Odysseia: Genetic Regulatory Feature Analysis with Interpretable Classification Machine Learning Models

Jack Yu[1,*1], **Jiawang Tao**[1,], **and Jie Wang**[1,*2]

[1]Center for Health Research, Guangzhou Institutes of Biomedicine and Health, Chinese Academy of Sciences, Guangzhou 510530, China
[*1]Correspondence: gyu17@alumni.jh.edu
[*2]Correspondence: wang_jie01@gibh.ac.cn

ABSTRACT

With rapid progress of robust single-cell transcriptome sequencing since last decade, numerous complex mechanisms underlying cell development has been revealed. Single-cell RNA sequencing (scRNA-seq) analysis is widely accepted as the main approach to define cell stages and phenotypes. As conversion of somatic cells into induced pluripotency cells succeeded, identification key genetic factors(GFs) with scRNA-seq for cell reprogramming in biological research and regenerative medicine fields gained increasing attention. Herein, we describe *Odysseia*, an interpretable machine learning classifier based single-cell gene expression profile(scGEP) analysis system, that assesses importances of genetic regulatory features in differentiating cell states(CSs). Furthermore, extracted factors, when combining with regulatory network analysis, can help to find key GFs in classifying CSs and possibly inducing CS conversions. Analyzed three published scRNA-seq datasets used to study divergent cell types, *Odysseia* correctly extracted GFs acclaimed to be capable of inducing CS conversions. Overall, *Odysseia* provides an automated alternative to obtain guidance information while explicating mechanism to engineer cellular phenotypes.

Introduction

Since cell reprogramming of mouse embryonic fibroblasts(MEF) to induced pluripotent stem cell(iPSC) succeed in 2006[1], several more CS transition methods have been established both *in vivo* and *in vitro*[2–4], spurring prospective development in biomedical area. However, reprogrammed cells may fail to strictly match *in vivo* counterparts' identities[5], and precise reprogramming toward cell subtypes or states remains impossible[6]. As gene expression profiles(GEPs) for each individual cell in query population became accessible with scRNA-seq, substantial progress in accurately defining cell subtypes and states with marker gene expression has been made[7,8]. Multiple computational methods, like *CellNet*[9], method of *D'Alessio et al*[10], and *Mogrify*[11], are available and proved informative of determining key transcription factors(TFs) for CS conversions via carrying out novel human cell conversion experiments. But limited generalized method could achieve similar goal while utilizing scRNA-seq based expression profiles instead of microarray based or Cap analysis gene expression (CAGE) based databases. Indeed, constructing scRNA-seq based gene expression database is possible to expand existing methods' applicability in predicting key TFs for CS conversions and decoding detailed mechanism behind cell reprogramming. Maintaining complicated database covering majority discovered CSs require massive work; hence, developing generalized scRNA-seq based predictive computational method for determining key GFs in CS conversion with existing methods remain challenging. In 2019, *SingleCellNet*[12] was published as successor of *CellNet* in classifying cellular samples with scRNA-seq data but not guiding CS conversion.

Here we present *Odysseia*, a scRNA-seq based expression profile analysis system utilizing interpretable machine learning classifiers, for guiding CS conversion method design and elucidating mechanisms behind CS conversions. Unlike previously described methods, *Odysseia* does not require any background expression database but searching for potential key GFs in converting one CS to another with only expression profiles labeled with binary CS categories as input. Genetic regulatory pathways(GRPs) significantly contributing to classifying CS categories will be addressed through assessing and interpreting classification models trained with input data. Then, key genes that likely induce CS conversion will be addressed through analyzing regulons constructed with the important GRPs described above. This design is based on the assumption that the key genes and GRPs may not have converged profiles in either expression or regulatory aspects. Thus, classification model designed to directly find CS deterministic genes or GRPs may fail to converge on a stable solution. One of the intuitive approach to solve the non-convergence issue is transforming previous question into classifying CSs, a converged and labeled profile among input data, with given scGEPs. Then, *Odysseia* analyzes the well-performing classifiers and find features, equivalent to GRPs, essential to correctly classify scGEPs.

However, if features associated with similar functions are sufficiently informative, CS classifiers could reach remarkably high accuracy with a small number of heavily-weighted features. As a result, potential key features will presumably be lost even classifiers demonstrated high accuracy over testing data, and output informativity will be impaired.

For example, cell differentiation or CS conversion is accompanied by a series of GEP and epigenetic changes, and these processes usually occur during the cell cycle.[13,14] As a result, GFs related to cell cycle or proliferation can be used as differential characteristics to classify CSs but are not key regulators of the differentiation process. If only these GFs are found after classifier analysis, system output's capability in guiding CS conversion may not be sufficient. Therefore, a particular classification model designed solving transformed question can still encounter feature lost issue.

In order to overcome feature lost issue, *Odysseia* applies multiple classification models with distinct architecture settings in each training iteration then selects well-performing classifiers for further assessment and interpretation. Through interpreting top-performing classifiers trained with different batches of input data, *Odysseia* enhances the feature extraction capability.

Further details and testing results are discussed in the remainder of this paper which is structured as follow:

- The **Method** section explains the overall system design.
- The **Result** section shows system performance on three existing datasets with high fidelity.
- At last, further experiments and potential improvements is discussed in the **Discussion** section.

Method

Briefly, *Odysseia* consists of four main steps as shown in Figure 1:

- ***Step 1***: Generate pseudo-celluar gene expression profiles(pseudo-cGEPs) and corresponding pseudo-cellular genetic regulatory network(pseudo-cGRN) with genetic regulatory network(GRN) reconstruction guidance.
- ***Step 2***: Train classification models with GRNs reconstructed with pseudo-cGEPs then select out well-performing models via accuracy evaluation.
- ***Step 3***: Interpret selected classifiers' correct predictions on all generated GRNs.
- ***Step 4***: Perform regulon based analysis on important GRPs found from classifer interpretation in step 3 and extract key genes indicated by the GRPs.

