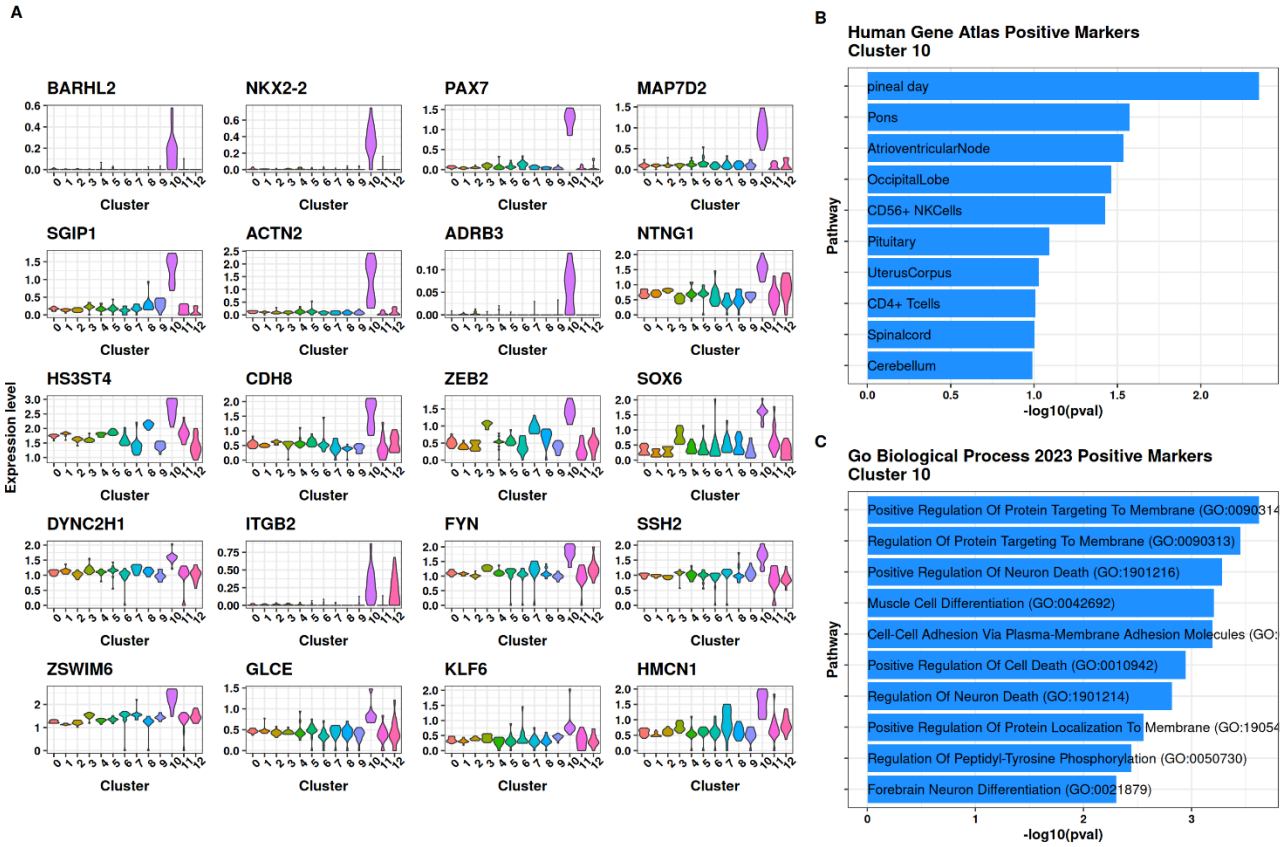
Supplementary Figure 9. Top 20 marker genes and gene set enrichment analysis of cluster 8. *(A) Violin plots illustrate the top 20 positive marker genes of cluster 8, with log-normalised gene expression plotted on the y-axis and the corresponding cell clusters on the x-axis. The marker gene names are positioned above each violin. (B) A bar plot depicting enriched pathways/terms for cluster 8, derived from gene set enrichment analysis using the Human Gene Atlas database. Here, the y-axis displays the cell types associated with the top 20 differentially expressed positive marker genes, as per the Human Gene Atlas database, while the x-axis represents the -log10 transformed p-value, indicating the significance of these genes in relation to the term. (C) Another barplot showcases enriched pathways or biological processes corresponding to the top 20 differentially expressed positive marker genes for cluster 8, based on the GO Biological Process 2023 database. The y-axis displays the enriched pathway or term, and the x-axis presents the -log10 scaled p-value.*



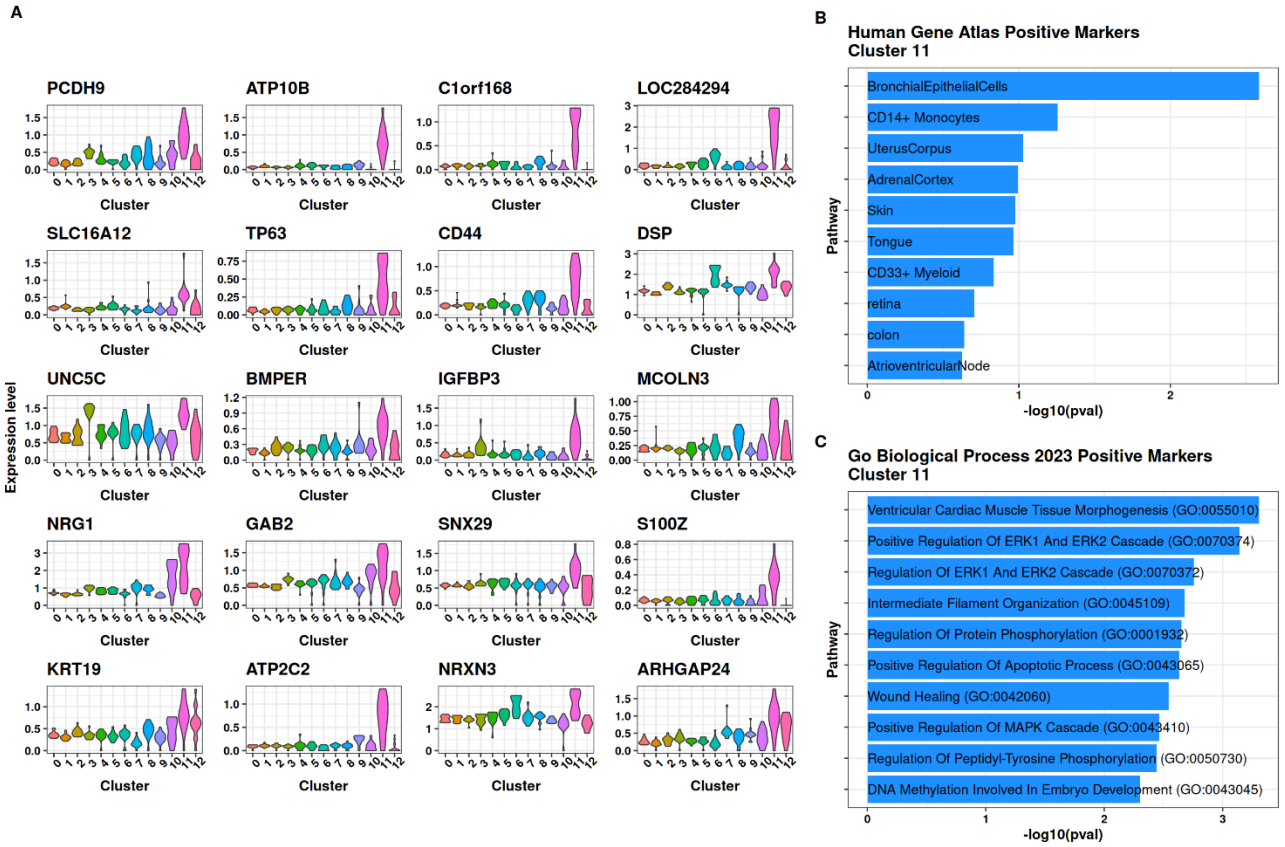
Supplementary Figure 10. Selected marker genes for cluster 8. *The figure illustrates violin plots depicting positively associated marker genes for Cluster 8. The x-axis represents the cluster number, while the y-axis shows normalised gene expression* [87], [88], [89], [90], [91].



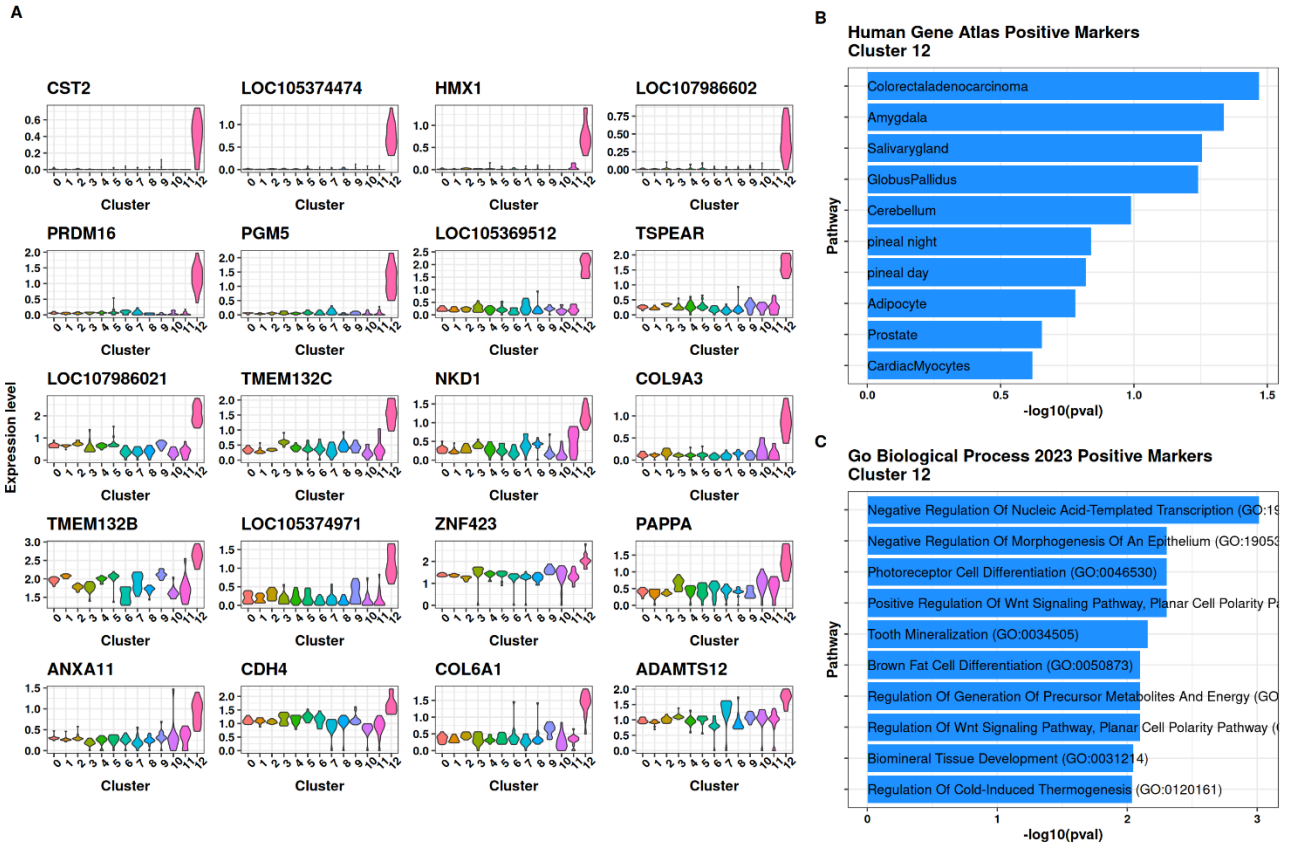
Supplementary Figure 11. Top 20 marker genes and gene set enrichment analysis of cluster 9. *(A) Violin plots display the top 20 positive marker genes of cluster 9, with log-normalised gene expression shown on the y-axis and corresponding cell clusters on the x-axis. The names of the marker genes are placed above each violin. (B) A bar plot illustrates the enriched pathways