Computational workflow for predicting transcriptional modulation to achieve therapeutically desired cellular conversion for regenerative medicine

Avani Mahadik and Vivek Singh[$]

avani.mahadik@thoughtworks.com, vivek.singh1@thoughtworks.com

Engineering for Research
Thoughtworks Technologies, Pune, India

[$] *Corresponding author*

Abstract

Regenerative medicine aims to develop methods for treating currently incurable diseases and pathological conditions. Its central idea is to leverage healthy cells to regenerate diseased or damaged cells, tissues or organs through the process of cellular conversion. The most common method to achieve this is by modulating the activity of specific transcription factors, which in turn, alters the transcriptional program of the cells resulting in their conversion to the desired target cell type. However, given the large number of protein-coding genes (~19,000) and transcription factors (~1,600) in humans and their complex interactions, it is challenging to identify the most suitable transcription factors for modulation.

In this study, we leverage three existing computational tools, namely, TransSynW, PAGA and SIGNET, to develop an analysis workflow to facilitate identifying suitable transcription factors for achieving desired cellular conversion. The pipeline helps in generating hypotheses on suitable transcription factors for modulation and also in inferring their mechanistic basis. We used this pipeline on a sample dataset of human hindbrain neuroepithelial stem cells (hNES) and midbrain medial floorplate progenitor (hProgFPM) cells to generate hypothesis for converting the former to the latter cell type.

The pipeline predicted the transcription factors for cellular conversion, highlighted their differential expression dynamics in starting and target cell types, and revealed their interactions and influence on the predicted gene regulatory network of the target cells. We believe these together can help researchers generate mechanistically founded hypotheses for achieving desired cellular conversions towards regenerative medicine.

Introduction

Regenerative medicine is an evolving branch of medicine aiming to address diseases without curative treatments [1]–[4].The central approach in regenerative medicine is to repair or replace the damaged and diseased cells, tissues and organs with healthy ones [5] .

The first step towards generation of complex tissues and organs is the generation of healthy cells of specific types. This can potentially be achieved by cellular conversion through: (1) directed cellular differentiation of pluripotent cells occurring naturally (e.g. embryonic or adult stem cells) [6], [7] , or developed artificially (e.g. iPSCs) [8], [9] ; and (2) by transdifferentiation of somatic cells directly to the target cells [10] .

Cells of a particular type maintain their identity and behavioral properties with the help of dynamic activities of transcription factors (TFs) and their corresponding target genes [11] . In line with this, cellular conversion from one type to another is shown to be achieved by modulating the activities of specific transcription factors [12], [13]. However, the challenge lies in identifying the correct TFs for modulation given that there are estimated 19,000 protein-coding genes [14] and ~1,600 TFs in humans [15] . Therefore, TF modulation for cellular conversions is typically based on prior knowledge, empirical intuition or purely trial and error, resulting in low efficiency of conversion [9] , and tumorigenicity [1] and immune rejection [1], [4] of the target cells in the body.

Technologies such as single cell RNA sequencing (scRNA-seq) provide transcriptome information at single cell resolution [16]. This offers a means to understand the transcriptional signatures of different cell types as well as their intermediate transient cellular states during cellular interconversion [17] and thus, help in identifying TFs whose modulation can bring about the desired cellular conversions [18] . In fact, several computational tools exist (e.g. TransSynW [19] , CellCartographer [20] , NETISCE [21] and scREMOTE [22]) that utilize scRNA alone or in combination with complementary biological data to predict the TF modulation required for desired cellular conversion. However, the hypotheses on TF modulation alone without understanding their molecular basis may be a black box, and hinder further experimentation.

To gain holistic understanding of TF modulation, their mechanistic basis and the downstream effects, existing computational tools can be leveraged in the form of analysis workflows. In this study, we propose one such analysis workflow to generate hypotheses on TF modulation for cellular conversion and understand their mechanistic basis with respect to the transcriptional regulatory network.

We demonstrate the usefulness of this workflow using the sample dataset used by Okawa *et al.* [23]. The dataset corresponds to human hindbrain hNES (Neuroepithelial stem cells) cell line (SAI2) [24] and human midbrain medial floorplate progenitor (hProgFPM) cells. The former was used as the starting cell type and the latter as the desired target cell type, thereby attempting cellular conversion from hindbrain progenitor cell line to a midbrain progenitor cell type.