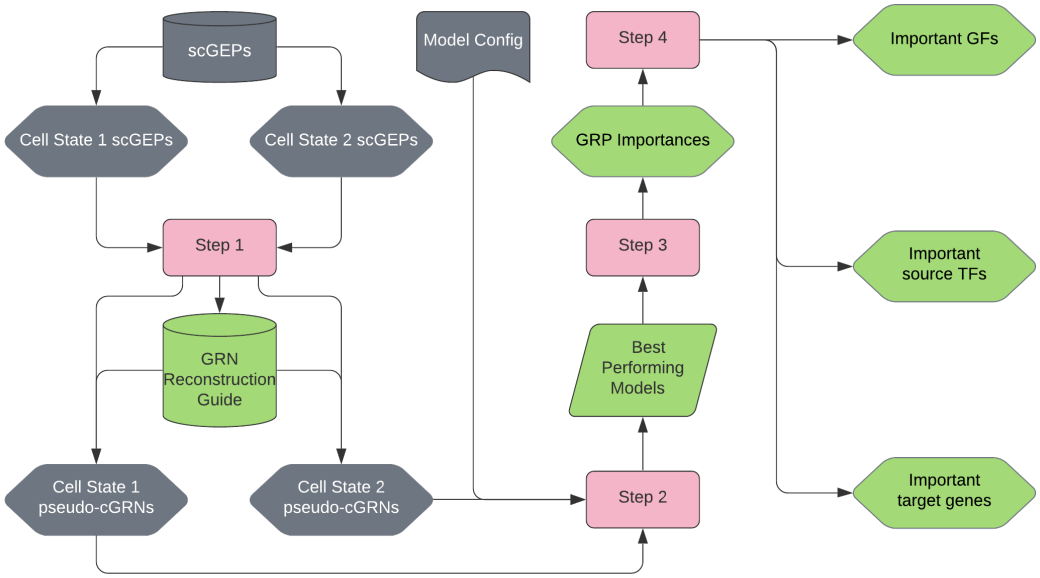

Figure 1. The overall workflow of *Odysseia*

Step 1: Data pre-processing

To reconstruct GRNs based on scGEPs, it is necessary to measure interaction strength among gene pairs. One of the widely adopted methods is Pearson's Correlation coefficient(PCC)[15,16]. Nevertheless, calculating PCC requires sequence of expression levels(ELs) for each gene amomg interaction rathar than single EL scRNA-seq can provide for one cell. Hence, in order to reconstruct cell-level GRNs, *Odysseia* segments scGEPs under same CS category into subsets with constant size to generate pseudo-cGEPs. More specifically, in this step, ELs of one gene from different cells are abstracted as a sequence of temporal changing ELs of corresponding gene in one pseudo cell.

When scGEPs scarece, sliding window algorithm(SWA) with customized window size and padding stride can be applied to ensure sufficient pseudo-cGEP for later classifier training and assessment processes. With SWA, *i-th* pseudo-cGEP can be obtained through:

$$SWA(i) = \left\{ x_j {}_{j=i*s}^{j+l} \right\}, j+l < N$$

Here N denotes total length of scGEPs gained from scRNA-seq, and scRNA-seq can be expressed as $\left\{ x_j {}_{j=0}^{N} \right\}$; l denotes window size; s denotes padding stride.

To reduce computational resource requirement, a meta-level process on determining gene pairs might form regulatory relationships in each pseudo-cGRN can dramatically improve efficiency from repeatedly testifying all possible gene pairs in every pseudo-cGEP. Assuming that pseudo-cGEPs have similar expression pattern with comprehensive GEP containing all scGEPs, the GRN reconstruction guidance can be created through analyzing all gene pairs in comprehensive GEP. The overall workflow to create GRN reconsturction guidance is summarized in Figure 2. In total, two sets of filters are used for validating a GRP:

1. **Gene level**:

   The filter set below determines binary CS category differentiability of single gene expression pattern.

   - **SD Filter**: The standard deviation(SD) of query gene's ELs must be significant. This filter aims to confirm query gene can both reach relatively high EL and possibly interact with other GFs. SD threshold is set to 1 by default, implying that only genes with expression status changing in different circumstances can be considered. In reality, SD threshold need to be adjusted in response to expression change weight and computational resource capacity.
   - **MWU Filter**: The expression distribution of selected gene across samples labeled with different CSs must be dissimilar. This filter aims to confirm that query gene's GEP is likely distinguishable across sample sets under binary CS categories. The Mann-Whitney U rank test(MWU) implemented by *SciPy*[17] will be performed on GEPs of selected gene in binary CS categories. The p-value for rejecting null hypothesis, that expression pattern underlying CS1 is the same as the expression pattern underlying CS2, is set to 0.05 by default.

2. **Pathway level**:

   The filter set below determines the biological interpretability and significance of a GRP.

   - **TF Filter**: The source input of a GRP must be a recorded TF. If not further specified, TF list will be retrieved from integrated *TRANSFAC*[18] datasets according to the species information in query.
   - **GRP Filter**: Regulatory target must have binding ability with source TF. If not further specified, regulatory gene list of given TF will be retrieved from integrated *GTRD*[19] datasets according to the species information in query.
   - **PCC Filter**: Expression correlation between source TF and target gene must be significant. The absolute value threshold of PCC is 0.2 by default, as calculated with *SciPy*[17].

Utilizing GRN reconstruction guidance, pseudo-cGRN will be reconstructed for each pseudo-cGEP. Each gene in the guidance, if also presents in pseudo-cGEP, need to pass the SD filter in pseudo-cGEP circumstance before validating correspond GRPs in the guidance. The pre-validated GRPs that also passes PCC filter will take part in eventual pseudo-cGRN.

Step 2: Classifier selection

According to given model config file, vanilla models will be initialized with specified model type and architecture parameters. *Odysseia* will perform accuracy tests systematically aiming to select out potentially well-performing classifiers for further analysis, while iterativly training classifiers with split pseudo-cGRN sets. The overall workflow is summarized in Figure 4. Currently, *Odysseia* supports three types of classification models:

- **SVC**: Support Vector Classifer implemented with *scikit-learn*.[20] All parameters adjustable with *scikit-learn*[20] are supported.

