bioRxiv preprint doi: https://doi.org/10.1101/2021.03.12.435089; this version posted March 12, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Discriminative feature of cells characterizes cell populations of interest by a small subset of genes

Takeru Fujii[1,2,†], Kazumitsu Maehara[1,†,*], Masatoshi Fujita[2], and Yasuyuki Ohkawa[1,*]

[1]Division of Transcriptomics, Medical Institute of Bioregulation, Kyushu University, 3-1-1 Maidashi, Higashi-ku, Fukuoka 812-0054, Japan

[2]Department of Cellular Biochemistry, Graduate School of Pharmaceutical Sciences, Kyushu University, 3-1-1 Maidashi, Higashi-ku, Fukuoka 812-0054, Japan

† These authors contributed equally to this work.

[*]Corresponding author: K.M. (kazumits@bioreg.kyushu-u.ac.jp), and Y.O. (yohkawa@bioreg.kyushu-u.ac.jp)

bioRxiv preprint doi: https://doi.org/10.1101/2021.03.12.435089; this version posted March 12, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

ABSTRACT

Statistical methods for detecting differences in individual gene expression are indispensable for understanding cell types. However, conventional statistical methods have faced difficulties associated with the inflation of *P*-values because of both the large sample size and selection bias introduced by exploratory data analysis such as single-cell transcriptomics. Here, we propose the concept of discriminative feature of cells (DFC), an alternative to using differentially expressed gene-based approaches. We implemented DFC using logistic regression with an adaptive LASSO penalty to perform binary classification for the discrimination of a population of interest and variable selection to obtain a small subset of defining genes. We demonstrated that DFC prioritized gene pairs with non-independent expression using artificial data, and that DFC enabled to characterize the muscle satellite cell population. The results revealed that DFC well captured cell-type-specific markers, specific gene expression patterns, and subcategories of this cell population. DFC may complement differentially expressed gene-based methods for interpreting large data sets.

bioRxiv preprint doi: https://doi.org/10.1101/2021.03.12.435089; this version posted March 12, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

INTRODUCTION

Organisms are composed of various cell types with specific functions. The cell types also include undifferentiated cells, such as stem cells and progenitor cells, or cells in transition during differentiation. Understanding cell types as the functional or structural units of an organism can confer a comprehensive understanding of the functions of organs and tissues, as well as their origins. The classification and definition of cell types has been based on the identification of specific marker genes that define the cell type. Marker genes have been identified by comprehensive analysis of gene expression. Statistical tests are particularly important in the identification of marker genes, for which evaluation of differences in the mean expression levels of genes is often used.

Cell-type-specific genes/proteins are responsible for cell-type-specific functions. Therefore, to identify marker genes, comparisons have been performed between the cell types of interest and control groups to extract specifically expressed genes. The use of differentially expressed genes (DEGs) is a widely accepted way of defining marker gene candidates, the validity of which has been confirmed by biological experiments. However, the risk of false positives is increased by the tens of thousands of statistical tests associated with comprehensive analysis. Therefore, correction methods for multiple testing, such as Benjamini–Hochberg's false discovery rate (FDR)[1], Storey's q-value[2,3], and Efron's local FDR[4] have been widely employed. Along with the increasing demand for multiple testing and correction methods, DEG detection methods in the field of biostatistics for high-dimensional data have also been developed. In particular, limma[5], using Bayesian statistics, edgeR[6], and DESeq1-2[7,8] have been developed to improve the statistical power (sensitivity) of DEGs while suppressing false positives. These methods have often been applied to cases in which only a small number of samples can be obtained because of the experimental scale and cost limitations[9].

bioRxiv preprint doi: https://doi.org/10.1101/2021.03.12.435089; this version posted March 12, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Nevertheless, with the development of comprehensive gene expression analysis, especially single-cell analysis[10], new challenges have arisen as a result of the rapid increase in the number of samples[11] and the involvement of exploratory analysis schemes. The presence of a large sample size along with the application of exploratory analysis inducing selection bias can lead to an overly small P-value, impeding the application of conventional methods to call differentially expressed genes (DEGs) for bulk RNA-seq[12]. In particular, because in principle the P-value decreases with increasing sample size, even minor variations are considered statistically significant with large sample sizes, resulting in the unnecessary expansion of candidate genes (e.g., the definition of the *t*-statistic is proportional to $\sqrt{n}$). Therefore, efforts to improve the ability to detect DEGs are still being made to adapt to scRNA-seq data with the characteristics of low coverage and large sample size. That is, one of the major problems to be solved in this field is gene prioritization, namely, the selection of a small list of genes that should be validated and interpreted as a priority.

Here, we propose discriminative feature of cells (DFC), an alternative approach to the use of DEGs for characterizing cell populations by discrimination and variable selection. We demonstrated that DFC succeeded in selecting a small set of genes that characterize a group of cells of interest, while avoiding the problem of a large candidate gene list due to the large sample size of scRNA-seq. DFC is also shown to have the potential to provide biological insights that are difficult to find using DEGs, such as detecting genes that characterize small subpopulations and specific compositions of genes that are functionally linked to each other.

bioRxiv preprint doi: https://doi.org/10.1101/2021.03.12.435089; this version posted March 12, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

RESULTS

We focused on the discriminative method to characterize a population of interest (POI) in cells (e.g., stem cell population), using scRNA-seq data from samples that contain a large number of cell types, such as tissues. In the discriminative method, each cell is considered as a data point in the gene expression space, and the boundary surface that separates the two groups is determined. The boundary surface is confined to a small dimension of the space by variable selection, which suggests the idea of characterizing cell populations by a selected subset of gene expression patterns, as DFC. (Fig. 1a). While the conventional concept of using a DEG-based approach involves the comparison of group means of individual gene expression levels in two cell populations, the POI and a control group, DFC obtains a group of genes that are useful to discriminate POI from the control group. Thus, DFC is supposed to provide a highly selective gene list of only the number of genes needed for discrimination, while simultaneously using information from all genes. In this study, we implemented the method for determining DFC using binary classification by logistic regression and variable selection by adaptive LASSO[13]. Specifically, we performed adaptive LASSO−logistic regression with the objective variable of belonging to the POI (1 or 0) and the explanatory variable of gene expression level; the genes whose weights were not 0 were considered as DFC.

