

PHOTOMICS ANALYSIS OF SAMPLES

CROSS-REFERENCE TO RELATED APPLICATIONS

5 [0001] This application claims the priority benefit of United States provisional patent application titled, "PHOTOMICS ANALYSIS OF SAMPLES," filed November 7, 2023, and having serial number 63/596,954. The subject matter of this related application is hereby incorporated herein by reference.

FIELD OF THE PRESENT DISCLOSURE

10 [0002] The present disclosure relates to methods for determining a transcriptomic profile for a sample, using machine learning and live cell imaging data. Related methods, systems and products are described.

BACKGROUND

15 [0003] The advent of the genomics era enabled biologists to measure entire classes of molecules by performing a single elegant, albeit destructive, experiment. Such experiments have enabled biologists to better define the concept of cell states, to accurately classify and enumerate the types of cells that exist in varied organisms, and to generate hypotheses about mechanisms by identifying genes that change in response to perturbations. However, for all of their power, omics measurements are almost entirely destructive.

20 [0004] This destructiveness stands in contrast to the gentleness of brightfield microscopy. Brightfield microscopy is one of the earliest forms of microscopy, relying on transmitted white light to relay information. However, much of the information present in images acquired through brightfield microscopy is cryptic, and required the use of dyes to reveal it in a form that made sense to the human eye. As a result, 25 microscopists developed new methods such as confocal microscopy that offer ways to make this cryptic information visible by pairing lasers with chemical and genetic labels. These microscopy methods do come with side-effects: the powerful and collated laser beams can damage the cells, reducing their viability and/or altering the behaviors observed. Moreover, microscopists have a small number of emission 30 channels available to them, limiting the number of measurements that can be readily made. The invention of optogenetic methods further compounded this problem, as these methods use these channels to control cellular behaviors, introducing an unwanted tension between measurement of the cell and control.

[0005] Artificial intelligence approaches have been leveraged to access some of the cryptic information offered by brightfield images. For example, Oukomol et al. (2018) have shown that it is possible to predict a fluorescence image labeling subcellular structures such as membranes and DNA (i.e. nuclei) from a transmitted light image, using a convolutional neural network. Thus, information such as virtual staining of e.g. organelles can be obtained from brightfield images. However, this information is still relatively limited.

[0006] Therefore, a need exists for improved systems and methods for characterizing samples comprising cells, which do not suffer from all of the drawbacks of the prior art.

SUMMARY

[0007] The present inventors recognised that brightfield images contain the information necessary to infer the distribution, shape and texture of subcellular structures, and postulated that it may be possible to build a machine learning pipeline that can perform cross-domain mapping between label free light microscopy images and transcriptomes to prove that such images contain sufficient information to infer molecular quantities of mRNA. In particular, they designed and demonstrated the use of an end-to-end pipeline that accurately infers perturbed bulk transcriptomes from single-cell unlabelled (brightfield) images. They further showed the generalizability of the method by measuring transcriptomes of exact replicate conditions, unseen perturbation dosages; and unseen perturbations. Finally, they showed proof-of-concept “photomics” experiments that showcase the power of multi-modal learning.

[0008] Thus, according to a first aspect of the disclosure, there is provided a computer-implemented method of determining a transcriptomic profile for a sample comprising one or more cells, the method comprising: obtaining one or more single cell images of the sample, wherein the images have been obtained using a live cell imaging technology; predicting a transcriptomic profile for one or more cells of the sample using a machine learning model configured to take as input one or more single cell images and produce as output a predicted transcriptomic profile for the one or more cells in the single cell images, wherein the machine learning model has been trained using training data comprising for each of a plurality of samples, a plurality of training single cell images and corresponding measured transcriptomic profiles, wherein a measured transcriptomic profile comprises a set of values each value being

indicative of the level of expression of a respective transcript, and a predicted transcriptomic profile comprises a plurality of values that are each indicative of the expression level of a plurality of transcripts.

5 **[0009]** The method of the first aspect may have any one or any combination of the following optional features.

10 **[0010]** The machine learning model may comprise: a first machine learning model configured to take as input a single cell image and produce as output an image embedding; and a second machine learning model configured to take as input an image embedding and produce as output a predicted transcriptomic profile. The first and second machine learning models may have been trained independently. The first machine learning model may comprise a machine learning model that has been trained in a supervised manner using training data comprising a plurality of single cell images and associated labels, wherein the labels are not derived from transcriptomic data. The first machine learning model may comprise a machine learning model that
15 has been trained as a part of a machine learning model configured to classify single cell images between a plurality of classes, the plurality of classes associated with respective cell types, cell cycle phases, staining patterns, morphologies, and/or experimental conditions. The first machine learning model may comprise a plurality of such machine learning models, the outputs of which are concatenated.

20 **[0011]** The predicted transcriptomic profile and the measured transcriptomic profile may be single cell transcriptomic profiles or bulk transcriptomic profiles. The machine learning model may be configured to determine an image embedding for each of a plurality of single cell images, and predict a transcriptomic profile for the plurality of cells in the plurality of single cell images using the plurality of image embeddings. The
25 machine learning model may be configured to obtain a summarised image embedding from the plurality of image embeddings, and predict a transcriptomic profile for the plurality of cells in the plurality of single cell images using the summarised image embedding.

30 **[0012]** Obtaining the one or more single cell images of the sample may comprise obtaining one or more images of the sample, each image comprising one or more cells, and obtaining one or more smaller segments of the one or more images each comprising a single cell. Obtaining one or more smaller segments of the one or more

- images each comprising a single cell may comprise analysing the one or more images using an image analysis algorithm configured to identify regions of an image that contain a whole cell or a part thereof that is unique to a single cell. Obtaining one or more smaller segments of the one or more images each comprising a single cell
- 5 may comprise performing cell segmentation or nuclear segmentation to identify regions of an image that contain individual cells or nuclei. The image analysis algorithm may be a machine learning model that has been trained to predict a stained image from a label free image using training comprising pairs of label free images and corresponding stained images, wherein the stained images comprise signal
- 10 associated with a label specific to a particular subcellular structure. Obtaining one or more smaller segments of the one or more images may comprise obtaining a smaller segment of the one or more images for each of at least one of one or more identified regions of the image. Each smaller segment may have a predetermined size. Each smaller segment may be centred on the centre of mass of an identified region.
- 15 **[0013]** The live cell imaging technology may be label free or may use one or more labels that are not transcriptomic labels. The live cell imaging technology may be selected from: optical microscopy and live fluorescence imaging. The live cell imaging technology may be brightfield microscopy, or phase contrast microscopy, differential interference contrast microscopy, quantitative label-free imaging with phase and
- 20 polarization (QLIPP), labelled fluorescence imaging or autofluorescence imaging. The live cell imaging technology may be brightfield imaging.
- [0014]** The method may further comprise determining a value for each of the transcripts of the plurality of transcripts from the predicted transcriptomic profile. The machine learning model may have been trained to predict training predicted
- 25 transcriptomic profiles derived from measured transcriptomic profiles corresponding to the plurality of training single cell images, wherein the training predicted transcriptomic profiles have been derived from the measured transcriptomic profiles using an embedding (e.g. dimensionality reduction) method. The method may comprise applying a reverse of the embedding method to a predicted transcriptomic
- 30 profile provided as output of the machine learning model, thereby obtaining a set of values each value being indicative of the level of expression of a respective transcript present in the measured transcriptomic profiles. The training predicted transcriptomic profiles may have been derived from the measured transcriptomic profiles using an

embedding method, and the method may comprise applying a reverse embedding to a predicted transcriptomic profile provided as output of the machine learning model.

