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

Jack Yu[1,*1] **and Jiawang Tao**[1,]

[1]Affiliation, department, city, postcode, country
[*1]Correspondence: gyu17@alumni.jh.edu

ABSTRACT

With rapid progress of robust single-cell transcriptome sequnecing since last decade, numerous complicate mechanism underlying cell development has been revealed. Yet, single-cell RNA sequencing (scRNA-seq) analysis is widely accepted as main approach to define cell stages and phenotypes. As conversion of somatic cells into induced pluripotency cells already succeeded, identification key genetic factors(GFs) with scRNA-seq for cell reprogramming in biological research and regenerative medicine fields has gained great 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 state(CS). Furthermore, combined with regulatory network analysis, extracted factors can help to find specific key GFs in implementing CS conversions. Analyzing 3 published cell reprogramming related reports studying diverge cell types, *Odysseia* correctly extracts genes acclaimed being capable to induce CS conversion. Overall, *Odysseia* provides an option to obtain guidance information while elucidating 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 could still fail to flawlessly match in vivo counterparts' identities[5], and precise reprogramming toward cell subtypes remain impossible[6]. As gene expression profiles for each individual cell in query population became accessible with scRNA-seq, substantial progress in accurately defining of cell subtypes with marker gene expression has been made[7,8]. Even though multiple computational methods, like *CellNet*[9], method of *D'Alessio et al*[10], and *Mogrify*[11], are already available and proved informativity of determining key transcription factors(TFs) for CS conversion via carrying out novel human cell conversion experiments, 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 cell subtype conversions and elucidating detailed mechanism behind cell reprogramming. But workload of maintaining complicated database covering majority discovered cell subtypes can be massive; 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] is 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 interpretablemachine learning classifiers, for not only guiding CS conversion method design but also elucidating mechanisms behind CS conversions. Unlike previous described methods, *Odysseia* does not require any background expression database but finding potential key GFs in conversing one CS to another with only expression profiles labeled with binary CS categories as input. Important genetic regulatory pathways(GRPs) being deterministic for classifying CS category will be addressed through assessing and interpreting trained classifiers. Then, key genes may induce CS conversion will be addressed through analyzing regulons constructed with important GRPs described previously. The main reason to keep this design is due to consideration on the assumption: 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 can potentially fail to converge on a stable solution. One of the intuitive approach to solve the non-convergence issue is transforming classification question into higher dimension. In our case, instead of asking which factors are CS deterministic, we can transform the question into determining CS category, which is supposed to be a converged and labeled profile among input data, with given GEP. Then, we can analysis the well-performing classifiers and determine which features, equivalent to GRPs, are essential for the classifiers to make correct decisions.

However, in this approach, if a few feature associating with similar functions can be informative enough, CS classifiers could reach remarkably high accuracy with limited features being heavily weighted. As a result, partial of potential key features will presumably be lost even classifiers being analyzed has demonstrated accuracy over testing data, and informativity of system output will be questionable. **For example, cell growth related genes can also be used as marker genes when separating cells not only in different types but also on distinct cell cycle stages. But, if only cell growth related genes are found after classifier analysis, system output's capability in guiding CS conversion is unlikely being sufficient.** Therefore, a particular classification model designed solving transformed question can still encounter feature lost issue.

In order to overcome feature lost issue as well, *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:

- **Method:** firstly, we will summarize the overall system design and describe how it can possibly avoid both non-convergence and feature lost issues.
- **Result:** secondly, we perform extensive analysis using diverse collection of sequencing data to demonstrate that *Odysseia* can be applied into real world problems, especially in finding potential GFs to perform cell reprogramming.
- **Discussion:** third, we discuss about how the outputs of this system can be utilized in further wet lab experiments and what further improvements can be done on this system.

Methods

Briefly, *Odysseia* consists of 4 main steps which can be summarized below and visualized as 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.

