https://doi.org/10.1093/nargab/lqac023

scREMOTE: Using multimodal single cell data to predict regulatory gene relationships and to build a computational cell reprogramming model

Andy Tran [1,2], Pengyi Yang [1,2,3], Jean Y.H. Yang [1,2] and John T. Ormerod [1,*]

[1] School of Mathematics and Statistics, The University of Sydney, Camperdown NSW 2006, Australia, [2] Charles Perkins Centre, The University of Sydney, Camperdown NSW 2006, Australia and [3] Children's Medical Research Institute, Westmead NSW 2145, Australia

Received June 25, 2021; Revised February 22, 2022; Editorial Decision February 25, 2022; Accepted March 10, 2022

ABSTRACT

Cell reprogramming offers a potential treatment to many diseases, by regenerating specialized somatic cells. Despite decades of research, discovering the transcription factors that promote cell reprogramming has largely been accomplished through trial and error, a time-consuming and costly method. A computational model for cell reprogramming, however, could guide the hypothesis formulation and experimental validation, to efficiently utilize time and resources. Current methods often cannot account for the heterogeneity observed in cell reprogramming, or they only make short-term predictions, without modelling the entire reprogramming process. Here, we present scREMOTE, a novel computational model for cell reprogramming that leverages single cell multiomics data, enabling a more holistic view of the regulatory mechanisms at cellular resolution. This is achieved by first identifying the regulatory potential of each transcription factor and gene to uncover regulatory relationships, then a regression model is built to estimate the effect of transcription factor perturbations. We show that scREMOTE successfully predicts the long-term effect of overexpressing two key transcription factors in hair follicle development by capturing higher-order gene regulations. Together, this demonstrates that integrating the multimodal processes governing gene regulation creates a more accurate model for cell reprogramming with significant potential to accelerate research in regenerative medicine.

INTRODUCTION

Cells generally begin their lives as a pluripotent stem cell that gradually differentiates into specialized cell fates over time. Once differentiated, cells usually have regulatory mechanisms to ensure that the cell maintains a stable state, reliably performing its required function. Recent advances in cell reprogramming have fundamentally altered our view of cell identity. Numerous experiments have established that overexpression of a few transcription factors (TFs) is sufficient to revert a differentiated cell to a pluripotent state or another specialized cell type.

These developments in cell reprogramming are significant for the field of regenerative medicine as it will facilitate the development of therapies to replenish cells our body can no longer produce. This opens the potential to regrow, repair or replace tissues and organs which may be damaged from age, disease, stress or trauma. For example, type 1 diabetes is the result of the loss of insulin-producing beta cells in the pancreas and recent experiments have shown that overexpressing the TFs *Pdx1* and *MafA* can reprogram pancreatic alpha cells into insulin-producing beta cells, effectively reversing type 1 diabetes in mice (1,2). Cell reprogramming has also been considered as a treatment for a wide variety of other diseases including Parkinson's disease (3,4), heart disease (5), spinal cord injury (6), macular degeneration (7), hearing loss (8), and aplastic anemia (9), among others.

Despite the overwhelming potential for cell reprogramming therapies to alleviate the world's disease burden, significant roadblocks have limited our ability to perform desired cell conversions. Cell reprogramming experiments are currently very slow and inefficient, taking several weeks and producing low quantities of the desired cell type (1%; (10)). Furthermore, reprogramming is often initiated by overexpressing a combination of TFs, but there is estimated to be more than 1,500 human TFs. Many successful combinations were historically determined through a trial and error approach which is time consuming and expensive, and may not even find an optimal combination (11).

These limitations for successful cell reprogramming could be addressed with a computational model to predict the outcome of a reprogramming experiment, even to some small degree of accuracy, which could guide the hypothe-

[*] To whom correspondence should be addressed. Email: john.ormerod@sydney.edu.au

ses to be experimentally validated. Advances in sequencing techniques have led to the generation of large multi-omics data sets that for the first time enable a systematic view into the regulatory processes in cells. In particular, this has facilitated the development of several computational methods for cell reprogramming, taking on a range of different approaches. These include differential expression (12–14), Boolean networks (15–17), dynamical systems (18,19), and regression (20). However, these methods have significant limitations. Firstly, most of them assume that the cell population is homogeneous and responds to perturbations in a fixed way. However, cell reprogramming experiments have resulted in very heterogeneous outcomes with many cell subpopulations, often dependent on the initial cell state (21,22). Furthermore, most methods are only able to make short term predictions of the effect of TF perturbations, which may not capture the entire cell reprogramming process which involves significant changes to the cell's identity. This motivates the need for a more holistic computational model for cell reprogramming.

Here, we present scREMOTE (single cell REprogramming MOdel Through cis-regulatory Elements), a computational method for cell reprogramming using data from simultaneous scRNA-seq and scATAC-seq. These data give us a more holistic view of the regulatory systems at the cellular level, allowing us to more accurately predict the downstream effect of the overexpression of transcription factors. We achieve this by first calculating a regulation potential, the ability of a TF to regulate a gene via cis-regulatory elements (CREs). We then build a linear regression model based on the gene expression and regulation potential, and demonstrate its applicability in predicting the effect of TF overexpression in murine hair follicle development. As the first model of its kind, using simultaneously sequenced multimodal data to model cell reprogramming, we also discuss the limitations of scREMOTE and avenues for future research.

MATERIALS AND METHODS

scREMOTE: a computational model to infer gene regulation and cell reprogramming

We present a novel computational model to infer gene regulation and cell reprogramming that leverages data from emerging multimodal single cell sequencing technologies. scREMOTE models four key components of gene regulation (Figure 1 A) as

(1) CRE accessibility, **A**, where TFs can only bind to regions of the genome that are accessible;
(2) TF motifs, **T**, where TFs need a matching motif in order to bind to a CRE;
(3) Chromatin conformation, **C**, where CREs need to be able to form a DNA loop with the promoter of the target gene; and
(4) Gene expression, **E**, which we expect to vary based on the previous three factors.

Ideally, we would want to measure all these components simultaneously in the same cell, but this is far beyond the capability of current single cell sequencing methods. Instead, we leverage on a series of recent techniques that are able to capture both the gene expression and CRE accessibility in the same cell (23–25). Fortunately, we can reasonably assume that the motifs a TF recognises remain the same between cells. Further, chromatin conformation will be relatively stable as it is restricted by physical constraints of the 3D genome organization into topologically associated domains (26).

The first step in scREMOTE is to estimate a regulation potential by integrating data from different modalities to model the regulation of each TF onto each gene through each CRE at the single cell level. Here, the regulatory effect of a TF onto a gene via a single CRE in a cell can be interpreted as the product of the three corresponding scores in **T**, **A** and **C**. That is for a regulatory potential to be positive, the CRE must be (1) enriched of the TF's motif, (2) accessible, and (3) able to form a DNA loop with the gene's promoter. We sum up the regulatory effect from all CREs to obtain an overall measure of regulation potential of a TF to a gene in a cell (Figure 1 B). It should be noted that the resulting array of regulation potentials will be rather sparse, as there are limited cases where all three conditions are met. See 'Regulation potential' for more details.

The second step of scREMOTE estimates how a cell will respond to a perturbation in TF expression. We achieve this by fitting a linear regression model with the cell's state, represented by its gene expression, as the response. We incorporate both the gene expression data and regulation potential into the predictor of our model (Figure 1 C) to better incorporate the multi-level nature of gene regulation. This way, in order for a coefficient to be significant, the TF requires both regulation potential and coexpression with the gene. The advantage of a linear model is that it allows for greater interpretability, and the estimated coefficients can be used for predictions (20). See 'Model fitting and evaluation' for more details.