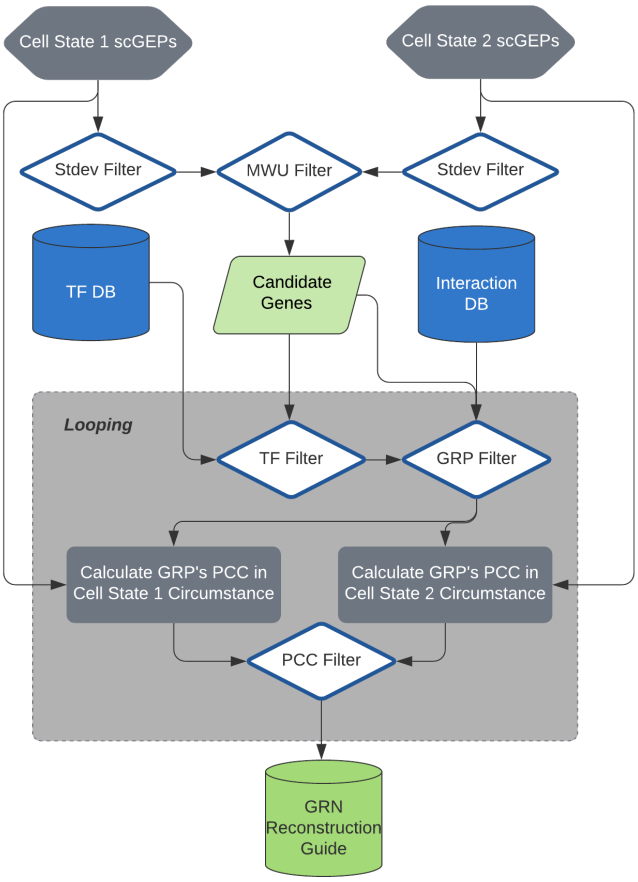

Figure 2. Workflow to create GRN reconstruction guidance from comprehensive GEPs. ***(1)*** Each gene in scGEP set either for CS1 or CS2 need to pass SD filter to confirm its expression pattern reacts with inconstant gene expression circumstances. ***(2)*** GEPs of genes passed SD filter will be inputted to MWU filter to confirm that its gene expression pattern is differentiable among binary CS categories. ***(3)*** All TFs in genes passed MWU filter will be selected out as potential regulatory source TF utilizing given TF database. ***(4)*** Using GF or protein-protein interaction databae, all genes passed MWU filter will be testified as target genes with potential TF found at previous step to form potential interacting gene pairs. ***(5)***, PCCs of potential interacting gene pairs found previously will be calculated under each CS category. A gene pair must have corresponding PCC under at least one CS category greater than the threshold to be considered as potential GRP.

- **GBTree**: Gradient Boosting Trees implemented with *XGBoost*.[21] All parameters adjustable with *XGBoost*[21] are supported.
- **CNN**: Convolutional Neural Network implemented with *Pytorch*[22].

  More specifically, the general architecture designs of *Odysseia*'s integrated CNNs are implemented referring to 1D-CNN and 2D-Hybrid-CNN applied in recent cancer type prediction study[23]. Unlike original 1D-CNN and 2D-Hybrid-CNN, we implemented the models with flexibilities on not only kernel related parameters but also others including amount of convolution layer set which consists of a convolution layer and adjacent max-pooling layer. An example of 1D-CNN with 2 convolution layer sets is illustrated in Figure 3.

  Other CNNs implemented with *Pytorch*[22] can also be analyzed with *Odysseia* but training protocol need to be clarifed by user in advance.

Eighty one model builds consisting of 1 SVC, 16 GBTrees, and 64 CNNs will be used to initialize vanilla models with default model config file. In each training iteration, each vanilla classifier will be fed with same set of pseudo-cGRNs as training samples and is primarily assessed according to prediction accuracy on untrained pseudo-cGRNs. By default setting, 70% of pseudo-cGRNs will be split into training set while 30% will be split into testing set at each iteration. Classifiers capable of passing accuracy threshold will be kept for further comparisons with other classifers trained with different sample sets in later

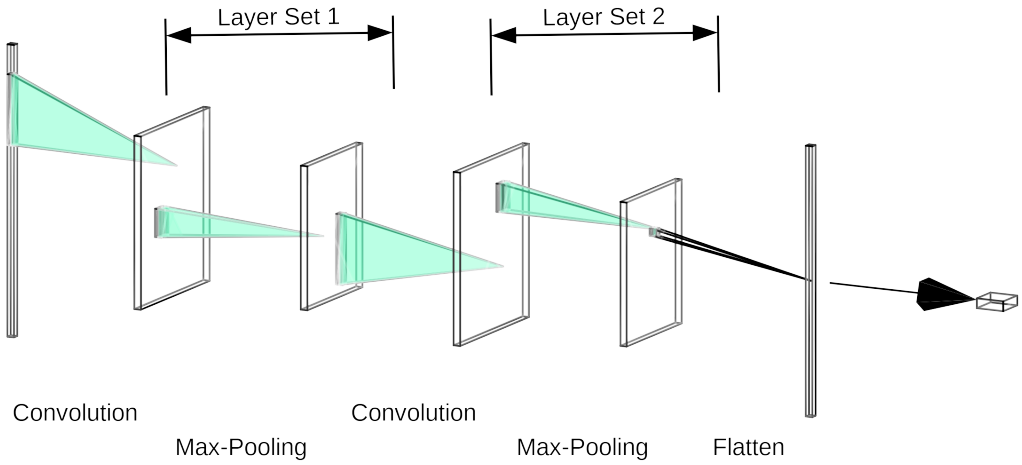

Figure 3. 1D-CNN with 2 convolution layer set.

iterations.

At the end of this step, all available pseudo-cGRNs will be joint to use as testing data, and all candidate classifers will be evaluated on their prediction accuracies of fully joint testing data. Considering that the real random choice can reach an accuracy of 50% in binary classification question, the accuracy threshold for either local accuracy test or fully joint data test is set to 90%.

