Supplementary Materials
Molecular Biology of the Cell
Wakui *et al*.

Supplementary Materials

Culture process and dataset

Figure S1. Overview of culture process. All cells were cultured for four days after passage. Phase-contrast imaging and bulk RNA-seq analysis were performed before passaging on Day 4 for each of the three stages.

Table S1. Dataset structure. Datasets of experiment A (A1-A3) were obtained from the samples of 15 clones derived from one donor, while those of experiment B (B1-B3) were obtained from the samples of 14 clones derived from 14 donors.

External validation process

As mentioned in section 3.3, we tried to validate our method on an external spatial transcriptomics dataset. We concluded that our method, using the VQ-VAE-based image feature extractor, performed at least equivalent to the end-to-end CNN model in gene expression prediction from image. In this supplemental section, we described the details of the validation process.

We first implemented an end-to-end model to predict expressions of 250 genes from small histopathologic image patches by following the description in the paper (He et al., 2020). DenseNet-121 (Huang et al., 2017) model (pretrained by ImageNet dataset) followed by a dense layer with 250 units was trained as a retest model using 224x224 image patches and corresponding gene expressions of the spots. We modified our model to solve the same task. We trained a VQ-VAE-2 model in advance and its encoder was used as an image feature extractor. Instead of SVR, a MLP model was trained as a gene expression predictor using 224x224 image patches and corresponding gene expressions of the spots. VQ-VAE-2 hyper parameters were not changed (K=64, D=64). The MLP model has 3 layers: (1) input layer with 128 units, (2) hidden layer with 1,024 units (same as the output dimensions of DenseNet-121), (3) output layer with 250 units (See Figure S2). Other training configurations below were applied both the retest model and our model:

- Loss: MSE
- Optimizer: stochastic gradient descent with learning rate of 1e-6 and momentum of 0.9
- Epochs: 50
- Batch size: 32
- Data augmentation: randomly rotating the image by 0, 90, 180 or 270°

Figure S2. Model training for external validation. The retest model (DenseNet-121) is trained in the end-to-end manner to minimize MSE for gene expression prediction. On the other hand, our model uses VQ-VAE-2 encoder as a feature extractor, and the simple MLP model is trained to minimize MSE.

Then we compared the results of the rest model and our method. We evaluated models by one-leave-out cross validation of 23 patients. As mentioned in section 3.3, the prediction performance for each gene is a median value of 23 correlation coefficients obtained by the cross validation. As shown in Table S2(A), although the ranges of top-5 prediction performances were smaller than the paper (left column), 3 of top-5 gene names were matched (p=3.8e-5) in both the retest model and our model. Table S2(B)

shows the prediction performance rankings of top-5 genes in the paper and 4 of top-6 gene names were matched except for XBP1.

Table S2. Results of ST-Net retest and our method validation. Values in () are median of correlation coefficients between measured and predicted gene expressions obtained by one-leave-out cross validation of 23 patients.

We could not reproduce the result of the previous study completely, but the trends are consistent. Therefore, we considered that we do not need to change our conclusion that our model works as well as the end-to-end model. On the other hand, the inadequate reproduction might be caused by newly implementing the retest program instead of using their official sources. We have noticed (but not confirmed so far) that potentially there were several mismatches between the original and ours such as 250 genes for prediction targets (they extracted 250 genes with highest mean expressions as prediction targets, but the entire list of them was not found.) and data loading (in their source, missing table data were referred in order to load spatial gene expressions and corresponding spot IDs.).

Restoration result by VQ-VAE2 model

Figure S3. Comparison between original images and restored images for each iPSC quality category. Color bar represents absolute error level in 8bit between the original and the restoration. A little difference between the images can be found mainly in high frequencies so that the restored images tend to have a smoother appearance than original images.

Training gene expression prediction models with different training patterns

To achieve a get prediction model, we have also performed the training of the model with different training patterns as shown in Table S3. However, the prediction performance could not be improved in both patterns.