The final step in scREMOTE is to perturb a TF's expression (or a combination of TF expressions), representing the process of TF overexpression, repression, or gene knockout. We incorporate the ability for pioneer transcription factors to open up inaccessible chromatin regions. This is because they have been strongly associated with the cell fate decision making process, as it allows more TFs to bind to the DNA, further regulating gene expression (27–29). Using the coefficients from the linear model, a change in TF expression would result in a change in the response, that is the gene expression. The predicted gene expression values of TFs can be refit into the model (Figure 1 D), calculating new values for the overall gene expression, and thus cell state. This process can iterated, representing the changes over time, until convergence, resulting in the final reprogrammed state.

Regulation potential

For a typical cell i, we can represent the regulatory potential, $\mathbf{R}_i$, a *gene* × *TF* matrix by

$$\mathbf{R}_i = \mathbf{C}\mathbf{A}_i\mathbf{T} \tag{1}$$

where $\mathbf{A}_i$ is a $CRE \times CRE$ matrix with the CRE accessibility scores for the ith cell along the diagonal, and ze-

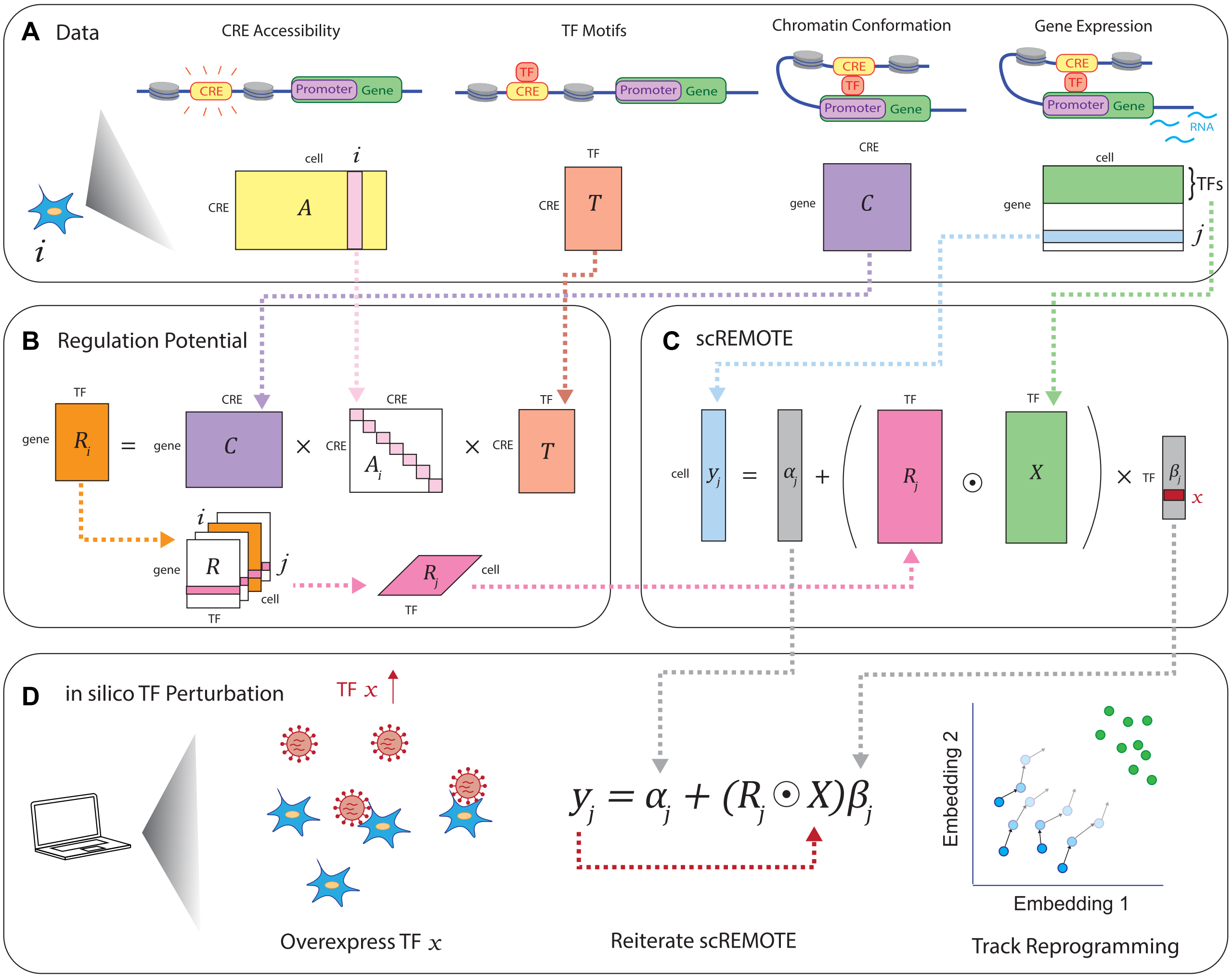

Figure 1. Schematic of scREMOTE. **(A)** The data inputs to scREMOTE. 1. CRE accessibility, a *CRE* × *cell* matrix, 2. TF motifs, a *CRE* × *TF* matrix, 3. Chromatin conformation, a *gene* × *CRE* matrix and 4. Gene expression, a *gene* × *cell* matrix, where the TFs are a subset of the genes. **(B)** Calculation of binding potential. The matrix A_i is created by placing the CRE accessibility scores for the *i*th cell along the diagonal, and zeroes elsewhere. This has the effect of summing the regulatory potential over all CREs. **(C)** Calculation of fitted coefficients. **(D)** In silico overexpression of TF *x*.

roes elsewhere. **C** and **T** can be either binary (representing the presence) or continuous (representing the degree) of chromatin conformation and TF motif enrichment respectively, where the continuous matrix would provide a more refined result when such information is available. This has the effect of summing the regulatory potential over all CREs. Calculating this for all cells, we end up with a *gene* × *TF* × *cell* array which we call **R** containing the regulation potential of all transcription factors to each gene in each cell.

To verify that our calculated regulation potential is capturing true regulatory relationships, we compared our calculated values to known TF-gene regulations from the following databases: TRRUST (30), hTFtarget (31), TFBSDB (32), RegNetwork (33) and MSigDB (34,35). As we filtered the data to the 1000 most highly expressed genes (see Data for details), we subsetted each data set to only those in our filtered list. We also consider a Combined database, which takes the union of interactions from the 5 other databases.

As our regulation potential is at a single cell resolution, we took the average over all cells to obtain a *gene* × *TF* matrix to compare it to these databases. If the regulation potential is accurate, we expect that the TF-gene regulations from the databases should have a greater regulation potential than a random subset of TF-gene regulations. By resampling 1 million random subsets of the same size as each database, we compute an empirical *p*-value as the probability that the mean regulation potential from a random sample is greater than the mean regulation potential of the database interactions. This is similar to the method used by Garcia-Alonso and colleagues, where gene expression is used as a reference to benchmark TF regulation databases (36). Here, our evaluation is in the reverse direction, using the databases as a ground truth to validate the regulation potential.

Model fitting and evaluation

For an individual gene *j*, we propose the following model to predict its expression.

$$\mathbf{y}_j = \alpha_j \mathbf{1} + (\mathbf{R}_j \odot \mathbf{X})\boldsymbol{\beta}_j + \boldsymbol{\varepsilon}_j, \tag{2}$$

where $\mathbf{y}_j$ is the gene expression of gene *j*, $\mathbf{X}$ is the gene expression of the TFs, and $\mathbf{R}_j$ is the regulation potential corresponding to gene *j*, α_j and $\boldsymbol{\beta}_j$ are the regression coefficients, $\boldsymbol{\varepsilon}_j$ is residual noise, and $\odot$ represents the Hadamard product (element-wise multiplication). Note that $\mathbf{R}_j$ can be interpreted as a slice corresponding to gene *j*, with dimensions *TF* × *cell* of the full regulation potential matrix $\mathbf{R}$ which has dimensions *gene* × *TF* × *cell* (Figure 1 B). This way, $\mathbf{R}_j$ can be interpreted as reweighting the gene expression values and the coefficients $\boldsymbol{\beta}_j$ represent the direct effect on gene *j*'s expression when perturbing a TF.