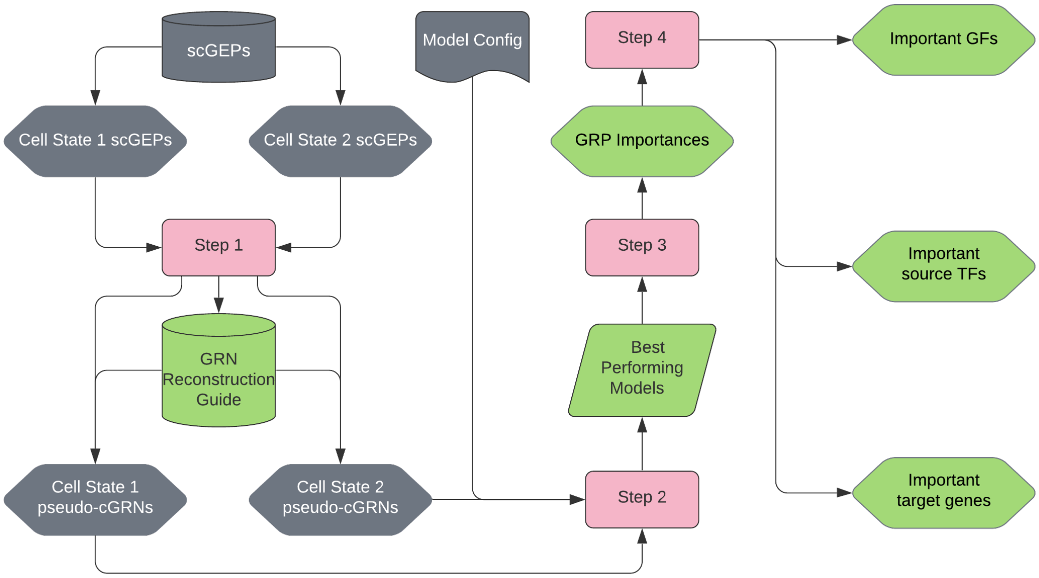

Figure 1. The overall workflow of *Odysseia*

Step 1: Data preprocessing

To reconstruct GRNs based on scGEPs, a computational method measurig itenaction strength among gene pairs is mandatory. One of the widely adopted methods is utilizing pearson correlation coefficient(PCC)[13,14]. Nevertheless, calculating PCC requires sequence of expression levels(ELs) for each gene amomg interaction rathar than single EL scRNA-seq can provide in one cell. Hence, in order to reconstruct cell-level GRNs, *Odysseia* segments scGEPs under same CS category into subsets with constant size to form pseudo CEPs, having gene ELs gained from different cells abstracted as sequence of altering ELs among single pseudo cell.

Uncer circumstances scGEPs being scarece, sliding window algorithm(SWA) with customized window size and padding stride can be applied to ensure pseudo-cGEP amount being sufficent for later classifier traning 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 for total length of scGEPs gained from scRNA-seq which can be expressed as $\left\{ {x_j}_{j=0}^{N} \right\}$; l denotes for window size; s denotes for padding stride.

Although pseudo-cGRNs will be reconstructed from pseudo-cGEPs respectively, a meta-level process on determining which gene pairs would form regulatory relationships in each pseudo-cGRN can dramatically reduce computational resouce requirement than repeatedly testifying all possible gene pairs in every pseudo-cGEP. Assuming 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 can be summarized as Figure 2. In total, two sets of filters must be passed for validating a GRP:

1. **Gene level**:

   The main purpose of this filter set is confirming whether a gene's expression pattern is differentiable among binary CS categories or not .

   - **Stdev Filter**: The standard deviation(stdev) of gene's EL must be significant. This filter is to confirm query gene not only can reach relatively high EL but also has expression pattern possibly interacting with other GFs. By default, the threshold is set to 1 implying that one gene's expression status must be changing in different circumstances to be consudered.

     However, stdev threshold would need to be adjusted significantly considering not only significant weight of expression change in studying data but also computational resouce capacity.