Table S3A shows the prediction performance when both Dataset A3 and B3 are used for training together. In this case, only the datasets of Timepoint 2 and 3 were used for the test. We expected to get more generalized models against the batches, however, it showed that the models can predict only one of the batches, same as Table3. As shown in Figure 5, the gene expression profiles of the batches are clearly separated, while the image features do not have clear differences. Therefore, the models could not find the relationships between image features and gene expressions that satisfy both batches.

Table S3B shows the prediction performance when either Dataset A1 or B1 is used for training. We tried this pattern because making models at an earlier time point is more beneficial. However, we could not get better results than those of the training pattern shown in table3. This could be because the models trained with the datasets of Timepoint1 cannot handle the image feature variation of Timepoint 3 which may be bigger than that of Timepoint 1.

Table S3. Prediction performance with different training patterns. A) Performance in case of training with the dataset of Timepoint 1. B) Performance in case of training with both A3 and B3. Only the genes that have R2>0.3 for the multiple test datasets are listed.

Predicting the genes that have small batch effects

We also investigated the prediction performance for the genes that have similar expression levels between Experiment A and B. We retrained the prediction models only for the genes that expression levels are within the same range, with 20% error, between the experiments. As a result of this gene selection, 93 genes have remained. However, no genes show a good prediction performance(R2>0.3) for the multiple timepoints.

Table S4. Prediction performance for the genes that do not show the batch effects. Only the genes that have R2>0.3 for the test datasets are listed.

VQ-VAE-2 model overview

Figure S4. VQ-VAE-2 encoder and decoder. We encoded an input patch into two feature maps with distinct resolutions (top- and bottom-level). Each of the feature maps with two different resolutions (top- and bottom-level) encoded by CNN is quantized elementwise by the nearest one of K distinct embedding vectors of D dimension, so the vector-quantized feature maps have integers from 1 to K, which represent indices of the embedding vectors. We decoded vector-quantized feature maps into a restored patch. The vectors for quantization and the parameters of the encoder and the decoder are all trained to minimize mean square error (MSE) between input and restored patch.

Reproducibility of gene expression measurement

Figure S5. Reproducibility of gene expression measurement. A pair of the culture cell samples was obtained from a single clone to measure the consistency of the gene expression measurement. Each point represents a specific gene. The graph shows that the gene expression levels are consistent between two different measurements with R2=0.9854.