and terms for cluster 9, obtained from gene set enrichment analysis using the Human Gene Atlas database. In this plot, the y-axis represents the cell types associated with the top 20 differentially expressed positive marker genes, as determined by the Human Gene Atlas, while the x-axis displays the -log10 transformed p-value, indicating the significance of these genes in relation to the term. (C) An additional bar plot highlights enriched pathways or biological processes related to the top 20 differentially expressed positive marker genes for cluster 9, based on the GO Biological Process 2023 database. The y-axis indicates the enriched pathway or term, and the x-axis presents the -log10 scaled p-value.*



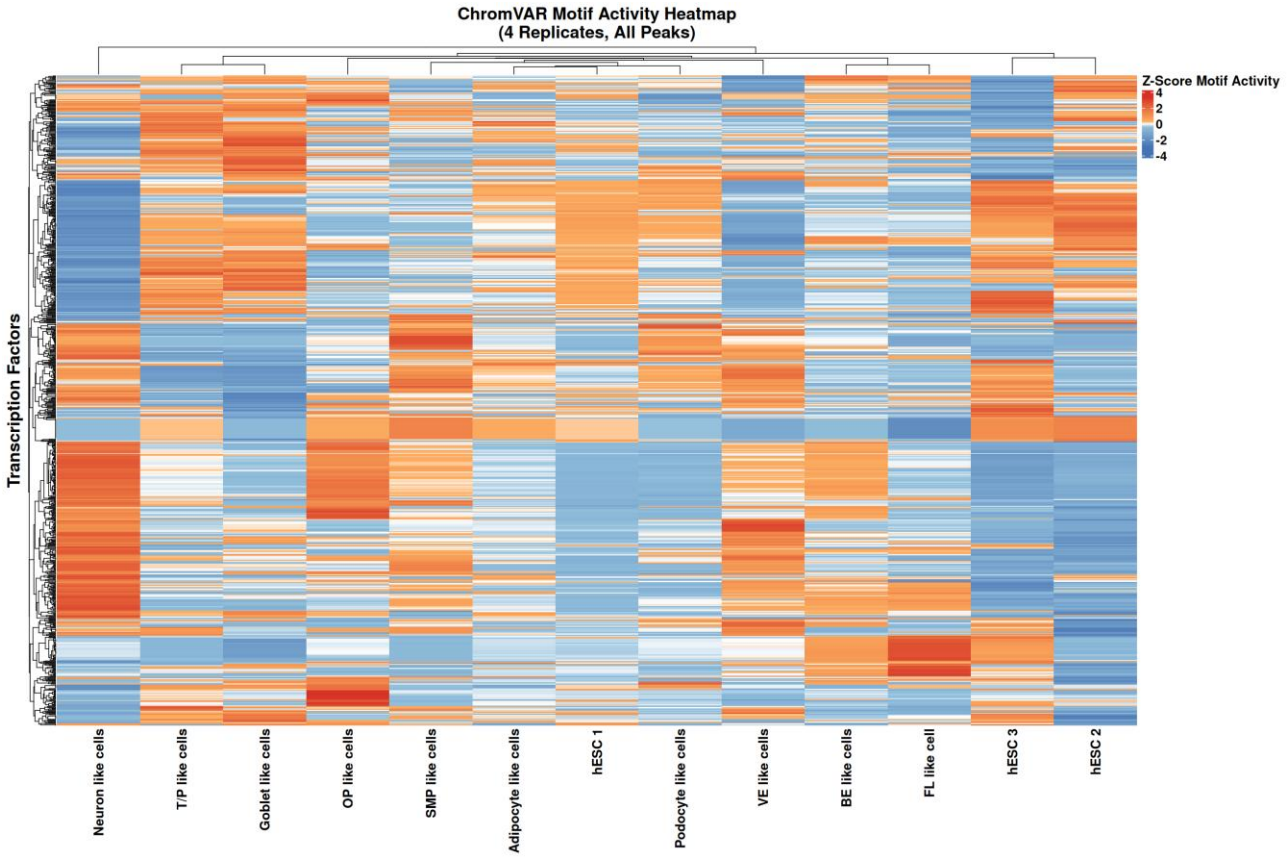
Supplementary Figure 12. Top 20 marker genes and gene set enrichment analysis of cluster 10. *(A) Violin plots display the top 20 positive marker genes from cluster 10, with log-normalised gene expression shown on the y-axis and corresponding cell clusters on the x-axis. The names of the marker genes are located above each violin. (B) A bar plot illustrates enriched pathways/terms for cluster 10, obtained from gene set enrichment analysis using the Human Gene Atlas database. In this plot, the y-axis represents cell types associated with the top 20 differentially expressed positive marker genes, as determined by the Human Gene Atlas database, while the x-axis indicates the -log10 transformed p-value, reflecting the significance of these genes in relation to the term. (C) An additional bar plot presents enriched pathways or biological processes associated with the top 20 differentially expressed positive marker genes for cluster 10, based on the GO Biological Process 2023 database. The y-axis features the enriched pathway or term, while the x-axis displays the -log10 scaled p-value.*