However, we found that in practice, the $\mathbf{R}_j$ matrix was extremely sparse, which could be attributed to the fact that each of the component matrices $\mathbf{T}$, $\mathbf{A}$, and $\mathbf{C}$ are already sparse due to the nature of sequencing techniques used. This causes the coefficients in (2) to have large bias in the fitted coefficients. To alleviate this issue, we consider adding a small constant to each entry in $\mathbf{R}_j$, similar to a pseudocount or fudge factor. This has the effect that when $\mathbf{R}_j$ values are all not available (i.e., 0), for a particular TF, the regression will rely on the available gene expression values only. However, if $\mathbf{R}_j$ values are available, i.e., not all 0 for a particular TF, then they will be incorporated into the regression. This gives us a new model

$$\mathbf{y}_j = \alpha_j \mathbf{1}_{n\times 1} + \left[(w\mathbf{1}_{n\times t} + (1-w)\mathbf{R}_j) \odot \mathbf{X}\right]\boldsymbol{\beta}_j + \boldsymbol{\varepsilon}_j, \tag{3}$$

where \$w\$ is a parameter that can be used to weight the influence of $\mathbf{R}_j$ when available. We choose to set \$w\$ = 0.1. We saw that when \$w\$ is small, there is minimal difference between different choices of \$w\$. This model was found to be the most effective when applied to experimental data, incorporating the regulation potential when available.

We compare our results to the Coexpression Model defined by:

$$\mathbf{y}_j = \alpha_j \mathbf{1} + \mathbf{X}\boldsymbol{\beta}_j + \boldsymbol{\varepsilon}_j, \tag{4}$$

which only uses gene expression data. When fitting the model in equation 3 and 4, we use ordinary least squares regression, using the `lm()` function in `R` (37). However, to address the situation of a TF regulating itself, that is when the response variable $\mathbf{y}_j$ is a TF, we set all the elements for the corresponding column in $\mathbf{X}$ to 0, as otherwise it would be perfectly equal to the response. We then change the fitted coefficient for the TF from 0 to 1, to encourage the TF expression level to stay constant when all other TF levels are constant, and so that external perturbations are fully captured by the model.

In silico perturbation with scREMOTE. To predict the effect of perturbing a TF (or a combination of TFs), we can perturb the values of $\mathbf{X}$, for example, by adding a constant to all values in the column corresponding to an overexpressed TF. The resulting changes in $\mathbf{y}_j$ represent the perturbed gene expression. To model the downstream effect of this perturbation, this process can be iterated, where the new values of the TF can be substituted into $\mathbf{X}$, which produces a new prediction for gene expression. This iterative process can be repeated for any number of time steps or until convergence. We chose to perform our simulations for 15 time steps, as this was generally enough iterations for perturbations to converge.

To incorporate the ability for pioneer transcription factors to open up inaccessible chromatin regions, we first consider the enhancer targets of the overexpressed TF. These are defined to be the CRE's with a score in the position weight matrix above a certain threshold, which we chose to be 0.2. We then increase the accessibility of these target CREs by adding a constant to the corresponding values in $\mathbf{R}_j$, and update the regulation potential, $\mathbf{R}_j$ in equation (3), before the perturbation.

To ensure the predicted gene expression values are biologically plausible, we can impose a minimum and/or maximum expression value. Here, any predicted expression below the minimum is replaced with the minimum value, and any predicted expression above the maximum is replaced with the maximum value. In practice, the minimum value will be 0, but the maximum would be harder to define since it would depend on the gene, cell type and sequencing depth. Some suggested ad hoc approaches for selecting a maximum could be the highest count observed in the entire data, or a few standard deviations above the mean for each gene. In our example, we imposed a minimum value of 0 and no maximum value, as predicted values seemed reasonable.

Marker gene analysis. Marker genes can be identified using a range of techniques, both supervised (38) and unsupervised (39,40). Since cell type labels are provided by Ma and colleagues, we choose to determine gene markers by using moderated *t*-tests implemented in the `limma` package (41) in `R`. By performing differential expression analysis between the IRS and Hair Shaft cells, we took the top three marker genes for each cell type ranked by *p*-value, but excluded *Gata3* and *Runx1* as they will be artificially overexpressed. We also consider the Spearman's rank correlation of the average expression for the 500 most highly expressed genes between the reprogrammed cells and the two target gene expressions. We should expect that a successful reprogramming model will cause the markers of the target cell type to increase in expression, and have a higher correlation with the target cell type.

Data

Chromatin conformation. The full mouse dataset was downloaded from the 4D Genome Database (42) on 21/10/2020. All coordinates were realigned from the mm9 genome to the mm10 genome using the LiftOver tool provided by the Human Genome Browser at UCSC (43). This list of chromatin interactions is filtered down to those which include gene promoters, determined as any interactions within 500bp of the transcription start site of a gene. Gene coordinates were downloaded from the Mouse Genome Informatics (MGI) website (44).

All chromatin regions which had an interaction with a promoter were considered a CRE. These regions were sorted into bins of length 1000bp which is now taken as our CRE list. We then construct **C**, as a binary matrix indicating a recorded connection between a CRE and a gene. Usually the chromatin conformation would be measured as a score representing the strength of the connection. However, as we are using a database, the data comes from many different experiments which would not be comparable. Our final matrix **C** is a *gene* × *CRE* matrix.

TF motifs. The affinity for a TF to bind to a CRE could be estimated using TF motif enrichment or ChIP-seq data. We chose to estimate **T** using TF motif enrichment, as TF motif data is readily available on the JASPAR database (45), whereas ChIP-seq databases can often be sparse and noisy (13). We note that only TFs present in these databases would be used in the analysis. If one wishes to include a particular TF (perhaps one that may be important for cell fate determination) which is not in the database, a baseline TF binding affinity or an estimate from ChIP-seq data may be used instead.

In our example, TF motifs were downloaded from the JASPAR database (8th release, 2020) (45) on 25/07/2020, using the full vertebrates position frequency matrices. From the CRE coordinates identified previously, the genomic sequences of our CREs were obtained using the `BSgenome.Mmusculus.UCSC.mm10` package on Bioconductor. TF motif enrichment was performed on each sequence using the AME function in the MEME Suite collection (46) with default settings. Only TFs that were highly enriched (marked as true positives) were kept, and their Position Weight Matrix score was normalized by dividing by the maximum value, so all values are between 0 and 1. This is then used as the corresponding value in **T**, a *CRE* × *TF* matrix.

scRNA-seq and scATAC-seq. The simultaneous scRNA-seq and scATAC-seq data with cell type labels from the SHARE-seq protocol was obtained from the authors upon request (S. Ma, personal communication, 23 September, 2020) (23). This data set is now available from GEO (Accession number: GSE140203). Due to the sparsity of the gene expression data, we only used the 1000 most highly expressed genes which were then $\log(x + 1)$ transformed. Our gene expression matrix is a *gene* × *cell* matrix.

The scATAC-seq data was then realigned to match our new CRE list in 1000bp bins. As the bin cutoffs did not match exactly, any observed scATAC-seq measurement that overlapped with our 1000bp CREs was considered as a count. Applying this criteria to all our CREs gives us **A**, a *CRE* × *cell* matrix.

We subsetted the data to only contain the cell types of interest: Hair Shaft (cuticle/cortex) cells, IRS cells, and two populations of TACs which were combined. Due to a large imbalance in the numbers of each cell type, we subsampled the the larger cell types so that they all have the same size. We repeated the in-silico cell reprogramming using a range of different subsamples, and observed similar results (Supplementary Figures S1, S2). All preprocessing steps and analysis were done in `R` and the code is available at https://github.com/SydneyBioX/scREMOTE.