The embedding method may be principal component analysis, and the reverse embedding method may be reverse PCA. A training predicted transcriptomic profile

5 may comprise the values of the first n principal components for a measured transcriptomic profile, wherein n is between 5 and 15, between 8 and 12, or 10.

Similarly, a predicted transcriptomic profile may comprise values for the first n principal components of a PCA space identified from the measured transcriptomic profiles.

10 **[0015]** The machine learning model may have been trained using training data comprising training single cell images and corresponding measured transcriptomics profiles for each of a plurality of samples exposed to different perturbations. A perturbation may be a physicochemical condition that applies a stress to the cells.

[0016] The first machine learning model may comprise a deep neural network, 15 such as a convolutional neural network. The second machine learning model may comprise a regression model, optionally a non-linear regression model and/or a regularised regression model. The second machine learning model may be a non-linear and/or regularised regression model, such as a Ridge regression model, a LASSO regression model, an elastic net model, a neural network or a regularised 20 neural network. The method may comprise performing any of the steps described in relation to the subsequent aspect.

[0017] According to a second aspect, there is provided a computer-implemented method of providing a tool for determining a transcriptomic profile of a sample comprising one or more cells; the method comprising: obtaining training data 25 comprising for each of a plurality of samples, a plurality of training single cell images and corresponding measured transcriptomic profiles, wherein a measured transcriptomic profile comprises a set of values each value being indicative of the level of expression of a respective transcript, and training a machine learning model to take as input one or more single cell images and produce as output a 30 predicted transcriptomic profile for the one or more cells in the single cell images, wherein a predicted transcriptomic profile comprises a plurality of values that are each indicative of the expression level of a plurality of transcripts.

[0018] The method according to the present aspect may have any of the features described in relation to the first aspect, such as e.g. any features related to the training and/or configuration of the machine learning model. The method of the second aspect may have any one or any combination of the following optional

5 features. Training the machine learning model may comprise training a first machine learning model configured to take as input a single cell image and produce as output an image embedding; and training a second machine learning model configured to take as input an image embedding produced by the trained first machine learning model and produce as output a predicted transcriptomic profile. The method may

10 comprise deriving a training predicted transcriptomic profile from each measured transcriptomic profile, and training the machine learning model may comprise training the machine learning model to predict a training predicted transcriptomic profile from one or more training single cell images.

[0019] According to a third aspect, there is provided a method of monitoring a cell

15 population in cell culture, the method comprising: optionally culturing a cell population and/or obtaining one or more images of the cell population during the cell culture process; and determining a transcriptomic profile of the cell population or one or more cells thereof, using a method according to any embodiment of the first aspect.

[0020] The method of the third aspect may have any one or any combination of the

20 following optional features.

[0021] The method may further comprise determining and/or implementing one or more control actions based on the transcriptomic profile(s). A control action may also be referred to as a protocol change.

[0022] A control action may be selected from: changing a cell culture parameter

25 (such as e.g. changing the cell culture medium by at least partial replacement of the culture medium or addition of one or more compounds or compositions, changing a physical parameter of the cell culture such as temperature, physical environment such as e.g. culture substrate, agitation, etc.), harvesting, passaging or fixing the cells, modifying a scheduled time of change of a cell culture parameter, and modifying a

30 scheduled change of a cell culture parameter.

[0023] The methods described herein find use in the context of obtaining cells that have a predetermined characteristic that can be defined in terms of omic (e.g. transcriptomic) characteristics (e.g. suitability for a particular downstream purpose), in the context of perturbation screening including in particular drug screening, including
5 for the purpose of characterising the effect of one or more perturbations on one or more omic (e.g. transcriptomic) characteristics of a cell population, for drug discovery and/or development.

[0024] Thus, also described herein according to a fourth aspect are methods of characterising and/or comparing one or more perturbations applied to a population of
10 cells, the method comprising: culturing a cell population in the presence of the one or more perturbations and/or obtaining one or more images of the cell population in culture in the presence of the one or more perturbations; determining a transcriptomic profile of the cell population or respective transcriptomic profiles of one or more cells thereof, using a method according to any embodiment of the first aspect; and
15 optionally comparing the one or more transcriptomic profile(s) determined for cells cultured in the presence of a first and second perturbation.

[0025] The method of the fourth aspect may have any one or any combination of the following optional features.

[0026] The perturbation may be a candidate therapeutic intervention, such as e.g.
20 a drug. The perturbation may be a test agent that is being tested for its effects on ageing or ageing related disorders.

[0027] According to any embodiment of any aspect, the one or more cells may be healthy and/or diseased cells. The one or more cells may comprise cells that have an ageing related disorder. The cells may be human cells. The cells may be primary
25 cells, such as e.g. cells from a subject to be treated for a disease or disorder.

[0028] The methods described herein are computer implemented unless context indicates otherwise. Indeed, the amount of data that forms part of any cell imaging data and transcriptomics data, and the process of training deep learning models is of a complexity that places the methods described herein far beyond the capability of
30 mental investigation.

[0029] According to a fifth aspect, there is provided a system comprising: at least one processor; and at least one non-transitory computer readable medium storing instructions that, when executed by the at least one processor, cause the at least one processor to perform the method of any embodiment of the first to fourth aspects.

5 [0030] According to a further aspect, there is provided a non-transitory computer medium or media comprising instructions that, when executed by at least one processor, cause the at least one processor to perform the method of any embodiment of any aspect described herein.

[0031] According to a further aspect, there is provided a computer program
10 comprising code which, when executed on a computer, causes the computer to perform the method of any embodiment of any aspect described herein.

[0032] Embodiments of the present disclosure will now be described by way of example and not limitation with reference to the accompanying figures. However, various further aspects and embodiments of the present invention will be apparent to
15 those skilled in the art in view of the present disclosure.

[0033] The present invention includes the combination of the aspects and preferred features described except where such a combination is clearly impermissible or is stated to be expressly avoided. These and further aspects and embodiments of the disclosure are described in further detail below and with
20 reference to the accompanying examples and figures.

BRIEF DESCRIPTION OF THE DRAWINGS

[0034] **Figure 1** is a flowchart illustrating a method of determining a transcriptomic profile according to an embodiment of the disclosure.

[0035] **Figure 2** is a flow diagram illustrating schematically a method of providing a
25 tool according to embodiments of the present disclosure.

[0036] **Figure 3** is a flow diagram illustrating schematically a method of performing a perturbation screen according to embodiments of the disclosure.

[0037] **Figure 4** illustrates schematically an exemplary system according to the disclosure.

[0038] **Figure 5** shows an overview of a method of providing a tool for determining a transcriptomic profile according to an embodiment of the disclosure used in the examples below. Bright field images (bf, 12,007 separate images in the illustrated example) were analysed using a machine learning model (here a V-Net) to obtain a nuclear segmentation map. Bounding boxes centered on each segmented nucleus were created, to obtain single cell image patches (275,399 patches of 224x224 pixels in the illustrated example). Image embeddings were obtained for each patch using a machine learning model with embedding dimension 512 in the illustrated example. The machine learning model is a model that has been trained for an auxiliary class that is not transcriptomic profile prediction, such as for example to classify single cell patches between cell type classes. Separately, transcriptomics data (here bulk RNA sequencing data) was obtained for each of 96 samples from which the 12,007 brightfield images were obtained. PCA was applied to the transcript level counts across all samples, and the first 10 principal components (PCs) were retained. Finally, various machine learning models were used to predict the values along these 10 PCs for a sample from the average embeddings across all patches that are derived from images of the sample. Transcript level values (counts) can then be obtained from this prediction by reverse PCA.