Results and Discussion

We propose an analysis workflow consisting of TransSynW [19] for generation of hypotheses on TF modulation for cellular conversion, PAGA [25] for understanding the TransSynW predicted genes in the context of cellular conversion process from starting cell type to the target cell type, and SIGNET [26] to understand the gene modulation in the context of gene regulatory network of the target cell type, and gain insight into the mechanistic basis of the gene modulation hypothesis (see Figure 1). We detail this workflow below using the sample dataset of hNES to hProgFPM conversion as an example.

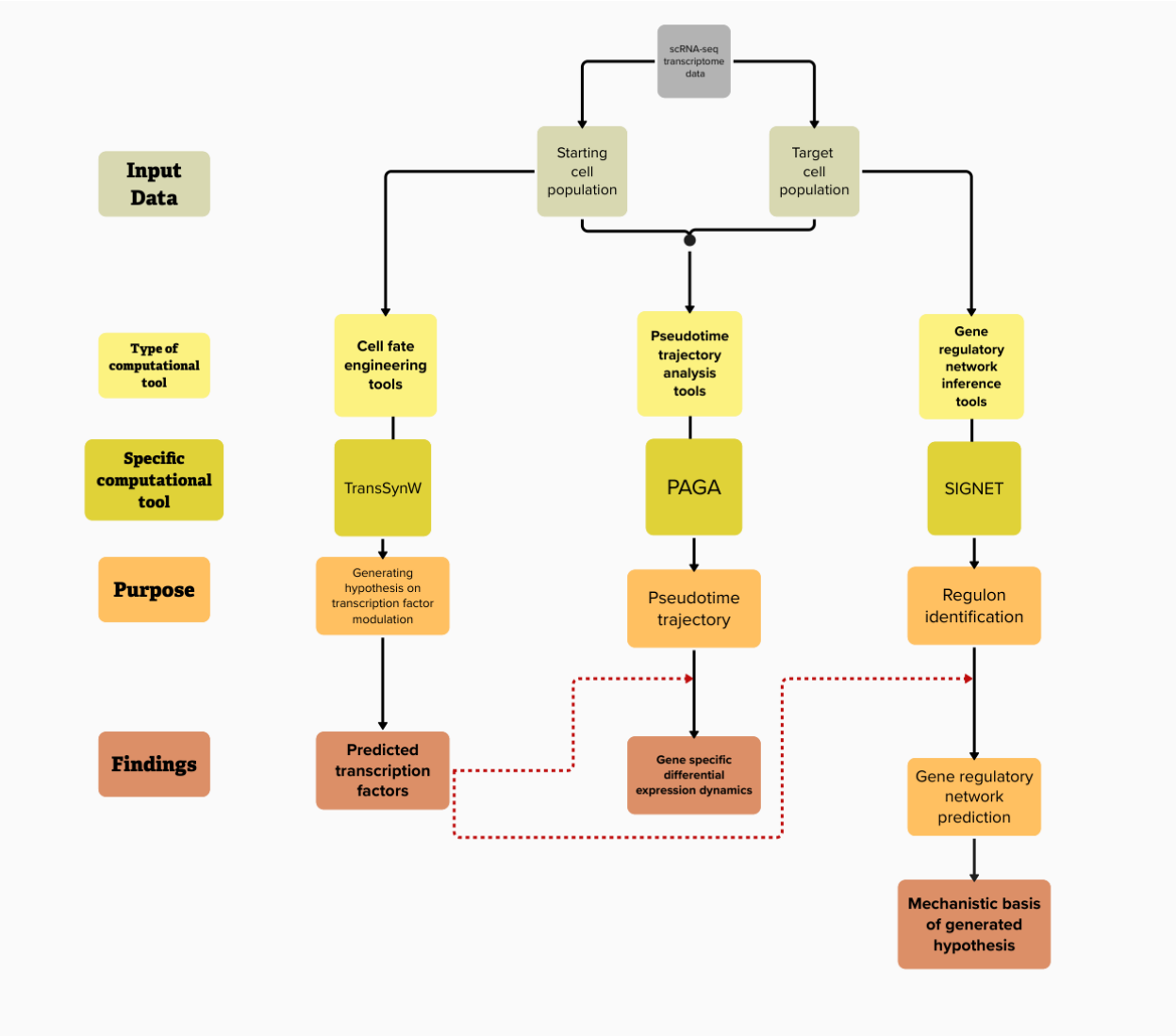

Figure 1: Proposed analysis workflow.