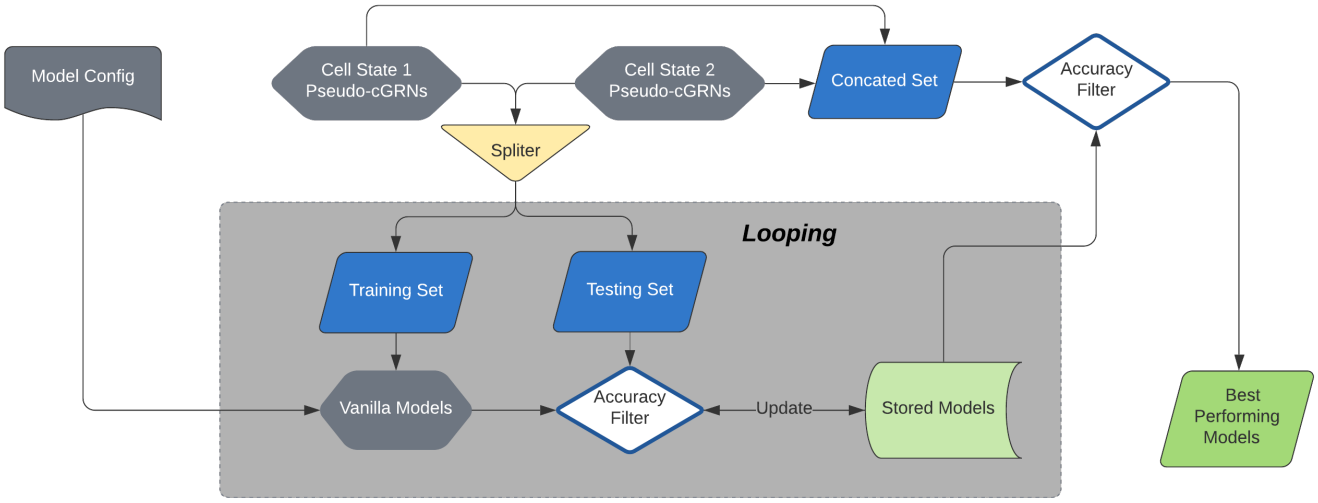

Figure 4. Workflow to generate well-performing classifers. ***(1)*** In each training iteration, input pseudo-cGRNs will be split into training set and testing set according to given ratio. By default, the ratio is set to 70% for training and 30% for testing. ***(2)*** According to given classification model config file, vanilla classifiers will be initialized and trained with training set. The predictive accuracy of classifiers will be pre-assessed with testing data. Classifers reaching accuracy threshold, which is set to 0.9 by default, will be kept for accuracy assessment in further training iterations. ***(3)*** After all training iterations, all pseudo-cGRNs will be concatenated as one testing set to assess accuracies of classifers being kept. Classifers reach accuracy threshold, set to 0.9 by default, will be passed to further analysis.

Step 3: Classifier interpretation

During the last accuracy assessment in Step 2, all classifer's predictions consisting with ground truths will be marked with corresponding input pseudo-cGRNs. Briefly, in this step, *Odysseia* interprets how well-performing classifiers make correct predictions on CS categories of given pseudo-cGRNs. Integrating interpretations of all well-performing classifers, the

generalized importance values of features, which are represented by GRPs in *Odysseia*'s scenario, in determining CS can be estimated.

Several packages already have internalized methods implemented for machine learning models to analyze feature importance. For example, GBTree implemented with *XGBoost*[21] can have feature importance approximated with average weight gain at each split involving the feature, and linear SVC implemented with *scikit-learn*[20] can have the importance estimated with feature coefficient.

Regarding the standard differences between these internalized methods, we utilize *softmax* function to normalize feature importance and define the normalized importance calculation function as:

$$T(Z) = \left\{ \frac{f(Z_i)}{\sum_{j=1}^{L} f(Z_j)} \right\}, i \in Z$$

Here Z is feature set; L is total number of features; $f(X)$ is internalized feature importance calculating function of classification model.

However, for numerous machine learning models, there is no internalized feature importance estimating method, for example, NN-based models implemented with Pytorch[22] or generalized method cannot be appropriately applied with selected models such as Kernelized SVC implement with *scikit-learn*[20]. Therefore, we apply the concept of The Shapley value[24] to build an universal approach for determining importance of each feature in any kind of machine learning model. Specific Shapley value calculation or approximating methods are implemented with *SHAP*[25] and applied to different types of classifiers as shown in Table1.

Model Genera	**Model Example**	**SHAP[25] Method**
Tree-Based	GBTree	Tree Explainer
NN-Based without Repetitive Layer Sets	CNN	Deep Explainer
NN-Based with Repetitive Layer Sets	Recurrent Neural Network	Gradient Explainer
Kernel Machines	SVC	Kernel Explainer
Linear Models	Logistic Regression	Linear Explainer

Table 1. Machine learning model generas with applicable Shapley value approximating methods

Let N denotes the total number of correct predictions, $\phi^i_{x,j}$ and $\phi^i_{y,j}$ denotes Shapley value of feature i when predicting sample j as type x or type y, the function determining importance values of feature set z can be expressed as:

$$T(Z) = softmax\left(\left\{ \sum_{j=1}^{N} \frac{\left|\phi^i_{x,j}\right| + \left|\phi^i_{y,j}\right|}{2} \right\}, i \in Z \right)$$

Following the analysis of all classifers, each feature's general importance value can be obtained through:

$$g(i) = \sum_{j=1}^{C} T_j(Z)[i]$$

Here i denotes i*th* feature or GRP, C denotes the total number of classifers being analyzed, $T_j(Z)[i]$ denotes i*th* feature's importance value calculated through analyzing j*th* classifier.

Step 4: Key genes extraction

With feature list ranked according to importance values, *Odysseia* reconstructs regulon for top-ranked GRPs and extracts influential genes among the regulon. If analoging regulon or GRN with the graph concept in discrete mathematics, one of the most common methods to analyze influence of a vertex, which is equivalent to a gene, is assessing correspond degree of the vertex. Although the regulatory source and target within a GRP can be determined by utilizing interaction database, such as *GTRD*[19] which shows regulatory relationship of interactors, numerous computational prediction methods, for example, *GRNBoost2*[26], and comprehensive interaction database, for example, *BioGRID*[27], could not provide faithful indication on regulatory directions. Consequently, based on choice of interaction database for *Odysseia*, the regulon reconstructed cannot be guaranteed to be a directed graph, but can surely be analyzed as an undirected graph. Therefore, in this step, *odysseia* primarily extracts genes with high degrees in the regulon, regardless of they are in-degree or out-degree. To limit total amount of output key genes, the default setting of *Odysseia* reconstructs regulon for top 100 GRPs, and genes with degree higher than two in the regulon will be considered influential. If regulatory relationships can be clarifed, the regulatory sources capable of regulating multiple key genes and the regulatory targets influenced by multiple key genes can also be extracted according to GRN reconstruction guidance.