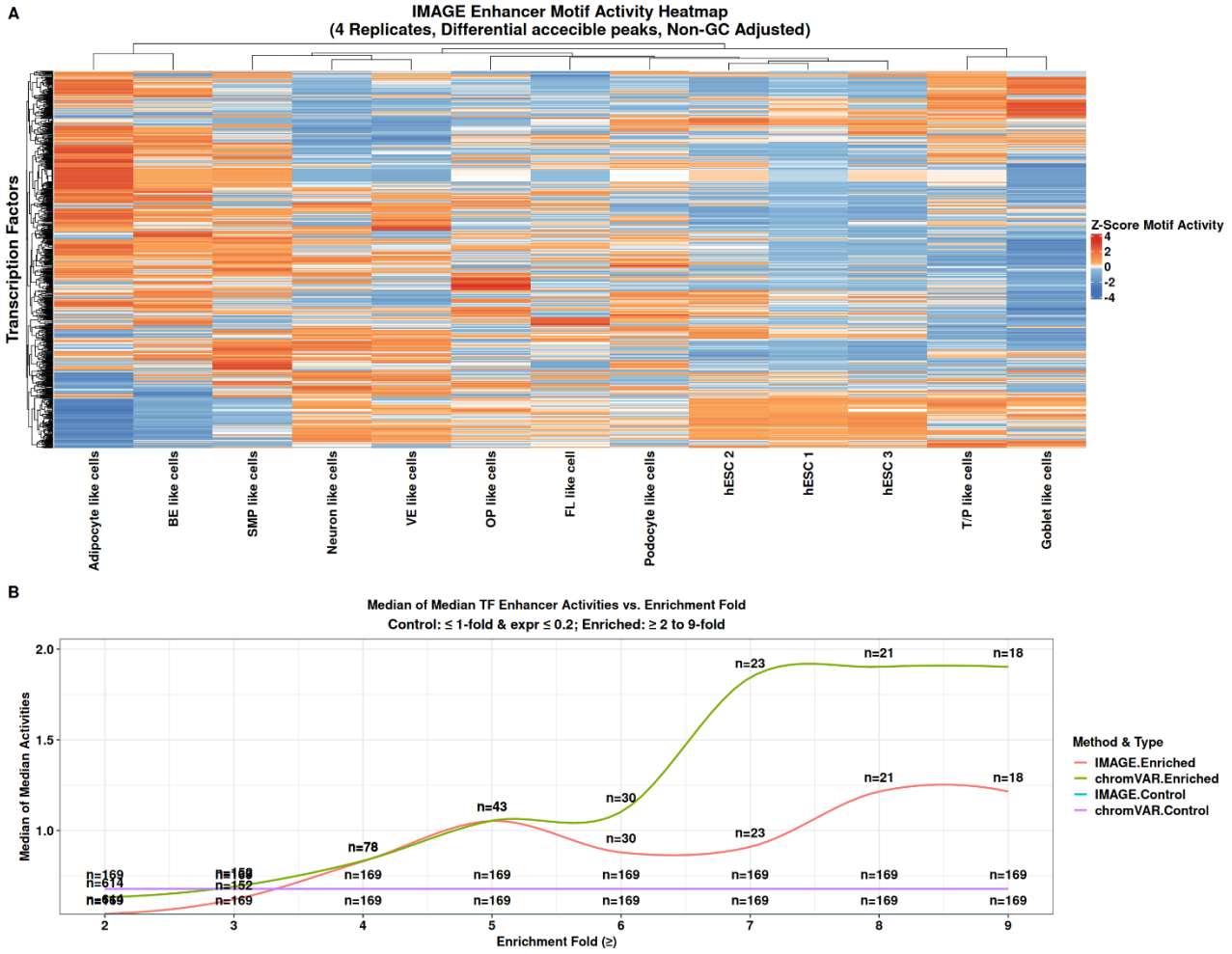
Supplementary Figure 13. Top 20 marker genes and gene set enrichment analysis of cluster 11. *(A) Violin plots display the top 20 positive marker genes of cluster 11, revealing log-normalised gene expression on the y-axis and the related cell clusters on the x-axis. The names of each marker gene are shown above the respective violin. (B) A bar plot presents the enriched pathways/terms for cluster 11, obtained from gene set enrichment analysis using the Human Gene Atlas database. In this plot, the y-axis represents the cell types associated with the top 20 differentially expressed positive marker genes from the Human Gene Atlas, while the x-axis indicates the -log10 transformed p-value, highlighting the significance of these genes in relation to the term. (C) Another bar plot highlights the enriched pathways or biological processes associated with the top 20 differentially expressed positive marker genes for cluster 11, utilising the GO Biological Process 2023 database. The y-axis shows the enriched pathway or term, while the x-axis displays the -log10 scaled p-value.*



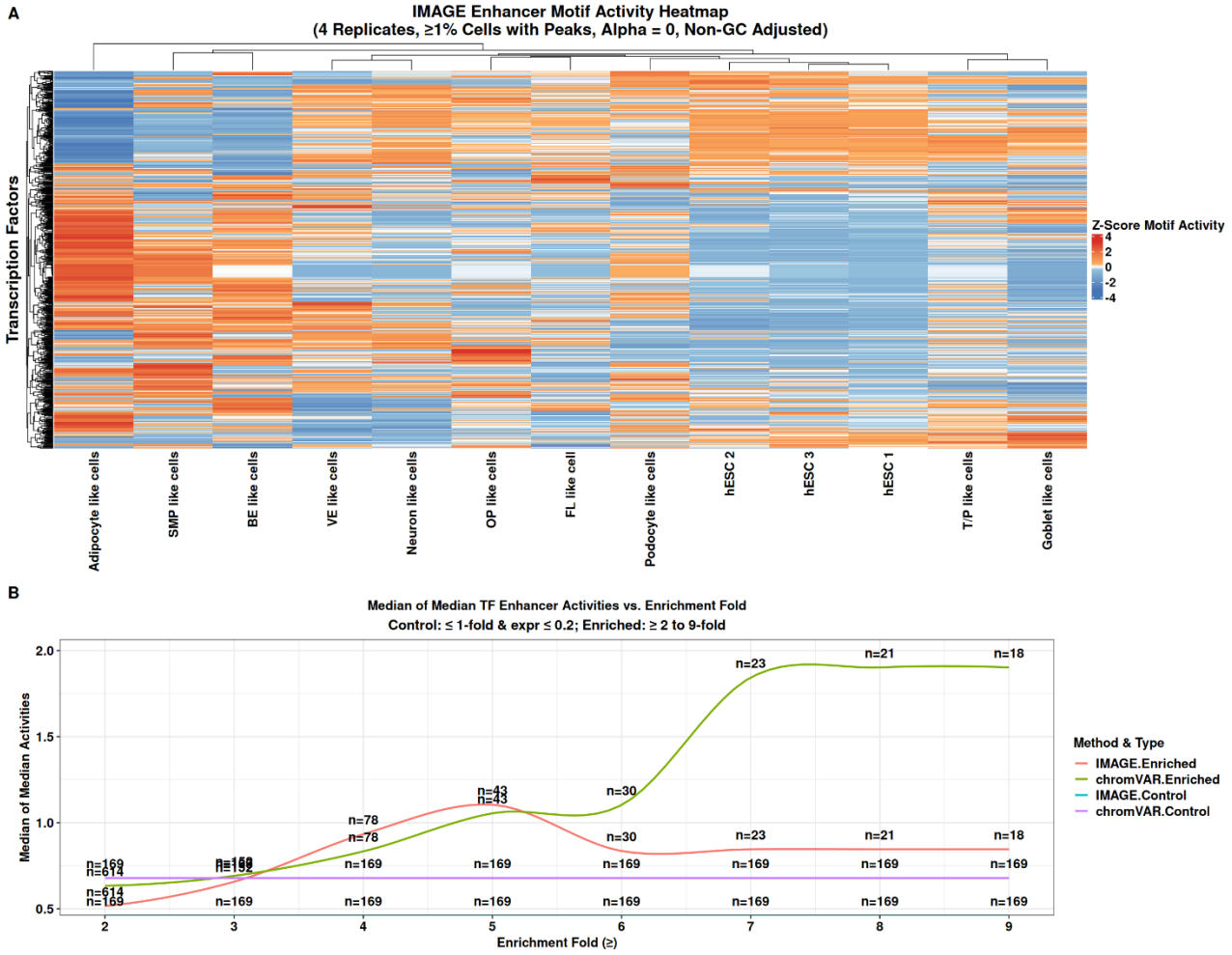
Supplementary Figure 14. Top 20 marker genes and gene set enrichment analysis of cluster 12. *(A) Violin plots of the top 20 positive marker genes of cluster 12, where the log-normalised gene expression is shown on the y-axis and the cell cluster in which they are expressed on the x-axis. The names of the marker genes are displayed above each violin plot. (B) Barplot of enriched pathways/terms in the gene set enrichment analysis using the human gene atlas database. On the y-axis, the cell type associated with the top 20 differentially expressed positive marker genes is displayed according to the Human Gene Atlas database, while the x-axis represents the -log10 transformed p-value, indicating