Hypothesis generation on transcription factors whose activity modulation can achieve desired cellular conversion - TransSynW

TransSynW is a cell fate engineering tool, and we used it in the workflow to generate hypotheses on the TFs whose activity modulation can achieve cellular conversion of starting cell type (hNES) to the target cell type (hProgFPM). It uses single cell transcriptome data (e.g. scRNA-seq) of starting and target cell type populations to predict a synergistic list of TFs whose activity modulation can result in the desired cellular conversion.

TransSyn predicts core TFs required for cellular conversion and marker genes for the target cell type. Core TFs consist of specific and pioneer transcription factors, whose activity modulations are hypothesized to bring about the cellular conversion. Specific TFs are predicted to be specifically expressed in the target cell type, and thus, are considered to be associated with cellular conversion and maintenance of target cell type identity. Pioneer TFs are generic in nature and consist of TFs previously reported to be involved in cell conversion protocols, for e.g. TFs associated with chromatin opening [19] . The marker genes (top10 shown by default) for the target cell type are predicted to enable assessing the performance of cell conversion experiments [19].

In the case of hNES to hProgFPM conversion, TransSynW predicted 8 core TFs (5 specific factors and 3 pioneer factors) (Table 1), whose activity modulations are hypothesized to bring about the desired cellular conversion. To gain insight into the relevance of these genes in the context of cellular conversion, we proceed to the next step of the workflow as detailed below. The marker genes predicted by TransSynW are shown in Table 2.

Table 1: Core transcription factors predicted by TransSynW

Gene	Fold Change	Type of Core Transcription Factors
OTX2	10.485	Specific
LMX1A	9.478	Specific
MSX1	6.99	Specific
DMRTA2	3.703	Specific
HEY1	3.182	Specific
SOX2	9.749	Pioneer
PBX1	5.604	Pioneer
SOX9	4.713	Pioneer

Table 2: Marker genes predicted by TransSynW. JSD Score: Jensen-Shannon Divergence score

Gene	JSD Score
DDC	78.61
NTN1	78.83
RND1	78.96
ELAVL4	78.97
PTPRO	79.03
APCDD1	79.04
RPS19	79.06
LDHA	79.17
PLOD2	79.17
EPHB2	79.24

Gaining insight into the differential expression dynamics of hypothesized transcription factors using pseudotime trajectory analysis - PAGA

We leveraged PAGA, a pseudotime trajectory inference tool, to understand the relevance of TransSynW predicted transcription factors in the context of cellular conversion from hNES to hProgFPM.

PAGA uses the starting cell type data (i.e. hNES) as the root (i.e. reference) and attempts to create a pseudotime trajectory representing the degree of conversion of each cell to the target cell type (hProgFPM) (see Figure 2). Overlaying the gene expression data of individual genes predicted by TransSynW on this pseudotime trajectory shows the difference in their expression as the cells transition from hNES to hProgFPM (see Figure 3, 4). Among the specific TFs, OTX2, LMX1A and MSX1 showed distinct change of expression between hNES and hProgFPM while DMRTA2 and HEY1 showed difference to a lesser extent. On the other hand, among the pioneer TFs, SOX2 and PBX1 showed no such change, while SOX9 showed minor differences in the expression between hNES and hProgFPM cells.

We hypothesized that this is due to the fact that the specific TFs are characteristic of a specific lineage, midbrain in the case of hProgFPM, and thus are specifically expressed in them. In contrast, the pioneer TFs are generic and thus, are expressed in pluri- or multi- potent cells (in our case, by both, hNES and hProgFPM cell populations) for bringing about transcriptional

alterations. To verify this, we studied the functional annotation of these genes in UniProt and corresponding Gene Ontology terms (see supplementary file). The specific TFs were indeed found to be involved in dopaminergic and midbrain development except for HEY1. On the other hand, pioneer TFs - SOX2 and PBX1 are involved in suppressing neuronal differentiation in line with their multipotent nature, and it is dynamic balance between the specific and pioneer TFs that possibly acts as a switch for neuronal differentiation. For SOX9, we did not find any annotation relating it to neuronal differentiation.