In the circumstance of finding genes potentially inducing CS conversion, referred as *Mogrify*[11], genes either CS deterministic or capable of regulating key genes shall gain attention. If a gene is not marked as important, it must be capable of regulating more than 2 listed key genes to be considered as an important regulatory interactor.

Results

In order to evaluate analytic power, we tested *Odysseia* with three previously published CS conversion related scRNA-seq datasets. Considering the differences in information availability and data size among datasets, each dataset is analyzed with adjusted process.

MEF and ESC

This dataset was used to study cell fate continuum when reprogramming MEF into iPSC, one of the most well-studied CS conversions.[28] To simulate the process of finding gene combination for reprogramming somatic cell into pluripotent stem cell, we used scGEPs-labeled MEFs and embryonic stem cells(ESCs) as input of *Odysseia*. Due to the limited amount of scGEPs, SWA with window size set to 10 and padding stride set to 1 was applied to generate 73 ESC pseudo-cGEPs and 65 MEF pseudo-cGEPs. SD threshold is set to 100, which resulted in GRN reconstruct guidance consisting 72440 GRPs, considering the computational load our testing device could accept. While all other parameters remained as default, *Odysseia* extracted 16 key genes and 9 common regulatory sources of key genes. Partial key genes which have already been reported in other studies are listed below, and complete output gene list is included in **Supplementary Table 1**.

Gene	Occurrence	Degree	Previous report
Nanog	16	0	Can reprogram MEF into iPSC with other GFs[29–31]
Esrrb	10	0	Can reprogram MEF into iPSC with other GFs[29,32–34]
Sall4	10	0	Can reprogram MEF into iPSC with other GFs[29,34]
Tfcp2l1	8	0	Key regulatory mediator supporting ESC identity[35,36]
Pou5f1 (Oct4)	7	0	Can reprogram MEF into iPSC with other GFs[1,30,31,37–39]
Fos	3	2	Subunit of AP-1 which can prohibit iPSC reprogramming[29,40]
Klf4	2	0	Can reprogram MEF into iPSC with other GFs[1,37–39]

Here, **Occurrence** refers to occurrences in top ranked GRPs; **Degree** refers to amount of regulating key genes; **Previous report** summarizes key findings from previous reports.

MEF and iMPC

The main purpose of analyzing this dataset is to assess *Odysseia*'s performance while stimulating natural cells with artificially induced cells. This dataset is retrieved from research project studying detailed processes of fibroblast to myocyte and myogenic progenitor cell(MPC) conversion.[41] For approaching muscle cell reprogramming, the output goal of this analysis is designed to find key genes for converting MEF into myoblast and further derived myotube can be stimulated with MyoD+ induced MPCs(iMPCs)[42]. Practically, only iMPC scGEPs with EL of Myod1 greater than 1, the lowest positive EL in dataset, were extracted as myoblasts and myotubes co-culture and analyzed with MEF scGEPs. Due to computational limitation on our device, pseudo-cGEPs were generated with every 100 scGEPs. As a result, in total of 99 MyoD+ iMPC and 62 MEFs pseudo-cGEPs were generated. SD threshold was also set to 5 to limit the size of GRN reconstruct guidance to be 3756 GRPs. Without any change on other default parameters, *Odysseia* extracted 6 key genes and 3 common regulatory sources of key genes. Partial analysis result is listed below under same format as mentioned in section above, and complete output gene list is included in **Supplementary Table 2**.

Gene	Occurrence	Degree	Previous report
Sox4	73	2	Myogenesis mediator[43,44]
Mef2c	23	3	Synergy factor of MyoD in the conversion of MEF into myoblast[45,46]
Myog	2	4	Marker of differentiating myoblast[47,48]
Tcf4	2	2	Marker of MEF in muscle connective tissue; Myogenesis mediator[49–51]
Mef2a	0	3	Synergy factor of MyoD in the conversion of MEF into myoblast[45,46]
Myod1	0	3	Can reprogram MEF into myoblast and myotube solely[41,52,53]

Radial glia and Neuron

Furthermore, contemplating feasible difficulty in obtaining purified scRNA-seq dataset of cell subtypes, we tested *Odysseia* when analyzing differences between purified and mixed CS data. The input dataset of this analysis consists of scGEPs of purified neuron and iPSC-derived radial glial cell, the progenitor cell of neuron, mixing with neuron. All of the data are retrieved from published dataset used to examine CS conversion from iPSC to neuron.[54] Similar with MEF and iMPC analysis, pseudo-cGEPs were also generated with every 100 scGEPs which result in a total of 90 neural co-culture and 72 purified neuron

pseudo-cGEPs. Without changes on other default parameters, 40441 GRPs were included in GRN reconstruct guidance, and *Odysseia* extracted 17 key genes and 19 common regulatory sources of key genes. Partial analysis result is listed below under same format as mentioned in section above, and complete output gene list is included in **Supplementary Table 3**.

