Predicting reprogramming-related gene expression from cell morphology in human induced pluripotent stem cells

Takashi Wakui, Mitsuru Negishi, Yuta Murakami, Shunsuke Tominaga, Yasushi Shiraishi, Anne Carpenter, Shantanu Singh, and Hideo Segawa

Corresponding author(s): Takashi Wakui, FUJIFILM

Review Timeline:

Submission Date:	2022-06-21
Editorial Decision:	2022-08-01
Revision Received:	2023-01-28
Editorial Decision:	2023-02-14
Revision Received:	2023-02-26
Accepted:	2023-02-27

Editor-in-Chief: Matthew Welch

Transaction Report:

(Note: With the exception of the correction of typographical or spelling errors that could be a source of ambiguity, letters and reports are not edited. The original formatting of letters and referee reports may not be reflected in this compilation.)

RE: Manuscript #E22-06-0215
TITLE: Predicting gene expression from cell morphology in human induced pluripotent stem cells

Dear Mr. Wakui:

Your manuscript has now been seen by 3 reviewers. As you can see from their comments, while they find the methods and concepts in the manuscript interesting and of value, two of the three reviewers bring up substantial concerns about the claims. In particular Reviewers 2 and 3 both point out the mixing of concepts between predicting iPSC quality and predicting gene expression. Based on my own reading of the paper, I agree with these two reviewers that while the paper demonstrates that iPSC quality can be predicted from cell morphology and gene expression can be predicted from cell morphology, the two concepts are not successfully connected together. Reviewer 3 particularly provides good suggestions about how to mitigate this major issue and I recommend the authors revise their manuscript taking into account all the reviewers' recommendations.

Sincerely,

Melike Lakadamyali
Monitoring Editor
Molecular Biology of the Cell

Dear Mr. Wakui,

The review of your manuscript, referenced above, is now complete. The Monitoring Editor has decided that your manuscript is not acceptable for publication at this time, but may be deemed acceptable after specific revisions are made, as described in the Monitoring Editor's decision letter above and the reviewer comments below.

A reminder: Please do not contact the Monitoring Editor directly regarding your manuscript. If you have any questions regarding the review process or the decision, please contact the MBoC Editorial Office (mboc@ascb.org).

When submitting your revision include a rebuttal letter that details, point-by-point, how the Monitoring Editor's and reviewers' comments have been addressed. (The file type for this letter must be "rebuttal letter"; do not include your response to the Monitoring Editor and reviewers in a "cover letter.") Please bear in mind that your rebuttal letter will be published with your paper if it is accepted, unless you haveopted out of publishing the review history.

Authors are allowed 180 days to submit a revision. If this time period is inadequate, please contact us at mboc@ascb.org.

Revised manuscripts are assigned to the original Monitoring Editor whenever possible. However, special circumstances may preclude this. Also, revised manuscripts are often sent out for re-review, usually to the original reviewers when possible. The Monitoring Editor may solicit additional reviews if it is deemed necessary to render a completely informed decision.

In preparing your revised manuscript, please follow the instruction in the Information for Authors (www.molbiolcell.org/info-for-authors). In particular, to prepare for the possible acceptance of your revised manuscript, submit final, publication-quality figures with your revision as described.

To submit the rebuttal letter, revised manuscript, and figures, use this link: Link Not Available

Please contact us with any questions at mboc@ascb.org.

Thank you for submitting your manuscript to Molecular Biology of the Cell. We look forward to receiving your revised paper.

Sincerely,

Eric Baker
Journal Production Manager
MBoC Editorial Office
mbc@ascb.org

Reviewer #1 (Remarks to the Author):

Review for "Predicting gene expression from cell morphology in human induced pluripotent stem cells"

Summary and Contributions: The authors propose a method for predicting gene expression from cell morphology using unsupervised feature extraction of live cell phase imaging data. This paper serves as a lovely proof-of-principle effort that illustrates the power of integrating imaging with sequencing data, which is bound to be an important part of the future of biological research. I also very much appreciate their taking care to highlight the limitations of their model and approach, avoiding the all-too-common glossy "sales pitch" style that many of these papers tend to adopt. While there are a few "holes" in explaining the rationale for certain strategies, I appreciate that the workflow was described in chunks with sufficient detail for potential non-experts to follow along, making this paper more suitable for a biological journal such as MBoC. Overall, this is a solid paper that deserves to be published with perhaps some minor revisions.

Minor points that could be clarified:
• What is their justification for using patches? Why not just learn on the entire image?
• What was their justification for model choice vs. other unsupervised models/Aes?
• Needs to be proofread a bit (missing parentheses, spelling errors etc).
• In intro, would be more useful to be more specific with placing the references that refer to specific studies. For example, in this section:

"Very recently, several studies have demonstrated that morphological features can be used to predict gene expression in various contexts (17-21For example, supervised deep learning models, such as a Convolutional Neural Network (CNN), were trained with datasets of pathological images and spatial gene expression profiles obtained by single-cell sequencing of the exact same sample to predict cancer gene expression from pathological images. The resulting models were able to accurately link morphology with gene expression, suggesting that a similar strategy could be extended beyond cancer and applied to the iPSC purification process. However, these strategies require expensive single-cell sequencing measurements, which limits their wide-spread adoption."

Which study are they citing in each sentence specifically?

• Figure 2 - the figure legend should indicate that Ni is number of images, and np is number of patches?
• It would be instructive to know what happens if you trained on both dataset A and B together as well to see if the resulting model generalizes better.

Reviewer #2 (Remarks to the Author):

The manuscript by Waku et al., describes development of a new method to use the VQ-VAE-2 unsupervised deep learning model to extract and reduce dimensionality of image features and use these to predict via SVR models the gene expression levels from images with the purpose of developing an improved workflow to identify clones of potential hiPSCs during reprogramming that are of high quality (e.g. high pluripotency/good reprogramming). There are several very promising components to this study. The new method seems to have very good image reconstruction potential and performed relatively well in a benchmark assesment of this method compared to CNN based methods. The data acquired e.g., expert-annotated image data and accompanying bulk RNAseq data for a total of 29 clones over 15 donors has all sorts of interesting potential for investigation and analysis. However, the current version of this manuscript suffers from a tension between development and assessment of a new model/method and the specific application that this model/method is being used for. As it stands now, I cannot recommend this manuscript for publication in this journal at this time. The need is beyond "major revision" or further analysis of particular parts of the manuscript story, but is a much more fundamental reassessment of what the goals of this manuscript are, which then would likely lead to an entirely different manuscritp requiring many additional layers of analysis and interpretation. I think the data iself and the types of methods these authors are capable of using together have great potential for a successful story but the current manuscript does not convey this potential yet.

To try to explain what I mean with "tension between possible goals of the manuscript" let me start by quoting two parts of the discussion (to avoid my own paraphrasing):

1) One goal was stated to be to develop a "strategy aimed at improving performance of the iPSC purification process by using cell morphology captured by phase-contrast microscopy to predict gene expression levels relating to stemness and pluripotency."

2) And another goal as stated is to develop "a new method to predict gene expression levels based on label-free images of cells that employs a classifier trained on image features extracted using an unsupervised deep learning model, VQ-VAE-2"

Small side note: Based on the current title of the manuscript, the expected goal of the mansucript is goal 2 above as there is no

mention of the key components of goal 1 related to improving performance of the iPSC purification process.

Related to goal 1:
This goal assumes that there is a link between gene expression levels and the ability to identify stemness/pluripotency quality of a clone that in the end has improved iPSC characteristics. However, this key link is not shown or described in this manuscript (unless I missed it, which if I did I apologize, althought that may mean readers could miss it too).

If the goal of this manuscript is indeed the first goal above, then there are obvious taks to be performed and assessed:
- what is the connection between the expert-annotations of the hiPSC images and the final desired outcome of a high quality, well reprogrammed hiPSC clone? If there is a clear connection then these annotations could be used in a classification type of task just based on these annotations to help acchive the main goal. If there is no connection between these annotations and the desired outcome, then first the authors need to establish what the link between the morphology seen in the clone images and that desired final outcome is? This could be explored via the VQ-VAE2 feature extraction approach or other methods. But it is key to create a link between what the authors are working on and that final outcome of identifying the right high quality reprogrammed clone.

- Let's assume that there is a connection between the annotated categories and the final outcome - here I will assume (but this may be incorrect since I could not figure this out from the current manuscript) that the "undifferentiated" image class is the one that leads to the best hiPSC clone "quality" outcome. In that case, what is the connection between the gene expression dataset and the expert-annotated images (no predictions needed)? Can the authors identify which gene combinations may be most useful for predicting these classes of images and is that sufficient to achieve their goal 1? How well does such a classification task achieve goal 1? And if it cannot do so well enough, how much of a performance improvement occurs when then instead taking the approach of using the VQ-VAE model to extract features and then the SVR models to predict gene expression?