[0039] **Figure 6** shows results of the method of Figure 5. In particular, the plot shows, for each of the 10 PCs, the percentage of the PC variable that could be explained by the best image-based predictive model described above (in particular, a fully connected neural network), i.e. the R^2 prediction performance (top), and the percentage of the variance in the data that is explained by the PC (bottom). The data shows that the method is able to predict approximately 70% of the variance associated with transcriptomes (obtained as a weighted sum of the percentage of variance explained for each PC, weighted by the R^2 for the respective PC), solely from label free images of samples.

[0040] **Figure 7** shows the architecture of a model used to obtain image embeddings according to embodiments of the disclosure. Specifically, by reference to the method illustrated on Figure 5. the model is used to extract an embedding for each single cell (bounding box) image by using the last layer of a Res-Net 18 classifier (as illustrated) trained on cell-type. In the illustrated example, the training data for the model comprises 322,879 single cell images (224x224 pixels) comprising

images of cells of one of two cell types (BJ fibroblasts or U2OS). The machine learning model has a feature space with 512 dimensions, which is fed to a pair of nodes representing the probability that a single cell image shows a cell of each of the two cell types. Images are allocated the class with highest probability. Accuracy is then evaluated on a hold out test set, and was found to be 99.6% in the illustrated example.

[0041] **Figure 8** shows results of prediction of a transcriptomic profile from brightfield images according to an example of the disclosure using the methods described in Figures 5 and 7. The table shows, for each of the PCs that were predicted, the top significantly enriched Gene Ontology (GO) term (separately for positive and negative genes) amongst the genes that contribute most to the PC loading, the percentage of variance explained by the PC, and the prediction performance obtained for the PC using the methods illustrated on Figure 5. The data shows that PC1 reflects biological differences in extracellular matrix production. These are likely to result in morphological features that are identifiable in the images and picked up by the image embeddings. PC3, a strong predictor of cycloheximide treatment, is associated with transcription of factors that open RNA Pol 1 promoters, which suggests that the cell is up-regulating ribosomes. This likely corresponds to changes in nucleolar morphology that are identifiable in the images, underlining the strong prediction performance for this PC.

[0042] **Figure 9** shows results of prediction of a transcriptomic profile from brightfield images according to an example of the disclosure using the methods described in Figures 5 and 7. **A.** The plot shows the mean R^2 score (point) and standard deviation around the mean (whiskers) obtained by regressing the transcriptome PCAs on image embeddings, using different machine learning models to obtain the predictions from the image embeddings. Each machine learning model has been trained using a fold of a cross-validation process (i.e. training the model on a subset of the training data and testing the performance on the remaining subset, where for each fold a different split of the training data is used), and the mean and standard deviation are calculated for the resulting ensemble of models. The R^2 score is the ratio of the total variance explained by the model ($SS_{res} = \sum_i (y_i - f_i)^2$ where y_i is an observed value and f_i is the corresponding prediction) relative to the total variance in the data to be predicted ($SS_{tot} = \sum_i (y_i - \bar{y})^2$ where $\bar{y}$ is the average of all

observations in the test set). **B.** Mean (point) and standard deviation (shaded envelope) of the R^2 score, separated by PC, for each of the models in A.

DETAILED DESCRIPTION

5 [0043] In describing the present invention, the following terms will be employed, and are intended to be defined as indicated below.

[0044] “and/or” where used herein is to be taken as specific disclosure of each of the two specified features or components with or without the other. For example “A and/or B” is to be taken as specific disclosure of each of (i) A, (ii) B and (iii) A and B, just as if each is set out individually herein.

10 [0045] As used herein, the terms “computer system” or “computer device” includes the hardware, software and data storage devices for embodying a system or carrying out a computer implemented method. For example, a computer system may comprise one or more processing units such as a central processing unit (CPU) and/or a graphical processing unit (GPU), input means, output means and data
15 storage, which may be embodied as one or more connected computing devices. Preferably the computer system has a display or comprises a computing device that has a display to provide a visual output display (for example in the design of the business process). The data storage may comprise RAM, disk drives or other computer readable media. The computer system may include a plurality of computing
20 devices connected by a network and able to communicate with each other over that network. For example, a computer system may be implemented as a cloud computer. The methods described herein may be provided as computer programs or as computer program products or computer readable media carrying a computer program which is arranged, when run on a computer, to perform the method(s)
25 described herein. The term “computer readable media” includes, without limitation, any non-transitory medium or media which can be read and accessed directly by a computer or computer system. The media can include, but are not limited to, magnetic storage media such as floppy discs, hard disc storage media and magnetic tape; optical storage media such as optical discs or CD-ROMs; electrical storage
30 media such as memory, including RAM, ROM and flash memory; and hybrids and combinations of the above such as magnetic/optical storage media.

[0046] The present disclosure provides methods to predict omics profiled from live cell images, using machine learning. Embodiments of the present disclosure relate to methods to predict transcriptomic profiles from live cell images, using machine learning.

5 **[0047]** Prediction of scRNAseq features (although not transcriptomic profiles as described herein) from images has been attempted in the past. For example, Jin et al. (2023) combined a cell picking system with single cell RNA sequencing and bright field live cell imaging. In particular, single cells were imaged in culture and a cell picking robot was used to isolate single cells and transfer them to individual wells for
10 RNA sequencing. The cells were then clustered into different cell types or subtypes using the RNA sequencing data, and a deep learning model was used to classify cells into these cell type classes (referred to as “transcriptome-Defined Cell Phenotypes”) using the brightfield images. Note that what is predicted here is not a transcriptomic profile but a classification between subgroups of cells that share transcriptomic
15 similarities. The same deep learning model was then trained to attempt to predict the expression level of each of the top 300 most variants genes (individually). The expression levels of most genes could not be accurately predicted, better than random predictions could only be obtained for 14 genes of the 300.

[0048] Wakui et al. (2023) investigated the relationship between gene expression
20 levels and morphology by analyzing live-cell, phase-contrast images and mRNA profiles collected during the process of purification of human induced pluripotent stem cells (hiPSCs). Multiple iPSC clones were imaged and gene expression analysis was done using targeted amplicon sequencing, a bulk RNA-seq strategy focused on a subset of genes, in this case 980 genes related to stemness and pluripotency. A
25 subset of 218 genes was selected based on having high mean expression coefficient of variation values (CV) across all the samples. Images were split into fixed size patches that tiled the image, comprising on average 30 cells. Each patch was analysed using a self-supervised deep learning model to obtain vector-quantized feature maps, from which feature vectors were extracted by counting the appearance
30 frequencies of each embedding vector from the feature maps for all patches. These were averaged across patches from the same image and a support vector regression model was trained to predict gene expression from these, separately for each gene. Statistically significant predictions could only be obtained for a selected set of genes,

and even for these the models trained on one batch of samples did not generalise to another batch of samples, even when increasing the diversity in terms of number of donors in the training batch.

[0049] By contrast, the present inventors have recognised that the number of variables to be predicted (e.g. transcriptomic variables, though other embodiments can predict metabolomic variables, proteomic variables or epigenomic variables) from imaging data would have to be limited in order to have a sufficiently constrained problem. However, instead of selecting specific genes based on expression level and/or variability and providing predictions for each of these, the present inventors adopted a fundamentally different approach of predicting a transcriptomic profile in one step, the predicted transcriptomic profile comprising a plurality of values that are each indicative of the expression levels of a plurality of genes. This can also be seen as predicting a transcriptomic profile comprising values for each of a plurality of metagenes, from which predictions for individual genes can be reverse engineered.