the significance of the differentially expressed positive marker genes associated with the term for cluster 12. (C) Barplot of enriched pathways or biological processes associated with the top 20 differentially expressed positive marker genes for cluster 12 according to the GO Biological Process 2023 database. The enriched pathway/term is shown on the y-axis, while the x-axis shows the -log10 scaled p-value.*



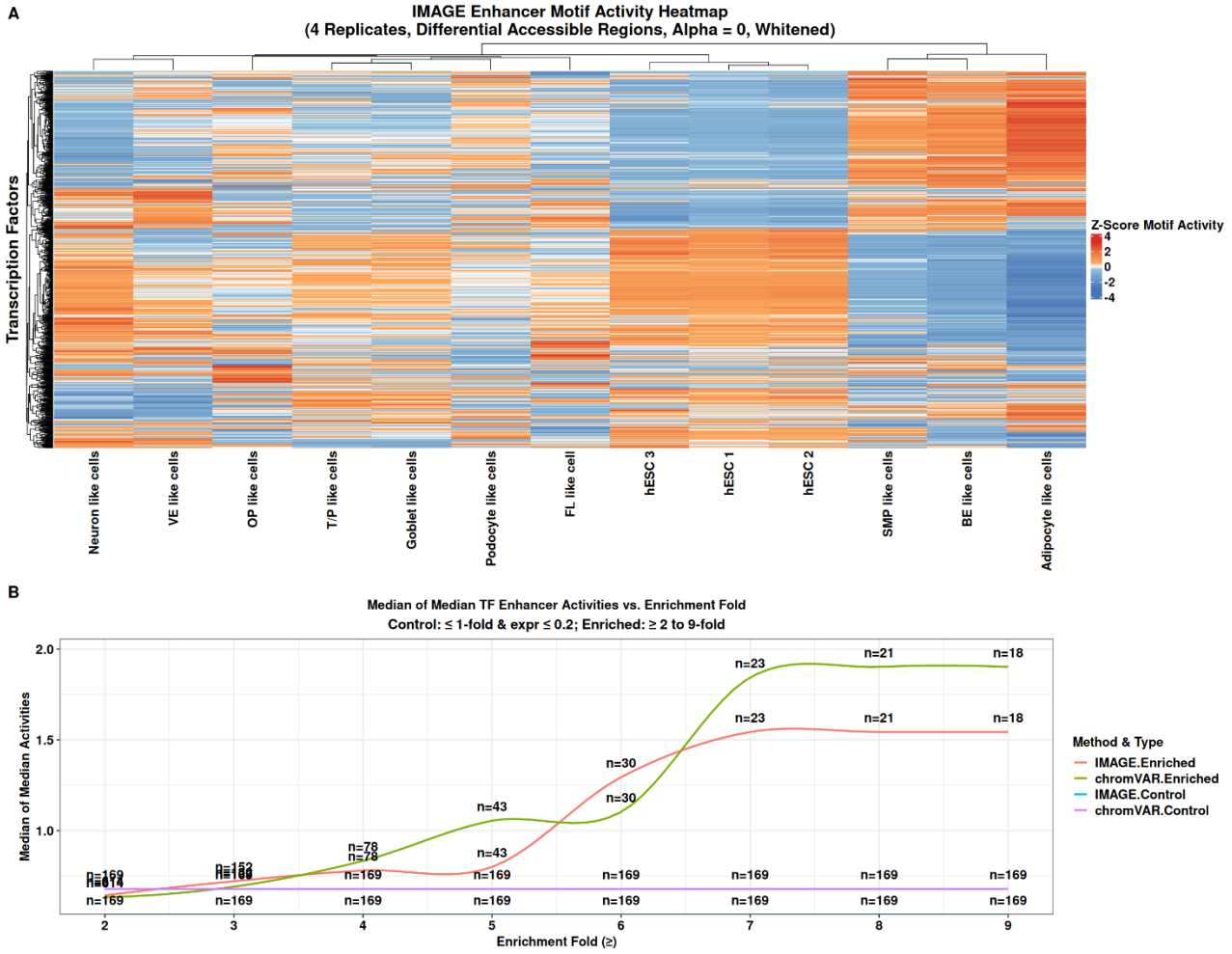
Supplementary Figure 15. ChromVAR motif activity for four replicates on all peaks. *This heatmap illustrates the predicted motif activities from chromVAR for all ATAC-seq peaks, aggregated over four replicates rather than at a single cell level. The x-axis shows each cell type, while the y-axis lists the transcription factors (TFs) (870 total). Each TF within each cell type is colour-coded based on the z-scored motif activity, displayed on the right side of the heatmap. A z-score reflects whether the motif activity is above or below the average for that particular motif, which is represented by a value of 0. Scores exceeding 0 signify higher than average activity, while scores below 0 indicate lower activity when compared across cell types for that specific TF motif. The dendrogram above the heatmap depicts the clustering of cell types based on their motif activities, while the dendrogram on the left shows the clustering of TFs according to their motif activities across various cell types.*



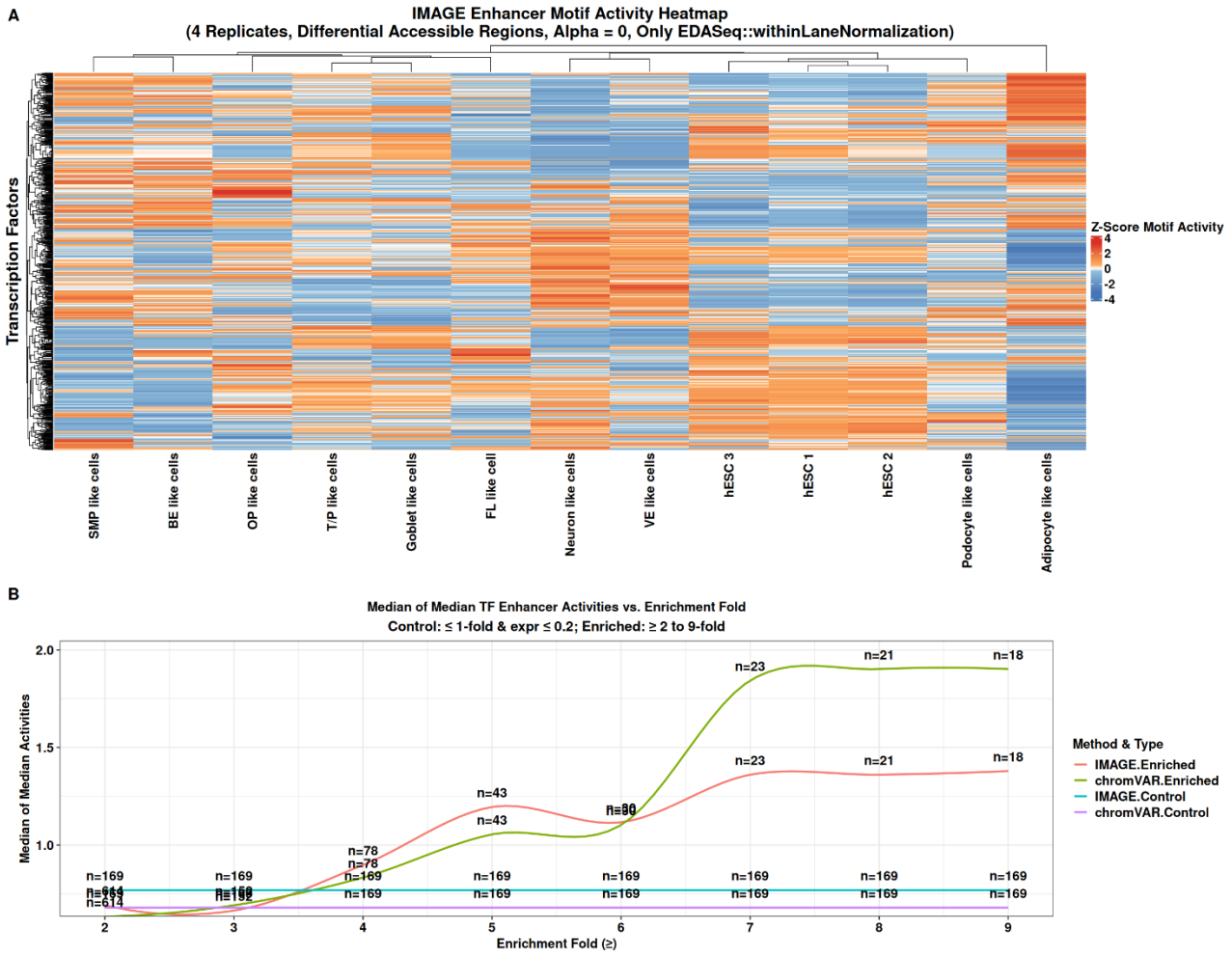
Supplementary Figure 16. Non-GC-adjusted enhancer motif activities for differentially accessible peaks and validation. *(A) The heatmap shows IMAGE-predicted enhancer motif activities, estimated using ridge regression (alpha = 0), for differential accessible peaks that haven't been GC-adjusted. The single-cell data was pseudobulked into four