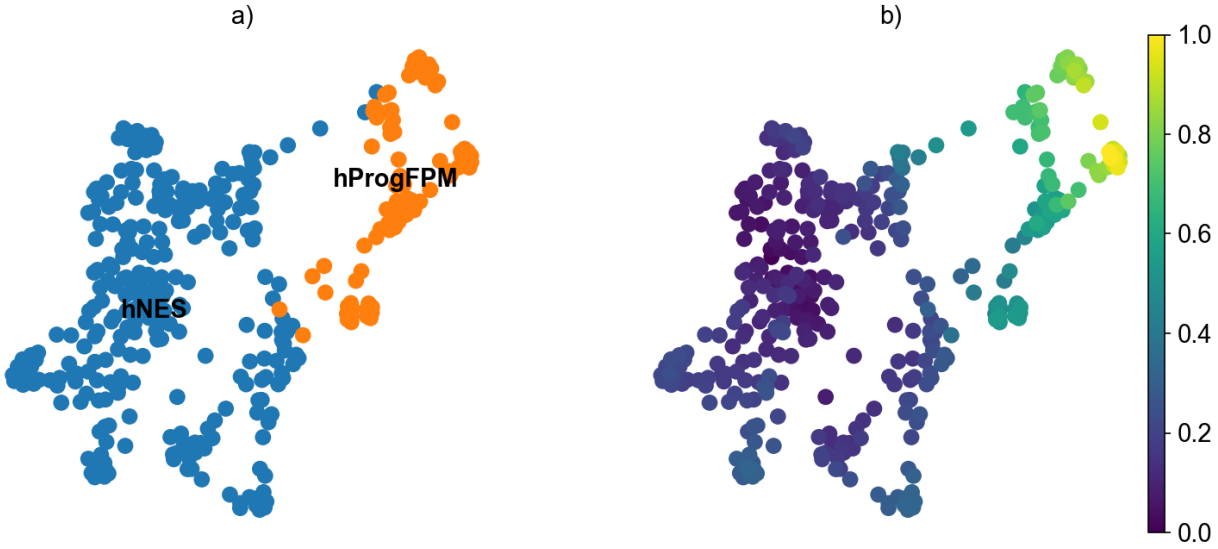

Figure 2: Pseudotime trajectory of hNES to hProgFPM conversion. a) Clustering of hNES and hProgFPM cells; Blue: hNES cells, Orange: hProgFPM cells; b) Pseudotime trajectory showing the progression from hNES cells (dark shades) to hProgFPM cells (lighter shades).

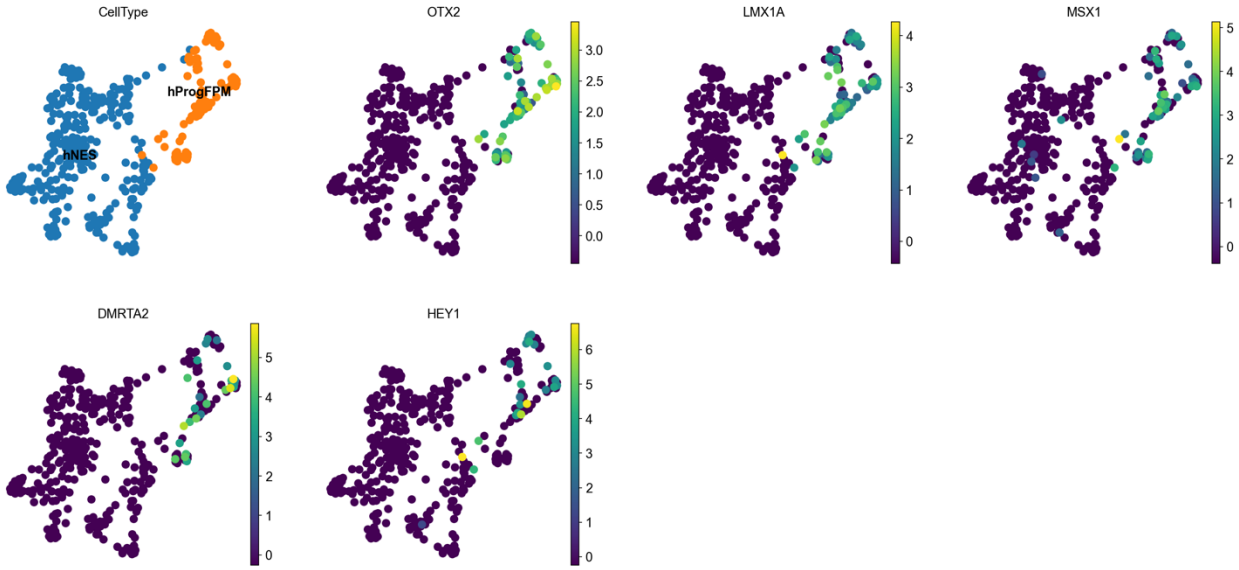

Figure 3: Differential expression dynamics of hypothesized specific transcription factors with reference to the pseudotime trajectory. Clustering of hNES and hProgFPM cells is provided for reference; Blue: hNES cells, Orange: hProgFPM cells.