   - **MWU Filter**: The expression distribution of selected gene across samples labeled with different CSs must be dissimilar. This filter is to confirm query gene's GEP is likely distinguishable across sample sets under binary CS categories. The Mann-Whitney U rank test(MWU) implemented by *SciPy*[15] will be performed on GEPs of selected gene in binary CS categories. By default, the p-value to reject null hypothesis, expression pattern underlying CS1 is the same as the expression pattern underlying CS2, is set to 0.05.

2. **Pathway level**:

   The main purpose of this filter set is confirming whether a GRP is both biologically interpretable and significant or not.

   - **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*[16] datasets according to the specie 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*[17] datasets according to the specie information in query.

   - **PCC Filter**: Expression correlation between source TF and target gene must be significant. The default setting of threshold is 0.2 for absolute value of Pearson correlation coefficient calculated with *SciPy*[15].

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

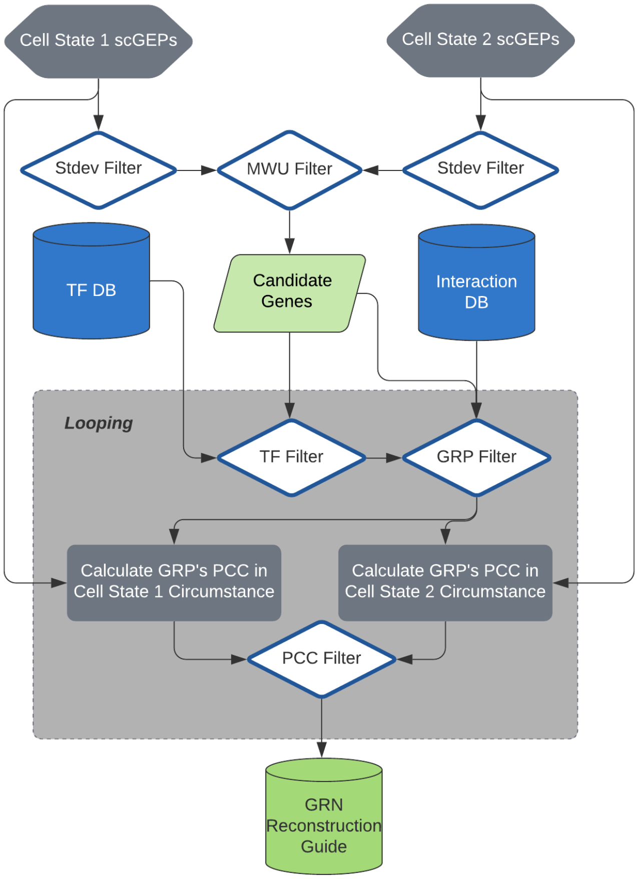

Figure 2. Workflow to create GRN reconstruction guidance from comprehensive GEPs. ***(1)*** Each gene in scGEP set either for CS1 or CS2 need to pass stdev filter confirming its expression pattern is reacting with inconstant gene expression circumstances. ***(2)*** GEPs of genes passed stdev filter will be input to MWU filter for confirming 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 correspond PCC under at least one CS category greater than threshold to be considered as potential GRP.

Step 2: Classifier selection

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

- **SVC**: Support Vector Classifer implemented with *scikit-learn*.[18] All parameters adjustable with *scikit-learn*[18] are supported.
- **GBTree**: Gradient Boosting Trees implemented with *XGBoost*.[19] All parameters adjustable with *XGBoost*[19] are supported.
- **CNN**: Convolutional Neural Network implemented with *Pytorch*[20].

  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[21]. Unlike original 1D-CNN and 2D-Hybrid-CNN, we implemented the models with flexibilities not only on kernel related parameters but also on others including amount of convolution layer set which consists a convolution layer and adjacent max-pooling layer. An example of 1D-CNN with 2 convolution layer sets can be illustrated as Figure 3.