replicates per cell type. The x-axis represents the cell types, while the y-axis shows each transcription factor (TF) (870). For each TF, the enhancer motif activity is colour-coded according to the z-scored enhancer motif activity of IMAGE, as indicated on the right-hand side of the heatmap. The z-score indicates whether the motif's activity is above or below its average activity across cell types, with the average activity defined as a value of zero. The dendrograms on the rows cluster the TFs based on their activity across cell types, and the dendrogram at the top of each column clusters the cell types according to their motif activity for each TF. (B) A line plot of the median of median z-score enhancer motif activities across different enrichment thresholds. The x-axis shows an increasing fold enrichment, while the y-axis indicates the median of median z-score enhancer motif activity values. Control transcription factors (TFs) are defined as having a fold enrichment of 0 or below and having a normalised expression equal to or below 0,2, while the enriched TFs are defined by their enrichment thresholds as indicated on the x-axis. The number of TFs (n) included in the analysis for each enrichment threshold is shown on the line plot. Each line is colour-coded according to whether it represents control or enriched TFs, and whether the activities were calculated by IMAGE or chromVAR, as indicated in the legend on the right-hand side of the plot.*



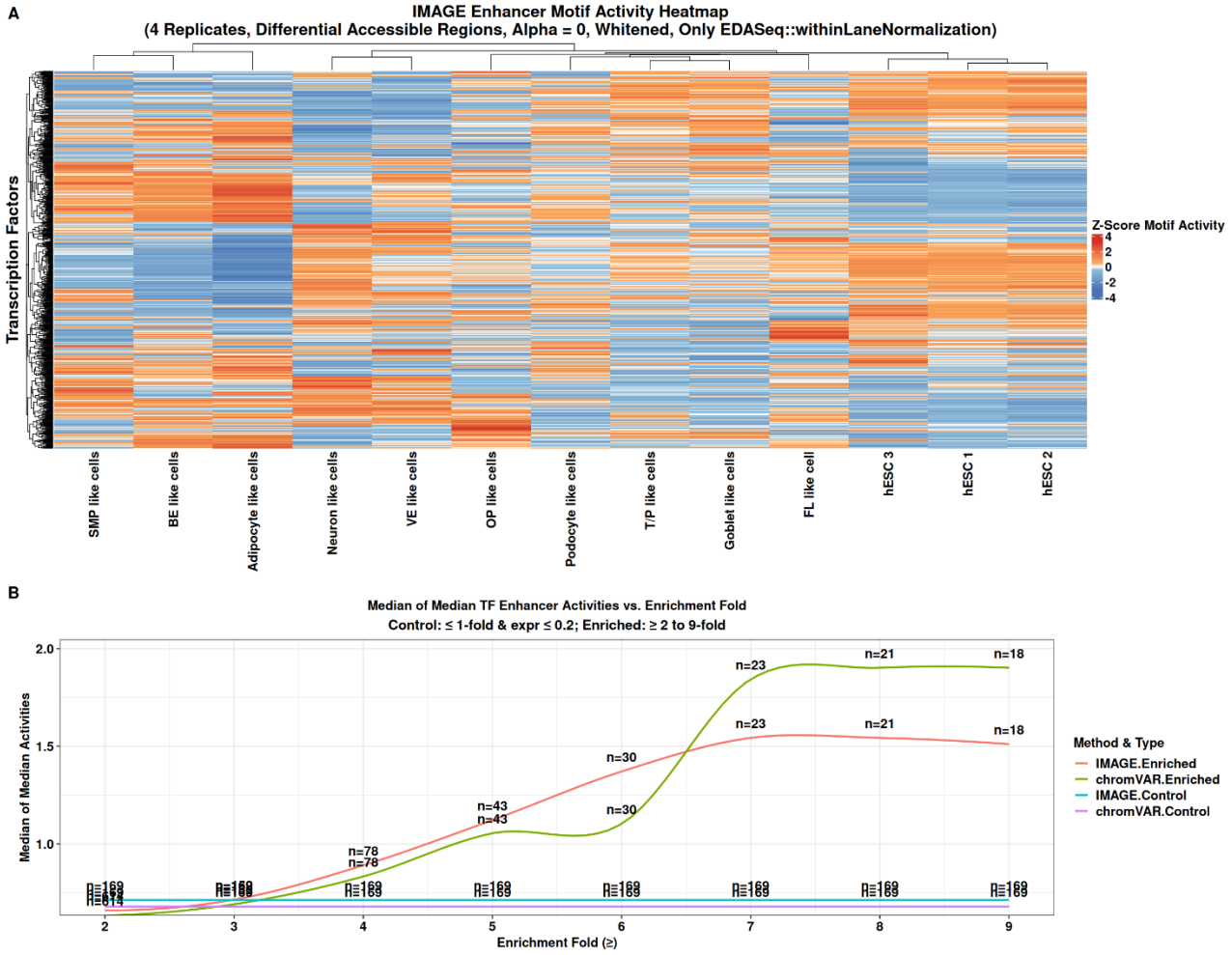
Supplementary Figure 17. Non-GC-adjusted enhancer motif activities for peaks detected in ≥1% of cells and validation. *(A) Panel A (A) Shows a heat-map of enhancer-motif activities predicted by IMAGE for peaks detected in ≥1% of cells that were not GC-corrected. These activities were inferred using ridge regression (α = 0). Single-cell profiles