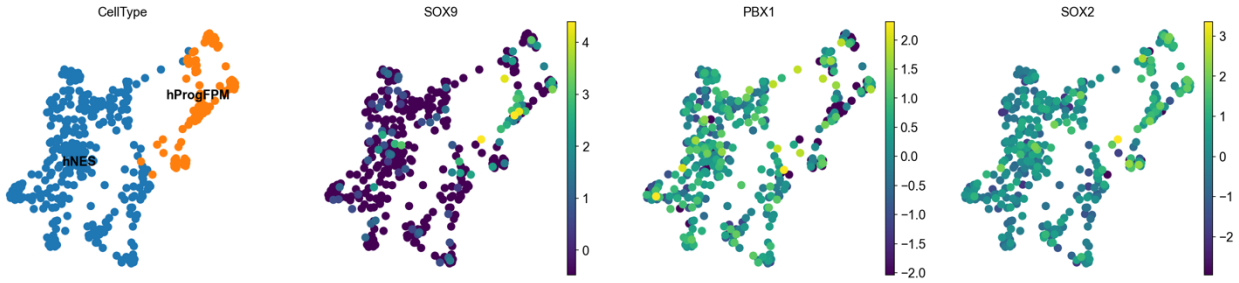

Figure 4: Differential expression dynamics of hypothesized pioneer transcription factors with reference to the pseudotime trajectory. Clustering of hNES and hProgFPM cells is provided for reference; Blue: hNES cells, Orange: hProgFPM cells.

Deciphering the mechanistic basis of transcription factors activity modulation - SIGNET

We used SIGNET to predict the gene regulatory network (GRN) of the target cell type (i.e. hProgFPM) that highlights the functional transcriptional regulation required for cellular functions (see Table 3). Assessing the TransSynW hypothesized TFs and marker genes in the context of this GRN revealed the likely molecular basis of their effect on cellular conversion. We used

TRRUST database to identify the corresponding TFs and the target genes of the genes obtained from TransSyn and SIGNET such that indirect interactions can also be captured.

Combining these genes and their corresponding regulatory TFs and the target genes in a network (see Figure 5) highlights the hypothesized TFs and marker genes in the context of hProgFPM GRN. For instance, (1) MSX1 (a specific TF) and SOX9 (a pioneer TF) predicted by TransSyn, seems to be jointly regulating the expression of HES1 (a TF in hProgFPM GRN), which in the downstream, regulates the expression of a marker gene, PTPRO; (2) Similarly, OTX2 (another specific TF) and SOX9 jointly regulates the expression of another marker gene, DDC in the downstream; and (3) NR2F6 and HMGA2 are two of the TFs predicted by SIGNET, and are shown to be regulating the expression of two of the pioneer TFs, SOX9 and SOX2, respectively. Overall, the network helps reveal the direct and indirect effects of modulation of the TransSyn hypothesized genes, their interaction with hProgFPM GRN, which together culminate into the regulation of few of the marker genes predicted by TransSynW.