First, we explored which characteristics of the POI are considered important by DFC. Because DFC could be related to multiple variables, it could potentially distinguish correlated gene pairs or mixed subpopulations. Therefore, we generated artificial scRNA-seq data assuming the above two scenarios of gene expression pattern that may be prioritized in DFC. We applied variable selection using adaptive LASSO–logistic regression on these artificial data to examine the characteristics of gene expression patterns of cell populations that tend to be DFC.

bioRxiv preprint doi: https://doi.org/10.1101/2021.03.12.435089; this version posted March 12, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

DFC calls a pair of genes with dependences in their expression

In the first case, we tested the possibility that adaptive LASSO would prioritize correlated gene pairs. As the gene expression pattern of a simulated cell population, the expression levels of four genes were generated from a normal distribution with equal differences in group means and identical variances (Fig. 1b). Therefore, all genes are equivalent as DEGs (i.e., they have the same P-value in the two-tailed t-test). However, only the gene pair X_3–X_4 is correlated (r=0.7) within the two groups, while all other pairs are uncorrelated (r=0). The expression of these hypothetical genes is shown in Figure 1c as a scatter plot. The process of variable selection (LASSO solution path) is shown in Figure 1d, indicating that, in the process of increasing the parameter λ, which adjusts the sparsity (the strength of variable selection), the correlated X_3–X_4 pair is finally selected (Fig. 1d). The correlated pair has the clearest boundary between the groups (Fig. 1c, bold box), and it can be intuited that the two selected variables are a useful variable pair for differentiating groups A and B. In addition, the accuracy of the discrimination using all four variables is 0.999, while the accuracy using only the two selected variables is 0.995, indicating that the model maintains adequate performance. These results indicate that gene pairs that are correlated within a group are likely to be selected as DFC. This suggests that DFC can select a set of genes that have direct or indirect dependences as useful features for discriminating groups.

Next, we examined the ability of DFC to detect mixed subpopulations. Here, we assume a situation that includes multiple subpopulations with different expression patterns in one group (Fig. 1e). As in the previous scenario, all genes are equivalent as DEGs, that is, they have the same variance, and the group means are the same for all variables (see Methods for details). Here, genes X_1 and X_2 are strongly expressed in only one-third of the cells in group B. Furthermore, the expression of X_1 and that of X_2 are mutually exclusive. Therefore, X_1 and X_2 are not statistically independent. In addition,

bioRxiv preprint doi: https://doi.org/10.1101/2021.03.12.435089; this version posted March 12, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

because group B is composed of multiple subpopulations, the distributions of gene expression of both X_1 and X_2 are multimodal (Fig. 1f). In this example, adaptive LASSO also prioritized the non-independent variable pair X_1 and X_2 (Fig. 1g). This suggests that DFC is generally prone to selecting non-independent pairs of genes. In this scenario, the condition for being in cell group A is the expression of both X_1 and X_2 at the same time, that is, the logical AND (&) relation, which cannot be realized by expressing X_1 or X_2 alone.

These results suggest that DFC may provide useful insights when considering cellular functions, not simply as a candidate list of differentially expressed genes, but as a set of genes that well defines characteristics of a cell population, including dependences such as correlations among multiple genes and mixtures of different populations.

bioRxiv preprint doi: https://doi.org/10.1101/2021.03.12.435089; this version posted March 12, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Fig1_Fujii

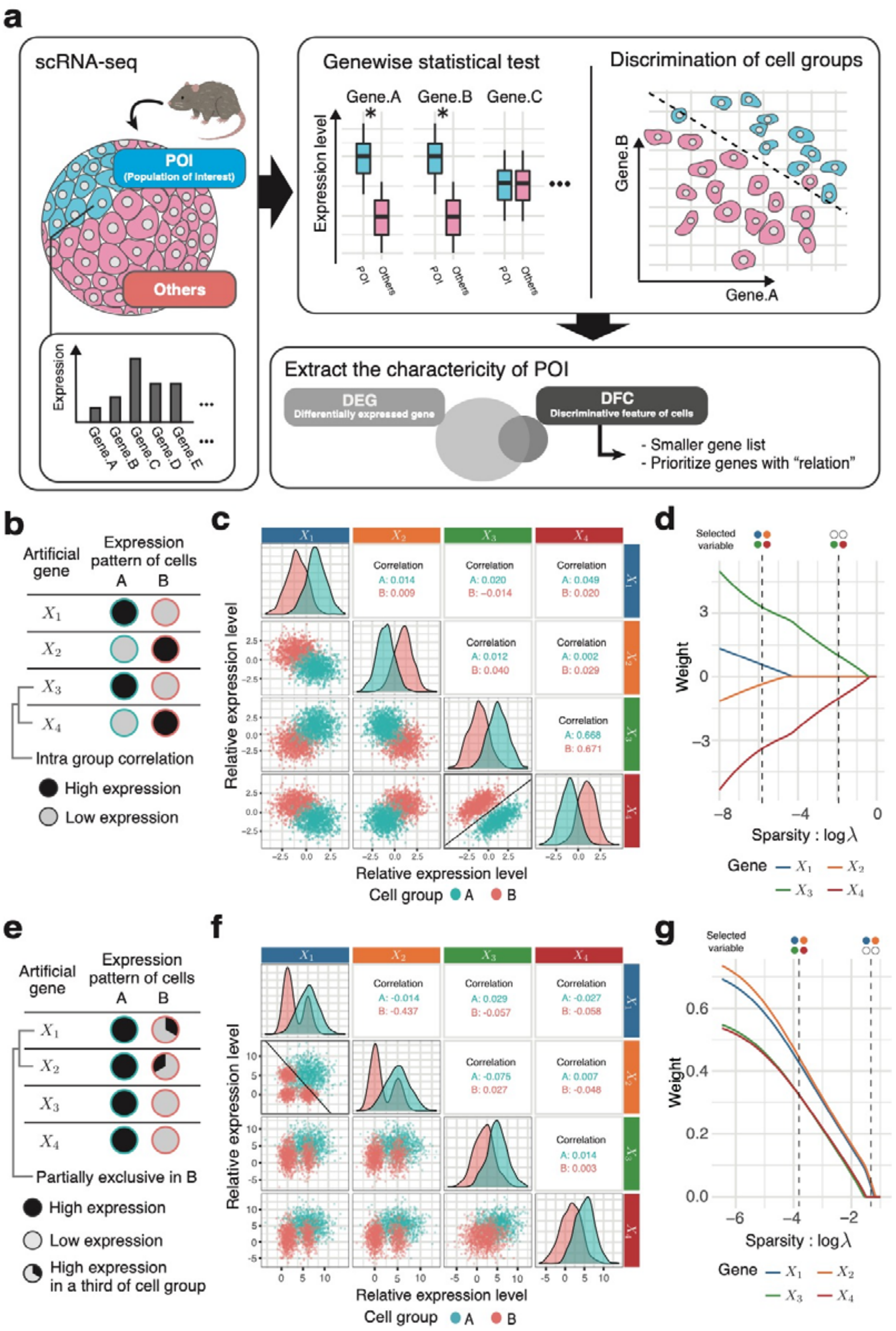


Fig. 1: The dependent pairs of gene expression selected as the DFC

bioRxiv preprint doi: https://doi.org/10.1101/2021.03.12.435089; this version posted March 12, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