The present inventors have found that this approach was able to reliably capture the majority of the variability present in the measured transcriptomic data (even when this is genome wide transcriptomic data, rather than panel transcriptomic data).

[0050] Additional improvements were further developed in embodiments of the disclosure, including the use of perturbation data to train the models, which results in more robust and generalisable predictions, the use of automatically segmented single cell patches (which increases the flexibility of the method both in terms of imaging set up and transcriptomic technology since the method is applicable to both single cell and bulk transcriptomic technologies) and also increases the information content in the images as largely empty patches are not created, and the use of a deep learning model trained in a supervised manner for a task unrelated to transcriptomic profile prediction for obtaining image embeddings from which transcriptomic profile predictions are made.

[0051] The present disclosure relates to methods that analyse images of cells obtained using live cell imaging. The expression "live cell imaging" refers to imaging modalities that are compatible with cell viability. These include optical microscopy (such as e.g. brightfield imaging, phase contrast imaging, differential interference contrast microscopy, quantitative label-free imaging with phase and polarization (QLIPP)) and fluorescence live cell imaging (including labelled fluorescence imaging

and autofluorescence imaging). The imaging modality may be a label-free imaging modality. Alternatively, non transcriptomics labels may be used, such as e.g. labels associated with specific subcellular structures (e.g. nuclear, membrane or mitochondrial labels), labels associated with specific proteins (e.g. proteins expressed with a fluorescent tag), etc. Transcriptomics labels are labels that are associated with specific transcripts and can be used to quantify the level of said transcripts in the cells. According to the methods of the present disclosure, such labels may be used in images used to train a machine learning model to predict a transcriptomic profile from an image that does not include such labels. In other words, the methods of the present disclosure predict transcriptomic profiles from imaging data that is not obtained using transcriptomics labels, although such imaging data may be used in the process of training a model (by providing ground truth transcriptomic data for training).

[0052] The present disclosure relates to methods that predict omics profiles, in particular transcriptomic profiles, and methods that provide tools to predict such profiles using data comprising measured profiles (e.g. measured transcriptomic profiles). An omics profile is a set of values that are indicative of the level of a plurality of proteins (proteomic profile), metabolites (metabolomic profile), or transcripts (transcriptomic profile) or the presence of epigenetic modifications at a plurality of epigenetic loci (epigenetic profile) A transcriptomic profile is a set of values that are indicative of the expression levels of a plurality of transcripts (or genes, depending on whether values are reported at the gene level or at the transcript level - the two terms are used interchangeably unless context indicates that reference specifically to a gene or a transcript of a gene is intended). A measured omic profile is a set of measured values, each value being indicative of the level of a protein (proteomic profile), metabolite (metabolomic profile), transcript (transcriptomic profile) or epigenetic modification (epigenetic profile) in a sample. A measured transcriptomic profile is a set of measured values, each value being indicative of the level of expression of a respective transcript in a sample. A transcriptomic profile according to the disclosure can be measured using any known transcriptomic technology, preferably a high throughput technology, i.e. a technology capable of measuring the expression level of multiple hundreds of genes in a sample simultaneously. Transcriptomic technologies as used herein may be panel-based technologies or genome wide technologies. Transcriptomic technologies may be single cell or bulk technologies. For example, technologies selected from single cell RNA sequencing

(scRNA-seq) bulk RNA sequencing (RNA-seq), expression arrays, spatially resolved single-cell transcriptome profiling technologies (e.g. MERFISH, smFISH, *in situ* sequencing, etc.), molecular counting assays, etc. may be used. In specific embodiments, a measured transcriptomic profile is obtained by RNA sequencing

5 (either bulk or single cell). Thus, a measured transcriptomic profile may comprise counts of reads associated with each of a plurality of transcripts. Transcriptomic data may comprise a plurality of sequencing reads, or information derived therefrom (e.g. counts of reads). Alternatively, transcriptomic data may comprise intensities

10 associated with each of a plurality of transcripts (e.g. in the case of expression arrays or MERFISH assays). Features of transcriptomic profiles described below should be understood to also refer to corresponding features for other types of omics profiles. A predicted transcriptomic profile comprises a plurality of values that are each indicative of the expression levels of a plurality of transcripts. Each value may be referred to as a summary value or metagene. An estimate of the expression level of each of the

15 plurality of transcripts may be derivable (e.g. reverse engineered) from the summary values. Measured transcriptomic profiles may be used to train one or more machine learning models to predict transcriptomic profiles. For this purpose, each measured transcriptomic profile in the training data may be transformed to a ground truth predicted transcriptomic profile which comprises a plurality of values that are each

20 indicative of the expression levels of a plurality of transcripts represented in the measured transcriptomic profile. In other words, methods described herein may comprise obtaining a ground truth predicted transcriptomic profile from a measured transcriptomic profile. This may be performed using an embedding approach (also referred to herein as a dimensionality reduction approach) which combines

25 information from multiple data points in the data. Any embedding approach, whether linear or non-linear, may be used. Linear embedding approaches suitable for use in the present context include principal component analysis (PCA), nonlinear matrix factorisation (NMF), singular value decomposition (SVD), sparse PCA, sparse NMF, and sparse SVD. Nonlinear approaches suitable for use in the present context include

30 any embedding generated by supervised or non-supervised non-linear learning methods, such as artificial neural networks including deep learning models such as generative adversarial networks (GANs), transformer-based models (including scGPT (Cui et al. 2023) and scFoundation (Hao et al. 2023)) and variational autoencoders. In embodiments, the embedding approach is principal component analysis (PCA). PCA

is computationally efficient and invertible, enabling complete inference of the transcriptome, rather than inference of a selected subset of the transcriptome. A predicted transcriptomic profile corresponding to a measured transcriptomic profile may comprise the values of the first n principal components for the measured transcriptomic profile, where n is an integer. For example, n may be any number between 3 and 15, between 5 and 15, or between 8 and 12. For example, n may be 10. Thus, methods of the disclosure may comprise performing a PCA on a training dataset comprising a plurality of measured transcriptomic profiles, thereby obtaining a corresponding plurality of ground truth predicted transcriptomic profiles each comprising the values of the first n principal components for a measured transcriptomic profile. The reverse of the process of obtaining ground truth predicted transcriptomic profiles that was applied to training data comprising a plurality of measured transcriptomic profiles may be used to obtain estimates of expression levels for individual transcripts from a predicted transcriptomic profile. For example, reverse PCA may be applied to a predicted transcriptomic profile, thereby obtaining an estimate of the expression level. Thus, an estimate of expression level may be obtained for each of a plurality of genes present in the training data that was used to train the model from which the predicted transcriptomic profile was obtained. A measured transcriptomic profile may be associated with a single cell (e.g. when single cell technologies such as MERFISH or single cell RNA-seq are used) or a plurality of cells (e.g. when bulk technologies such as bulk RNAseq are used). A pseudo-bulk transcriptomic profile may be obtained from a plurality of single cell transcriptomic profiles using methods known in the art, such as e.g. by averaging multiple single cell transcriptomic profiles. Similarly, a predicted transcriptomic profile may be associated with a single cell or a plurality of cells. For example, when single cell transcriptomic profiles are available for training of the machine learning models described herein, the models can be trained to predict a single cell transcriptomic profile from a single cell image. In other embodiments, the machine learning models described herein may be trained to predict a bulk transcriptomic profile from a plurality of single cell images. In such embodiments, a transcriptomic profile may be predicted from a summarised image embedding obtained by combining a plurality of image embeddings from respective single cell images. For example, a summarised image embedding may be obtained by averaging image embeddings obtained for a plurality of single cell images. The plurality of single cell images may comprise at least 10 images, at least

20 images or at least 100 images. Thus, a summarised image embedding may be obtained which summarises information about at least 10 cells, at least 20 cells, or at least 100 cells. The number of cells that best captures the diversity in a population may depend on the sample, such as e.g. the number of cells in the sample, the
5 diversity in the sample, etc. The plurality of single cell images may be obtained from the same original image or from a plurality of images of the same sample. In other words, a plurality of frames of a sample may be taken, and one or more single cell images may be obtained from each of the plurality of frames.