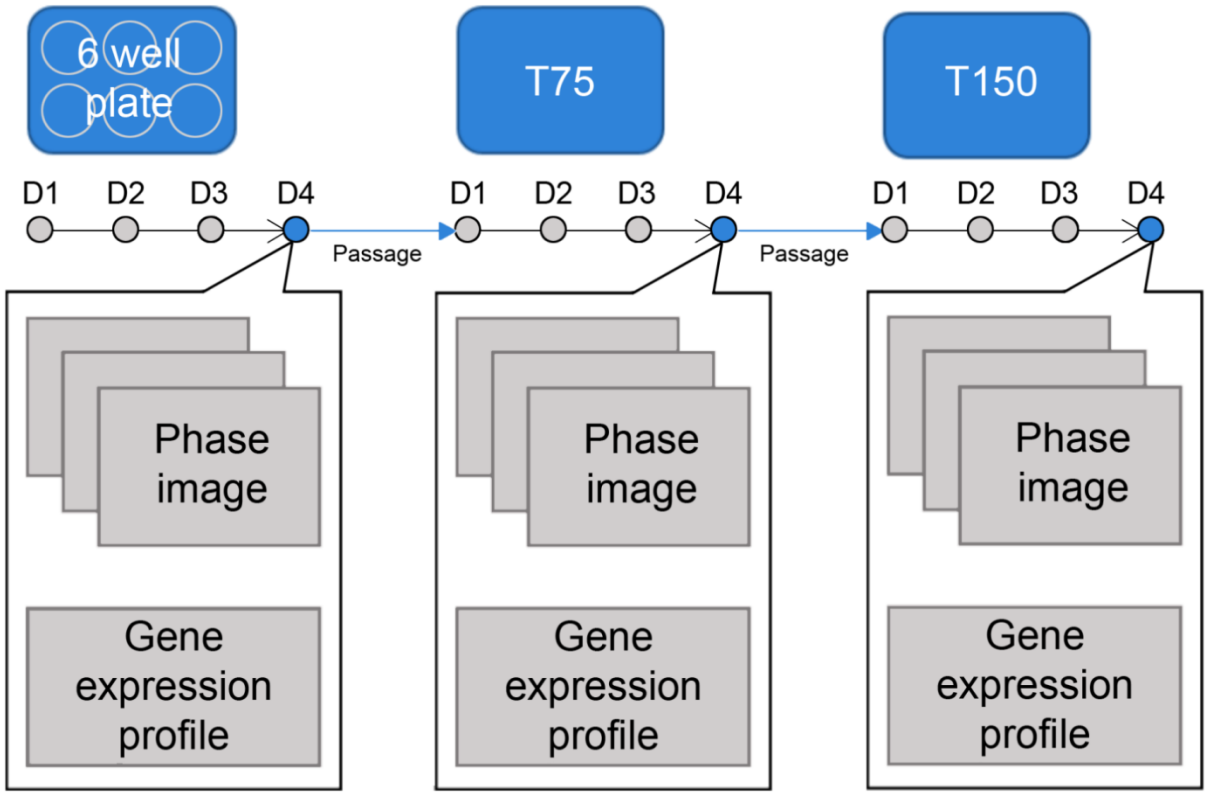

Figure S1

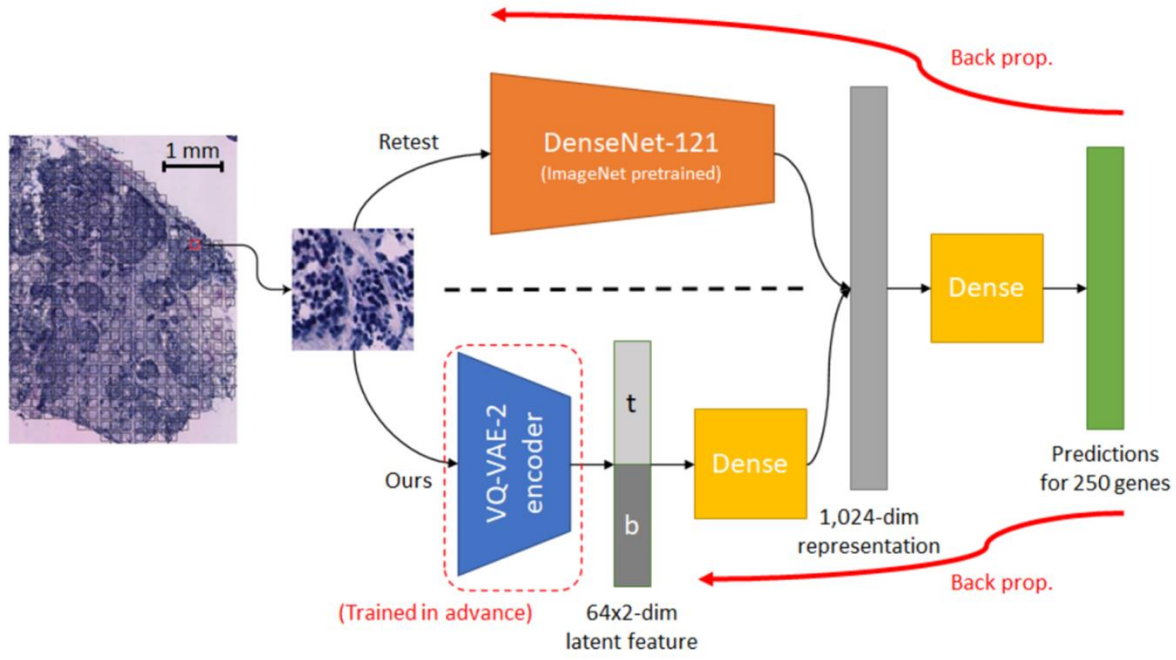

Figure S2

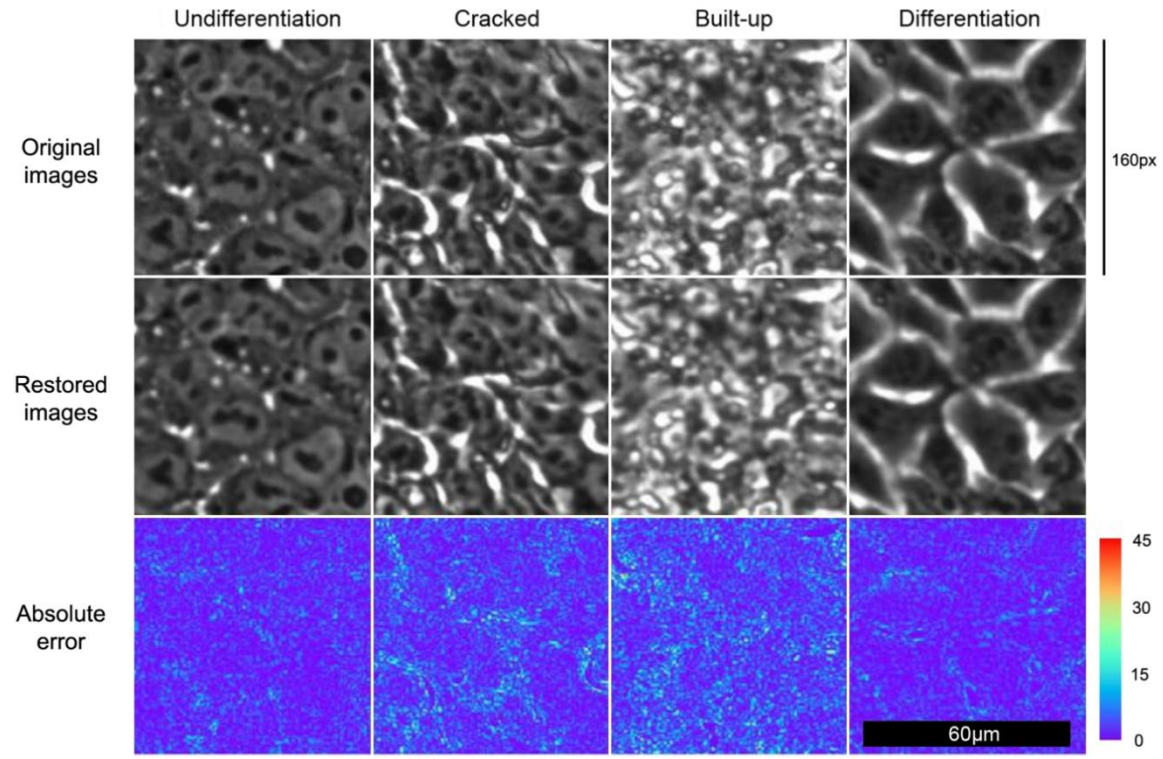
Undifferentiation
Cracked
Built-up
Differentiation
Original images
Restored images
Absolute error
160px
45
30
15
0
60μm