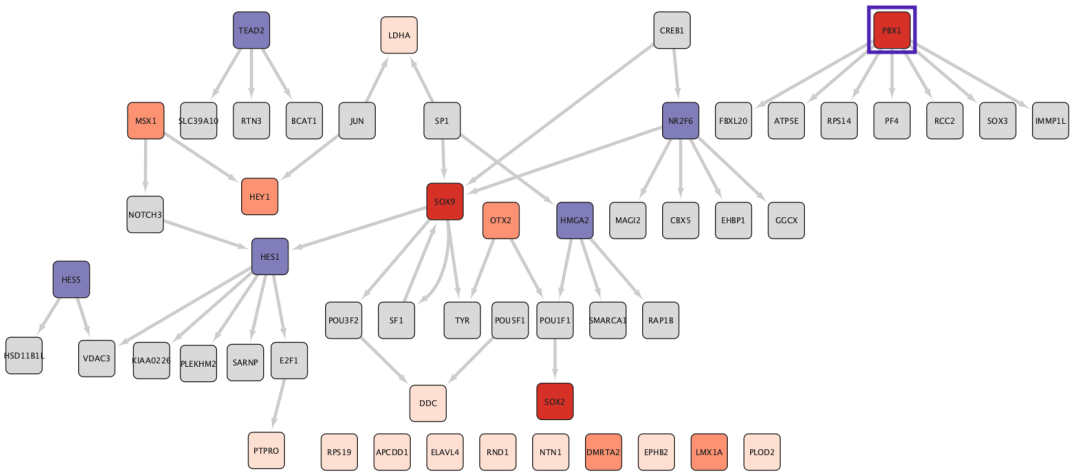

Figure 5: Transcriptional regulatory network based on the genes obtained from TransSynW and SIGNET. Red: Pioneer transcription factors, Orange: Specific transcription factors, Light Orange: Marker genes, Purple: Transcription factors predicted by SIGNET, Red with Purple box: Common transcription factors between hypothesized genes and SIGNET, Grey: Additional interacting genes obtained from TRRUST database.

Table 3: Regulons predicted by SIGNET for target cell type, hProgFPM.

Transcription Factor	Target Genes
HES1	PLEKHM2, SARNP, KIAA0226, VDAC3
HES5	HSD11B1L, VDAC3
HMGA2	RAP1B, SMARCA1
NR2F6	CBX5, EHBP1, GGCX, MAGI2, SOX9
PBX1	RCC2, IMMP1L, FBXL20, ATP5E, RPS14
TEAD2	RTN3, BCAT1, SLC39A10

Discussion

Efficient cellular conversion of therapeutically suitable starting cell types (e.g. stem cells, skin fibroblasts) to the desired target cell types (e.g. dopaminergic neurons, cardiomyocytes, pancreatic beta cells [27]) lies at the center of regenerative medicine. We proposed an analysis workflow that leverages three disparate computational tools to generate not just hypotheses of gene modulation for cellular conversion, but also infer their mechanistic basis.

Although, scientific literature is replete with examples of using multiple tools to address their specific use cases in stem cell research and regenerative medicine, to the best of our knowledge, such defined workflows are not present. We believe that the proposed workflow is reusable and can help researchers formalize new and mechanistically founded hypotheses for achieving desired cellular conversions.

It is important to note that the presented workflow is one of the possible workflows, and many more such workflows can be devised to address this or other challenges in regenerative medicine field. We would like to draw parallels to genomic data analysis workflows that have streamlined such analyses and have contributed greatly in accelerating the progress of the field.

Conclusion

In this study, we leveraged diverse computational tools together to navigate through high-dimensional transcriptomic data and propose hypotheses on transcription factors modulation for achieving desired cellular conversion. We believe that the proposed workflow would benefit researchers in generating mechanistically founded hypotheses that addresses the common pitfalls of cellular conversion including low efficiency of conversion, and the risk of tumorigenicity

and immunogenicity in the target cells. Furthermore, we believe integrating already existing resources (computational tools are an example) can in general provide deeper and diverse insights than individual tools. Thus, the proposed workflow can be further extended with other combinations of tools, prior biological knowledge, etc. based on their suitability to specific use-cases.

Materials and Method

scRNA-seq based gene expression data of starting and target cells

The data used in the study were obtained from NCBI GEO: hNES from GSE114670 and hProgFPM from GSE76381. In both of them, gene expression values are expressed as counts of detected mRNA molecules from each gene in each cell.

The hNES and hProgFPM datasets consist of 15212 genes X 288 samples (i.e. cells) and 19531 genes X 1977 samples respectively. The genes common in both of these datasets were identified (= 15036), and used for further analysis.

TransSynW

TransSynW was set up locally using code available at: TransSynW
scRNA-seq data of starting and target cells were provided separately as input to TransSynW. The target cell scRNA-seq data file is appended with starting cell data, which serves as the background cluster required by TransSynW.

PAGA

PAGA [25] requires the scRNA-seq data in the form of a samples X genes matrix as input. For this, the starting and target cell scRNA-seq data were merged, creating a matrix of 370 samples *(288 hNES + 82 hProgFPM)* X 15036 genes.