Related to goal 2:
However, if the goal of the manuscript is the second one then there are entirely different tasks to perform
- the benchmark comparison section of the paper then becomes a key part of the paper and a thorough analysis/comparison based on taht dataset (and others like it) would be more of a goal. The dataset within this manuscript would also be of value but a deeper thorough analysis would be required to understand the link between gene expression and morphology, not simply with the purpose of identifying "better performing clones". This would likely not be a revision of the current manuscript but instead an entirely different manuscript but still based on this dataset.

Some figure specific comments
In addition to the above general comments about the manuscript, I also include some more specific comments below in case they are helpful, although some of these might not be applicable in a fully rewritten manuscript.

- Several figures need some readjusting/rearrangement. If there are separate figure legends then these should be separately numbered figures. If they are instead a multi-panel figure then that is how they should be presented. For example FIgure 4 A and B are currently split - so are these 2 figures? Same for Fig6

- It seems like Fig1 belongs near Figure 5 not at the very start of the story with then no mention of the gene expression analysis till Fig5

- Fig 3 - if the VQ-VAE-2 model was trained on A3 why is this the one with the largest spread in MSE/not the best one? And why do the two 1st timepoint samples show so much lower MSE/less variability overall? Is there a significance to thiss result

- Fig4A: Perhaps call these categories "undifferentiated" and "differentiated" instead of "undifferentiation" and "differentiation" to be consistent grammatically with the use of "cracked" and "built up". Also how do you define "differentiated" morphology-wise? The other categories have a description to explain what makes them that category but differentiated doesn't.
- as part of Fig4, the statement was made that "The image features were split 3:1 and used for training and test respectively". I do not fully understand that statement - how were image features split 3:1? Do you mean images were split into train test? Or features? And which images? The 1024 patches from one single A3 image? Or something else? And what about performance across all the images? More clarity would be appreciated.

- Fig5: There are clear gene expression differences between many clones from one donor (A's) and one clone from many donors (B). It would be useful to see whether there is donor to donor variation within B (e.g., coloring each of the 15 patients) vs clone to clone variation within one donor (A) to understand more of what the source of the variabilty is. The manuscript seems to be conclusing that the variation is due to reprogramming batch dependent variation but it is not clear to me why it could not be donor to donor variation as well?

Table2 is especially hard to follow.

- Fig6: I did not understand the statement "Nevertheless, training with 14 donors (as compared to single-donor Experiment A) did show an advantage: it yielded stronger ability to predict genes within the experimental batch, especially for dataset B1" as it

seemed that the R2 values were worse. More clarification would be appreciated.

There was a lot of emphasis on the link between image classes and the predicted gene expression. Was there a reason for the focus on only the predictions as opposed to an analysis of the link between image classes and the actual gene expression? This seems to be a key missing analysis for the paper. Without this how can the authors assess whether the model predictions are actually useful?

Related to the MBoC Author Submission Checklist: I marked no because there are at least two instances where the details of the methodology were not clear to me. Given that I cannot recommend this manuscript for publication in this journal at this time, and that I commented on at least one poitn of methodology confusion in my comments above, I have not gonet through and checked each of the components in the checklist closely. If a new version of the manuscript is submitted and I end up being a reviewer for that manuscript, I will do so at that time.

Reviewer #3 (Remarks to the Author):

In this manuscript, Wakui and colleagues demonstrate a new approach for using images of human induced pluripotent stem cells (hiPSCs) to determine their quality and predict their gene expression. They describe their workflow, highlighting the advantage of using an autoencoder for feature extraction and the use of bulk RNA sequencing to minimize costs. They validate their method by showing they can predict gene expression in publicly available breast cancer spatial transcriptomics data with similar performance as the end-to-end convolutional neural network model DenseNet-121. By using the autoencoder-extracted features and a Support Vector Machine they are able to accurately predict the quality of the iPSCs. They are also able to predict the gene expression of iPSCs using the auto encoder combined with a Support Vector Regression model trained per gene of interest in each batch, but they see significant differences in performance when they build a model to one batch and apply it to the other.

Overall, this is an interesting manuscript. This method shows the power of using simple phase contrast images to extract biological information. The ability to train a generalizable model to determine the quality of iPSC cells should make it much more efficient to identify which iPSC cells to keep. It is also notable that by using phase contrast images and bulk RNA seq this method minimizes costs and therefore should be very accessible for anyone trying to predict gene expression from images. Finally, although the batch effects seen in the gene expression makes it hard to make a generalizable model, it potentially shows an interesting biological phenomenon that cells can accomplish the same phenotype with different gene expression states.

My main concern with this paper is the intertwining of the concepts of predicting the iPSC quality and predicting gene expression. The paper motivates this work by saying gene expression is the best predictor of the quality of iPSCs therefore it would be ideal to extract gene expression from images. This paper is able to independently predict iPSC quality and gene expression from images, however it does not show that by predicting gene expression from images they can effectively identify good iPSCs. In fact, it appears that adding the prediction of gene expression would make it more difficult to classify iPSCs as the gene expression data has batch effects while the image data does not. It would be helpful to understand better whether the RNA-seq itself has batch effects or whether the batch effects that the authors discuss are coming from the batches of iPSC.

Therefore our suggestion is to clearly separate the demonstrated ability to predict the quality of iPSCs and the ability to predict the gene expression of iPSCs based on images. If the authors want to demonstrate their ability to predict iPSC quality by passing through gene expression then I suggest they train a model to use the extracted gene expression data to define the quality of iPSC cells. The effectiveness of this model should be compared to the one presented in this paper that goes directly from extracted image features to classifying iPSC quality.

It is also generally confusing that the authors claim that inspection of cell morphology is a less accurate predictor of iPSC quality, but then use that metric as the standard throughout the paper. This could be addressed also through separating these concepts.

A separate concern is that the authors claim their model required less labeled data for the same performance as state of the art methods (in section 2.7 and in discussion). The paper does not make it clear what evidence they have for this claim. I suggest explaining specifically what labels are used in the state of the art and which are not needed for this model. If the same labeled data is needed, but just less of it, I suggest comparing the performance of the proposed model compared to the state of the art trained on different amounts of labeled data, while still specifying what the labels are.

Some minor suggestions:
In the discussion the authors claim that their method can be used to train a model early in the iPSC differentiation process. If the purpose of this model is to be able to identify good quality iPSC cells early to avoid using resources in culturing bad iPSC cells, then I would suggest models be trained on the earliest time point in the data, not the latest (which I believe is the data used here). This would allow the authors to substantiate the claim in the discussion that these models can be trained on early time points and applied later in the experiment. The authors should modify the wording in the discussion to be clear about what they

have accomplished in this paper and the proposed use case.

When training the model for gene expression prediction in iPSCs, the authors select the genes to use based on their expression level and the variation, however on the breast cancer histopathology images the mean top expressing genes were used. It could be beneficial to explain why the genes used in the model were selected differently for the different data sets.

While performing the analysis on variable genes could identify genes that show differences in iPSC quality, they could also be identifying many genes that are just variable between batches for other reasons (as they potentially do here). Is it possible to remove genes that show batch effects? This could be done by performing a differential expression between batches and then removing those genes from the analysis.

Are batch effects seen in the images too when comparing Experiment A vs. B? It would be useful to know the source of batch effects here.

Figure 5 is unclear as shown. It would be helpful if the authors could add label titles (gene expression or image features) to each of the hierarchical graphs, x- and y-axis labels, in addition to a more complete caption that would aid in comprehension. The x-axis text (assumed to be sample labels) is indecipherable as shown. It would be more useful to put color blocks or explicit clone/sample labels. Also, adding time points to the clone labels may be interesting.

Response to review

Wakui, et al.

We appreciate the careful thought and time that went into the review of our paper and hope that the reviewers and editor find our responses and changes to the paper to be satisfactory!

Because it can be a bit tricky to follow the logic and time has passed since review, we provide this quick guide to remind the editor and reviewers about our paper.

It is already known that:
- Expert annotation of images can predict differentiation (daily used in practice)
- mRNA levels can predict differentiation (uncommonly used, for developing novel protocols)

So our paper addresses:
- can raw images predict mRNA, in the context of iPSCs?