(a) Different concepts for gene selection of DFC and DEG. The common goal is to extract a set of genes that characterizes the population of interest. A DEG-based approach involves a list of genes with statistically significant differences between the studied groups. In contrast, DFC-based approach involves a subset of genes that distinguish between two populations. DFC is expected to feature a small set of genes selected by taking into account the relationships among genes. (b–d) Artificially generated data set in which DFC has priority over DEG; case 1: correlation. (b) Schematic of the synthesized data design. Only the pair X_3 and X_4 has intra-group correlation; the other pairs are independent. All variables have the same variance, and the differences in means are the same for all pairs (see Methods for details). (c) Pairs that are easier to classify are given priority to become DFCs. The lower triangle shows the plot of each pair of variables; the diagonal elements show the distribution of each variable and the upper triangle shows the correlation coefficient within the cluster of each two variables. The separating boundary of the selected variable pair X_3 and X_4 is shown as solid line. (d) The process of selecting discriminative variables; solution path. This indicates transition of the weights (partial regression coefficients) of each variable when regularization parameter λ (sparsity) is varied. (e–g) Synthesized data set in which DFC has priority over DEG; case 2: exclusive. (e) Schematic of the synthesized data design. In one-third of the group A cells, the expression of X_1 and that of X_2 are mutually exclusive. The variances and the means of variables are designed as in case 1. In other words, this simulates a logical product relationship such that cells that express X_1 and X_2 simultaneously are equivalent to the population of group A. (f) An example of logical relationships of case 2, shown in a scatter plot as in (c). (g) The solution path in case 2 as shown in (d).

bioRxiv preprint doi: https://doi.org/10.1101/2021.03.12.435089; this version posted March 12, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Small gene set of DFC determined by a unique criterion that differs from DEG

Next, to demonstrate the practical applicability of DFC, we performed scRNA-seq data analysis and compared the yielded gene lists of DEG and DFC. The data were obtained from scRNA-seq data of mouse tibialis anterior (TA) muscle tissue injured by notexin at days 0, 2, 5, and 7 during the regeneration of skeletal muscle[14]. Skeletal muscle satellite cells (MuSCs) are known to be an essential cell population for muscle regeneration, and are activated upon muscle injury and undergo multiple progenitor cell stages until differentiating into myofibers. In this process of muscle regeneration, there is also a self-renewing state in which MuSCs again transition to a quiescent state and refill the MuSC pool[15,16]. Therefore, we attempted to characterize heterogeneous cell populations of MuSCs with multiple transient states by DFC, which has been difficult to capture by a DEG-based approach.

Figure 2a shows the procedure for defining the POI and the determination of DFC. First, public data of scRNA-seq (GSE143437)[14] were obtained. Then, the genes expressed in very few cells (fewer than 10 cells) were filtered (see Methods for details). Next, UMAP was used to visualize the data in two dimensions. The 12th cluster obtained by Louvain clustering (Fig. 2b) was designated as the POI for this analysis, and all other clusters were designated as the control group (Fig. 2c). Cluster 12 was selected according to the expression levels of *Pax7*, a marker gene for MuSCs, and *Myod1*, a transcription factor that functions in progenitor cells and activated satellite cells. The specific expression of the genes indicated that cluster 12 is the satellite/progenitor cell cluster (Fig. 2d, e).

Next, to elucidate the differences in the criteria for selecting genes in DFC and DEG in the data, we compared their gene lists. For DEGs, the criterion of FDR < 0.01 in DESeq2 was used. As a result of applying this criterion, the number of DEGs was 7,495, which was nearly half of the total (42%) of

bioRxiv preprint doi: https://doi.org/10.1101/2021.03.12.435089; this version posted March 12, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

17,730 mouse genes. Among the DFCs, most of the genes (96/108) overlapped with the DEGs, but there were also 12 DFC-specific genes (Fig. 2f). Next, we examined whether genes in DFC could be obtained by adjusting the gene selection criteria such as the P-value and log2 fold change in DEGs. Figure 2g shows the position of genes selected as DFC among all DEGs by a volcano plot. The large sample size of scRNA-seq and the statistical test on the clusters, which were also determined using the same scRNA-seq data, resulted in the overly small FDRs. In addition, the genes selected as DFC among the DEGs were scattered irregularly. The results suggested that genes in DFC were selected independently of the DEG criteria. Similarly, in the MA plot (Fig. 2h), genes in DFCs were found to be scattered among the DEGs, indicating that the DFC selection was also independent of the gene expression levels. These results indicate that DFC selects genes according to its own criteria and also selects genes that are more useful for discrimination of the POI among the DEG candidates.

bioRxiv preprint doi: https://doi.org/10.1101/2021.03.12.435089; this version posted March 12, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