were combined into four pseudobulk replicates per cell type, which are arranged along the x-axis. The 870 transcription factors (TFs) are listed on the y-axis. For each TF-by-cell-type pair, the colour indicates the z-scored activity produced by IMAGE (scale bar at the right). A z-score of 0 represents the TF's mean activity across all cell types, while positive or negative values indicate activity above or below that mean. The row dendrograms cluster TFs with similar activity patterns, and the column dendrogram groups cell types with comparable motif-activity profiles. Panel B (B) Tracks how the median-of-medians z-scored motif activity changes as the motif-enrichment threshold increases. The x-axis shows the fold-enrichment cut-off, and the y-axis reports the resulting median-of-medians z-score. "Control" TFs have fold-enrichment ≤ 0 and normalised expression ≤ 0,2, while "enriched" TFs meet the progressively higher thresholds indicated on the x-axis. The number of TFs included at each threshold (n) is labelled on the curve. Line colours distinguish (i) control versus enriched groups and (ii) whether activities were computed with IMAGE or with chromVAR, as detailed in the legend.*



Supplementary Figure 18. Double normalised whitened enhancer motif activities for differentially accessible peaks and validation. *(A) Panel A shows a heatmap of whitened enhancer-motif activities predicted by IMAGE for double-normalised differentially accessible peaks. These activities were inferred using ridge regression (α = 0). Single-cell profiles were combined into four pseudobulk replicates per cell type along the x-axis. The 870 transcription factors (TFs) are listed on the y-axis. The colour for each TF-by-cell-type pair indicates the z-scored activity from IMAGE (scale bar at right). A z-score of 0 represents the mean activity across