[0053] The methods described herein use images of samples comprising cells.
10 Such samples may be cell cultures, tissue cultures, organoid culture or multicellular organism culture that can be imaged live (e.g. cultures of *C. elegans* or *Drosophila*). A “cell culture” refers to a bioprocess whereby live cells are maintained in an artificial environment such as a cell culture dish. A cell culture dish may be a plate, a flask, or any other type of container that is compatible with the acquisition of images of a cell
15 culture while the cell culture is underway. The methods and systems described herein are applicable to bioprocesses that use any types of cells that can be maintained in culture, whether eukaryotic or prokaryotic. In embodiments, the cells are eukaryotic cells, preferably animal cells. In embodiments, the cells are mammalian cells. For example, the cells may be human, mouse, or rat cells. In general, any human or non-
20 human model animal may be a source of cells or tissues that can be analysed according to the disclosure. A tissue culture refers to a bioprocess whereby a plurality of cells organised as a tissue either in 2D or 3D are maintained in an artificial environment such as a cell culture dish, or in an environment in which the tissue can be subject to live tissue imaging (including *in vivo*). Cells may be primary cells or cell
25 lines. The samples may be *in vitro* samples.

[0054] In embodiments, the cells in a sample analysed or used to provide training data may be subject to one or more perturbations. Data obtained from samples that have been subject to one or more perturbations may be referred to as “perturbation data”. Such data may comprise paired live cell images and transcriptomic profiles (or
30 transcriptomic data from which transcriptomic profiles can be derived). A perturbation is a physicochemical condition or genetically-encoded alteration that cells in a sample are exposed to, which applies a stress to the cells. A perturbation may comprise any non-physiological condition, such as e.g. a non-physiological pH, salinity, pressure,

temperature, starvation, the presence of a drug (e.g. a small or a large molecule), or the alteration of cellular physiology by genetic modification (e.g. knock out or knock in of a sequence resulting in the production of a new protein or the lack of or reduced production of a native protein, e.g. CRISPR KO screen). It is particularly

- 5 advantageous for the training data used to train one or more machine learning models according to the disclosure to comprise data from a plurality of samples of cells that have been exposed to different perturbations. In other words, it is advantageous for the training data to comprise images of samples and corresponding transcriptomic profiles, wherein the samples have been exposed to a plurality of perturbations.
- 10 Indeed, this may result in a richer data set as at least some of the perturbations can be expected to have an effect on the cell morphology and the transcriptome, and triggering these different effects may enable to learn relationships between the transcriptome and cell morphology.

- [0055]** The methods of the present disclosure make use of and/or train machine
- 15 learning models. The term “machine learning model” refers to a mathematical model that has been trained to predict one or more output values based on input data, where training refers to the process of learning, using training data, the parameters of the mathematical model that result in a model that can predict outputs values that satisfy an optimality criterion or criteria. A machine learning model may be trained in a
- 20 supervised or unsupervised manner. In the case of supervised learning, training typically refers to the process of learning, using training data, the parameters of the mathematical model that result in a model that can predict outputs values with minimal error compared to comparative (known) values associated with the training data (where these comparative values are commonly referred to as “labels” or
- 25 “ground truth”). In the case of unsupervised learning, the parameters of the model are learned from unlabelled data, for example by learning a latent representation of input data from which the input data can be reconstructed. In embodiments, the machine learning models described herein are trained in a supervised manner. The term “machine learning algorithm” or “machine learning method” refers to an algorithm
- 30 or method that trains and/or deploys a machine learning model. Regression models can be seen as machine learning models. Conversely, some machine learning models can be seen as regression models in that they capture the relationship between a dependent variable (the values that are being predicted) and a set of independent variables (the values that are used as input to the machine learning

model, from which the machine learning model makes a prediction). A machine learning model as described herein may be used to obtain image embeddings and/or predict a transcriptomic profile from image embeddings. In embodiments, a first machine learning model is used and/or trained to obtain image embeddings and a
5 second machine learning model is used and/or trained to predict a transcriptomic profile from image embeddings. In other embodiments, an end-to-end machine learning model may be used and/or trained to obtain image embeddings and predict a transcriptomic profile from the image embeddings. The first model may be referred to herein as an "image model". The second model may be referred to herein as a
10 "photomic model", by reference to the fact that it "translates" information from an image (photography) into corresponding transcriptomic information (a predicted transcriptomic profile). When a single machine learning model is used, this may comprise a first part that obtains an image embedding (which may be referred to as an image model) and a second part that predicts a transcriptomic profile from an
15 image embedding (which may be referred to as a photomics model). Thus, an image model and a photonics model may be trained independently or together as part of a single model trained to predict a transcriptomic profile from an input image. Advantageously, the training of the image model and the photonics model may be decoupled, i.e. an image model may be trained for a first task, and then a photonics
20 model may be trained to predict a transcriptomic profile from image embeddings produced by the trained image model.

[0056] An image embedding is a representation of an input image in a learned latent space that is of lower dimension than the input image while attempting to capture information in the image that is relevant to a particular task. The task may be
25 reconstructing the image (in the case of an image model that is trained in an unsupervised manner), predicting a label associated with the image or individual regions/pixels of an image (in the case of an image model that is trained in a supervised manner). An image model may be a deep neural network, such as a convolutional neural network (CNN). For example, an image model may be a ResNet
30 (e.g. ResNet-18, ResNet-50, etc.), a variational autoencoder (e.g. V-net, a U-Net), a vision transformer, a Swin transformer, etc. In embodiments, the image model is a CNN, such as ResNet-18. In embodiments, the image model is trained in a supervised manner. Specifically, the image model may be trained in a supervised manner for an auxiliary task. An "auxiliary task" is a task that is used to train an image

model in a supervised manner, that is different from the prediction of a transcriptomic profile. In other words, an auxiliary task is the prediction of a label associated with images of cells, the labels being derived from information that is not obtained by transcriptomic analysis of the cells. For example, an image model may be trained to

5 predict a characteristic of a cell, from an image of the cell, wherein the characteristic is selected from: the cell type, an experimental condition (e.g. a perturbation that the cell was exposed to, a disease status associated with the cells), a morphology of the cells, a cell cycle phase (e.g. determining whether the cells are in the G1 phase), the presence of a label associated with expression of a particular marker (also referred to

10 as "staining pattern"), a cell division rate, a cell density (number of neighbours within a predetermined distance of the nucleus of a cell), etc. The characteristic that is predicted is advantageously one that is not so trivial that very high prediction accuracies can be obtained with a small number of features. Indeed, this may result in a loss of features that would ultimately lead to poor generalisability to the final

15 prediction task (omics profile prediction). However, the prediction would also advantageously be achievable with better than random accuracy, as otherwise this may indicate that the model was not able to learn informative features. The above-mentioned characteristics are believed to strike a good balance in terms of being complex enough to learn features that will be informative in the final prediction task,

20 while being achievable with better than random accuracy. In embodiments, an image model may be trained to predict a characteristic of a cell including a cell type. Thus, the training data used to train a machine learning model as described herein may comprise data for a plurality of cells from different cell types. Each cell represented in a training image may be associated with a cell type label. These may be obtained by

25 manual or automatic annotation (e.g. when samples comprise cells from different cell types), or may be known from the experimental conditions (e.g. when samples comprise cells from a single, known cell type). Automatic annotation may use a machine learning model that has been previously trained on another training dataset. Manual annotation may be more reliable and hence preferable. Similar principles can

30 be applied when using an image model trained to predict other characteristics of a cell, such as e.g. cell cycle phase or the presence of a label associated with presence of a particular marker (e.g. expression of a particular protein or transcript). In the former case, each cell represented in a training image may be associated with a cell cycle phase label. In the latter case, each cell represented in a training image may be

associated with a label indicating whether the cell is positive or negative for the label associated with presence of the marker, a label indicating an extent of presence of the label associated with presence of the marker (e.g. a measured intensity value corresponding to the presence of the label), or a plurality of binary or continuous labels associated with respective regions (including e.g. individual pixels) of the image.