Visualization

PCA is the dimension reduction technique in this study for the visualization of all simulation results as we believe it is more appropriate than other sophisticated dimension reduction techniques like tSNE or UMAP commonly used in single cell research (47). This is because PCA enables the projection of simulated data onto the same embedding as the original data, which can be used to track the changes over time. Other methods like tSNE or UMAP do not allow additional points to be projected into the embedding, unless the entire embedding is recalculated at each time point, but this would cause all cells to change positions, which cannot track the effect over time as in Supplementary Figures S1–S2. As expected, we found that PC1 is strongly correlated with the total read count in each cell (Figure 2 B), and it does not help to distinguish between different cell clusters (Figure 2 A). However, the combination of PC2 and PC3 shows a clear separation between the TAC, IRS and Hair Shaft clusters (Figure 2 C) so we use this for all visualizations.

RESULTS

Regulation potential captures transcription factor to gene regulations

To demonstrate the applicability of scREMOTE, we investigate the hair follicle developmental system, as it is a natural differentiation system in the adult skin and has been profiled with simultaneous scRNA-seq and scATAC-seq (23). Furthermore, the ability to reprogram hair cells in the inner ear could be used as a cure to permanent hearing loss (48). We chose to estimate **T** using TF motif enrichment and we estimated **C** using data from the 4D Genome Database (42). See *Materials and Methods* for more details.

We first verify that our calculated regulation potential is capturing true regulatory relationships, by comparing our calculated values to known TF-gene regulations. There are a variety of databases that record known and predicted regulations, such as TRRUST (30), hTFtarget (31), TFBSDB (32), RegNetwork (33) and MSigDB (34,35). As expected, we find that the TF-gene regulations from these databases have a greater regulation potential than random subsets of TF-gene pairs. By resampling 1 million random subsets for each database, we see that all databases are significantly enriched with interactions containing a high regulation potential, implying that our regulation potential captures true regulatory TF-gene relationships (Table 1).

scREMOTE predicts the outcome of cell reprogramming experiments

We now show how scREMOTE can be used to perform an in silico TF overexpression experiment. In the hair follicle developmental system, Transit-Amplifying Cells (TACs) differentiate into either the Inner Root Sheath (IRS) or Hair Shaft lineages. *Gata3* has long been identified as a reprogramming TF for the IRS lineage (49,50) and also *Runx1* for the Hair Shaft lineage (51–53). Thus, we expect that an accurate model will predict that an overexpression of *Gata3* will reprogram the TACs towards the IRS cells (Figure 3

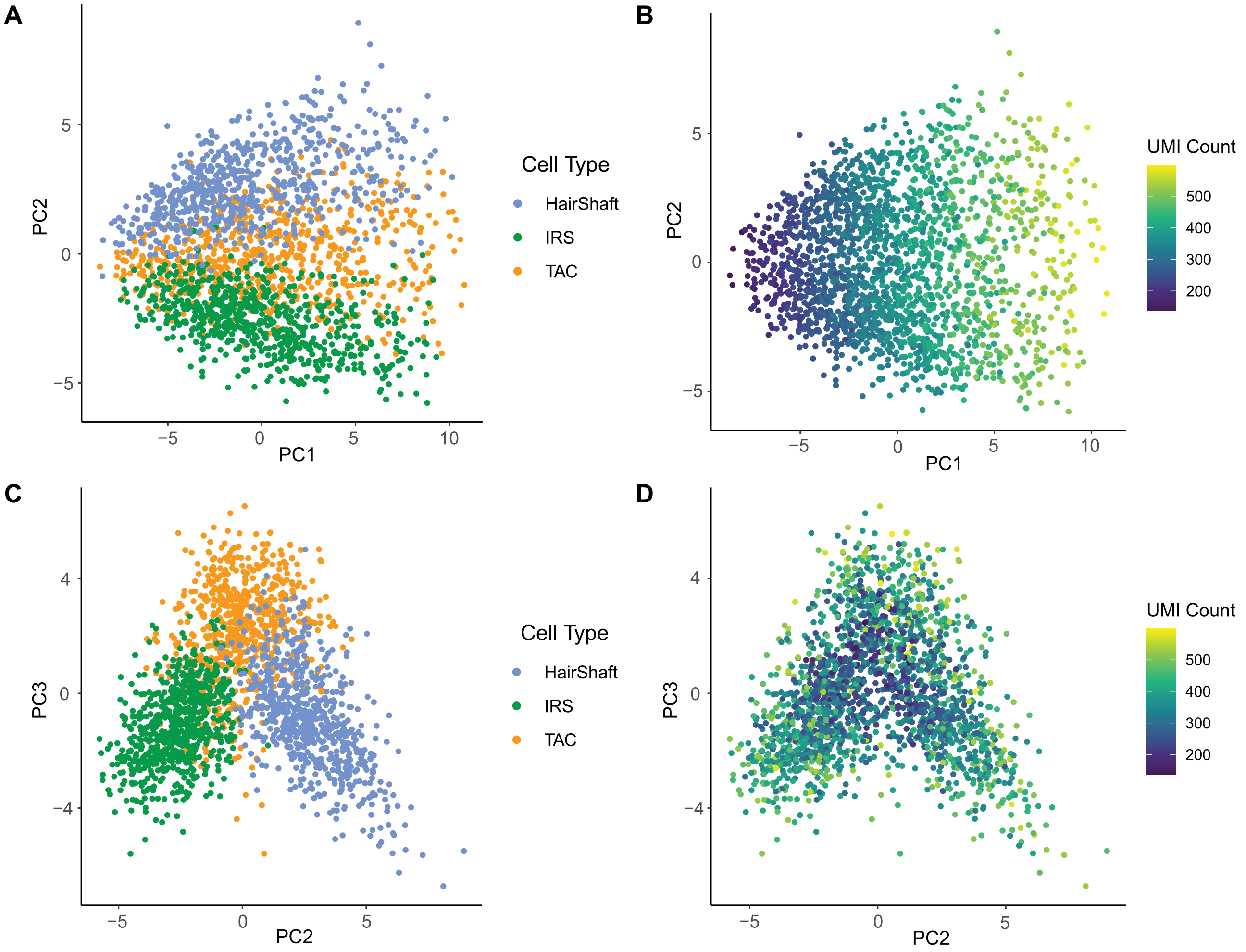

Figure 2. Data visualization with PCA. **(A)** Plot with PC1 and PC2, colored by cell type. **(B)** Plot with PC1 and PC2, colored by read count. **(C)** Plot with PC2 and PC3, colored by cell type. **(D)** Plot with PC2 and PC3, colored by read count.

Table 1. Results of testing regulation potential with TF-gene regulation databases. The empirical *p*-value is calculated to be the probability that the mean regulation potential from a random sample (of the same size as the database) of TF-gene pairs is greater than those from the database

Database	Number of Regulations	Empirical p-value
TRRUST	43	0.03
hTFtarget	10294	$<1 \times 10^{-6}$
TFBSDB	5253	$<1 \times 10^{-6}$
RegNetwork	729	3.9×10^{-5}
MSigDB	695	$<1 \times 10^{-6}$
Combined	14112	$<1 \times 10^{-6}$

A), and that an overexpression of *Runx1* will reprogram the TACs towards the Hair Shaft cells (Figure 3 B). If the perturbed cells converge to a different cluster, the reprogramming may have been unsuccessful, or the model may not have captured all the regulatory dynamics in the cell. To demonstrate the value of multimodal data in scREMOTE, we compare it to an equivalent model which only uses gene expression, which we call the Coexpression Model, as it would only detect linear coexpression patterns.