all cell types, while positive or negative values indicate activity above or below that mean. Row dendrograms cluster TFs with similar activity patterns, while column dendrograms group cell types with comparable motif-activity profiles. Panel B tracks changes in median-of-medians z-scored motif activity as the motif-enrichment threshold increases. The x-axis shows the fold-enrichment cut-off, and the y-axis reports the resulting median-of-medians z-score. "Control" TFs have fold-enrichment ≤ 0 and normalised expression ≤ 0,2, while "enriched" TFs meet higher thresholds on the x-axis. The number of TFs included at each threshold (n) is labelled on the curve. Line colours distinguish control versus enriched groups and whether activities were computed with IMAGE or chromVAR, as detailed in the legend.*



Supplementary Figure 19. IMAGE enhancer motif activity and validation for only within-lane GC normalisation. *Panel (A) shows a heat map of predicted enhancer-motif activities by IMAGE for the GC-corrected differentially accessible peaks (within lane normalisation only). Activities were estimated using ridge regression (α = 0) after pooling single-cell data into four pseudobulk replicates per cell type (x-axis). The y-axis lists 870 transcription factors (TFs). For each TF–cell-type pair, the colour indicates the z-scored IMAGE activity (see scale bar): 0 is the overall mean across cell types, with positive and negative values indicating higher- and lower-than-average activity, respectively. The row dendrogram groups TFs with similar activity patterns, while the column dendrogram clusters cell types with comparable motif-activity profiles. Panel (B) depicts the change in median-of-medians z-scored motif activity as the motif-enrichment cut-off increases. The x-axis represents the fold-enrichment threshold, and the y-axis shows the