Gene	Occurrence	Degree	Previous report
Ets2	48	12	Cell cycle regulator need to be repressed for primary neurogenesis[55,56]
Isl1	23	4	Key gene for motor neuron development[57–59]
Ascl1	16	13	Key neurogenesis regulator[54,60]; Can induce neuronal reprogramming solely[61,62]
Nr3c1	0	9	Cell cycle inhibitor during embryonic neurogenesis[63]
Rbpj	0	8	Key mediator of Notch signaling in neurogenesis[64–66]
Neurod1	0	7	Key neurogenesis regulator[54,67–70]

Even though *Odysseia* is unable to guarantee a gene combination inducing CS conversion, numerous previously discovered key genes can be successfully extracted, which indicated gene combination known capable to convert CSs. For example, when analyzing MEF with iPSC, the top three genes having most occurrences among important GRPs are determined as key factors to reprogram MEF into iPSC with 7F combination[29]. With Jdp2, which is selected to block AP-1 activity also as Fos indicated, the combination of Jdp2, Nanog, Esrrb, and Sall4 is demonstrated as a minimum 4F functional set to perform cell reprogramming[29]. While Sall4 indicated interaction between Sox2 and Oct4[71], main components among Yamanaka factors[1,39] can also be extracted with *Odysseia*'s analysis result. Between MEF and muscle cell, multiple works demonstrated capability of Myod1, which is successfully extracted by *Odysseia*, in artificially induce CS conversion.[52,53] For neurogenesis analysis, Ascl1, included as key factor in several neuronal conversion methods[61,62], is marked as key gene with the most consequential regulatory influence by *Odysseia*. It is noteworthy that other factors extracted with *Odysseia* may also play an important role in CS conversions but have not yet been testified with adequate experiments.

Discussion

As scRNA-seq technology becomes increasingly accessible and applicable, an efficient and robust analytical system capable of extracting key GFs differentiating CSs will also become highly demanded. In this paper, we showed *Odysseia*'s applicability in analyzing scRNA-seq derived data for further CS-related research. According to test cases so far, key genes significant enough to induce CS conversions appear to have either high occurrence in top ranked GRPs or high regulatory influence on other key genes. Without any previous knowledge on studying CSs, Gene Ontology (GO) enrichment analysis can potentially narrow down key cellular functions changed in different CSs.

However, further tests and optimizations are necessary to confirm *Odysseia*'s applicability on guiding frontier researches. While more resources become available, other classification models and interpretation methods will be added and tested with current version on larger scales. Further optimized model selection process will also be critical to improve system's overall performance.

With result achieved so far, we anticipate that *Odysseia* will become useful in providing insightful advices on not only learning differences between CSs and cell subtypes, but also finding key GFs to induce CS conversions unachievable so far.

References

1. Takahashi, K. & Yamanaka, S. Induction of pluripotent stem cells from mouse embryonic and adult fibroblast cultures by defined factors. *Cell* **126**, 663–676, DOI: 10.1016/j.cell.2006.07.024 (2006).

2. Buganim, Y. *et al.* Direct reprogramming of fibroblasts into embryonic sertoli-like cells by defined factors. *Cell Stem Cell* **11**, 373–386, DOI: 10.1016/j.stem.2012.07.019 (2012).

3. Qian, L. *et al.* In vivo reprogramming of murine cardiac fibroblasts into induced cardiomyocytes. *Nature* **485**, 593–598, DOI: 10.1038/nature11044 (2012).

4. Sekiya, S. & Suzuki, A. Direct conversion of mouse fibroblasts to hepatocyte-like cells by defined factors. *Nature* DOI: 10.1038/nature10263 (2011).

5. Earley, A. M., Burbulla, L. F., Krainc, D. & Awatramani, R. Identification of ascl1 as a determinant for human ipsc-derived dopaminergic neurons. *Sci. Reports* **11**, DOI: 10.1038/s41598-021-01366-4 (2021).

6. Wang, H., Yang, Y., Liu, J. & Qian, L. Direct cell reprogramming: Approaches, mechanisms and progress. *Nat. Rev. Mol. Cell Biol.* **22**, 410–424, DOI: 10.1038/s41580-021-00335-z (2021).

7. Baron, M. *et al.* A single-cell transcriptomic map of the human and mouse pancreas reveals inter- and intra-cell population structure. *Cell Syst.* **3**, DOI: 10.1016/j.cels.2016.08.011 (2016).

8. Treutlein, B. *et al.* Reconstructing lineage hierarchies of the distal lung epithelium using single-cell rna-seq. *Nature* **509**, 371–375, DOI: 10.1038/nature13173 (2014).

9. Cahan, P. *et al.* Cellnet: Network biology applied to stem cell engineering. *Cell* **158**, 903–915, DOI: 10.1016/j.cell.2014.07.020 (2014).

10. D'Alessio, A. C. *et al.* A systematic approach to identify candidate transcription factors that control cell identity. *Stem Cell Reports* **5**, 763–775, DOI: 10.1016/j.stemcr.2015.09.016 (2015).

11. Rackham, O. J. *et al.* A predictive computational framework for direct reprogramming between human cell types. *Nat. Genet.* **48**, 331–335, DOI: 10.1038/ng.3487 (2016).

12. Tan, Y. & Cahan, P. Singlecellnet: A computational tool to classify single cell rna-seq data across platforms and across species. *Cell Syst.* **9**, DOI: 10.1016/j.cels.2019.06.004 (2019).

13. Dalton, S. Linking the cell cycle to cell fate decisions. *Trends Cell Biol.* **25**, 592–600, DOI: 10.1016/j.tcb.2015.07.007 (2015).

14. Engström, Y. Cell cycle regulators control stemness and differentiation. *BioEssays* **43**, 2100123, DOI: 10.1002/bies.202100123 (2021).

15. Liu, L.-y. D., Hsiao, Y.-C., Chen, H.-C., Yang, Y.-W. & Chang, M.-C. Construction of gene causal regulatory networks using microarray data with the coefficient of intrinsic dependence. *Bot. Stud.* **60**, DOI: 10.1186/s40529-019-0268-8 (2019).

16. Song, L., Langfelder, P. & Horvath, S. Comparison of co-expression measures: Mutual information, correlation, and model based indices. *BMC Bioinforma.* **13**, DOI: 10.1186/1471-2105-13-328 (2012).

17. Virtanen, P. *et al.* SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nat. Methods* **17**, 261–272, DOI: 10.1038/s41592-019-0686-2 (2020).

18. Matys, V. Transfac(r) and its module transcompel(r): Transcriptional gene regulation in eukaryotes. *Nucleic Acids Res.* **34**, DOI: 10.1093/nar/gkj143 (2006).

19. Kolmykov, S. *et al.* GTRD: an integrated view of transcription regulation. *Nucleic Acids Res.* **49**, D104–D111, DOI: 10.1093/nar/gkaa1057 (2020). https://academic.oup.com/nar/article-pdf/49/D1/D104/35364856/gkaa1057.pdf.