And we find that:
- Yes it can! though a model must be retrained on each batch or donor
- We also note that raw images can predict expert annotation of images

Our unsupervised deep learning model has two steps:
1) extracts morphology features from each image, using a modified VQ-VAE-2 model as a self-supervised feature extraction method
2) Support Vector Regression (SVR) model to predict gene expression based on the image features.

Reviewer 1

1 (Minor, relating to section 2.2 in the manuscript)

What is their justification for using patches? Why not just learn on the entire image?

We used patches mainly because of the memory constraints of the VQ-VAE2 model. The size of a single image is 5120px in each dimension, and there are over 600 images for a single well. This is too huge to apply VQ-VAE2 directly. On the other hand, the size of each patch is 160px each dimension that can contain 100 cells. The variation of the cell quality inside a patch is negligibly small and can be represented by latent vector of the VQ-VAE2 model. We have added this rationale to section 2.2.

2 (Minor, relating to section 2.2 in the manuscript)

What was their justification for model choice vs. other unsupervised models/AEs?

The advantage of VQ-VAE is vector quantization. As indicated in the paper, the present dataset is highly unbalanced in the number of gene profiles and images, and image features need to be aggregated in order to connect the image data with gene expression data. This aggregation requires statistical procedures such as averaging, but with models other than VQ-VAE, it is difficult to preserve feature distribution information within a sample because the features themselves are averaged. VQ-VAE can easily show which and how many

features are present in a sample, so VQ-VAE was used this time. We have added this rationale to section 2.2 (“Vector quantization inherent in VQ-VAE-2 model…”).

3 (Minor)

Needs to be proofread a bit (missing parentheses, spelling errors etc).

We have carefully checked the manuscript.

4 (Minor, relating to the Introduction in the manuscript)

"Very recently, several studies have demonstrated that morphological features can be used to predict gene expression in various contexts (17-21. For example, supervised deep learning models, such as a Convolutional Neural Network (CNN), were trained with datasets of pathological images and spatial gene expression profiles obtained by single-cell sequencing of the exact same sample to predict cancer gene expression from pathological images. The resulting models were able to accurately link morphology with gene expression, suggesting that a similar strategy could be extended beyond cancer and applied to the iPSC purification process. However, these strategies require expensive single-cell sequencing measurements, which limits their wide-spread adoption."

Which study are they citing in each sentence specifically?

We have updated all such references and appreciate the comment.

5 (Minor, relating to Fig2 in the manuscript)

Figure 2 - the figure legend should indicate that Ni is number of images, and np is number of patches?

Thank you for noticing this, we have added the information.

6 (Minor, relating to Table 3 in the manuscript)

It would be instructive to know what happens if you trained on both dataset A and B together as well to see if the resulting model generalizes better.

We had previously performed this analysis and found it did not improve performance, likely because the characteristics of the datasets differ too much. We include this result now as Table S4.

Reviewer 2

To try to explain what I mean with "tension between possible goals of the manuscript" let me start by quoting two parts of the discussion (to avoid my own paraphrasing):

1) One goal was stated to be to develop a "strategy aimed at improving performance of the iPSC purification process by using cell morphology captured by phase-contrast microscopy to predict gene expression levels relating to stemness and pluripotency."

2) And another goal as stated is to develop "a new method to predict gene expression levels based on label-free images of cells that employs a classifier trained on image features extracted using an unsupervised deep learning model, VQ-VAE-2"

Small side note: Based on the current title of the manuscript, the expected goal of the manuscript is goal 2 above as there is no mention of the key components of goal 1 related to improving performance of the iPSC purification process.

We appreciate the reviewer noting this lack of clarity. While Goal 1 is indeed the ultimate goal of iPSC development, it is indeed not the immediate goal of this paper; instead we are only focused on Goal 2. Currently, the development of good iPSC cells is not an activity with a clear performance goal, but rather a continuous activity of culture protocol improvement aiming to produce higher quality cells. In this activity, gene expression information, which directly indicates cell differentiation status is used as a helpful indicator, and biologists improve protocols based on gene expression information. However, gene expression measurement is not suited as a means of cell monitoring due to its destructive nature and cost. Therefore, as an alternative, visual evaluation has been used until now as a simple quality control method; its adoption by practitioners attests to its success to date. However, current morphology-based evaluation has also problems such as unclear biological relationships and coarse indicators that are rounded to a few labels. In summary, we recognize that both gene expression measurement and the visual evaluation method are insufficient for further improvement activities in the future due to their restrictions. We believe that there should be a new evaluation method that overcomes those restrictions having the features of both methods - biologically meaningful and non-destructive, to understand cell culture more precisely. With this motivation, the goal of this research is to show that iPSC gene expression levels can be predicted based on cell morphology. We believe that we have made a good first step, as a basic research study, by showing that there is a possibility of prediction for some genes. We have adjusted the introduction to make it clear that our presented work aims at Goal 2 and is only a first step towards Goal 1.

2 (Major, relating to the Introduction and Discussion in the manuscript)

The Goal 1 assumes that there is a link between gene expression levels and the ability to identify stemness/pluripotency quality of a clone that in the end has improved iPSC characteristics. However, this key link is not shown or described in this manuscript.

We have revised the introduction section to clarify the past literature that demonstrates such a link, where gene expression levels for pluripotency and undifferentiation are commonly used as quality markers for good iPSCs:

1) Brown, Matthew E., et al. "Derivation of induced pluripotent stem cells from human peripheral blood T lymphocytes." PloS one 5.6 (2010): e11373.
2) Mack, A. A., Kroboth, S., Rajesh, D., & Wang, W. B. (2011). Generation of induced pluripotent stem cells from CD34+ cells across blood drawn from multiple donors with non-integrating episomal vectors. PloS one, 6(11), e27956.
3) Wakao, Shohei, et al. "Morphologic and gene expression criteria for identifying human induced pluripotent stem cells." PloS one 7.12 (2012): e48677.

We have also elaborated in the Discussion section that we do not *prove* a link between particular genes' expression and iPSC purification but what we have done is a foundation for future work that could solidify the relationships previously identified in the literature.

3 (Major, relating to the Introduction and Discussion in the manuscript)
If the goal of this manuscript is indeed the first goal above, then there are obvious tasks to be performed and assessed:

- what is the connection between the expert-annotations of the hiPSC images and the final desired outcome of a high quality, well reprogrammed hiPSC clone? If there is a clear connection then these annotations could be used in a classification type of task just based on these annotations to help achieve the main goal. If there is no connection between these annotations and the desired outcome, then first the authors need to establish what the link between the morphology seen in the clone images and that desired final outcome is? This could be explored via the VQ-VAE2 feature extraction approach or other methods. But it is key to create a link between what the authors are working on and that final outcome of identifying the right high quality reprogrammed clone.

3-1 **What is the connection between the expert-annotations of the hiPSC images and the final desired outcome of a high quality, well reprogrammed hiPSC clone?**

We have revised the introduction section to clarify the past literature that demonstrates this ("there is also a qualitative relationship between cell quality and morphology…"). The use of expert iPSC morphological evaluation in the iPSC development process has been documented in the following literature:

1) González, F., Boué, S. & Belmonte, J. Methods for making induced pluripotent stem cells: reprogramming à la carte. Nat Rev Genet 12, 231–242 (2011). https://doi.org/10.1038/nrg2937
2) Healy, Lyn, and Ludmila Ruban. Atlas of human pluripotent stem cells in culture. Springer, 2014.
3) Wakui, Takashi, et al. "Method for evaluation of human induced pluripotent stem cell quality using image analysis based on the biological morphology of cells." Journal of Medical Imaging 4.4 (2017): 044003.
4) Wakao, Shohei, et al. "Morphologic and gene expression criteria for identifying human induced pluripotent stem cells." PloS one 7.12 (2012): e48677.
5) Kato, Ryuji, et al. "Parametric analysis of colony morphology of non-labelled live human pluripotent stem cells for cell quality control." Scientific reports 6.1 (2016): 1-12.
6) https://www.stemcell.com/assessing-morphology-of-hpscs.html
7) https://www.youtube.com/watch?v=HriqimbQm-c

The general relationship between the expert-annotations and iPSC quality is well established. However, the classification of visual evaluation is rough and its relevance to gene expression is unclear, so we believe that its performance and functionality is insufficient for further cell improvement, motivating our study.