resulting median-of-medians z-score. Control TFs have fold-enrichment ≤ 0 and normalised expression ≤ 0.2; enriched TFs meet higher thresholds on the x-axis. Each point on the curve is labelled with the number of included TFS (n). Line colours differentiate control versus enriched sets and indicate whether activities were derived from IMAGE or chromVAR, as explained in the legend.*



Supplementary Figure 20. Only within lane normalised whitened enhancer motif activities for differentially accessible peaks and validation. *(A) Panel A displays a heatmap of whitened enhancer-motif activities predicted by IMAGE, limited to within-lane normalised differentially accessible peaks. These activities were inferred using ridge regression (α = 0). The x-axis shows single-cell profiles combined into four pseudobulk replicates per cell type, while the y-axis represents the 870 transcription factors (TFs). Each TF-by-cell-type pair's colour indicates the z-scored activity from IMAGE (scale bar at right). A z-score of 0 represents the mean activity across all cell types; positive or negative values indicate activity above or below that mean. Row dendrograms cluster TFs with similar activity patterns, and column dendrograms group cell types with comparable motif-activity profiles. Panel B tracks changes in median-of-medians z-scored motif activity as the motif-enrichment threshold increases. The x-axis shows the fold-enrichment cut-off, and the y-axis reports the resulting median-of-medians z-score. "Control" TFs have fold-enrichment ≤ 0 and normalised expression ≤ 0,2, while "enriched" TFs meet higher thresholds on the x-axis. The number of TFs included at each threshold (n) is labelled on the curve. Line colours distinguish control versus enriched groups and whether activities were computed with IMAGE or chromVAR, as detailed in the legend.*