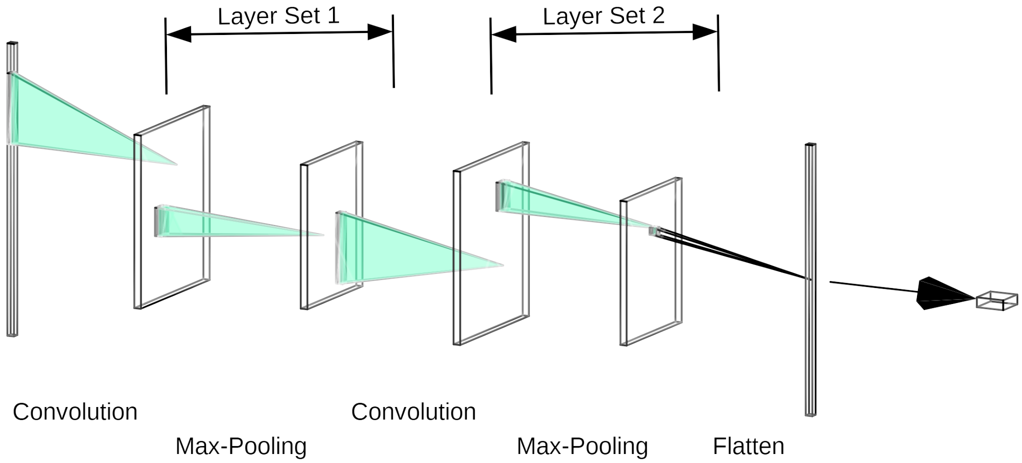

Figure 3. 1D-CNN with 2 convolution layer set.

Other CNNs implemented with *Pytorch*[20] can also be analyzed with *Odysseia* but training protocol need to be clarifed by user in advance.

By default model config file, 81 model builds consisting of 1 SVC, 16 GBTrees, and 64 CNNs will be used to initialize vanilla models. In each training iteration, each vanilla classifier will be fed with same set of pseudo-cGRNs as training samples and primarily assessed according to prediction accuracy on untrained pseudo-cGRNs. By default setting, at each iteration, 70% of pseudo-cGRNs will be split into training set while 30% will be split into testing set. The classifiers capable of passing accuracy threshold will be kept for further comparisons with other classifers trained with different sample sets in afterward iterations.

At the end of this step, all available pseudo-cGRNs will be joint as testing data, and all candidate classifers will be evaluated on their prediction accuracy of fully joint testing data. By default, considering the true randomly choosing 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%.

Step 3: Classifier interpretation

During last accuracy assessment in Step 2, all classifer's predictions consisting with ground facts 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 GRPs in *Odysseia*'s scenario, in determining CS can be approached.

To analysis feature importance, several packages already have internalized methods implemented for machine learning models. For example, GBTree implemented with *XGBoost*[19] can have feature importance approached through calculating average weight gain at each split involving the feature, and linear SVC implemented with *scikit-learn*[18] can have the importance approached 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 approaching method, for example NN based models implemented with Pytorch[20], or generalized method cannot be appropriately applied with selected models, such as Kernelized SVC implement with *scikit-learn*[18]. Therefore, we apply the concept of The Shapley value[22] to build an universal approach for determining importance of each feature in any kind of machine learning model. Specific Shapley value calculation or approaching methods are implemented with *SHAP*[23] and applied to different types of classifiers as shown in Table1.

Let N denoting total number of correct predictions, $\phi^i_{x,j}$ and $\phi^i_{y,j}$ denoting 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:

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

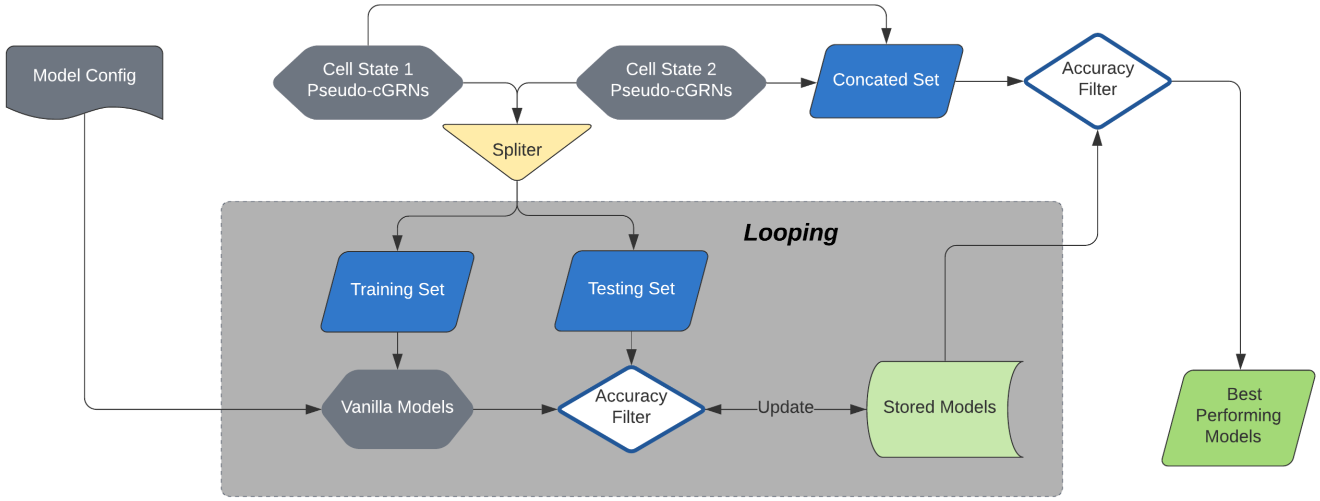

Figure 4. Workflow to generate well-performing classifers. ***(1)*** In each training iteration, input pseudo-cGRNs will be split into training set and testing set according to given ratio. By default, the ratio is set to 70% for training and 30% for testing. ***(2)*** Accoridng to given classification 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 further accuracy assessment in afterward training iterations. ***(3)*** After all training iterations, all pseudo-cGRNs will be concated as one testing set to assess accuracies of classifers being kept. Classifers can still reach accuracy threshold, which is also set to 0.9 by default, will be passed to further analysis.

Model Genera	**Model Example**	**SHAP[23] 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 approaching methods

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

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

Here i denoting i th feature or GRP, C denoting total number of classifers being analyzed, $T_j(Z)[i]$ denoting 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 a 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 analysis influence of a vertex, which is equivalent with a gene, is assessing correspond degree of the vertex. Although the regulatory source and target within a GRP can be determined if utilizing interaction database, such as *GTRD*[17], showing regulatory relationship of interactors, numerous computational prediction method, *GRNBoost2*[24] for example, and comprehensive interaction database, *BioGRID*[25] for example, 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 undirected graph. Therefore, in this step, *odysseia* primarily extracts genes with high degrees in the regulon regardless in-degree or out-degree. To limit total amount of output key genes, the default setting of *Odysseia* is reconstructing regulon for top 100 GRPs, and genes with degree higher than 2 in the regulon will be considered as influential. If regulatory relationships can be clarifed, the regulatory sources being capable to regulate multiple key genes and the regulatory targets being influenced by mulitple key genes can also be extracted according to GRN reconstruction guidance.

In the circumstance of finding genes can potentially induce CS conversion, referring *Mogrify*[11], genes being either CS deterministic or capable of regulating key genes shall gain attention. By default setting, if a gene is not marked important, it must be capable to regulate more than 2 listed key genes to be considered as an important regulatory interactor.

Results

In order to evaluate analytics power, we tested *Odysseia* with 3 previously published CS conversion related scRNA-seq datasets. Considering the differences in information availability and data size among datasets, each dataset was analyzed with adjusted process:

- **MEF and ESC**: This dataset has been used to study cell fate continuum when reprogramming MEF into iPSC, one of the most well-studied CS conversions.[26] 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. Stdev threshold was set to 100, which resulted in GRN reconstruct guidance consisting 72440 GRPs, in concerning computational load our testing device could accept. While all other parameters remained at 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 in below, and complete output gene list is included as **Supplementary Table 1**.