4 (Major, relating to the Introduction and Discussion in the manuscript)

Let's assume that there is a connection between the annotated categories and the final outcome - here I will assume (but this may be incorrect since I could not figure this out from the current manuscript) that the "undifferentiated" image class is the one that leads to the best hiPSC clone "quality" outcome. In that case, what is the connection between the gene expression dataset and the expert-annotated images (no predictions needed)? Can the authors identify which gene combinations may be most useful for predicting these classes of images and is that sufficient to achieve their goal 1? How well does such a classification task achieve goal 1? And if it cannot do so well enough, how much of a performance improvement occurs when then instead taking the approach of using the VQ-VAE model to extract features and then the SVR models to predict gene expression?

4-1 What is the connection between the gene expression dataset and the expert-annotated images (no predictions needed)?

We apologize because we may not understand the meaning of this question but we will attempt to clarify. We describe some of the connections in the paper below. This paper partially shows the relationship between

expert annotated images and gene expression with the comparison of the annotations and the stained images which represent the gene expression. By measuring gene expression for iPSC colonies partially sampled from the cultures, it has been shown that several well-known genes relating to pluripotency correlate qualitatively to the expert annotation toward the sampled colonies. However, unfortunately it is not possible to evaluate the connection between the annotations and the gene expression for our datasets, because our datasets consist of bulk level gene expression, with too few separate samples to do this effectively; in other words our dataset lacks the power to determine which genes' expression best correspond to images annotated as "good".

1) Wakui, Takashi, et al. "Method for evaluation of human induced pluripotent stem cell quality using image analysis based on the biological morphology of cells." Journal of Medical Imaging 4.4 (2017): 044003.

4-2 Can the authors identify which gene combinations may be most useful for predicting these classes of images and is that sufficient to achieve their goal 1?

No, such gene combinations have not been found yet quantitatively. As mentioned above, we think that there may be a qualitative relationship, but not enough correspondence to predict the classes. We also believe that the conventional visual evaluation method has worked well so far, however that is insufficient for more precise protocol improvements to be made in the future.

4-3 How well does such a classification task achieve goal 1? How much of a performance improvement occurs when then instead taking the approach of using the VQ-VAE model to extract features and then the SVR models to predict gene expression?

We think that such a classification task is basically working well until today, as we can see in many iPSC studies. However, because biologists are seeking better protocols mainly based on gene expression measurements, we don't think the visual evaluation method is sufficient to improve the iPSC culture protocol further. We believe that a quantitative evaluation method linking to biological meaning would be necessary. For this purpose, our method will allow for more direct modification of culture protocols compared to conventional visual evaluation-based methods, since our method makes morphology evaluation more strictly correlated with gene expression levels. In addition, because our method has less restrictions on measurement cost compared to RNA-seq and other genetic measurement methods, this may allow us the visualization of expression status with finer time steps and spatial heterogeneity of gene expression levels. We added the sentence into the introduction to mention this background.

5 (Major, relating to the Introduction in the manuscript)

Related to goal 2:
However, if the goal of the manuscript is the second one then there are entirely different tasks to perform - the benchmark comparison section of the paper then becomes a key part of the paper … deeper thorough analysis would be required to understand the link between gene expression and morphology, not simply with the purpose of identifying "better performing clones". This would likely not be a revision of the current manuscript but instead an entirely different manuscript but still based on this dataset.

As mentioned above, the purpose of this study is to find a new means of cell evaluation that can predict gene expression levels from iPSC morphology. The proposed use case, which we have now elaborated in the Discussion, is to train the model on earlier stages where both images and gene expression are captured, then use the model to predict gene expression at later stages, eliminating the need to repeatedly capture gene expression data in a destructive manner, providing a means for continuous improvement of iPSC cell culture

protocols. In this respect, we believe that our research has made a good first step as a fundamental study. We revised the Introduction section accordingly. Also, we elaborate in the Discussion section on what would be needed to achieve Goal 1.

6 (Minor, relating to Fig4A, B and Fig6 in the manuscript)

Several figures need some readjusting/rearrangement. If there are separate figure legends then these should be separately numbered figures. If they are instead a multi-panel figure then that is how they should be presented. FIgure 4 A and B are currently split - so are these 2 figures? Same for Fig6

We have revised all the figures accordingly.

7 (Minor, relating to FIg3 in the manuscript)

If the VQ-VAE-2 model was trained on A3 why is this the one with the largest spread in MSE/not the best one? And why do the two 1st timepoint samples show so much lower MSE/less variability overall?

As written in section 4.1, the earlier timepoints use smaller culture plates (6-well plates rather than flasks). This issue might have occurred because of this plate type difference, because larger plates usually cause bigger morphological variation that cannot be fully restored by the AE model trained with a small portion of the dataset.

8 (Minor, relating to Fig4 in the manuscript)

How do you define "differentiated" morphology-wise? The other categories have a description to explain what makes them that category but differentiated doesn't.

Good catch. "Differentiated" is defined as dark and flat-looking, differentiated- cells-like morphology with a lot of cytoplasm, as opposed to compact iPSC cells with little cytoplasm, which are considered high quality. We have added this description to Fig.4A, and a reference to our prior work:

1) Wakui, Takashi, et al. "Method for evaluation of human induced pluripotent stem cell quality using image analysis based on the biological morphology of cells." Journal of Medical Imaging 4.4 (2017): 044003.

9 (Minor, relating to Fig4 in the manuscript)

The statement was made that "The image features were split 3:1 and used for training and test respectively". I do not fully understand that statement - how were image features split 3:1?

The image features are divided image-wise. Image feature vectors obtained from a single image are treated as a single data set as is.

10 (Minor, relating to Fig5 in the manuscript)

It would be useful to see whether there is donor to donor variation within B (e.g., coloring each of the 15 patients) vs clone to clone variation within one donor (A) to understand more of what the source of the variability is.
The manuscript seems to be concluding that the variation is due to reprogramming batch dependent variation but it is not clear to me why it could not be donor to donor variation as well?

Great idea! We updated Figure 5 to more clearly show which donor each sample derived from. We suspect the variation is caused mainly by the batch effect, because the gene expression profile of the clone #16 in Experiment B, which is derived from the same donor as the clone #1-15 in Experiment A, belongs to the different cluster than the cluster of #1-15 in Figure 5 (top). However, we cannot conclude only the batch-effect is the cause, eliminating the possibility of donor-difference, because we cannot compare for the other clones. We also mentioned this fact in Section 4.1 and Discussion.

11 (Minor)

Table2 is especially hard to follow.

The purpose of Table 2 is to show that our model can predict gene expression levels with a similar performance to ST-Net. As we got the result that three of the top five predicted genes are common with ST-Net, we concluded that our method is working properly to predict genes. We corrected the description of Table 2 to make this point clear.

12 (Minor, relating to Fig6 in the manuscript)

I did not understand the statement "Nevertheless, training with 14 donors (as compared to single-donor Experiment A) did show an advantage

We apologize for confusing you. We removed the sentence from the manuscript because it wasn't necessary.

13 (Minor, relating to the Introduction and Discussion in the manuscript)

Was there a reason for the focus on only the predictions as opposed to an analysis of the link between image classes and the actual gene expression?

As shown in the papers below, the connection between some undifferentiated gene-expression markers and image class is known qualitatively. In addition, as mentioned above, because this manuscript aims to quantitatively predict gene expression levels, we aimed to predict gene expression level directly from images, rather than going through the intermediate of a few coarse categories judged by human interpretation, which will surely lose some of the power of the original morphological information.

1) Wakui, Takashi, et al. "Method for evaluation of human induced pluripotent stem cell quality using image analysis based on the biological morphology of cells." Journal of Medical Imaging 4.4 (2017): 044003.
2) Wakao, Shohei, et al. "Morphologic and gene expression criteria for identifying human induced pluripotent stem cells." PloS one 7.12 (2012): e48677.
3) Kato, Ryuji, et al. "Parametric analysis of colony morphology of non-labelled live human pluripotent stem cells for cell quality control." Scientific reports 6.1 (2016): 1-12.

Reviewer 3

1 (Major, relating to the Introduction and Discussion in the manuscript)

My main concern with this paper is the intertwining of the concepts of predicting the iPSC quality and predicting gene expression. The paper motivates this work by saying gene expression is the best predictor of the quality of iPSCs therefore it would be ideal to extract gene expression from images. This paper is able to independently predict iPSC quality and gene expression from images, however it does not show that by predicting gene expression from images they can effectively identify good iPSCs. In fact, it appears that adding the prediction of gene expression would make it more difficult to classify iPSCs as the gene expression data has batch effects while the image data does not. It would be helpful to understand better whether the RNA-seq itself has batch effects or whether the batch effects that the authors discuss are coming from the batches of iPSC.

1-1 This paper is able to independently predict iPSC quality and gene expression from images, however it does not show that by predicting gene expression from images they can effectively identify good iPSCs.