20. Pedregosa, F. *et al.* Scikit-learn: Machine learning in Python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011).

21. Chen, T. & Guestrin, C. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, 785–794 (2016).

22. Paszke, A. *et al.* Pytorch: An imperative style, high-performance deep learning library. In Wallach, H. *et al.* (eds.) *Advances in Neural Information Processing Systems 32*, 8024–8035 (Curran Associates, Inc., 2019).

23. Mostavi, M., Chiu, Y.-C., Huang, Y. & Chen, Y. Convolutional neural network models for cancer type prediction based on gene expression - bmc medical genomics (2020).

24. Roth, A. E. & Shapley, L. S. The shapley value : essays in honor of lloyd s. shapley. *Economica* **101**, 123 (1991).

25. Lundberg, S. M. & Lee, S.-I. A unified approach to interpreting model predictions. In *Proceedings of the 31st international conference on neural information processing systems*, 4768–4777 (2017).

26. Moerman, T. *et al.* Grnboost2 and arboreto: Efficient and scalable inference of gene regulatory networks. *Bioinforma. (Oxford, England)* **35**, DOI: 10.1093/bioinformatics/bty916 (2018).

27. Stark, C. Biogrid: A general repository for interaction datasets. *Nucleic Acids Res.* **34**, DOI: 10.1093/nar/gkj109 (2006).

28. Guo, L. *et al.* Resolving cell fate decisions during somatic cell reprogramming by single-cell rna-seq. *Mol. Cell* **73**, DOI: 10.1016/j.molcel.2019.01.042 (2019).

29. Wang, B. *et al.* Induction of pluripotent stem cells from mouse embryonic fibroblasts by jdp2-jhdm1b-mkk6-glis1-nanog-essrb-sall4. *Cell Reports* **27**, DOI: 10.1016/j.celrep.2019.05.068 (2019).

30. Yu, J. *et al.* Induced pluripotent stem cell lines derived from human somatic cells. *Science* **318**, 1917–1920, DOI: 10.1126/science.1151526 (2007).

31. Wang, Y. *et al.* Reprogramming of mouse and human somatic cells by high-performance engineered factors. *EMBO reports* **12**, 373–378, DOI: 10.1038/embor.2011.11 (2011).

32. Huang, D. *et al.* Lif activated jak signaling determines esrrb expression during late-stage reprogramming. *Biol. Open* DOI: 10.1242/bio.029264 (2017).

33. Benchetrit, H. *et al.* Direct induction of the three pre-implantation blastocyst cell types from fibroblasts. *Cell Stem Cell* **24**, DOI: 10.1016/j.stem.2019.03.018 (2019).

34. Iseki, H. *et al.* Combined overexpression of jarid2, prdm14, esrrb, and sall4a dramatically improves efficiency and kinetics of reprogramming to induced pluripotent stem cells. *Stem Cells* **34**, 322–333, DOI: 10.1002/stem.2243 (2015).

35. Wang, X. *et al.* The transcription factor tfcp2l1 induces expression of distinct target genes and promotes self-renewal of mouse and human embryonic stem cells. *J. Biol. Chem.* **294**, 6007–6016, DOI: 10.1074/jbc.ra118.006341 (2019).

36. Ye, S., Li, P., Tong, C. & Ying, Q.-L. Embryonic stem cell self-renewal pathways converge on the transcription factor tfcp2l1. *The EMBO J.* **32**, 2548–2560, DOI: 10.1038/emboj.2013.175 (2013).

37. Nemajerova, A., Kim, S. Y., Petrenko, O. & Moll, U. M. Two-factor reprogramming of somatic cells to pluripotent stem cells reveals partial functional redundancy of sox2 and klf4. *Cell Death Differ.* **19**, 1268–1276, DOI: 10.1038/cdd.2012.45 (2012).

38. Shi, Y. *et al.* Induction of pluripotent stem cells from mouse embryonic fibroblasts by oct4 and klf4 with small-molecule compounds. *Cell Stem Cell* **3**, 568–574, DOI: 10.1016/j.stem.2008.10.004 (2008).

39. Maekawa, M. *et al.* Direct reprogramming of somatic cells is promoted by maternal transcription factor glis1. *Nature* **474**, 225–229, DOI: 10.1038/nature10106 (2011).

40. Li, D. *et al.* Chromatin accessibility dynamics during ipsc reprogramming. *Cell Stem Cell* **21**, 819–833.e6, DOI: https://doi.org/10.1016/j.stem.2017.10.012 (2017).

41. Kim, I. *et al.* Integrative molecular roadmap for direct conversion of fibroblasts into myocytes and myogenic progenitor cells. *bioRxiv* DOI: 10.1101/2021.08.20.457151 (2021).

42. Bar-Nur, O. *et al.* Direct reprogramming of mouse fibroblasts into functional skeletal muscle progenitors. *Stem Cell Reports* **10**, 1505–1521, DOI: https://doi.org/10.1016/j.stemcr.2018.04.009 (2018).

43. Jang, S.-M. *et al.* Sox4-mediated caldesmon expression facilitates skeletal myoblast differentiation. *J. cell science* **126**, DOI: 10.1242/jcs.131581 (2013).

44. Liu, S. *et al.* Genome architecture mediates transcriptional control of human myogenic reprogramming. *iScience* **6**, 232–246, DOI: 10.1016/j.isci.2018.08.002 (2018).

45. Taylor, M. V. & Hughes, S. M. Mef2 and the skeletal muscle differentiation program. *Semin. Cell Dev. Biol.* **72**, 33–44, DOI: 10.1016/j.semcdb.2017.11.020 (2017).

46. Liu, N. *et al.* Requirement of mef2a, c, and d for skeletal muscle regeneration. *Proc. Natl. Acad. Sci.* **111**, 4109–4114, DOI: 10.1073/pnas.1401732111 (2014). https://www.pnas.org/content/111/11/4109.full.pdf.

47. Andrés, V. & Walsh, K. Myogenin expression, cell cycle withdrawal, and phenotypic differentiation are temporally separable events that precede cell fusion upon myogenesis. *J. Cell Biol.* **132**, 657–666, DOI: 10.1083/jcb.132.4.657 (1996).