[0057] Further, the image model may comprise a plurality of image models, each trained for a different auxiliary task, and the outputs of which are combined (e.g. concatenated) to obtain an image embedding for transcriptomic prediction.

10 **[0058]** An image model may take as input a single cell image, also referred to as a “single cell patch” or simply “patch”. A single cell image may be obtained or may have been obtained from an image comprising a plurality of cells, using a cell segmentation process. A cell segmentation process may comprise identifying one or more regions of an image that each include a single cell, and defining a bounding box for each such
15 identified region. The area within said bounding box may then be selected as a single cell image. A bounding box may be an area that has a predetermined size. Identifying one or more regions of an image that each include a single cell may comprise using a machine learning model trained to identify regions of an image that include a single cell. The machine learning model may have been trained to identify regions of an
20 image that contain a whole cell or a part thereof that is unique to a single cell, such as a nucleus. Such a machine learning model may be referred to as a cell segmentation model (regardless of whether the model segments single cells or a part thereof, such as e.g. nuclei, Golgi apparatus, etc), or a virtual staining model (as identifying regions of an image corresponding to specific parts of cell may be equivalent to predicting an
25 image that would have been obtained using stains specific for these parts). Thus, identifying one or more regions of an image that each include a single cell may comprise performing nuclear segmentation. Thus, a cell segmentation model may be a virtual staining model that has been trained to predict, from a label-free image, a corresponding image labelled with a nuclear stain. The same principles can be
30 applied to any other cellular part that can be used to identify single cells. A bounding box may be defined as a fixed size area that is centred on the centre of mass of a segmented cell or part thereof (e.g. a segmented nucleus). The size of the bounding box may be defined in terms of a number of pixels, typically using the same width and

height dimensions (i.e. square bounding boxes). The size of a bounding box may be a predetermined size, defined depending on the expected size of the cells (in terms of numbers of pixels) in the images that are analysed. This may in turn depend on the resolution and magnification of the images, as well as the cell types that are analysed.

- 5 A cell segmentation model may be a deep neural network, such as a CNN, trained in a supervised manner using images of cells where regions of the images corresponding to single cells or parts thereof have been annotated. A cell segmentation model may be a deep neural network, such as a CNN, trained in a supervised manner using pairs of unlabelled and labelled images, wherein specific
- 10 cellular structures have been labelled in the labelled images. Such a model may be referred to as a virtual staining model. The specific cellular structure may include the nucleus (e.g. using labelled images labelled with a DNA stain such as Hoechst, DAPI or hematoxylin), the cytoplasm (e.g. using labelled images labelled with an actin stain such as phalloidin, eosin, etc.), mitochondria (e.g. using labelled images labelled with
- 15 a mitochondrial stain such as mitotracker – where mitochondrial staining enables cytoplasm segmentation), the Golgi complex (e.g. using labelled images labelled with ceramide dyes, Golgi fluorescent fusion proteins or wheat-germ agglutinin), the endoplasmic reticulum (e.g. using labelled images labelled with an ER stain, such as concanavalin A). For example, a cell segmentation model may be a V-Net, U-Net,
- 20 ResNet, etc. Any encoder-decoder model or generative adversarial network may be used for the purpose of cell segmentation including by virtual staining. For example, a cell segmentation model may be a model as described in Guo et al. 2020 or Rivenson et al. 2020. Guo et al. 2020 describes a deep neural network (multi-channel 2.5D U-Net) trained to predict fluorescence images of diverse cell and tissue structures from
- 25 QLIPP images. In particular, the authors demonstrate the prediction of fluorescence associated with F-actin, enabling cellular segmentation. Rivenson et al. 2020 describe a convolutional neural network trained using a generative adversarial-network model can transform wide-field autofluorescence images of unlabelled tissue sections into images that are equivalent to the bright-field images of histologically stained versions
- 30 of the same samples. Commonly used histological stains such as hematoxylin and eosin enable nucleus and cytoplasm segmentation. Alternatively, a cell segmentation model may be a virtual staining model trained to predict stained images from corresponding brightfield images in two steps comprising a first step using a first loss (e.g. a pixel-based loss such as mean squared error calculated based on differences

between pixel intensity values for predicted and ground truth stained images) and a second step using a second loss, where the second loss is based on shape-based and/or texture-based attributes of the predicted and ground truth stained images (e.g. a loss that quantifies the difference between a segmentation mask for predicted
 5 stained images and corresponding ground truth stained images, and/or a loss that quantifies perceived similarity between predicted and ground truth stained images, for example using the structural similarity index or a perceptual loss). The structural similarity index measure (SSIM) is a metric used to quantify perceived quality of digital images (see Wang et al. 2004).

10 **[0059]** A photomic model may take as input an image embedding and produce as output a omic profile, specifically a transcriptomic profile. An image embedding may be a feature vector comprising a predetermined number of features that is higher than a predetermined number of features in a transcriptomic profile to be predicted. For example, a transcriptomic profile comprising between 5 and 15 summary values may
 15 be predicted from an image embedding comprising between approximately 100 and approximately 2000 features. Thus, an image embedding may comprise a number of features between 128 and 2048, between 256 and 2048, between 512 and 2048, between 128 and 1024, between 256 and 1024, or between 512 and 1024. For example, an image embedding may comprise about 512 features. A photomic model
 20 may be a linear or non-linear regression model. A regression model is any machine learning model that can be used to learn a relationship between one or more response variables (e.g. values of a transcriptomic profile) from one or more predictive variables (e.g. features of an image embedding). The relationship may be linear or non-linear. In embodiments, the photomic model is a non-linear regression
 25 model. A non-linear regression model may advantageously be able to capture non-linear transcriptomic behaviours. A photomic model may be a regularised regression model. A regularised model may be L1-regularised and/or L2-regularised. Regularisation comprises including one or more terms in the learning objective used to train a machine learning model, which penalise model complexity. Regularisation
 30 may be used to prevent overfitting of the training data. In L1 regularisation one or more penalty terms are used which depend on the absolute value of the magnitude of the coefficients of the model. In L2 regularisation one or more penalty terms are used which depend on the squared magnitude of the coefficients of the model. L1-regularisation eliminates features of lower importance. Thus, L1 regularisation may be

useful when the predictive variables comprise redundant or uninformative variables. By contrast, L2-regularisation reduces the contribution of less informative variables. In embodiments, a photomic model is a L2-regularised regression model.