Figure S3

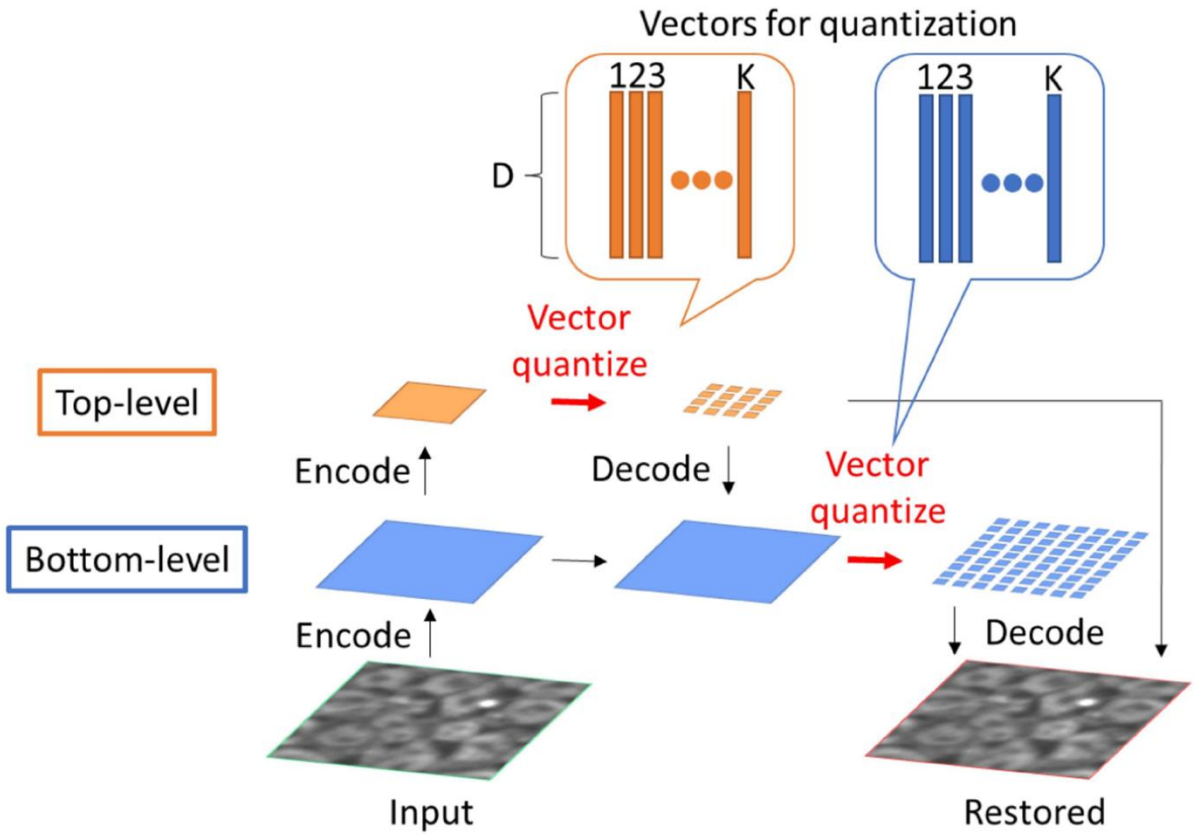

Figure S4

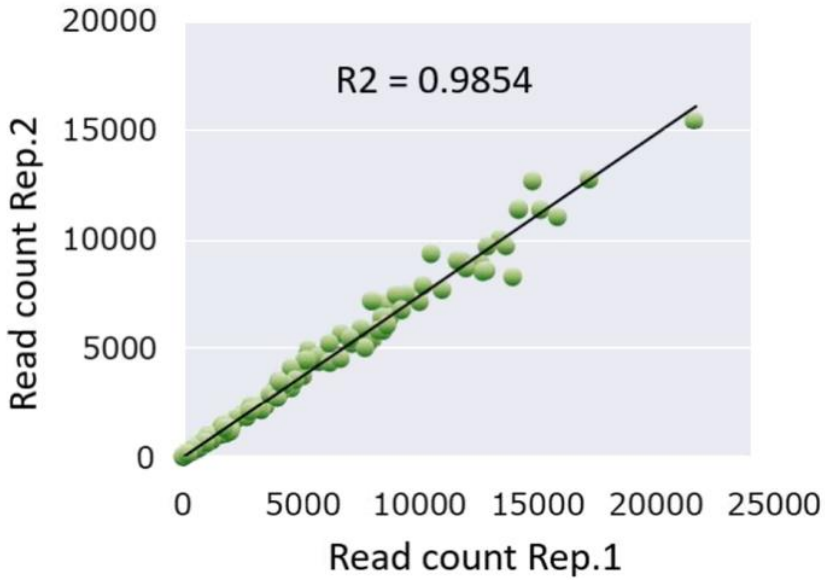

Figure S5

Experiment	Clones	Timepoint 1 (6 well plate, D4)	Timepoint 2 (T75, D4)	Timepoint 3 (T150, D4)
A (1 donor)	15 clones (Clone #1-15)	Dataset A1	Dataset A2	Dataset A3
B (14 donors)	14 clones (Clone #16-29)	Dataset B1	Dataset B2	Dataset B3