Figure 3 reveals the prediction of both scREMOTE and the Coexpression Model for overexpressing *Gata3* (Figure 3 C) and *Runx1* (Figure 3 D) in the TACs. We can see that when *Gata3* is overexpressed, both scREMOTE and the Coexpression Model make accurate short term predictions (early and middle time points), perturbing the cells towards the IRS cell fate. However, we see that only scREMOTE produces an accurate long term prediction (late time point). Likewise, when *Runx1* is overexpressed, we can see that both models make accurate short term predictions, perturbing the cells towards the Hair Shaft cell fate. But again, we see that only scREMOTE produces an accurate long term prediction. We believe that this is because scREMOTE is, to some extent, capturing the regulatory dynamics driving cell reprogramming whereas the Coexpression Model is limited by the highly correlated nature of gene expression. Animations of the entire reprogramming process can be found in Supplementary Figures S1– S2. We see that in some cases, reprogrammed cells converge to the expected cell cluster, however occasionally, the cells are perturbed in the right direction, but a distinct cluster is formed. We believe that this is because scREMOTE has not captured the entire regulatory dynamics, involving undetected TFs and genes, and

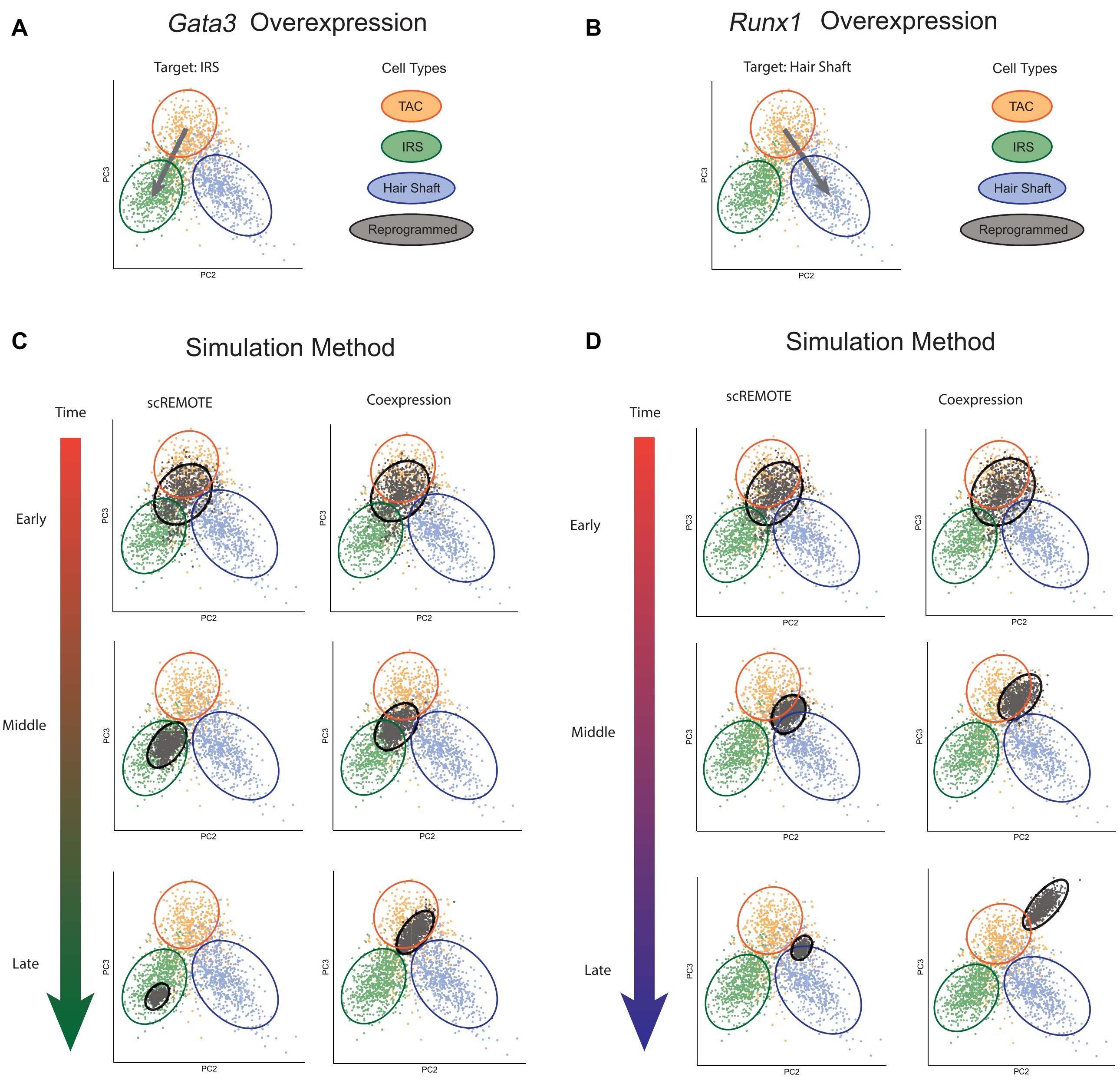

Figure 3. Perturbation of TFs in hair follicle development. **(A)** Expected result of *Gata3* overexpression. **(B)** Expected result of *Runx1* overexpression. **(C)** Simulated result of *Gata3* overexpression with both scREMOTE and Coexpression Model. **(D)** Simulated result of *Runx1* overexpression with both scREMOTE and Coexpression Model.

other factors like microRNAs and DNA methylation. This means that the predicted cell cluster may not match the true cell cluster as seen in the data.

Computationally reprogrammed cells show markers of cell identity

To verify the fidelity of the scREMOTE predicted cell state to the true target cell state, we tracked the expression of several marker genes over time. Here, we should expect that successful reprogramming will have the markers of the target cell type increase in expression and the markers of the opposing cell type to potentially decrease in expression.

In the overexpression of *Gata3* (Figure 4 A), we see that the IRS markers all increase in expression but the hair shaft marker *Lef1* decreases, and the other hair shaft markers *Trps1* and *Kcnh1* increase to a small extent. We also show that the average gene expression of the reprogrammed cells have a higher Spearman's rank correlation with the IRS cells compared to the Hair Shaft cells ($\rho = 0.63$ vs $\rho = 0.36$). Similarly, in the overexpression of *Runx1* (Figure 4 B), we see that all Hair Shaft markers increase in expression and all IRS markers decrease in expression and reprogrammed cells have a higher Spearman's rank correlation with the Hair Shaft cells compared to the IRS cells ($\rho = 0.35$ vs $\rho = 0.26$). This demonstrates that the predicted cell state from scRE-

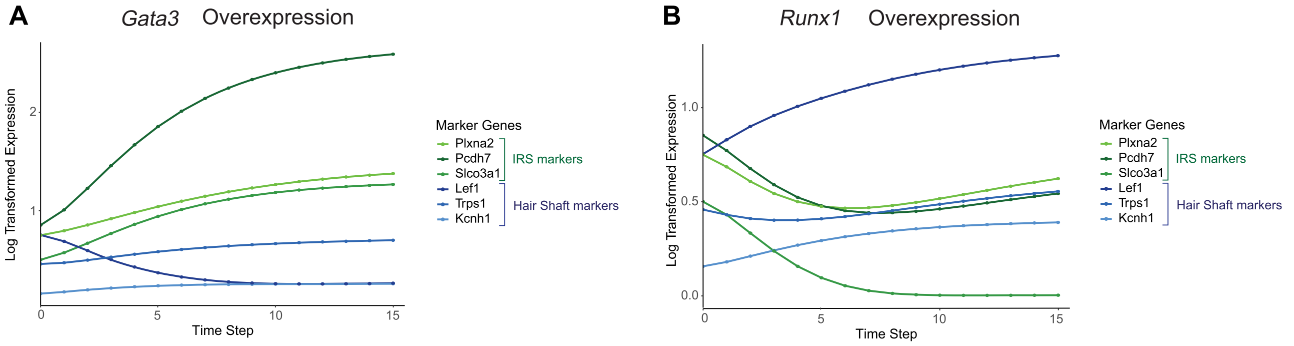

Figure 4. Tracking marker genes. **(A)** Marker genes during *Gata3* overexpression. Genes colored in green represent IRS markers and genes colored in blue represent Hair Shaft markers. **(B)** Marker genes during *Runx1* overexpression.