[0060] A photomic model may be selected from: a simple linear regression model, a LASSO (least absolute shrinkage and selection operator) regression model, a multi-task LASSO regression model, an elastic net model, a ridge regression model, a random forest regression model, and a neural network (such as e.g. a fully connected neural network, or a regularised neural network, such as e.g. using the lockout method, Valdes et al. 2021). A LASSO regression model is an L1-regularised linear regression model. A multi-task LASSO (see scikit-learn.org/stable/modules/linear_model.html#multi-task-lasso) is a linear regression mode trained with a mixed L1-L2 norm. Elastic net is a linear regression model trained with both L1 and L2-norm regularisation of the coefficients. A Ridge regression is a linear regression model trained with L2-regularisation. A random forest regression model is a regression model that trains an ensemble of decision trees. Regularisation is also possible in neural networks such as e.g. using the lockout method (Valdes et al., 2021). In embodiments, the photomic model is a Ridge regression model or a neural network. In embodiments, the photomic model is a neural network, for example a fully connected neural network.

20 ***Methods of Determining Transcriptomic characteristics***

[0061] **Figure 1** is a flow diagram showing, in schematic form, a method of determining a transcriptomic profile according to embodiments of the disclosure. At optional step 10, a sample comprising one or more cells is cultured. In other words, cells are maintained in a culture environment compatible with cell viability and live cell imaging. At optional step 12, one or more images of the sample are obtained during the culture process using a live cell imaging technology. The images may be 2D images or 2.5D or 3D images. When 2.5D or 3D images are obtained (i.e. each image comprising a plurality of images corresponding to different depths, also referred to as a “z-stack”), 2D images may be obtained from these by selecting an image corresponding to a single plane of each z-stack (e.g. an image corresponding to the focal plane). At optional step 14, one or more single cell images of the sample are obtained, from the images obtained at step 12. This may be comprise obtaining smaller segments of the images obtained at step 12, each comprising a single cell.

This may be performed by analysing the images obtained at step 12 using an image analysis algorithm configured to identify regions of an image that contain a whole cell or a part thereof that is unique to a single cell (e.g. a machine learning model trained to identify regions of an image that contain a whole cell or a part thereof that is unique to a single cell). Obtaining one or more smaller segments of the one or more images each comprising a single cell comprises performing cell segmentation or nuclear segmentation to identify regions of an image that contain individual cells or nuclei, and obtaining a smaller segment of the one or more images for each of at least one of one or more identified regions of the image. Each smaller segment may have a predetermined size. Each smaller segment may be centred on the centre of mass of an identified region (e.g. when the identified region is a nucleus or a whole cell / cytoplasm). In other words, single cell images may be obtained by defining a bounding box for an identified region. Steps 10 to 14 are optional because the methods described herein can start from images or single cell images of a sample that have been previously acquired and received from a computing device, memory or database. Therefore, obtaining one or more single cell images or one or more images of a sample may comprise receiving the one or more images from a computing device, imaging device, data store or user interface. The method may comprise acquiring one or more images of a sample using imaging means. The images may be single cell images or images comprising one or more cells. A data store may be a public or private database.

[0062] At step 16, a transcriptomic profile is predicted for one or more cells of the of the sample using a machine learning model configured to take as input one or more single cell images and produce as output a predicted transcriptomic profile for the one or more cells in the single cell images. The machine learning model has been trained using training data comprising for each of a plurality of samples, a plurality of training single cell images and corresponding measured transcriptomic profiles, wherein a measured transcriptomic profile comprises a set of values each value being indicative of the level of expression of a respective transcript, and a predicted transcriptomic profile comprises a plurality of values that are each indicative of the expression level of a plurality of transcripts. Further details in relation to training of the machine learning model will be described by reference to Figure 2 below. Step 16 may comprise a step 16A of processing the single cell image with a first machine learning model configured to take as input a single cell image and produce as output

an image embedding, and a step 16C of processing the image embedding obtained at step 16A with a second machine learning model configured to take as input an image embedding and produce as output a predicted transcriptomic profile. The predicted transcriptomic profile and the measured transcriptomic profile may be single cell transcriptomic profiles or bulk transcriptomic profiles. In embodiments where the predicted transcriptomic profile and the measured transcriptomic profile are bulk transcriptomic profile, step 16 may comprise optional step 16B of obtaining a summarised image embedding from a plurality of image embeddings, each corresponding to one of a plurality of single cell images (e.g. an average of the plurality of embeddings corresponding to single cell images). Thus, a summarised image embedding may be obtained as the average of a plurality of image embeddings. The summarised image embedding may then be processed at step 16C. Alternatively, single cell images may be processed at steps 16A and 16C, resulting in a plurality of predicted transcriptomic profiles (one for each single cell image), and a summarised predicted transcriptomic profile may be obtained (e.g. an average of the plurality of predicted transcriptomic profiles).

[0063] At step 18, a value may be determined for each of the transcripts of the plurality of transcripts from the predicted transcriptomic profile. This step is optional because the predicted transcriptomic profile comprising values that represent multiple transcripts is also usable in itself, for example to compare multiple samples or conditions. The machine learning model used at step 16 may have been trained to predict training predicted transcriptomic profiles derived from measured transcriptomic profiles corresponding to the plurality of training single cell images, wherein the training predicted transcriptomic profiles have been derived from the measured transcriptomic profiles using an embedding (e.g. dimensionality reduction) method. In such embodiments, when step 18 is used, the step may comprise applying a reverse of the embedding method to a predicted transcriptomic profile provided as output of the machine learning model, thereby obtaining a set of values each value being indicative of the level of expression of a respective transcript present in the measured transcriptomic profiles. The embedding method may be PCA and the reverse embedding method may be reverse PCA.

[0064] Any embodiment of a method as described by reference to Figure 1 may further comprise training the machine learning model, training the first machine

learning model and/or training the second machine learning model, for example as described by reference to Figure 2. At optional step 20, results of any one or more of steps 12 to 18, and/or any information derived therefrom are provided to a user, for example through a user interface. Results of the methods described herein may

5 comprise one or more of: one or more predicted transcriptomic profile, a value indicative of the level of expression of a transcript derived from a predicted transcriptomic profile, a trained machine learning model, the values of parameters (e.g. architecture and weights) of a trained machine learning model. Information derived from the results of the method may comprise any information that can be

10 derived from transcriptomic information, as known in the art. Steps 12 to 18 may be repeated any number of times during the cell culture, thereby providing a time lapse transcriptomic profile for the cells in the cell culture.

[0065] **Figure 2** is a flow diagram that illustrates in schematic form, a method of providing a tool for determining a transcriptomic profile according to embodiments of

15 the disclosure. At step 200, training data is obtained comprising for each of a plurality of samples, a plurality of training single cell images and corresponding measured transcriptomic profiles, wherein a measured transcriptomic profile comprises a set of values each value being indicative of the level of expression of a respective transcript. Step 200 may comprise any of optional steps 210 to 218. These steps are optional

20 because the method may also start from training data that has been previously acquired and that is being received from a computing device or retrieved from a memory or data store at step 200.