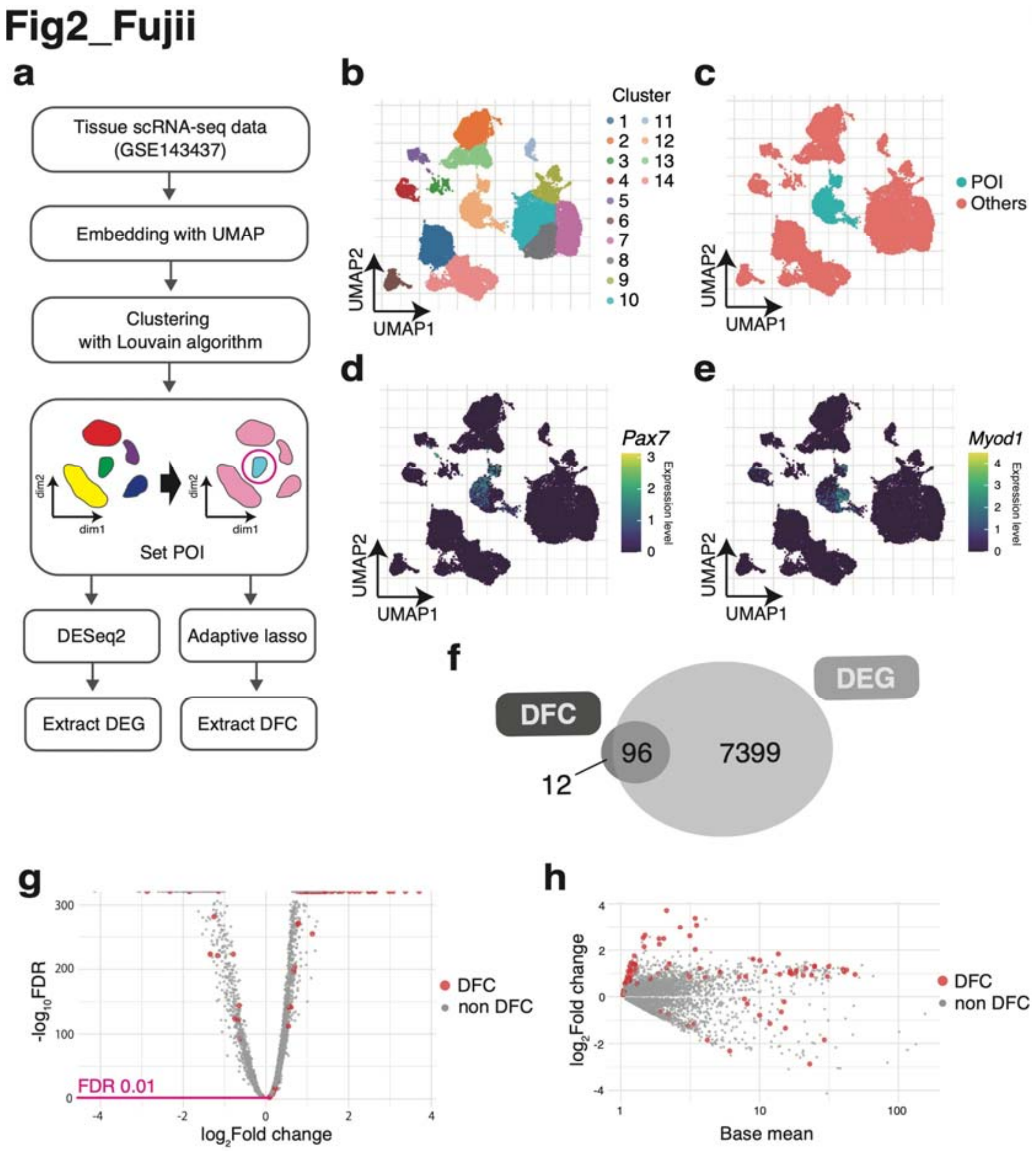


Fig. 2: Smaller gene set of DFC was selected by a unique selection criterion

(a) Procedure of extracting DEG and DFC from scRNA-seq data. (b–e) The determined POI is compared with all other cell clusters in the muscle tissue. Embedding the scRNA-seq data into two-dimensional space with UMAP. (b) The clusters determined by the Louvain algorithm. The 12th cluster corresponds to the cluster of muscle stem cells and progenitors. (c) The 12th cluster is set as the POI, and the other clusters are assigned as the control group, “Others.” (d, e) Single-cell

bioRxiv preprint doi: https://doi.org/10.1101/2021.03.12.435089; this version posted March 12, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

expression levels for *Pax7* and *Myod1*. (f) A part of the DEGs are selected as DFC. Venn plot indicating the overlap of DEGs and DFC. (g) Genes in DFC not selected by the DEGs' criteria. Volcano plot of DEGs and (h) MA plot of DEGs.

bioRxiv preprint doi: https://doi.org/10.1101/2021.03.12.435089; this version posted March 12, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

DFC is useful to identify the combinatorial patterns of gene expression and minor subpopulations

Next, we investigated whether DFC can extract the biological function of the POI. To evaluate this, we interpreted how genes in DFC help to discriminate the POI by referring to known marker genes of MuSC or muscle progenitors. First, we classified the DFC genes into three groups according to the specificity of expression in each cluster: The “Strong” feature refers to the genes expressed in 25% or more of the cells in one or two clusters. The “Weak” feature refers to the genes expressed in three or more clusters, as depicted in Figure 3a. The “Niche” feature is defined as genes with minor expression in less than 25% of the cells in all clusters (Fig. 3b for the annotation of clusters, Fig. S1 for the original annotation by the authors of scRNA-seq data and S2-3 for highlighted expression of all genes in DFC on the UMAP visualization).

Genes belonging to the Strong feature included many of the genes known as markers of MuSCs and activated satellite cells (Fig. 3c, Fig. S3a). M-cadherin (*Cdh15*)[17] was expressed almost universally in the POI. The results of GO enrichment analysis using only the Strong feature showed that many genes are related to skeletal muscle cells in muscle tissue (Fig. 3d). In addition, *Myf5* is a representative feature of the group of cells that are not represented by *Myog* and *Myod1* in the POI (Fig. S3a). Thus, it can be interpreted that *Myf5* plays a different role from other myogenic regulatory factors[18]. These results indicate that DFC can select genes that correspond to single biomarkers, similar to DEG.