MOTE has an elevated expression of marker genes, verifying the validity of the simulated cell reprogramming.

Interestingly, in the *Runx1* overexpression, we observe a reversal in the expression of the IRS markers *Plxna2* and *Pcdh7* which decrease and then increase, and also for the Hair Shaft marker *Trps1* which decreases and then increases. This suggests that scREMOTE is able to capture, to some extent, higher order gene regulations where the downstream targets of *Runx1* are causing a reversal in the trend, allowing scREMOTE to make accurate long term predictions. In contrast, the Coexpression Model is only able to predict the immediate effects of the perturbation, which may not capture the downstream effects that result in successful cell conversion.

Further, we tested the ability of scREMOTE to model the overexpression of a combination of TFs, *Runx1* and *Lef1*. We consider this as *Lef1* has been implicated to be a driver of the Hair Shaft cell fate (23). However, we found that there was a minor difference in the outcome compared to the overexpression of *Runx1* alone (Supplementary Figure S4). We suspect that this may be due to the overexpression of *Runx1* already leading to the upregulation of *Lef1* (Figure 4 B), which is consistent with experimental literature (52,54).

DISCUSSION

Understanding the regulatory dynamics in a cell is a complex yet important challenge, especially in the context of cell reprogramming, which results in large changes to the cell's identity. Here, we present scREMOTE, a model for long-term predictions of TF perturbations at the single cell level that extends on existing algorithms (20). Integrating simultaneous scRNA-seq and scATAC-seq data provides a more comprehensive view of the regulatory dynamics occurring in each cell. By aggregating the regulatory effect through each CRE, a regulation potential is calculated between each TF and gene. We then combined the regulatory potential and TF expression to construct a linear model for gene expression. By iteratively updating gene expression, we are able to predict the long term effects of TF perturbation. We demonstrated scREMOTE on experimental data, revealing that it can successfully model the cell reprogramming process, and capture higher order levels of gene regulation.

scREMOTE, like any computational model, is based on a set of assumptions that, to varying degrees, reflect the underlying biology. Here, we approximate the regulatory effect of each TF to be linear and additive, whereas in reality, TFs often work in combinations and in complex relationships (55). Although, in our current implementation, we used ordinary least squares linear regression, the scREMOTE framework could be easily extended to include regularization like LASSO if there are too many TFs, or if multicollinearity is a concern. Furthermore, it could be extended to more advanced models, such as deep learning models like a multi-layer perceptron, possibly capturing non-linear relationships. The choice of these extended models will be motivated by the specific data structure. However, in our illustration, we found that after filtering, only 35 TFs were considered highly expressed. These TFs were not strongly correlated with each other (Supplementary Figure S3), and thus regularization is unlikely to have a significant impact for our data.

Multimodal single cell sequencing technologies are still in their infancy so there is very limited data to evaluate scREMOTE under a practical setting. For our validation, we required a source cell type that could be reprogrammed into at least one (but ideally more) target cell types which show clear separation when visualized with PCA, where the reprogramming is driven by different key TFs that are sufficiently expressed. Despite this limitation of our ability to validate the scREMOTE workflow more widely, we believe that the currently available data gives us a convincing example to support the applicability of scREMOTE. In particular, it will be challenging to evaluate the performance of scREMOTE in its ability to model lowly expressed TFs, which may be very important for cell reprogramming. This challenge is due to the bias in the estimated coefficients as a consequence of most high throughput sequencing technologies which leads to a high proportion of dropouts (56). To date, the single cell research community has employed a variety of imputation methods to provide a partial solution to this issue, but there is no consensus on the most appropriate or optimal approach (57). Going forward, with the constant improvement in these sequencing technologies and their increasing accessibility (24,58), we expect that in the future, cell reprogramming predictions from scREMOTE will become more applicable and extensions to the algorithm can be developed to produce more accurate and generalizable results.

The chromatin conformation data would ideally be measured in the same cells as the other modalities, however this is not feasible with current sequencing protocols. Also, chromosome conformation capture techniques like Hi-C are currently expensive to run with multiple complex experimental steps (59), and so it is difficult to obtain this type of data. Further complicating this component is that Hi-C has very low resolution (up to 10kb) (60) making it difficult to use for the precision required to model CRE-gene interactions. We bypassed these issues by using a database of measured interactions (42) from a range of chromatin conformation techniques which we used as a baseline measure for the chromatin conformation in all cells. As chromatin capture technologies improve and become cheaper, we will be able to collect more Hi-C data, extending the applicability of scREMOTE.

The TF motifs data is dependent on publicly available databases which are currently incomplete, for example the JASPAR database currently contains 592 profiles for mouse TFs out of an estimated 1640 (45). This limits the applicability of scREMOTE to model TFs which may be important for cell fate determination but whose binding profile has not yet been characterized. However, with the regular update of these databases (45), scREMOTE will have continued expansion of the number of TFs that could be incorporated.

In summary, our method is the first to our knowledge that simulates cell reprogramming experiments by modeling gene regulatory systems at the single cell level through the integration of matched scRNA-seq and scATAC-seq data. The ability of scREMOTE to model the biological mechanisms behind cell reprogramming at the single cell level would lead to its increased applicability over earlier methods. As the first model of its kind, we see large potential for the algorithm to be extended and improved, as data quality and availability increases. We hope that this will contribute to our understanding of the role of gene regulation in cell identity and accelerate research in regenerative medicine by allowing researchers to screen candidate combinations in silico before doing wet lab validation.

DATA AVAILABILITY

All code written in support of this publication is publicly available at https://github.com/SydneyBioX/scREMOTE.

SUPPLEMENTARY DATA

Supplementary Data are available at NARGAB online.

ACKNOWLEDGEMENTS

We thank all of our colleagues at the Sydney Precision Bioinformatics Alliance for their support and intellectual engagement.

FUNDING

The following sources of funding for each author, and for the manuscript preparation, are gratefully acknowledged: Australian Research Council Discovery Project grant (DP170100654) to JYHY and JO, National Health and Medical Research Council Investigator Grant (1173469) to PY, Australian Research Council Discovery Project grant (DP210100521) to JO, and Research Training Program Tuition Fee Offset to AT. The funding sources had no role in the study design; in the collection, analysis, and interpretation of data, in the writing of the manuscript, and in the decision to submit the manuscript for publication.

Conflict of interest statement. None declared.