	He, B. et al[20]	Re-test (DenseNet)	Ours
1	DDX5 (0.52)	HSP90AB1 (0.325)	FASN (0.390)
2	ACTG1 (0.50)	FASN (0.324)	DDX5 (0.372)
3	FASN (0.50)	GNAS (0.319)	HSP90AB1 (0.343)
4	GNAS (0.49)	ACTG1 (0.317)	TPT1 (0.314)
5	XBP1 (0.43)	FN1 (0.315)	GNAS (0.309)

A) Genes with top-5 prediction performance

Genes	He, B. et al[20]	Re-test (DenseNet)	Ours
DDX5	1 (0.52)	6 (0.290)	2 (0.372)
ACTG1	2 (0.50)	4 (0.317)	6 (0.297)
FASN	3 (0.50)	2 (0.324)	1 (0.390)
GNAS	4 (0.49)	3 (0.319)	5 (0.309)
XBP1	5 (0.43)	21 (0.225)	61 (0.189)

B) Prediction performance ranks of the top-5 genes in the previous study

A

		Experiment A (15 clones, 1 donor)			Experiment B (14 clones, 14 donors)		
		Dataset A1	Dataset A2	Dataset A3	Dataset B1	Dataset B2	Dataset B3
Case 1 Trained with A1	BHLHE40	0.383	0.460	0.468	NS	NS	NS
	SIPA1L2	0.826	0.760	0.481	NS	NS	NS
Case 2 Trained with B1	SIPA1L2	0.552	0.526	0.502	0.753	NS	NS

B

		Experiment A (15 clones, 1 donor)			Experiment B (14 clones, 14 donors)		
		Dataset A1	Dataset A2	Dataset A3	Dataset B1	Dataset B2	Dataset B3
Case 3 Trained with A3 and B3	ABHD8	0.545	0.481	0.682	0.452	NS	0.391
	ENPP2	0.553	0.380	0.526	0.439	NS	0.362
	KLF4	0.595	0.663	0.783	0.115	NS	0.659
	SIPA1L2	0.364	0.445	0.711	NS	NS	NS
	ROR2	NS	0.453	0.615	0.685	NS	0.749
	NRXN1	NS	NS	0.526	0.418	0.351	0.481
	CCDC167	NS	NS	0.833	0.271	0.476	0.935
	COL11A1	NS	NS	0.752	0.407	0.343	0.526
	DNER	NS	NS	NS	0.565	0.431	0.759

		Experiment A (15 clones, 1 donor)			Experiment B (14 clones, 14 donors)		
		Dataset A1	Dataset A2	Dataset A3	Dataset B1	Dataset B2	Dataset B3
Case 1 Trained with A3	**ALCAM**	0.445	NS	0.676	NS	NS	NS
	IFITM2	0.732	NS	0.968	NS	NS	NS
	SFRP2	0.684	NS	0.716	NS	NS	NS
	ADA	NS	NS	0.859	NS	NS	0.672
	DLL3	NS	NS	0.431	NS	NS	0.392
Case 2 Trained with B3	**ALCAM**	0.660	NS	NS	NS	NS	0.649
	MAP2K1	NS	0.692	NS	NS	NS	0.841
	IFITM2	NS	NS	NS	0.587	NS	0.878
	OTX2	NS	NS	NS	0.473	NS	0.441
	PSMA1	NS	NS	NS	0.480	NS	0.885
	SEMA3A	NS	NS	NS	0.458	NS	0.822
	COL11A1	NS	NS	NS	NS	0.577	0.937
	FN1	NS	NS	NS	NS	0.384	0.642