The Weak feature (Fig. 3e, Fig. S2) included Kai (*Cd82*)[19–21] and a group of genes used as quiescent satellite cell markers such as syndecan-4 (*Sdc4*). In contrast, the negative LASSO weight of *Col6a2*, which is selectively expressed in fibro/adipogenic progenitors (FAPs), helps to characterize the POI as non-FAPs (Fig. 3e). Furthermore, a notable property of the Weak feature

bioRxiv preprint doi: https://doi.org/10.1101/2021.03.12.435089; this version posted March 12, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

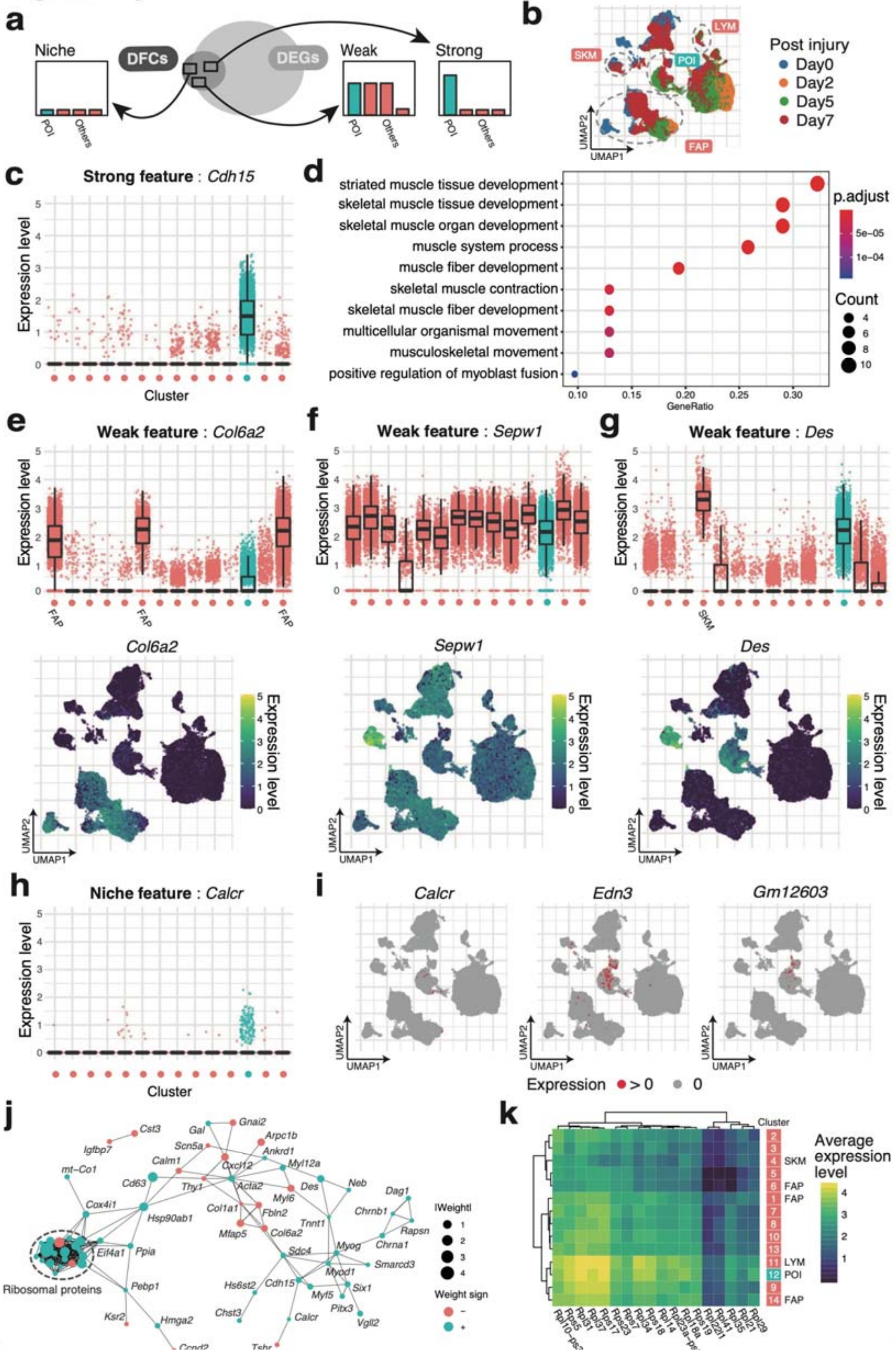
involved the combined expression pattern using multiple genes. For example, *Des* is expressed in both a part of the POI and the mature SKM population, which is a part of Others. This means that *Des* itself cannot characterize the POI (Fig. 3g, Fig. S2). Therefore, to exclude the character of mature SKM from the POI, the condition of *Sepw1*(–), which is significantly expressed in mature SKM, is additionally imposed (Fig. 3f). Thus, DFC can be used as a molecular marker to identify specific cell types for immunostaining and cell sorting, for example, *Des*(+)*Sepw1*(–)*Col6a2*(–).

The Niche feature detects minor subpopulations scattered within the POI (Fig. 3h). These genes in DFC are difficult to prioritize in the list of DEG (i.e., they tend to have larger P-values). In fact, *Calcr*, which is ranked lower than 5,000th in the order of *P*-values in the DEGs, is known to be expressed transiently in a quiescent state. We also detected *Edn3*[22], which has a prominent localization on day 7 after injury in the POI, and *Gm12603* (*Wincr1*; WNT-induced noncoding RNA), a gene expressed in a different cell group from *Calcr* and *Edn3*.

These binary combinations (+/–) of multiple genes and minor cell groups would appear as the representative cases demonstrated in Figure 1f. Therefore, DFC can characterize the best combinations of genes for determining cell type and even small subpopulations caused by transient expression or state changes, even using a small list of genes, and even upon comparison with the heterogeneous control group of cells in tissue.

bioRxiv preprint doi: https://doi.org/10.1101/2021.03.12.435089; this version posted March 12, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Fig3_Fujii



bioRxiv preprint doi: https://doi.org/10.1101/2021.03.12.435089; this version posted March 12, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Fig. 3: Biological significance of genes in DFC revealed by the discriminative ability

(a) According to the specificity of the expression, the genes in DFC are classified into three groups. The three groups are named Strong (specific to 1–2), Weak (>2), and Niche features (none of them). (b) The data are from samples collected at 0, 2, 5, and 7 days after skeletal muscle injury. In addition, the clusters of fibro/adipogenic progenitor (FAP), mature skeletal muscle (SKM), and lymphocytes (LYM) are shown. (c) Genes specifically expressed in the POI are assigned to the Strong features. For the Strong feature *Cdh15*, its expression level for each cluster shown in Fig. 2b is plotted. The medians, 25/75th percentiles, and 1.5 IQR (inter-quartile range) are employed to draw the box plots. (d) The Strong features contain many genes that act as markers of skeletal muscle. The results of GO enrichment analysis for the Strong features. GOs are ordered by the proportion of their inclusion in the Strong features. (e) Genes expressed in some clusters are assigned to the Weak features. For the Weak feature *Col6a2*, its expression level is plotted in the upper panel, and the single-cell expression level on UMAP visualization is plotted in the lower panel. (f, g) The expression levels of *Sepw1* and *Des*, two of the Weak features, are plotted as (e). (h) Genes with low expression levels that are expressed in a minor subpopulation of the POI are assigned to the Niche feature. For the Niche feature *Calcr*, its expression level is plotted as (c). (i) Cells expressing Niche features (*Calcr*, *Edn3*, and *Gm12603*) are highlighted. (j) DFCs have the property of capturing interrelated genes. STRING is used to connect related DFCs. (k) Ribosomal proteins are a notable example of Weak features in DFC that are difficult to interpret in the context of binary combinations. Eighteen ribosomal protein genes in DFC are averaged in each cluster as a heat map.