REFERENCES

1. Xiao,X., Guo,P., Shiota,C., Zhang,T., Coudriet,G. M., Fischbach,S., Prasadan,K., Fusco,J., Ramachandran,S., Witkowski,P. *et al.* (2018) Endogenous Reprogramming of Alpha Cells into Beta Cells, Induced by Viral Gene Therapy, Reverses Autoimmune Diabetes. *Cell Stem Cell*, **22**, 78–90.
2. Furuyama,K., Chera,S., van Gurp,L., Oropeza,D., Ghila,L., Damond,N., Vethe,H., Paulo,J.A., Joosten,A.M., Berney,T. *et al.* (2019) Diabetes relief in mice by glucose-sensing insulin-secreting human α-cells. *Nature*, **567**, 43–48.
3. Barker,R.A., Parmar,M., Studer,L. and Takahashi,J. (2017) Human Trials of Stem Cell-Derived Dopamine Neurons for Parkinson's Disease: Dawn of a New Era. *Cell Stem Cell*, **21**, 569–573.
4. Parmar,M., Grealish,S. and Henchcliffe,C. (2020) The future of stem cell therapies for Parkinson disease. *Nat. Rev. Neurosci.*, **21**, 103–115.
5. Aguirre,A., Sancho-Martinez,I. and Izpisua Belmonte,J. (2013) Reprogramming toward Heart Regeneration: Stem Cells and Beyond. *Cell Stem Cell*, **12**, 275–284.
6. Khazaei,M., Ahuja,C.S. and Fehlings,M.G. (2017) Induced Pluripotent Stem Cells for Traumatic Spinal Cord Injury. *Front. Cell Dev. Biol.*, **4**, 152.
7. Chichagova,V., Hallam,D., Collin,J., Zerti,D., Dorgau,B., Felemban,M., Lako,M. and Steel,D.H. (2018) Cellular regeneration strategies for macular degeneration: past, present and future. *Eye*, **32**, 946–971.
8. Bermingham-McDonogh,O. and Reh,T. (2011) Regulated Reprogramming in the Regeneration of Sensory Receptor Cells. *Neuron*, **71**, 389–405.
9. Melguizo-Sanchis,D., Xu,Y., Taheem,D., Yu,M., Tilgner,K., Barta,T., Gassner,K., Anyfantis,G., Wan,T., Elango,R. *et al.* (2018) iPSC modeling of severe aplastic anemia reveals impaired differentiation and telomere shortening in blood progenitors. *Cell Death Dis.*, **9**, 128.
10. Omole,A.E. and Fakoya,A.O.J. (2018) Ten years of progress and promise of induced pluripotent stem cells: historical origins, characteristics, mechanisms, limitations, and potential applications. *PeerJ*, **6**, e4370.
11. Takahashi,K. and Yamanaka,S. (2016) A decade of transcription factor-mediated reprogramming to pluripotency. *Nat. Rev. Mol. Cell Biol.*, **17**, 183–193.
12. Rackham,O.J.L., Firas,J., Fang,H., Oates,M.E., Holmes,M.L., Knaupp,A.S., Suzuki,H., Nefzger,C.M., Daub,C.O., Shin,J.W. *et al.* (2016) A predictive computational framework for direct reprogramming between human cell types. *Nat. Genetics*, **48**, 331–335.
13. Qin,Q., Fan,J., Zheng,R., Wan,C., Mei,S., Wu,Q., Sun,H., Brown,M., Zhang,J., Meyer,C.A. *et al.* (2020) Lisa: inferring transcriptional regulators through integrative modeling of public chromatin accessibility and ChIP-seq data. *Genome Biol.*, **21**, 32.
14. Xu,Q., Georgiou,G., Frölich,S., van der Sande,M., Veenstra,G., Zhou,H. and van Heeringen,S. (2021) ANANSE: an enhancer network-based computational approach for predicting key transcription factors in cell fate determination. *Nucleic Acids Res.*, **49**, 7966–7985.
15. Lang,A.H., Li,H., Collins,J.J. and Mehta,P. (2014) Epigenetic Landscapes Explain Partially Reprogrammed Cells and Identify Key Reprogramming Genes. *PLOS Comput. Biol.*, **10**, e1003734.
16. Okawa,S., Nicklas,S., Zickenrott,S., Schwamborn,J. and del Sol,A. (2016) A Generalized Gene-Regulatory Network Model of Stem Cell Differentiation for Predicting Lineage Specifiers. *Stem Cell Rep.*, **7**, 307–315.