Gene	Occurrence	Degree	Previous report
Nanog	16	0	Can reprogram MEF into iPSC with other GFs[27–29]
Esrrb	10	0	Can reprogram MEF into iPSC with other GFs[27,30–32]
Sall4	10	0	Can reprogram MEF into iPSC with other GFs[27,32]
Tfcp2l1	8	0	Key regulatory mediator supporting ESC identity[33,34]
Pou5f1 (Oct4)	7	0	Can reprogram MEF into iPSC with other GFs[1,28,29,35–37]
Fos	3	2	Subunit of AP-1 which can prohibit iPSC reprogramming[27,38]
Klf4	2	0	Can reprogram MEF into iPSC with other GFs[1,35–37]

  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 peformance 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.[39] 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 both can be stimulated with MyoD+ induced MPCs(iMPCs)[40]. Practically, only iMPC scGEPs having 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 the 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. Stdev threshold was also set to 5 to limit the size of GRN reconstruct guidance being 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 as **Supplementary Table 2**.

Gene	Occurrence	Degree	Previous report
Sox4	73	2	Myogenesis mediator[41,42]
Mef2c	23	3	Synergy factor of MyoD in the conversion of MEF into myoblast[43,44]
Myog	2	4	Marker of differentiating myoblast[45,46]
Tcf4	2	2	Marker of MEF in muscle connective tissue; Myogenesis mediator[47–49]
Mef2a	0	3	Synergy factor of MyoD in the conversion of MEF into myoblast[43,44]
Myod1	0	3	Can reprogram MEF into myoblast and myotube solely[39,50]

- **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 scGEPs of purified neuron and radial glial cell, progenitor cell of neuron, mixed with neuron. All the data are retrieved from published dataset used to examine CS conversion from iPSC to neuron.[51] Similar with MEF and iMPC analysis, pseudo-cGEPs were also generated with every 100 scGEPs which result in 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 as **Supplementary Table 3**.

Gene	Occurrence	Degree	Previous report
Ets2	48	12	Cell cycle regulator need to be repressed for primary neurogenesis[52,53]
Isl1	23	4	Key gene for motor neuron development[54–56]
Ascl1	16	13	A lot[51]
Nr3c1	0	9	Cell cycle inhibitor during embryonic neurogenesis[57]
Rbpj	0	8	Key mediator of Notch signaling in neurogenesis[58–60]
Neurod1	0	7	A lot[51]

Even though *Odysseia* is unable to guarantee a gene combination inducing CS conversion, numerous previously discovered key genes can be successfully extracted and indicate gene combination known capable to convert CSs. For example, when analyzing MEF with iPSC, the top 3 genes having most occurrences among important GRPs are determined as key factors to reprogram MEF into iPSC with 7F combination[27]. With Jdp2 which is selected to block AP-1 activity also as Fos indicating, the combination of Jdp2, Nanog, Esrrb, and Sall4 is demonstrated as a minimum 4F functional set to perform cell reprogramming[27]. While Sall4 indicating interaction between Sox2 and Oct4[61], main components among Yamanaka factors[1,37] can also be extracted with *Odysseia*'s analysis result.

Discussion

Considering the variability of available types of data in different studies, $Odysseia's$ modules not only can be adjusted upon available resources but also, in order to improve biological interpretability on outputs, provide processed generalized background datasets, such as $GTRD$[17] for finding potential target genes of selected transcription factors when building *GRN* reconstruction guidance. Admittedly, generalized background datasets can still fail to provide substantial support in boundary cases like studying specie is not covered in the dataset. To achieve optimized biological interpretability, substituting generalized dataset with case specific data would be essential.

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. 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).

14. 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).

15. 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).

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

17. 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.

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

19. 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).

20. 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).

21. 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).

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

23. 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).

24. 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).

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

26. 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).

27. 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).

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

29. 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).

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

31. 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).

32. 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).

33. 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).

34. 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).

35. 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).

36. 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).

37. 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).

38. 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).

39. 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).

40. 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).

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

42. 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).

43. 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).

44. 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.

45. 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).

46. 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).

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

48. 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).

49. 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).

50. 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).

51. 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).

52. 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).

53. 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).

54. 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).

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

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

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

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

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

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

61. 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

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 start with 'mef' as MEF Samples with label start 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