Scanpy was used for PAGA analysis, and the steps are based on its tutorial . Briefly, dimensionality reduction was performed using Principal Component Analysis, followed by construction of neighborhood graph of the samples. The data was further denoised using Diffusion map. Starting cell type (i.e. hNES) data is used as root for computing the pseudotime trajectory. The gene expression values of TransSynW predicted genes were overlaid on the trajectory to visualize the expression changes of each gene in the context of the trajectory.

SIGNET

Signet [26] requires scRNA-seq data in the form of samples X genes matrix as input, and list of transcription factors based on the species (*Homo sapiens* in our case). The steps followed were in accordance with the SIGNET tutorial. In brief, SIGNET predicts the relationship between TF and non-transcription factor genes (NTF), and then uses RcisTarget for motif analysis and thereby filtering out the false positive ones.

Acknowledgments

The authors would like to thank Dr. Mariana Ribeiro for addressing our TransSynW related queries. We would also like to thank Justin Jose, Divye Singh and Anuja Chaitanya for valuable discussions and helping us in troubleshoot technical issues with installation and usage of the tools.

Conflict of interest

All authors are affiliated with Thoughtworks.

References

[1] W. Zakrzewski, M. Dobrzyński, M. Szymonowicz, and Z. Rybak, “Stem cells: Past, present, and future,” *Stem Cell Research and Therapy*, vol. 10, no. 1. BioMed Central Ltd., Feb. 26, 2019. doi: 10.1186/s13287-019-1165-5.

[2] R. Hodson, “Stem cells,” *Nature*, vol. 597, no. 7878, pp. S5–S5, Sep. 2021, doi: 10.1038/d41586-021-02620-5.

[3] J. Oswald and P. Baranov, “Regenerative medicine in the retina: from stem cells to cell replacement therapy,” *Ther Adv Ophthalmol*, vol. 10, p. 251584141877443, Jan. 2018, doi: 10.1177/2515841418774433.

[4] S. Petrus-Reurer, M. Romano, S. Howlett, J. L. Jones, G. Lombardi, and K. Saeb-Parsy, “Immunological considerations and challenges for regenerative cellular therapies,” *Communications biology*, vol. 4, no. 1. NLM (Medline), p. 798, Jun. 25, 2021. doi: 10.1038/s42003-021-02237-4.

[5] S. L. Waters, L. J. Schumacher, and A. J. El Haj, “Regenerative medicine meets mathematical modelling: developing symbiotic relationships,” *npj Regenerative Medicine*, vol. 6, no. 1. Nature Research, Dec. 01, 2021. doi: 10.1038/s41536-021-00134-2.

[6] X. Bai, “Stem Cell-Based Disease Modeling and Cell Therapy,” *Cells*, vol. 9, no. 10. NLM (Medline), Sep. 29, 2020. doi: 10.3390/cells9102193.

[7] W. Zakrzewski, M. Dobrzyński, M. Szymonowicz, and Z. Rybak, “Stem cells: Past, present, and future,” *Stem Cell Research and Therapy*, vol. 10, no. 1. BioMed Central Ltd., Feb. 26, 2019. doi: 10.1186/s13287-019-1165-5.

[8] G. Cossu *et al.*, “Lancet Commission: Stem Cells and Regenerative Medicine.”

[9] L. E. Wadkin, S. Orozco-Fuentes, I. Neganova, M. Lako, A. Shukurov, and N. G. Parker, “The recent advances in the mathematical modelling of human pluripotent stem cells,” *SN Applied Sciences*, vol. 2, no. 2. Springer Nature, Feb. 01, 2020. doi: 10.1007/s42452-020-2070-3.

[10] C. Mollinari, J. Zhao, L. Lupacchini, E. Garaci, D. Merlo, and G. Pei, “Transdifferentiation: a new promise for neurodegenerative diseases,” *Cell Death and Disease*, vol. 9, no. 8. Nature Publishing Group, Aug. 01, 2018. doi: 10.1038/s41419-018-0891-4.

[11] I. A. Mellis *et al.*, “Responsiveness to perturbations is a hallmark of transcription factors that maintain cell identity in vitro,” *Cell Syst*, vol. 12, no. 9, pp. 885-899.e8, Sep. 2021, doi: 10.1016/j.cels.2021.07.003.

[12] K. Takahashi and S. Yamanaka, “Induction of Pluripotent Stem Cells from Mouse Embryonic and Adult Fibroblast Cultures by Defined Factors,” *Cell*, vol. 126, no. 4, pp. 663–676, Aug. 2006, doi: 10.1016/j.cell.2006.07.024.