bioRxiv preprint doi: https://doi.org/10.1101/2021.03.12.435089; this version posted March 12, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

DFC extracts genes harboring functional associations

Finally, we attempted to elucidate more complex functional associations between genes from the obtained DFC. The artificial data in Figure 1 show that DFC selects the pair of genes with dependences among DEG-equivalent genes. This suggests that DFCs may tend to contain the network of many-to-many gene relations that forms the unique characteristics of each cell type.

First, we examined the correlation matrix of expression levels for the genes in DFCs, and found that the POI showed a more distinct hierarchical structure than the Others group (Fig. S4). We further evaluated the functional associations of 108 genes by STRING[23] (see Fig. 3j, Fig. S5–6 for details). The 104 genes in DFC were contained in the STRING database and had 257 association edges, which is a larger number than would occur by chance (87 edges), indicating that these are a set of genes that are strongly related to each other (PPI enrichment p-value $< 10^{-6}$).

Next, we attempted to interpret the network by referring to the characteristics of the adaptive LASSO−logistic regression. First, we examined the correspondence with the weights of adaptive LASSO. We found that the two clusters of genes related to the function of FAPs with negative weights and the clusters of genes related to the activation of MuSCs with positive weights, such as *MyoD* and *Myog*, were linked via the proteoglycan *Sdc*4 (Fig. 3j). Among them, the most remarkable result was obtained for the ribosomal protein-coding genes, which formed a dense cluster of ribosome subunit components. All of these genes are included in the Weak feature. The expression patterns of these genes are not clearly segregated into clusters and are difficult to interpret by binary combinations (Fig. S2), suggesting that they reflect a certain composition of ribosomal protein-coding genes that may have a critical function in MuSCs and progenitors. To confirm the discriminative ability of the genes, we extracted only the ribosomal protein-coding genes in the DFC, and visualized them by PCA. The results confirmed that the subset of DFC was sufficient to

bioRxiv preprint doi: https://doi.org/10.1101/2021.03.12.435089; this version posted March 12, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

discriminate the POI from the Others (Fig. S7a, b). In more detail, the composition of these genes was characterized by higher overall expression levels compared with the Others, with *Rpl31* and *Rps37* being particularly highly expressed (Fig. 3k). In contrast, the POI has a ribosomal protein profile closest to that of lymphocytes (LYM), but the negative weight of *Rpl34* appears to act to differentiate the POI from the lymphocyte population. Furthermore, these ribosomal protein-coding genes clearly captured the temporal changes after muscle injury in the POI (Fig. S7c, d). This suggests that a certain composition of ribosomal protein-coding genes is the critical factor in stem cell functions, such as self-renewal[24–26], and supports their role in muscle regeneration[27]. In conclusion, DFC can extract a small set of genes that characterize the POI, including functional associations between genes.

bioRxiv preprint doi: https://doi.org/10.1101/2021.03.12.435089; this version posted March 12, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

DISCUSSION

In this paper, we have proposed a new concept of characterizing a cell population of interest, which is an alternative to the DEG-based approach that uses lists of genes with differences in expression between groups. Our method, can be termed a discriminative approach, has potential applications in the task of cell characterization. In particular, given the recent developments of high-throughput biological measurements, the statistical models based on discrimination can be effective for the increased sample size of scRNA-seq (capable cell number).

To select a small number of genes to characterize a cell, as in a DEG-based approach, rather than to determine the cell type itself, a variable selection procedure was employed to select a small set of genes that are effective for discrimination. Variable selection is a methodology that selects a small number of $M < N$ optimal combinations of variables from N input variables, while preserving the predictive performance of the statistical model. Several methods of variable selection with discriminative models have been developed, such as SVM[28], and logistic regression with LASSO penalty. Among them, LASSO–logistic regression is a method that can construct an interpretable linear model and perform variable selection in one step. However, the gene clusters obtained by these discriminative methods have been mostly used as gene signatures in cell-type classification (e.g., a cell is normal or malignant[29]), and no attempt has been made to interpret these gene signatures themselves biologically.

In this study, we compared two different population groups: the POI, which is specified after nonlinear dimensionality reduction and clustering, and the rest of the population. In this paper, this comparison was assumed to be the most frequently used procedure for profiling unknown cell populations using scRNA-seq data. However, DE analysis after clustering has been criticized for introducing selection bias, which results in excessively low P-values[12]. This exploratory data

bioRxiv preprint doi: https://doi.org/10.1101/2021.03.12.435089; this version posted March 12, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

analysis of scRNA-seq makes the proper use of P-values more difficult, while our method bypasses the use of P-values. In addition, it can be assumed that a control group including heterogeneous populations will increase the variance in the group and lead to large P-values. For this reason, DE may miss subtle changes of state or fail to discover minor subpopulations within the cell groups of interest. In other words, calling DEGs may not be the best strategy for exploratory discovery in a mixed cell population represented by tissue. Furthermore, a simple two-group comparison of one vs. others is practical enough and thus is one of the major advantages of our method. The reason why this easy comparison works well is that DFC combines multiple Weak/Niche features in order to improve discriminative performance and pick up even small populations. The advantage comes from the linearity of the model adapted in DFC; that is, the results can be interpreted as a superposition of features, as described above.