We hope that our guide introducing our response above clarifies the overall goals, and we have edited the paper to make this more clear throughout. As answered for Reviewer 2 above, the development of good iPSC cells is not an activity with a clear performance goal, but rather a continuous activity of culture protocol improvement aiming to produce higher quality cells. In this activity, gene expression information, which directly indicates cell inner activity, is used as a helpful indicator, and biologists improve protocols based on gene expression information. However, gene expression measurement is inappropriate as a means of cell monitoring due to its destructive nature and cost. Therefore, as an alternative, visual evaluation has been used until now as a simple quality control method. So far, we think visual evaluation has worked well, as evidenced by the success of iPSCs to date. However, we recognize that current morphology-based evaluation is insufficient for further improvement activities in the future due to problems such as unclear biological relationships and coarse indicators that are rounded to a few labels. With this motivation, the goal of this research is to show that iPSC gene expression levels can be predicted based on cell morphology. We believe that we have made a good first step as a basic research by showing that there is a possibility of prediction for some genes.

1-2 In fact, it appears that adding the prediction of gene expression would make it more difficult to classify iPSCs as the gene expression data has batch effects while the image data does not.

We thank the reviewer for pointing out that this was not clear in the paper. In actual operation, a large number of iPSC clones are established in a single reprogramming process and accumulated in a working cell bank for subsequent cell production. In addition, iPSC culturing during the purification process and the early differentiation process is a long-term process. Thus, we believe that it is still valuable to train models for each reprogramming batch (making batch effects not a concern in practice). Our method is a quantitative and biologically meaningful improvement upon existing qualitative visual evaluation, so it will be useful for improving culture protocols by providing more detailed information of cell behavior.

1-3 It would be helpful to understand better whether the RNA-seq itself has batch effects or whether the batch effects that the authors discuss are coming from the batches of iPSC.

We evaluated the reproducibility of RNA-seq we used, and confirmed that the method can measure gene expression with R^2=0.9854 between the different measurements (see Figure S5).

2 (Major, relating to Table 1 in the manuscript)

Therefore our suggestion is to clearly separate the demonstrated ability to predict the quality of iPSCs and the ability to predict the gene expression of iPSCs based on images. If the authors want to demonstrate their ability to predict iPSC quality by passing through gene expression then I suggest they train a model to use the extracted gene expression data to define the quality of iPSC cells. The effectiveness of this model should be compared to the one presented in this paper that goes directly from extracted image features to classifying iPSC quality.

We apologize for the confusion. Unfortunately, the proposal would be a bit circular, because iPSC quality is currently defined by either mRNA markers or by image-based qualitative evaluation (the latter being the case in this study). As written above in this letter, the goal of this research is to show that iPSC gene expression levels can be predicted based on cell morphology. The relationship between iPSC morphology and gene expression levels has not been well-studied even though the cell quality has been monitored for long by visual inspection. Also, the quality categories used in this study are those used in the actual cell manufacturing process, however we think that this is not enough for further improvements of cell culture protocols in the future, and need another methodology that enables us to see the activity inside cells non-destructively. For these reasons, we believe that our approach is a step in the right direction.

3 (Major, relating to the Introduction, Discussion, and section 2.4 in the manuscript)

It is also generally confusing that the authors claim that inspection of cell morphology is a less accurate predictor of iPSC quality, but then use that metric as the standard throughout the paper. This could be addressed also through separating these concepts.

It is true that visual evaluation is insufficient because of coarseness of the defined quality classes and its unclear relationship toward biological meaning. However, because visual evaluation has been the standard method generally used in the iPSC culture process, we employed it as an iPSC quality indicator. We believe that this study makes morphology-based evaluation methods more well-validated and useful to improve iPSC protocols by evaluating cells more quantitatively and in a more biologically meaningful way.

6 (Major, relating to section 2.7 in the manuscript)

A separate concern is that the authors claim their model required less labeled data for the same performance as state of the art methods (in section 2.7 and in discussion). The paper does not make it clear what evidence they have for this claim. I suggest explaining specifically what labels are used in the state of the art and which are not needed for this model. If the same labeled data is needed, but just less of it, I suggest comparing the performance of the proposed model compared to the state of the art trained on different amounts of labeled data, while still specifying what the labels are.

We apologize for the error - the amount of data employed in this study is the same as that of ST-Net; we have corrected this in the manuscript.

7 (Minor, relating to the Discussion in the manuscript)

In the discussion the authors claim that their method can be used to train a model early in the iPSC differentiation process. If the purpose of this model is to be able to identify good quality iPSC cells early to avoid using resources in culturing bad iPSC cells, then I would suggest models be trained on the earliest time point in the data, not the latest (which I believe is the data used here). This would allow the authors to substantiate the claim in the discussion that these models can be trained on early time points and applied later in the experiment. The authors should modify the wording in the discussion to be clear about what they have accomplished in this paper and the proposed use case.

Thank you for your thoughtful suggestion, we added the training result using the early time point into the supplementary materials. Interestingly, as you can see in Table S3A, this did not give us better results. This could be caused by the difference of culture vessels between the time points. Timepoint1 uses the smallest plate type in this study that leads to the least morphological variation in a plate. Training with smaller morphological variation causes the incompatibility toward samples cultured in larger plates that have bigger variation. Therefore, we employed the datasets of Timepoint 3 in this study as the target dataset to be used in training. In the actual purification process, as culture continues with T50 for a while after Timepoint 3, we believe our method is still beneficial to monitor the culture. We also addressed the proposed use case in the discussion as well.

9 (Minor, relating to section 2.6 and Supl. Table 1 in the manuscript)

When training the model for gene expression prediction in iPSCs, the authors select the genes to use based on their expression level and the variation, however on the breast cancer histopathology images the mean top expressing genes were used. It could be beneficial to explain why the genes used in the model were selected differently for the different data sets.

To demonstrate the consistency of prediction between the previous study and our method, we showed that the gene sets with high prediction accuracy are similar between the methods, regarding the result of ST-Net as a ground truth. In contrast, for iPSCs, it is not possible to define a ground truth, so we selected target genes according to variation and level. We have now clarified this in the description of Table 2.

10 (Minor, relating to section 2.7 and the Discussion in the manuscript)

While performing the analysis on variable genes could identify genes that show differences in iPSC quality, they could also be identifying many genes that are just variable between batches for other reasons (as they potentially do here). Is it possible to remove genes that show batch effects? This could be done by performing a differential expression between batches and then removing those genes from the analysis.

Thank you for your suggestion. We tried to reanalyze the data removing the genes that show batch effects (Table S4). In this analysis, we retrained the SVR models with the same procedure after removing the genes that have different expression levels between Experiment A and B. However, the batch effects were not resolved unfortunately. This may suggest that cell morphology cannot be determined by a few genes and varies depending on the balance of multiple gene expressions. In other words, even if targeting the specific genes that do not have the batch effects, it is challenging to predict gene expression level consistently on both batches since the batch effects in the other genes change the relationship between the genes and the morphology.

11 (Minor, relating to section 2.5 and Fig5 in the manuscript)

Are batch effects seen in the images too when comparing Experiment A vs. B? It would be useful to know the source of batch effects here.

No - the bottom side of Figure 5 shows that the image features do not have batch effects between Experiment A and B. We have added this point to the figure legend.

12 (Minor, relating to Fig5 in the manuscript)

Figure 5 is unclear as shown. It would be helpful if the authors could add label titles (gene expression or image features) to each of the hierarchical graphs, x- and y-axis labels, in addition to a more complete caption that would aid in comprehension. The x-axis text (assumed to be sample labels) is indecipherable as shown. It would be more useful to put color blocks or explicit clone/sample labels. Also, adding time points to the clone labels may be interesting.

Thank you for the suggestion, we have added the labels and the titles as suggested.

RE: Manuscript #E22-06-0215R
TITLE: "Predicting gene expression from cell morphology in human induced pluripotent stem cells"

Dear Mr. Wakui:

The reviewers are mostly satisfied with the revisions. However, some of the clarifications in the response letter were not efficiently incorporated into the text. Reviewer 2 has made specific suggestions about including these clarifications into the text. This will help the reader with clarity and I suggest that the authors include further clarification in the text following Reviewer 2's suggestion. Reviewer 2 also suggested a more appropriate title that better reflects the scope of the work. Please reconsider the title according to this suggestion. Please incorporate these suggestions into the text and provide a point by point list of the incorporated changes. The paper will not be sent back to the reviewers, however, before we can accept the paper, the title change and further clarification in the text should be incorporated.