[13] S. L. Wild and D. Tosh, “Molecular mechanisms of transcription factor mediated cell reprogramming: Conversion of liver to pancreas,” *Biochemical Society Transactions*, vol. 49, no. 2. Portland Press Ltd, pp. 579–590, Apr. 01, 2021. doi: 10.1042/BST20200219.

[14] A. Piovesan, F. Antonaros, L. Vitale, P. Strippoli, M. C. Pelleri, and M. Caracausi, “Human protein-coding genes and gene feature statistics in 2019,” *BMC Res Notes*, vol. 12, no. 1, Jun. 2019, doi: 10.1186/s13104-019-4343-8.

[15] S. A. Lambert *et al.*, “The Human Transcription Factors,” *Cell*, vol. 172, no. 4. Cell Press, pp. 650–665, Feb. 08, 2018. doi: 10.1016/j.cell.2018.01.029.

[16] T. Chen *et al.*, “Single-cell Sequencing in the Field of Stem Cells,” *Curr Genomics*, vol. 21, no. 8, pp. 576–584, Jun. 2020, doi: 10.2174/1389202921999200624154445.

[17] P. Zhang *et al.*, “Single-cell RNA sequencing to track novel perspectives in HSC heterogeneity,” *Stem Cell Research and Therapy*, vol. 13, no. 1. BioMed Central Ltd, Dec. 01, 2022. doi: 10.1186/s13287-022-02718-1.

[18] B. Li and G. C. Hon, “Single-Cell Genomics: Catalyst for Cell Fate Engineering,” *Frontiers in Bioengineering and Biotechnology*, vol. 9. Frontiers Media S.A., Oct. 18, 2021. doi: 10.3389/fbioe.2021.748942.

[19] M. M. Ribeiro, S. Okawa, and A. del Sol, “TransSynW: A single-cell RNA-sequencing based web application to guide cell conversion experiments,” *Stem Cells Transl Med*, vol. 10, no. 2, pp. 230–238, Feb. 2021, doi: 10.1002/sctm.20-0227.

[20] E. Appleton, J. Tao, G. Fonseca, S. Liu, C. Glass, and G. Church, “MACHINE-GUIDED CELL-FATE ENGINEERING”, doi: 10.1101/2022.10.14.512279.

[21] L. Marazzi, M. Shah, S. Balakrishnan, A. Patil, and P. Vera-Licona, “NETISCE: a network-based tool for cell fate reprogramming,” *NPJ Syst Biol Appl*, vol. 8, no. 1, Dec. 2022, doi: 10.1038/s41540-022-00231-y.

[22] A. Tran, P. Yang, J. Y. H. Yang, and J. T. Ormerod, “scREMOTE: Using multimodal single cell data to predict regulatory gene relationships and to build a computational cell reprogramming model,” *NAR Genom Bioinform*, vol. 4, no. 1, Mar. 2022, doi: 10.1093/nargab/lqac023.

[23] S. Okawa *et al.*, “Transcriptional synergy as an emergent property defining cell subpopulation identity enables population shift,” *Nat Commun*, vol. 9, no. 1, Dec. 2018, doi: 10.1038/s41467-018-05016-8.

[24] J. Tailor *et al.*, “Stem cells expanded from the human embryonic hindbrain stably retain regional specification and high neurogenic potency,” *Journal of Neuroscience*, vol. 33, no. 30, pp. 12407–12422, 2013, doi: 10.1523/JNEUROSCI.0130-13.2013.

[25] F. A. Wolf *et al.*, “PAGA: graph abstraction reconciles clustering with trajectory inference through a topology preserving map of single cells,” *Genome Biol*, vol. 20, no. 1, pp. 1–9, Mar. 2019, doi: 10.1186/s13059-019-1663-x.

[26] Q. Luo, Y. Yu, and X. Lan, “Problem Solving Protocol SIGNET: single-cell RNA-seq-based gene regulatory network prediction using multiple-layer perceptron bagging,” *Brief Bioinform*, vol. 23, pp. 1–14, 2022, doi: 10.1093/bib/bbab547.

[27] K. Horisawa and A. Suzuki, “Direct cell-fate conversion of somatic cells: Toward regenerative medicine and industries,” *Proceedings of the Japan Academy Series B: Physical and Biological Sciences*, vol. 96, no. 4. Japan Academy, pp. 131–158, 2020. doi: 10.2183/PJAB.96.012.