As an implementation of the concept of DFC, we employed the framework of binary classification with logistic regression and variable selection with adaptive LASSO. In addition to its several beneficial statistical properties (e.g., consistency in variable selection), adaptive LASSO has been shown to have superior practical performance among the improved versions of original LASSO[30]. Although there are many methods for variable selection, such as best subset selection (L0)[31], random forest[32], and SVM[28], in this study, we did not provide the benchmark tests of each method. However, we believe that how the mathematical properties of each method are utilized in the various scenarios of scRNA-seq data analysis is an important topic. In this paper, we have discussed the usefulness of the discriminative method when the dependence among all genes is included. Indeed, Ntranos et al. shed light on a discriminative approach that uses logistic regression to detect isoform-level gene expression changes from scRNA-seq data[33]. We also found the usefulness of the discriminative approach, especially in the analysis of tissue scRNA-seq data where gene expression correlations

bioRxiv preprint doi: https://doi.org/10.1101/2021.03.12.435089; this version posted March 12, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

and subpopulations within the same population are expected to be mixed. The further development of methods that focus on the interpretation of large-scale data is anticipated.

MATERIALS & METHODS

Setting POI of scRNA-seq data

Normalized count matrix and the annotations of cells of scRNA-seq data were downloaded from GEO (GSE143437). We filtered out genes that were expressed (> 0) in fewer than 10 cells. The UMAP visualization was performed using the *uwot* R package (version 0.1.10)[34]. In the embedded two-dimensional space, the clusters were determined by the Louvain method[35] implemented in the *igraph* R package[36] (version 1.2.6). In our analysis of scRNA-seq data from muscle tissue, the POI was set as the cluster in which the majority of cells expressed both *Pax7* (53.3%) and *Myod1* (42.6%).

Adaptive LASSO–logistic regression

The adaptive LASSO–logistic regression was performed using the *glmnet*[37] (version 4.1) R package. To perform the adaptive LASSO, we followed the two steps of parameter estimation: fit ridge and then fit LASSO regression with the penalty factor. The penalty factor was set to be $1/|\boldsymbol{\beta}_{\text{ridge}}|$, where $\boldsymbol{\beta}_{\text{ridge}}$ is estimated by ridge regression in the first step. The sparsity parameter λ was determined by 10-fold cross-validation (*cv.glmnet)* of binomial deviance. To reduce the computational cost in real scRNA-seq data, we obtained 30% subsamples of cells from each cluster in the estimation of $\boldsymbol{\beta}_{\text{ridge}}$.

bioRxiv preprint doi: https://doi.org/10.1101/2021.03.12.435089; this version posted March 12, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Differential expression analysis

A raw count matrix was downloaded from GEO (GSE143437). The Wald test was performed using the *DESeq2*[38] (version 1.28.1) R package to identify genes that were differentially expressed (DEGs) between the POI and the Others. The parameters were used as the default settings. Genes with FDR < 1% were considered significantly differentially expressed.

Synthetic data generation

The artificial data set consists of randomly generated data points (cells) with four variables (genes). We set the variables as the *equivalent genes* in terms of DE. Because the statistical significance of a DEG that is estimated by the *z*- or *t*-statistic is uniquely determined by variances and the difference of means between groups, we only modified the relationships of the genes, while maintaining the variance and difference of means. Specifically, the two cases of DE-equivalent genes were generated as follows. Case I: Correlated expression. All four variables follow the Gaussian distribution with constant variance $\sigma^2 = 1$ and difference of means $|\mu_A - \mu_B| = 2$, where A and B indicate groups of cells (each of 1,000 cells). All pairs of variables are independent ($r=0$), except for the pair (X_3, X_4) having a strong correlation $r=0.7$ in both groups. Case II: Heterogenous population. All four variables have the same variance σ^2 and the same group means $\mu_{A,}$ μ_B. We set X_1 and X_2 of group B as having an exclusive relationship. We divided group B into three subgroups (B1–3: each of 333 cells). Group B1 expresses X_1, group B2 expresses X_2, and group B3 expresses none of them. The others are independent Gaussian

bioRxiv preprint doi: https://doi.org/10.1101/2021.03.12.435089; this version posted March 12, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

variables. To equalize the variance and means in group B to those in the others, we used the mixture distribution as the marginal distribution of X_1 and X_2 in B:

$$f = pg(\mu_1, \sigma_1^2) + (1-p)g(\mu_2, \sigma_2^2),$$

where g is the Gaussian probability density function and p is the proportion of subgroup relative to the size of group B. In general, the mean and variance of f are calculated as follows:

$$\begin{aligned} \mu &= p\mu_1 + (1-p)\mu_2 \\ \sigma^2 &= p\sigma_1^2 + (1-p)\sigma_2^2 + p(1-p)(\mu_1 - \mu_2)^2. \end{aligned}$$

We used the parameters: $\sigma_1{}^2 = \sigma_2{}^2 = 1$, $\mu_1 = 0$, $\mu_2 = 5$, $p = 2/3$, and hence $\mu_B = 5/3$ and $\sigma^2 = 59/9$ for all variables. We set $\mu_A = 5$.

Code Availability

The codes used for the DFC extraction and analysis are available at: https://github.com/tfwis/DFC

ACKNOWLEDGMENTS

Computations were carried out using the computer resources offered under the category of Intensively Promoted Projects by the Research Institute for Information Technology at Kyushu University. This work was supported in part by JST PRESTO JPMJPR2026 to K.M.; JST CREST JPMJCR16G1 to Y.O.; MEXT/JSPS KAKENHI JP19H04970, JP19H03158, and JP20H05393 to K.M.; and JP18H04802, JP18H05527, JP19H05244, JP17H03608, JP20H00456,

bioRxiv preprint doi: https://doi.org/10.1101/2021.03.12.435089; this version posted March 12, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

and JP20H04846 to Y.O. The authors thank Edanz (https://en-author-services.edanz.com/ac) for editing the English text of a draft of this manuscript.

AUTHOR CONTRIBUTIONS

T.F. and K.M. analyzed the data. K.M. performed statistical analysis. T.F., K.M., M.F., and Y.O. wrote the paper. All authors read and approved the final manuscript.

REFERENCES

1. Benjamini, Y. & Hochberg, Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *J. R. Stat. Soc. Ser. B* **57**, 289–300 (1995).

2. Storey, J. D. & Tibshirani, R. Statistical significance for genomewide studies. *Proc. Natl. Acad. Sci. U. S. A.* **100**, 9440–9445 (2003).

3. Storey, J. D. The positive false discovery rate: a Bayesian interpretation and the q -value. *Ann. Stat.* **31**, 2013–2035 (2003).