Sincerely,
Melike Lakadamyali
Monitoring Editor
Molecular Biology of the Cell

Dear Mr. Wakui,

The review of your manuscript, referenced above, is now complete. The Monitoring Editor has decided that your manuscript requires minor revisions before it can be published in Molecular Biology of the Cell, as described in the Monitoring Editor's decision letter above and the reviewer comments (if any) below.

A reminder: Please do not contact the Monitoring Editor directly regarding your manuscript. If you have any questions regarding the review process or the decision, please contact the MBoC Editorial Office (mboc@ascb.org).

When submitting your revision include a rebuttal letter that details, point-by-point, how the Monitoring Editor's and reviewers' comments have been addressed. (The file type for this letter must be "rebuttal letter"; do not include your response to the Monitoring Editor and reviewers in a "cover letter.") Please bear in mind that your rebuttal letter will be published with your paper if it is accepted, unless you have opted out of publishing the review history.

Authors are allowed 15 days to submit a revision. If this time period is inadequate, please contact us immediately at mboc@ascb.org.

In preparing your revised manuscript, please follow the instruction in the Information for Authors (www.molbiolcell.org/info-for-authors). In particular, to prepare for the possible acceptance of your revised manuscript, submit final, publication-quality figures with your revision as described.

To submit the rebuttal letter, revised version, and figures, please use this link (please enable cookies, or cut and paste URL):
Link Not Available

Authors of Articles and Brief Communications whose manuscripts have returned for minor revision ("revise only") are encouraged to create a short video abstract to accompany their article when it is published. These video abstracts, known as Science Sketches, are up to 2 minutes long and will be published on YouTube and then embedded in the article abstract. Science Sketch Editors on the MBoC Editorial Board will provide guidance as you prepare your video. Information about how to prepare and submit a video abstract is available at www.molbiolcell.org/science-sketches. Please contact mboc@ascb.org if you are interested in creating a Science Sketch.

Thank you for submitting your manuscript to Molecular Biology of the Cell. Please do not hesitate to contact this office if you have any questions.

Sincerely,

Eric Baker
Journal Production Manager
MBoC Editorial Office
mbc@ascb.org

Reviewer #1 (Remarks to the Author):

The authors did a great job addressing reviewer concerns and the paper is now ready for publication!

Reviewer #2 (Remarks to the Author):

Overall, the authors have clarified my main points of confusion, and refocused the framing of their paper around their main goal of developing a method to predict gene expression from bright-field image data for the eventual purpose of tracking the expression of a handful of key genes via bright-field images to improve hiPSC reprogramming protocols in the future. I have included point by point comments and suggestions for the authors to hopefully improve the ease of understanding their manuscript further for the reader but all of these are only suggestions and not required.
As stated by the authors, they "have made a good first step, as a basic research study, by showing that there is a possibility of prediction for some genes" and they have adjusted their abstract, introduction, and discussion accordingly to emphasize the specific and initial nature of their results as well as their potential going forward. The core result of this study is, therefore, the actual model they developed, its validation on a previously published dataset, and then its initial application and assessment on their goal application. All of this is technically sound and with some minor revisions as recommended below, also understandable by the reader. I must, however, rely on the editor to determine whether all of this together is sufficient for publication in MBOC vs. in a more technically focused method development journal.
I have included my comments in line below with a ">>" at the start of my comments.

Reviewer 2
To try to explain what I mean with "tension between possible goals of the manuscript" let me start by quoting two parts of the discussion (to avoid my own paraphrasing):
1) One goal was stated to be to develop a "strategy aimed at improving performance of the iPSC purification process by using cell morphology captured by phase-contrast microscopy to predict gene expression levels relating to stemness and pluripotency."
2) And another goal as stated is to develop "a new method to predict gene expression levels based on label-free images of cells that employs a classifier trained on image features extracted using an unsupervised deep learning model, VQ-VAE-2"
Small side note: Based on the current title of the manuscript, the expected goal of the manuscript is goal 2 above as there is no mention of the key components of goal 1 related to improving performance of the iPSC purification process.
We appreciate the reviewer noting this lack of clarity. While Goal 1 is indeed the ultimate goal of iPSC development, it is indeed not the immediate goal of this paper; instead we are only focused on Goal 2. Currently, the development of good iPSC cells is not an activity with a clear performance goal, but rather a continuous activity of culture protocol improvement aiming to produce higher quality cells. In this activity, gene expression information, which directly indicates cell differentiation status is used as a helpful indicator, and biologists improve protocols based on gene expression information. However, gene expression measurement is not suited as a means of cell monitoring due to its destructive nature and cost. Therefore, as an alternative, visual evaluation has been used until now as a simple quality control method; its adoption by practitioners attests to its success to date. However, current morphology-based evaluation has also problems such as unclear biological relationships and coarse indicators that are rounded to a few labels. In summary, we recognize that both gene expression measurement and the visual evaluation method are insufficient for further improvement activities in the future due to their restrictions. We believe that there should be a new evaluation method that overcomes those restrictions having the features of both methods - biologically meaningful and non-destructive, to understand cell culture more precisely. With this motivation, the goal of this research is to show that iPSC gene expression levels can be predicted based on cell morphology. We believe that we have made a good first step, as a basic research study, by showing that there is a possibility of prediction for some genes. We have adjusted the introduction to make it clear that our presented work aims at Goal 2 and is only a first step towards Goal 1.
>>Thank you for the clarification. To restate to make sure I have understood - the goal of the paper is goal 2 - to develop a new method to predict gene expression levels based on label-free images of cells..." with the application being related to eventually a strategy to improve iPSC purification. I have adjusted my expectations around this clarified goal in second round of comments below.
>> I agree with your assessment that "We believe that we have made a good first step, as a basic research study, by showing that there is a possibility of prediction for some genes" and your abstract is consistent with this sentiment. The title however suggests something much larger scale - not the predicted expression of a few key genes that could help monitor reprogramming quality in a quantitative manner but a general ability to predict overall gene expression profiles. Some adjustment to incorporate the application and/or that this is an initial feasibility study and/or to emphasize the now quantitative nature of the image-based monitoring would be helpful.

2 (Major, relating to the Introduction and Discussion in the manuscript)
The Goal 1 assumes that there is a link between gene expression levels and the ability to identify stemness/pluripotency quality of a clone that in the end has improved iPSC characteristics. However, this key link is not shown or described in this manuscript.
We have revised the introduction section to clarify the past literature that demonstrates such a link, where gene expression

levels for pluripotency and undifferentiation are commonly used as quality markers for good iPSCs:
1. 1) Brown, Matthew E., et al. "Derivation of induced pluripotent stem cells from human peripheral blood T lymphocytes." PloS one 5.6 (2010): e11373.
2. 2) Mack, A. A., Kroboth, S., Rajesh, D., & Wang, W. B. (2011). Generation of induced pluripotent stem cells from CD34+ cells across blood drawn from multiple donors with non-integrating episomal vectors. PloS one, 6(11), e27956.
3. 3) Wakao, Shohei, et al. "Morphologic and gene expression criteria for identifying human induced pluripotent stem cells." PloS one 7.12 (2012): e48677.
We have also elaborated in the Discussion section that we do not *prove* a link between particular genes' expression and iPSC purification but what we have done is a foundation for future work that could solidify the relationships previously identified in the literature.
>>Thank you for this clarification. I agree that if you have elaborated this way in the discussion then this helps clarify this point in the paper. However, I could not find this point in the discussion. Perhaps I missed it?