48. Dumont, N. A. & Rudnicki, M. A. Characterizing satellite cells and myogenic progenitors during skeletal muscle regeneration. *Methods Mol. Biol.* 179–188, DOI: 10.1007/978-1-4939-6788-9_12 (2017).

49. Mathew, S. J. *et al.* Connective tissue fibroblasts and tcf4 regulate myogenesis. *Development* **138**, 371–384, DOI: 10.1242/dev.057463 (2011).

50. Contreras, O., Rebolledo, D. L., Oyarzún, J. E., Olguín, H. C. & Brandan, E. Connective tissue cells expressing fibro/adipogenic progenitor markers increase under chronic damage: Relevance in fibroblast-myofibroblast differentiation and skeletal muscle fibrosis. *Cell Tissue Res.* **364**, 647–660, DOI: 10.1007/s00441-015-2343-0 (2016).

51. Fry, C. S., Kirby, T. J., Kosmac, K., McCarthy, J. J. & Peterson, C. A. Myogenic progenitor cells control extracellular matrix production by fibroblasts during skeletal muscle hypertrophy. *Cell Stem Cell* **20**, 56–69, DOI: 10.1016/j.stem.2016.09.010 (2017).

52. Choi, J. *et al.* Myod converts primary dermal fibroblasts, chondroblasts, smooth muscle, and retinal pigmented epithelial cells into striated mononucleated myoblasts and multinucleated myotubes. *Proc. Natl. Acad. Sci.* **87**, 7988–7992, DOI: 10.1073/pnas.87.20.7988 (1990).

53. Chakraborty, S. *et al.* A crispr/cas9-based system for reprogramming cell lineage specification. *Stem cell reports* **3**, 940–947 (2014).

54. Earley, A. M., Burbulla, L. F., Krainc, D. & Awatramani, R. Identification of ascl1 as a determinant for human ipsc-derived dopaminergic neurons. *Sci. Reports* **11**, DOI: 10.1038/s41598-021-01366-4 (2021).

55. Janesick, A., Wu, S. C. & Blumberg, B. Retinoic acid signaling and neuronal differentiation. *Cell. Mol. Life Sci.* **72**, 1559–1576, DOI: 10.1007/s00018-014-1815-9 (2015).

56. Janesick, A. *et al.* Erf and etv3l are retinoic acid-inducible repressors required for primary neurogenesis. *Development* **140**, 3095–3106, DOI: 10.1242/dev.093716 (2013).

57. Pfaff, S. L., Mendelsohn, M., Stewart, C. L., Edlund, T. & Jessell, T. M. Requirement for lim homeobox gene isl1 in motor neuron generation reveals a motor neuron–dependent step in interneuron differentiation. *Cell* **84**, 309–320 (1996).

58. Liang, X. *et al.* Isl1 is required for multiple aspects of motor neuron development. *Mol. Cell. Neurosci.* **47**, 215–222 (2011).

59. Zhou, M. *et al.* Reprogramming astrocytes to motor neurons by activation of endogenous ngn2 and isl1. *Stem Cell Reports* **16**, 1777–1791 (2021).

60. Raposo, A. A. *et al.* Ascl1 coordinately regulates gene expression and the chromatin landscape during neurogenesis. *Cell reports* **10**, 1544–1556 (2015).

61. Chanda, S. *et al.* Generation of induced neuronal cells by the single reprogramming factor ascl1. *Stem cell reports* **3**, 282–296 (2014).

62. Robinson, M. *et al.* Transdifferentiating astrocytes into neurons using ascl1 functionalized with a novel intracellular protein delivery technology. *Front. Bioeng. Biotechnol.* 173 (2018).

63. Shin, J. *et al.* Single-cell rna-seq with waterfall reveals molecular cascades underlying adult neurogenesis. *Cell stem cell* **17**, 360–372 (2015).

64. Imayoshi, I. & Kageyama, R. The role of notch signaling in adult neurogenesis. *Mol. neurobiology* **44**, 7–12 (2011).

65. Fujimoto, M. *et al.* Rbp-j promotes neuronal differentiation and inhibits oligodendroglial development in adult neurogenesis. *Dev. biology* **332**, 339–350 (2009).

66. Tanigaki, K. & Honjo, T. Two opposing roles of rbp-j in notch signaling. *Curr. topics developmental biology* **92**, 231–252 (2010).

67. Pataskar, A. *et al.* Neurod1 reprograms chromatin and transcription factor landscapes to induce the neuronal program. *The EMBO journal* **35**, 24–45 (2016).

68. Gao, Z. *et al.* Neurod1 is essential for the survival and maturation of adult-born neurons. *Nat. neuroscience* **12**, 1090–1092 (2009).

69. Boutin, C. *et al.* Neurod1 induces terminal neuronal differentiation in olfactory neurogenesis. *Proc. Natl. Acad. Sci.* **107**, 1201–1206 (2010).

70. Roybon, L. *et al.* Neurogenin2 directs granule neuroblast production and amplification while neurod1 specifies neuronal fate during hippocampal neurogenesis. *PloS one* **4**, e4779 (2009).

71. Tanimura, N., Saito, M., Ebisuya, M., Nishida, E. & Ishikawa, F. Stemness-related factor sall4 interacts with transcription factors oct-3/4 and sox2 and occupies oct-sox elements in mouse embryonic stem cells. *J. Biol. Chem.* **288**, 5027–5038, DOI: 10.1074/jbc.m112.411173 (2013).

Author contributions statement

J.Y.: Methodology, Software, Writing- Original draft preparation **J.T.**: Writing - Review & Editing **J.W.**: Project administration

Additional information

All scRNA-seq datasets are retrieved from Gene Expression Omnibus(GEO) as described in Key Resource Table below.

Key Resource Table

Resource	Source	Usage
MEF and ESC scRNA-seq data	GEO Accession: GSE103221	Samples with label started with 'mef' as MEF Samples with label started with 'esc' as ESC
MEF and iMPC scRNA-seq data	GEO Accession: GSE169054	GSM5175907 as iMPC GSM5643793 as MEF
Radial glia and Neuron scRNA-seq data	GEO Accession: GSE185275	GSM5609927 as neural co-culture GSM5609930 as purified neurons