4. Efron, B., Tibshirani, R., Storey, J. D. & Tusher, V. Empirical bayes analysis of a microarray experiment. *J. Am. Stat. Assoc.* **96**, 1151–1160 (2001).

5. Ritchie, M. E. *et al.* Limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res.* **43**, e47 (2015).

6. MD Robinson, D. M. G. S. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26**, 139–140 (2010).

bioRxiv preprint doi: https://doi.org/10.1101/2021.03.12.435089; this version posted March 12, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

7. Anders, S. & Huber, W. Differential expression analysis for sequence count data. *Genome Biol.* **11**, R106 (2010).

8. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* **15**, 550 (2014).

9. Schurch, N. J. *et al.* How many biological replicates are needed in an RNA-seq experiment and which differential expression tool should you use? *RNA* **22**, 839–851 (2016).

10. Soneson, C. & Robinson, M. D. Bias, robustness and scalability in single-cell differential expression analysis. *Nat. Methods* **15**, 255–261 (2018).

11. Svensson, V., Vento-Tormo, R. & Teichmann, S. A. Exponential scaling of single-cell RNA-seq in the past decade. *Nature Protocols* vol. 13 599–604 (2018).

12. Zhang, J. M., Kamath, G. M. & Tse, D. N. Valid Post-clustering Differential Analysis for Single-Cell RNA-Seq. *Cell Syst.* **9**, 383-392.e6 (2019).

13. Zou, H. The Adaptive Lasso and Its Oracle Properties. (2006) doi:10.1198/016214506000000735.

14. De Micheli, A. J. *et al.* Single-Cell Analysis of the Muscle Stem Cell Hierarchy Identifies Heterotypic Communication Signals Involved in Skeletal Muscle Regeneration. *Cell Rep.* **30**, 3583-3595.e5 (2020).

15. Motohashi, N. & Asakura, A. Muscle satellite cell heterogeneity and self-renewal. *Frontiers in Cell and Developmental Biology* vol. 2 (2014).

16. Lazure, F. *et al.* Myf6/MRF4 is a myogenic niche regulator required for the maintenance of the muscle stem cell pool. *EMBO Rep.* **21**, 1–19 (2020).

17. Wróbel, E., Brzóska, E. & Moraczewski, J. M-cadherin and β-catenin participate in differentiation of rat satellite cells. *Eur. J. Cell Biol.* **86**, 99–109 (2007).

bioRxiv preprint doi: https://doi.org/10.1101/2021.03.12.435089; this version posted March 12, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

18. Conerly, M. L., Yao, Z., Zhong, J. W., Groudine, M. & Tapscott, S. J. Distinct Activities of Myf5 and MyoD Indicate Separate Roles in Skeletal Muscle Lineage Specification and Differentiation. *Dev. Cell* **36**, 375–385 (2016).

19. Uezumi, A. *et al.* Cell-Surface Protein Profiling Identifies Distinctive Markers of Progenitor Cells in Human Skeletal Muscle. *Stem Cell Reports* **7**, 263–278 (2016).

20. Barruet, E. *et al.* Functionally heterogeneous human satellite cells identified by single cell RNA sequencing. *Elife* **9**, (2020).

21. Camps, J. *et al.* Interstitial Cell Remodeling Promotes Aberrant Adipogenesis in Dystrophic Muscles. *Cell Rep.* **31**, (2020).

22. Fukada, S. *et al.* Molecular Signature of Quiescent Satellite Cells in Adult Skeletal Muscle. *Stem Cells* **25**, 2448–2459 (2007).

23. Szklarczyk, D. *et al.* STRING v11: Protein-protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. *Nucleic Acids Res.* **47**, D607–D613 (2019).

24. Fortier, S., MacRaea, T., Bilodeau, M., Sargeant, T. & Sauvageau, G. Haploinsufficiency screen highlights two distinct groups of ribosomal protein genes essential for embryonic stem cell fate. *Proc. Natl. Acad. Sci. U. S. A.* **112**, 2127–2132 (2015).

25. Khajuria, R. K. *et al.* Ribosome Levels Selectively Regulate Translation and Lineage Commitment in Human Hematopoiesis. *Cell* **173**, 90-103.e19 (2018).

26. Guthikonda, P. K. *et al.* Polymorphic dynamics of ribosomal proteins gene expression during somatic cell reprogramming and their differentiation in to specialized cells-types. *bioRxiv* 114868 (2017) doi:10.1101/114868.

27. Figueiredo, V. C. & McCarthy, J. J. Regulation of ribosome biogenesis in skeletal muscle hypertrophy. *Physiology* vol. 34 30–42 (2019).

bioRxiv preprint doi: https://doi.org/10.1101/2021.03.12.435089; this version posted March 12, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

28. Guyon, I., Weston, J., Barnhill, S. & Vapnik, V. Gene selection for cancer classification using support vector machines. *Mach. Learn.* **46**, 389–422 (2002).

29. Lopes, M. B. & Vinga, S. Tracking intratumoral heterogeneity in glioblastoma via regularized classification of single-cell RNA-Seq data. *BMC Bioinformatics* **21**, 1–12 (2020).

30. Freijeiro-González, L., Febrero-Bande, M. & González-Manteiga, W. A critical review of LASSO and its derivatives for variable selection under dependence among covariates. (2020).

31. Bertsimas, D., King, A. & Mazumder, R. Best subset selection via a modern optimization lens. *Ann. Stat.* **44**, 813–852 (2016).

32. Deng, H. & Runger, G. Gene selection with guided regularized random forest. *Pattern Recognit.* **46**, 3483–3489 (2013).

33. Ntranos, V., Yi, L., Melsted, P. & Pachter, L. A discriminative learning approach to differential expression analysis for single-cell RNA-seq. *Nat. Methods* **16**, 163–166 (2019).

34. McInnes, L., Healy, J. & Melville, J. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. *arXiv* (2018).

35. Blondel, V. D., Guillaume, J. L., Lambiotte, R. & Lefebvre, E. Fast unfolding of communities in large networks. *J. Stat. Mech. Theory Exp.* **2008**, (2008).

36. CSARDI & G. The igraph software package for complex network research. *InterJournal Complex Syst.* **1695**, (2006).

37. Friedman, J., Hastie, T. & Tibshirani, R. Regularization paths for generalized linear models via coordinate descent. *J. Stat. Softw.* **33**, 1–22 (2010).

38. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* **15**, 550 (2014).