3 (Major, relating to the Introduction and Discussion in the manuscript)
If the goal of this manuscript is indeed the first goal above, then there are obvious tasks to be performed and assessed:
- what is the connection between the expert-annotations of the hiPSC images and the final desired outcome of a high quality, well reprogrammed hiPSC clone? If there is a clear connection then these annotations could be used in a classification type of task just based on these annotations to help achieve the main goal. If there is no connection between these annotations and the desired outcome, then first the authors need to establish what the link between the morphology seen in the clone images and that desired final outcome is? This could be explored via the VQ-VAE2 feature extraction approach or other methods. But it is key to create a link between what the authors are working on and that final outcome of identifying the right high quality reprogrammed clone.
3-1 What is the connection between the expert-annotations of the hiPSC images and the final desired outcome of a high quality, well reprogrammed hiPSC clone?
We have revised the introduction section to clarify the past literature that demonstrates this ("there is also a qualitative relationship between cell quality and morphology..."). The use of expert iPSC morphological evaluation in the iPSC development process has been documented in the following literature:
1. 1) González, F., Boué, S. & Belmonte, J. Methods for making induced pluripotent stem cells: reprogramming à la carte. Nat Rev Genet 12, 231-242 (2011). https://doi.org/10.1038/nrg2937
2. 2) Healy, Lyn, and Ludmila Ruban. Atlas of human pluripotent stem cells in culture. Springer, 2014.
3. 3) Wakui, Takashi, et al. "Method for evaluation of human induced pluripotent stem cell quality using image analysis based on the biological morphology of cells." Journal of Medical Imaging 4.4 (2017): 044003.
4. 4) Wakao, Shohei, et al. "Morphologic and gene expression criteria for identifying human induced pluripotent stem cells." PloS one 7.12 (2012): e48677.
5. 5) Kato, Ryuji, et al. "Parametric analysis of colony morphology of non-labelled live human pluripotent stem cells for cell quality control." Scientific reports 6.1 (2016): 1-12.
6. 6) https://www.stemcell.com/assessing-morphology-of-hpscs.html
7. 7) https://www.youtube.com/watch?v=HriqimbQm-c
>>Since the goal of the manuscript is no longer the first one, this comment is no longer applicable and I appreciate that you revised the introduction to clarify.
The general relationship between the expert-annotations and iPSC quality is well established. However, the classification of visual evaluation is rough and its relevance to gene expression is unclear, so we believe that its performance and functionality is insufficient for further cell improvement, motivating our study.
>> Thank you for this clarification. As it currently stands, section 2.4 is difficult to understand - the intro does mention your concerns about the visual evaluation morphological classes yet then you do evaluate these in light of your model features. Later on in the discussion the reasons for this are explained. I think a little additional context at the start of section 2.4 (e.g. what its motivation is within the context of this study) would be helpful to the reader and help avoid derailing them down similar thought processes that both reviewer 3 and I seem to have had, which led both of us to recommend directly predicting/analyzing the morphological classes from the gene expression (which I now understand is not the goal of this study). Or perhaps this section could be in the supplemental with a brief statement in the results somewhere to make it clear it is somewhat peripheral to the main goal of the study?

4 (Major, relating to the Introduction and Discussion in the manuscript)
Let's assume that there is a connection between the annotated categories and the final outcome - here I will assume (but this may be incorrect since I could not figure this out from the current manuscript) that the "undifferentiated" image class is the one that leads to the best hiPSC clone "quality" outcome. In that case, what is the connection between the gene expression dataset and the expert-annotated images (no predictions needed)? Can the authors identify which gene combinations may be most useful for predicting these classes of images and is that sufficient to achieve their goal 1? How well does such a classification task achieve goal 1? And if it cannot do so well enough, how much of a performance improvement occurs when then instead taking the approach of using the VQ-VAE model to extract features and then the SVR models to predict gene expression?
4-1 What is the connection between the gene expression dataset and the expert-annotated images (no predictions needed)?
We apologize because we may not understand the meaning of this question but we will attempt to clarify. We describe some of the connections in the paper below. This paper partially shows the relationship between expert annotated images and gene

expression with the comparison of the annotations and the stained images which represent the gene expression. By measuring gene expression for iPSC colonies partially sampled from the cultures, it has been shown that several well-known genes relating to pluripotency correlate qualitatively to the expert annotation toward the sampled colonies. However, unfortunately it is not possible to evaluate the connection between the annotations and the gene expression for our datasets, because our datasets consist of bulk level gene expression, with too few separate samples to do this effectively; in other words our dataset lacks the power to determine which genes' expression best correspond to images annotated as "good".
>> Thank you for your clarification. You did understand my suggestion, which was to explore the direct connection between the gene expression and the expert annotated classes (also the suggestion of reviewer 3). Clarifying that this is not the goal of the study (as you have done in the introduction) is helpful. Additionally, your explanation above of why this is not feasible with this dataset is also very helpful - perhaps this could be incorporated somehow as a future direction with clear explanation of how the current dataset would need to be adjusted to make it feasible.
1) Wakui, Takashi, et al. "Method for evaluation of human induced pluripotent stem cell quality using image analysis based on the biological morphology of cells." Journal of Medical Imaging 4.4 (2017): 044003.

4-2 Can the authors identify which gene combinations may be most useful for predicting these classes of images and is that sufficient to achieve their goal 1?
No, such gene combinations have not been found yet quantitatively. As mentioned above, we think that there may be a qualitative relationship, but not enough correspondence to predict the classes. We also believe that the conventional visual evaluation method has worked well so far, however that is insufficient for more precise protocol improvements to be made in the future.
>>Understood. Also, this comment no longer relevant since goal 1 is not the goal of the study.

4-3 How well does such a classification task achieve goal 1? How much of a performance improvement occurs when then instead taking the approach of using the VQ-VAE model to extract features and then the SVR models to predict gene expression?
We think that such a classification task is basically working well until today, as we can see in many iPSC studies. However, because biologists are seeking better protocols mainly based on gene expression measurements, we don't think the visual evaluation method is sufficient to improve the iPSC culture protocol further. We believe that a quantitative evaluation method linking to biological meaning would be necessary. For this purpose, our method will allow for more direct modification of culture protocols compared to conventional visual evaluation-based methods, since our method makes morphology evaluation more strictly correlated with gene expression levels. In addition, because our method has less restrictions on measurement cost compared to RNA-seq and other genetic measurement methods, this may allow us the visualization of expression status with finer time steps and spatial heterogeneity of gene expression levels. We added the sentence into the introduction to mention this background.
>>Understood. Also, this comment is no longer relevant since goal 1 is not the goal of the study.

5 (Major, relating to the Introduction in the manuscript)
Related to goal 2:
However, if the goal of the manuscript is the second one then there are entirely different tasks to perform - the benchmark comparison section of the paper then becomes a key part of the paper ... deeper thorough analysis would be required to understand the link between gene expression and morphology, not simply with the purpose of identifying "better performing clones". This would likely not be a revision of the current manuscript but instead an entirely different manuscript but still based on this dataset.
As mentioned above, the purpose of this study is to find a new means of cell evaluation that can predict gene expression levels from iPSC morphology. The proposed use case, which we have now elaborated in the Discussion, is to train the model on earlier stages where both images and gene expression are captured, then use the model to predict gene expression at later stages, eliminating the need to repeatedly capture gene expression data in a destructive manner, providing a means for continuous improvement of iPSC cell culture protocols. In this respect, we believe that our research has made a good first step as a fundamental study. We revised the Introduction section accordingly. Also, we elaborate in the Discussion section on what would be needed to achieve Goal 1.
>>Thank you for clarifying that goal 2 is indeed the goal of this study.

6 (Minor, relating to Fig4A, B and Fig6 in the manuscript)
Several figures need some readjusting/rearrangement. If there are separate figure legends then these should be separately numbered figures. If they are instead a multi-panel figure then that is how they should be presented. FIgure 4 A and B are currently split - so are these 2 figures? Same for Fig6
We have revised all the figures accordingly. 7 (Minor, relating to FIg3 in the manuscript)
If the VQ-VAE-2 model was trained on A3 why is this the one with the largest spread in MSE/not the best one? And why do the two 1st timepoint samples show so much lower MSE/less variability overall?
As written in section 4.1, the earlier timepoints use smaller culture plates (6-well plates rather than flasks). This issue might have occurred because of this plate type difference, because larger plates usually cause bigger morphological variation that cannot be fully restored by the AE model trained with a small portion of the dataset.
>>Thank you for the clarification. This difference in degree of variation due to cell culture vessel size also came up as a reason for why perhaps training on the first timepoint itself was not so useful (item 7 from reviewer 3). However, in the paper it is only

mentioned in section 4.1 but with no mention of the implications of these vessel sizes on the data in the paper. More information related to the differences due to the vessel size is recommended, including clarification that the third timepoint is appropriate for model training because subsequent timepoints in the actual hiPSC purification process would remain in that size vessel.

8 (Minor, relating to Fig4 in the manuscript)
How do you define "differentiated" morphology-wise? The other categories have a description to explain what makes them that category but differentiated doesn't.
Good catch. "Differentiated" is defined as dark and flat-looking, differentiated- cells-like morphology with a lot of cytoplasm, as opposed to compact iPSC cells with little cytoplasm, which are considered high quality. We have added this description to Fig.4A, and a reference to our prior work:
1) Wakui, Takashi, et al. "Method for evaluation of human induced pluripotent stem cell quality using image analysis based on the biological morphology of cells." Journal of Medical Imaging 4.4 (2017): 044003.
>>Thank you.