17. Heydari,T., Langley,M.A., Fisher,C., Aguilar-Hidalgo,D., Shukla,S., Yachie-Kinoshita,A., Hughes,M., McNagny,K.M. and Zandstra,P.W. (2021) IQCELL: A platform for predicting the effect of gene perturbations on developmental trajectories using single-cell RNA-seq data. bioRxiv doi: https://doi.org/10.1101/2021.04.01.438014, 03 April 2021, preprint: not peer reviewed.
18. Del Vecchio,D., Abdallah,H., Qian,Y. and Collins,J.J. (2017) A Blueprint for a Synthetic Genetic Feedback Controller to Reprogram Cell Fate. *Cell Systems*, **4**, 109–120.
19. Ronquist,S., Patterson,G., Muir,L.A., Lindsly,S., Chen,H., Brown,M., Wicha,M.S., Bloch,A., Brockett,R. and Rajapakse,I. (2017) Algorithm for cellular reprogramming. *Proc. Nat. Acad. Sci.*, **114**, 11832–11837.
20. Kamimoto,K., Hoffmann,C.M. and Morris,S.A. (2020) CellOracle: Dissecting cell identity via network inference and in silico gene perturbation. bioRxiv doi: https://doi.org/10.1101/2020.02.17.947416, 21 April 2020, preprint: not peer reviewed.
21. Weinreb,C., Rodriguez-Fraticelli,A., Camargo,F.D. and Klein,A.M. (2020) Lineage tracing on transcriptional landscapes links state to fate during differentiation. *Science*, **367**, 758–772
22. Kong,W., Biddy,B.A., Kamimoto,K., Amrute,J.M., Butka,E.G. and Morris,S.A. (2020) CellTagging: combinatorial indexing to simultaneously map lineage and identity at single-cell resolution. *Nat. Protoc.*, **15**, 750–772.
23. Ma,S., Zhang,B., LaFave,L.M., Earl,A.S., Chiang,Z., Hu,Y., Ding,J., Brack,A., Kartha,V.K., Tay,T. *et al.* (2020) Chromatin Potential Identified by Shared Single-Cell Profiling of RNA and Chromatin. *Cell*, **183**, 1103–1116.
24. Yan,R., Gu,C., You,D., Huang,Z., Qian,J., Yang,Q., Cheng,X., Zhang,L., Wang,H., Wang,P. *et al.* (2021) Decoding dynamic epigenetic landscapes in human oocytes using single-cell multi-omics sequencing. *Cell Stem Cell*, **28**, 1641–1656
25. Cao,J., Cusanovich,D.A., Ramani,V., Aghamirzaie,D., Pliner,H.A., Hill,A.J., Daza,R.M., McFaline-Figueroa,J.L., Packer,J.S., Christiansen,L. *et al.* (2018) Joint profiling of chromatin accessibility and gene expression in thousands of single cells. *Science*, **361**, 1380–1385.
26. McArthur,E. and Capra,J.A. (2021) Topologically associating domain boundaries that are stable across diverse cell types are evolutionarily constrained and enriched for heritability. *The American Journal of Human Genetics*, **108**, 269–283.
27. Zaret,K.S. and Carroll,J.S. (2011) Pioneer transcription factors: establishing competence for gene expression. *Genes Dev.*, **25**, 2227–2241.
28. Iwafuchi-Doi,M. and Zaret,K.S. (2014) Pioneer transcription factors in cell reprogramming. *Genes Dev.*, **28**, 2679–2692.
29. Mayran,A. and Drouin,J. (2018) Pioneer transcription factors shape the epigenetic landscape. *J. Biol. Chem.*, **293**, 13795–13804.
30. Han,H., Cho,J.-W., Lee,S., Yun,A., Kim,H., Bae,D., Yang,S., Kim,C.Y., Lee,M., Kim,E. *et al.* (2018) TRRUST v2: an expanded reference database of human and mouse transcriptional regulatory interactions. *Nucleic Acids Res.*, **46**, D380–D386.
31. Zhang,Q., Liu,W., Zhang,H.-M., Xie,G.-Y., Miao,Y.-R., Xia,M. and Guo,A.-Y. (2020) hTFtarget: A Comprehensive Database for Regulations of Human Transcription Factors and Their Targets. *Genomics, Proteomics & Bioinformatics*, **18**, 120–128.
32. Plaisier,C.L., O'Brien,S., Bernard,B., Reynolds,S., Simon,Z., Toledo,C.M., Ding,Y., Reiss,D.J., Paddison,P.J. and Baliga,N.S. (2016) Causal Mechanistic Regulatory Network for Glioblastoma Deciphered Using Systems Genetics Network Analysis. *Cell Syst.*, **3**, 172–186.
33. Liu,Z.-P., Wu,C., Miao,H. and Wu,H. (2015) RegNetwork: an integrated database of transcriptional and post-transcriptional regulatory networks in human and mouse. *Database*, **2015**, bav095.
34. Subramanian,A., Tamayo,P., Mootha,V.K., Mukherjee,S., Ebert,B.L., Gillette,M.A., Paulovich,A., Pomeroy,S.L., Golub,T.R., Lander,E.S. *et al.* (2005) Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Nat. Acad. Sci.*, **102**, 15545.
35. Kolmykov,S., Yevshin,I., Kulyashov,M., Sharipov,R., Kondrakhin,Y., Makeev,V.J., Kulakovskiy,I.V., Kel,A. and Kolpakov,F. (2021) GTRD: an integrated view of transcription regulation. *Nucleic Acids Res.*, **49**, D104–D111.
36. Garcia-Alonso,L., Holland,C.H., Ibrahim,M.M., Turei,D. and Saez-Rodriguez,J. (2019) Benchmark and integration of resources for the estimation of human transcription factor activities. *Genome Res.*, **29**, 1363–1375.
37. R Core Team (2015) *R: A Language and Environment for Statistical Computing R Foundation for Statistical Computing*. Vienna, Austria.
38. Costa-Silva,J., Domingues,D. and Lopes,F.M. (2017) RNA-Seq differential expression analysis: An extended review and a software tool. *PLOS ONE*, **12**, e0190152.
39. Townes,F.W., Hicks,S.C., Aryee,M.J. and Irizarry,R.A. (2019) Feature selection and dimension reduction for single-cell RNA-Seq based on a multinomial model. *Genome Biol.*, **20**, 295.
40. Su,K., Yu,T. and Wu,H. (2021) Accurate feature selection improves single-cell RNA-seq cell clustering. *Brief. Bioinformat.*, **22**, bbab034.
41. Ritchie,M.E., Phipson,B., Wu,D., Hu,Y., Law,C.W., Shi,W. and Smyth,G.K. (2015) limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res.*, **43**, e47.
42. Teng,L., He,B., Wang,J. and Tan,K. (2015) 4DGenome: a comprehensive database of chromatin interactions. *Bioinformatics*, **31**, 2560–2564.
43. Kent,W.J., Sugnet,C.W., Furey,T.S., Roskin,K.M., Pringle,T.H., Zahler,A.M. and Haussler,D. (2002) The human genome browser at UCSC. *Genome Res.*, **12**, 996–1006.
44. Eppig,J.T. (2017) Mouse Genome Informatics (MGI) Resource: Genetic, Genomic, and Biological Knowledgebase for the Laboratory Mouse. *ILAR J.*, **58**, 17–41.
45. Fornes,O., Castro-Mondragon,J.A., Khan,A., van der Lee,R., Zhang,X., Richmond,P.A., Modi,B.P., Correard,S., Gheorghe,M., Baranašiç,D. and et,al. (2019) JASPAR 2020: update of the open-access database of transcription factor binding profiles. *Nucleic Acids Res.*, **48**, D87–D92.
46. McLeay,R.C. and Bailey,T.L. (2010) Motif Enrichment Analysis: a unified framework and an evaluation on ChIP data. *BMC Bioinformatics*, **11**, 165.
47. Kulkarni,A., Anderson,A.G., Merullo,D.P. and Konopka,G. (2019) Beyond bulk: a review of single cell transcriptomics methodologies and applications. *Curr. Opin. Biotech.*, **58**, 129–136.
48. Menendez,L., Trecek,T., Gopalakrishnan,S., Tao,L., Markowitz,A.L., Yu,H.V., Wang,X., Llamas,J., Huang,C., Lee,J. *et al.* (2020) Generation of inner ear hair cells by direct lineage conversion of primary somatic cells. *eLife*, **9**, e55249.
49. Kaufman,C.K., Zhou,P., Pasolli,H.A., Rendl,M., Bolotin,D., Lim,K.C., Dai,X., Alegre,M.L. and Fuchs,E. (2003) GATA-3: an unexpected regulator of cell lineage determination in skin. *Genes Dev.*, **17**, 2108–2122.
50. Kurek,D., Garinis,G.A., van Doorninck,J.H., van der Wees,J. and Grosveld,F.G. (2007) Transcriptome and phenotypic analysis reveals Gata3-dependent signalling pathways in murine hair follicles. *Development*, **134**, 261–272.
51. Raveh,E., Cohen,S., Levanon,D., Negreanu,V., Groner,Y. and Gat,U. (2006) Dynamic expression of Runx1 in skin affects hair structure. *Mech. Develop.*, **123**, 842–850.
52. Osorio,K.M., Lee,S.E., McDermitt,D.J., Waghmare,S.K., Zhang,Y.V., Woo,H.N. and Tumbar,T. (2008) Runx1 modulates developmental, but not injury-driven, hair follicle stem cell activation. *Development*, **135**, 1059–1068.
53. Hoi,C.S.L., Lee,S.E., Lu,S.-Y., McDermitt,D.J., Osorio,K.M., Piskun,C.M., Peters,R.M., Paus,R. and Tumbar,T. (2010) Runx1 directly promotes proliferation of hair follicle stem cells and epithelial tumor formation in mouse skin. *Mol. Cell. Biol.*, **30**, 2518–2536.
54. Li,Q., Lai,Q., He,C., Fang,Y., Yan,Q., Zhang,Y., Wang,X., Gu,C., Wang,Y., Ye,L. *et al.* (2019) RUNX1 promotes tumour metastasis by activating the Wnt/β-catenin signalling pathway and EMT in colorectal cancer. *J. Exp. Clin. Cancer Res.*, **38**, 334.
55. Caramori,G., Nucera,F., Coppolino,I., Bello,F.L., Ruggeri,P., Ito,K., Di Stefano,A. and Adcock,I.M. (2020) Transcription Factors. In: *Reference Module in Biomedical Sciences*, Elsevier.
56. Wu,Y. and Zhang,K. (2020) Tools for the analysis of high-dimensional single-cell RNA sequencing data. *Nat. Rev. Nephrol.*, **16**, 408–421.
57. Hou,W., Ji,Z., Ji,H. and Hicks,S.C. (2020) A systematic evaluation of single-cell RNA-sequencing imputation methods. *Genome Biol.*, **21**, 218.

58. Bakken,T.E., Jorstad,N.L., Hu,Q., Lake,B.B., Tian,W., Kalmbach,B.E., Crow,M., Hodge,R.D., Krienen,F.M., Sorensen,S.A. *et al.* (2020) Evolution of cellular diversity in primary motor cortex of human, marmoset monkey, and mouse. bioRxiv doi: https://doi.org/10.1101/2020.03.31.016972, 01 April 2020, preprint: not peer reviewed.
59. Yardımcı,G.G., Ozadam,H., Sauria,M.E.G., Ursu,O., Yan,K.-K., Yang,T., Chakraborty,A., Kaul,A., Lajoie,B.R., Song,F. *et al.* (2019) Measuring the reproducibility and quality of Hi-C data. *Genome Biol.*, **20**, 57.
60. Hong,H., Jiang,S., Li,H., Du,G., Sun,Y., Tao,H., Quan,C., Zhao,C., Li,R., Li,W. *et al.* (2020) DeepHiC: A generative adversarial network for enhancing Hi-C data resolution. *PLOS Comput. Biol.*, **16**, e1007287.