9 (Minor, relating to Fig4 in the manuscript)
The statement was made that "The image features were split 3:1 and used for training and test respectively". I do not fully understand that statement - how were image features split 3:1?
The image features are divided image-wise. Image feature vectors obtained from a single image are treated as a single data set as is.
>>Thank you for this clarification. Has this been incorporated into the text? Perhaps just adding "the image feature vectors were split 3:1" would suffice?

10 (Minor, relating to Fig5 in the manuscript)
It would be useful to see whether there is donor to donor variation within B (e.g., coloring each of the 15 patients) vs clone to clone variation within one donor (A) to understand more of what the source of the variability is.
The manuscript seems to be concluding that the variation is due to reprogramming batch dependent variation but it is not clear to me why it could not be donor to donor variation as well?
Great idea! We updated Figure 5 to more clearly show which donor each sample derived from. We suspect the variation is caused mainly by the batch effect, because the gene expression profile of the clone #16 in Experiment B, which is derived from the same donor as the clone #1-15 in Experiment A, belongs to the different cluster than the cluster of #1-15 in Figure 5 (top). However, we cannot conclude only the batch-effect is the cause, eliminating the possibility of donor-difference, because we cannot compare for the other clones. We also mentioned this fact in Section 4.1 and Discussion.
>>Thank you, this is very helpful.

11 (Minor)
Table2 is especially hard to follow.
The purpose of Table2 is to show that our model can predict gene expression levels with a similar performance to ST-Net. As we got the result that three of the top five predicted genes are common with ST-Net, we concluded that our method is working properly to predict genes. We corrected the description of Table 2 to make this point clear.
>>Thank you, this is very helpful.

12 (Minor, relating to Fig6 in the manuscript)
I did not understand the statement "Nevertheless, training with 14 donors (as compared to single-donor Experiment A) did show an advantage
We apologize for confusing you. We removed the sentence from the manuscript because it wasn't necessary. 13 (Minor, relating to the Introduction and Discussion in the manuscript)
>>Thank you.

Was there a reason for the focus on only the predictions as opposed to an analysis of the link between image classes and the actual gene expression?
As shown in the papers below, the connection between some undifferentiated gene-expression markers and image class is known qualitatively. In addition, as mentioned above, because this manuscript aims to quantitatively predict gene expression levels, we aimed to predict gene expression level directly from images, rather than going through the intermediate of a few coarse categories judged by human interpretation, which will surely lose some of the power of the original morphological information.
1. 1) Wakui, Takashi, et al. "Method for evaluation of human induced pluripotent stem cell quality using image analysis based on the biological morphology of cells." Journal of Medical Imaging 4.4 (2017): 044003.
2. 2) Wakao, Shohei, et al. "Morphologic and gene expression criteria for identifying human induced pluripotent stem cells." PloS one 7.12 (2012): e48677.
3. 3) Kato, Ryuji, et al. "Parametric analysis of colony morphology of non-labelled live human pluripotent stem cells for cell quality control." Scientific reports 6.1 (2016): 1-12.

Reviewer #3 (Remarks to the Author):

The reviewers have addressed the comments regarding the manuscript.

Response to review (2nd)

Wakui, et al.

We appreciate the careful thought and time that went into the review of our paper. We hope that these changes adequately address your concerns and meet your expectations.

Reviewer 2

1. I agree with your assessment that "We believe that we have made a good first step, as a basic research study, by showing that there is a possibility of prediction for some genes" and your abstract is consistent with this sentiment. The title however suggests something much larger scale - not the predicted expression of a few key genes that could help monitor reprogramming quality in a quantitative manner but a general ability to predict overall gene expression profiles. Some adjustment to incorporate the application and/or that this is an initial feasibility study and/or to emphasize the now quantitative nature of the image-based monitoring would be helpful.

Thank you for your feedback on our manuscript. We have carefully considered your comments and have changed the title to clarify the scope of this study.

Before: Predicting gene expression from cell morphology in human induced pluripotent stem cells
Revised: Predicting reprogramming-related gene expression from cell morphology in human induced pluripotent stem cells

3. Thank you for this clarification. As it currently stands, section 2.4 is difficult to understand - the intro does mention your concerns about the visual evaluation morphological classes yet then you do evaluate these in light of your model features. Later on in the discussion the reasons for this are explained. I think a little additional context at the start of section 2.4 (e.g. what its motivation is within the context of this study) would be helpful to the reader and help avoid derailing them down similar thought processes that both reviewer 3 and I seem to have had, which led both of us to recommend directly predicting/analyzing the morphological classes from the gene expression (which I now understand is not the goal of this study). Or perhaps this section could be in the supplemental with a brief statement in the results somewhere to make it clear it is somewhat peripheral to the main goal of the study?

Thank you for your suggestion. We added the following sentence into section 2.4 to inform the purpose of the verification in this section.

"As mentioned in the introduction section, these categories have no clear link to underlying gene expression, but they likely contain some information that is useful for predicting gene expression levels."

4-1. Thank you for your clarification. You did understand my suggestion, which was to explore the direct connection between the gene expression and the expert annotated classes (also the suggestion of reviewer 3). Clarifying that this is not the goal of the study (as you have done in the introduction) is helpful. Additionally, your explanation above of why this is not feasible with this dataset is also very helpful - perhaps this could be incorporated somehow as a future direction with clear explanation of how the current dataset would need to be adjusted to make it feasible.

We added the following sentence to the Discussion section in order to clarify the limitation of our dataset. We hope this addition will help to understand the limitation. To investigate the relationship between the annotations and gene expression directly, spatial gene expression datasets could be useful because the annotation classes indicate the positional quality of cells.

"It is difficult to examine the relationship effectively with a limited number of samples between gene expression and the quality categories, since our bulk-level gene expression dataset lacks spatial information. However, since there exists a relationship between image features and categories, as well as between gene expression and image features, our results suggest that the method we developed could also assist in linking image features with both gene expression and the quality categories. This would further facilitate visual inspection and iPSC purification, and make the process more reliable."

7. Thank you for the clarification. This difference in degree of variation due to cell culture vessel size also came up as a reason for why perhaps training on the first timepoint itself was not so useful (item 7 from reviewer 3). However, in the paper it is only mentioned in section 4.1 but with no mention of the implications of these vessel sizes on the data in the paper. More information related to the differences due to the vessel size is recommended, including clarification that the third timepoint is appropriate for model training because subsequent timepoints in the actual hiPSC purification process would remain in that size vessel.

Thank you for your consideration. We added the following sentences to answer these questions.

4.1 "T150 flasks will continue to be used for a while thereafter in the actual purification process."

4.3 "Furthermore, since the cells in Timepoint 3 are cultured in larger vessels compared to Timepoints 1 and 2, there tends to be greater variation in cell morphology, making dataset A3 considered more suitable for learning a wider range of cell morphologies."

9. Thank you for this clarification. Has this been incorporated into the text? Perhaps just adding "the image feature vectors were split 3:1" would suffice?

Thank you for your thoughtful suggestions! We changed the relevant sentence in section 2.4 as you suggested.

" The image feature vectors were split 3:1 for training and test respectively. "

RE: Manuscript #E22-06-0215RR
TITLE: "Predicting reprogramming-related gene expression from cell morphology in human induced pluripotent stem cells"

Dear Mr. Wakui:

I am pleased to accept your manuscript for publication in Molecular Biology of the Cell.

Sincerely,
Melike Lakadamyali
Monitoring Editor
Molecular Biology of the Cell

Dear Mr. Wakui:

Congratulations on the acceptance of your manuscript.

A PDF of your manuscript will be published on MBoC in Press, an early release version of the journal, within 10 days. The date your manuscript appears at www.molbiolcell.org/toc/mboc/0/0 is the official publication date. Your manuscript will also be scheduled for publication in the next available issue of MBoC.

Within approximately four weeks you will receive a PDF page proof of your article.

Would you like to see an image related to your accepted manuscript on the cover of MBoC? Please contact the MBoC Editorial Office at mboc@ascb.org to learn how to submit an image.

Authors of Articles and Brief Communications are encouraged to create a short video abstract to accompany their article when it is published. These video abstracts, known as Science Sketches, are up to 2 minutes long and will be published on YouTube and then embedded in the article abstract. Science Sketch Editors on the MBoC Editorial Board will provide guidance as you prepare your video. Information about how to prepare and submit a video abstract is available at www.molbiolcell.org/science-sketches. Please contact mboc@ascb.org if you are interested in creating a Science Sketch.

We are pleased that you chose to publish your work in MBoC.

Sincerely,

Eric Baker
Journal Production Manager
MBoC Editorial Office
mbc@ascb.org