[0066] At optional step 210, a plurality of samples each comprising one or more cells are cultured. In other words, cells are maintained in a culture environment

25 compatible with cell viability and live cell imaging. The samples may comprise a plurality of samples exposed to different perturbations, where a perturbation can be a physicochemical condition that applies a stress to the cells. At optional step 212, one or more images of each sample are obtained during the culture process using a live cell imaging technology. At optional step 214, one or more single cell images of the

30 sample are obtained, from the images obtained at step 212. Each of these steps may have any of the features described by reference to the corresponding steps in the method of Figure 1. At optional step 216, transcriptomic data is acquired corresponding to each of the samples. At optional step 218, a training predicted

transcriptomic profile is derived from each measured transcriptomic profile obtained at step 218, using an embedding method. An embedding method may be a dimensionality reduction method. Thus, the machine learning model trained at step 220 as explained further below may be trained to predict training predicted transcriptomic profiles derived at step 218 from measured transcriptomic profiles obtained at step 216 corresponding to the plurality of training single cell images obtained at step 214, using an embedding (e.g. dimensionality reduction) method. The embedding method may be PCA. A training predicted transcriptomic profile obtained at step 218 may comprise the values of the first n principal components for a measured transcriptomic profile, wherein n is between 5 and 15, between 8 and 12, or 10.

[0067] At optional step 219, ground truth non-transcriptomic labels are obtained for each of a plurality of the single cell images obtained at step 214. These ground truth labels are used to train a first machine learning model as described further below, in embodiments in which such a model is used. The single cell images used for this may be a subset of those obtained at step 214, which may (but does not need to) overlap (partially or fully) with the single cell images used to obtain image embeddings from which a second machine learning model is trained as will be explained further below.

[0068] At step 220, a machine learning model is trained using the training data obtained at step 200. The machine learning model is trained to take as input one or more single cell images and produce as output a predicted transcriptomic profile for the one or more cells in the single cell images, wherein a predicted transcriptomic profile comprises a plurality of values that are each indicative of the expression level of a plurality of transcripts.

[0069] Step 220 may comprise a step 220A of training a first machine learning model configured to take as input a single cell image and produce as output an image embedding, and step 220C of training a second machine learning model configured to take as input an image embedding produced by the trained first machine learning model and produce as output a predicted transcriptomic profile. Steps 220A and steps 220C may be performed independently, i.e. the first and second machine learning models have been trained independently. In other words, the parameters of the first and second machine learning models may be determined through separate training steps where the performance of the respective model is evaluated in order to identify

parameters that are optimal according to a respective loss function for a respective task. The first machine learning model comprises a machine learning model that has been trained in a supervised manner using training data comprising a plurality of single cell images and associated labels obtained at step 219, wherein the labels are

5 not derived from transcriptomic data. For example, the first machine learning model may be trained to classify single cell images between a plurality of classes, the plurality of classes associated with respective cell types, cell cycle phases, staining patterns, morphologies, and/or experimental conditions, using training data comprising single cell images and corresponding labels obtained at step 219. Note

10 that the first machine learning model may include a plurality of machine learning models each trained for a respective task, and each producing a respective image embedding. These respective image embeddings may be concatenated or otherwise summarised to obtain the embeddings used as input to the second machine learning model.

15 **[0070]** As explained by reference to Figure 1, the predicted transcriptomic profile and the measured transcriptomic profile are single cell transcriptomic profiles or bulk transcriptomic profiles. When bulk transcriptomic profiles are used, the machine learning model may be configured to determine an image embedding for each of a plurality of single cell images, and predict a transcriptomic profile for the plurality of

20 cells in the plurality of single cell images using the plurality of image embeddings. This may be achieved by, for each training image, at optional step 220B, obtaining a summarised image embedding from a plurality of image embeddings, each corresponding to one of a plurality of single cell images obtained from the training image (e.g. an average of the plurality of embeddings corresponding to single cell

25 images). Thus, a summarised image embedding may be obtained as the average of a plurality of image embeddings. The summarised image embedding may then be used to train the second machine learning model at step 220C. Alternatively, single cell images may be processed at steps 220A and 220C, resulting in a plurality of predicted transcriptomic profiles (one for each single cell image), and a summarised

30 predicted transcriptomic profile may be obtained (e.g. an average of the plurality of predicted transcriptomic profiles) for comparison with a corresponding training transcriptomic profile obtained from bulk transcriptomic data.

[0071] At optional step 230, results of any one or more of steps 200 to 220, and/or any information derived therefrom are provided to a user, for example through a user interface. Results of the methods described herein may comprise one or more of: one or more predicted transcriptomic profile, the first and/or second trained machine learning models, the values of parameters (e.g. architecture and weights) of a trained machine learning model. Information derived from the results of the method may comprise information associated with the training, such as e.g. prediction accuracy for the first and/or second models.

Applications

10 [0072] The methods described herein find use in any context where transcriptomics information about cells can be used to characterise the cells, and particularly in any such context where it is beneficial to characterise the transcriptome of the cells in a non-destructive manner, e.g. while the cells are alive / maintained in culture.

15 [0073] In particular, the methods described herein find use in the context of producing cells that have a predetermined characteristic that can be defined in terms of transcriptomic characteristics (e.g. suitability for a particular downstream purpose), for perturbation screening including in particular drug screening, including for the purpose of characterising the effect of one or more perturbations on one or more transcriptomic characteristics of a cell population, for drug discovery and/or development.

[0074] Thus, also described herein are methods of characterising and/or comparing one or more perturbations applied to a population of cells, the method comprising: culturing a cell population in the presence of the one or more perturbations and/or obtaining one or more images of the cell population in culture in the presence of the one or more perturbations; determining a transcriptomic profile of the cell population or respective transcriptomic profiles of one or more cells thereof, using a method according to a method as described herein. The method may comprise determining one or more transcriptomic profile(s) of a cell population in culture in the presence of a first perturbation and one or more transcriptomic profile(s) of a cell population in culture in the presence of a second perturbation. The method may comprise comparing the one or more transcriptomic profile(s) determined for cells cultured in the presence of the first and second perturbations.

[0075] The methods find particular use in the context of longitudinal characterisation of cells, in the presence or absence of perturbations. Thus, also described herein are methods comprising: obtaining one or more images of a cell population in culture, optionally in the presence of one or more perturbations, at a first time point and at a second time point; and determining a transcriptomic profile of the cell population or respective transcriptomic profiles of one or more cells thereof, using a method according to a method as described herein, at the first time point and the second time point. The method may comprise comparing the one or more transcriptomic profile(s) determined for the cells at the first and second time points.

10 **[0076]** The above methods find applications in the context of characterising perturbations and/or samples based on transcriptomics information, including but not limited to the context of aging and rejuvenation treatment.

[0077] Thus, the present disclosure also provides methods of observing the effects of one or more test agents on aging or disease in human cells. These methods may comprise combining the test agent(s) with human cells (e.g. for specified period of time such as at least one day, one week or one month), and then determining transcriptomic profile of the human cells using the methods described herein. The determining may be performed after a plurality of specified time periods, such as e.g. after one day, two days, one week, etc. The transcriptomic profiles determined may be referred to as "observations". The methods may compare the observations from cells exposed to the test agent with observations obtained using the methods described herein for control cells not exposed to the test agent or a control or reference value such that effects of the test agent on aging or disease of the human cells is observed. Optionally, the test agent is a compound having a molecular weight less than 3,000, 2,000, 1,000 or 500 g/mol, an antibody, a polypeptide, a polynucleotide, an antisense compound (e.g., ASO, siRNA) or the like. The cells may be fibroblasts. The cells may be human cells.

[0078] Also described herein are methods of observing the effects of one or more test agents on aging or disease in human cells and comparing the effects of the one or more test agents to the effects of a different test agent, using the methods described here. For example, the methods of the disclosure can be used to compare effects of a rejuvenating agent, e.g., OSKM, with the effects of a test agent, to

determine whether the test agent has similar properties to the rejuvenating agent, e.g., OSKM.

5 [0079] Also described herein are methods of observing biomarkers in a sample, eg, skin sample, that correlate with an age of an individual, the method comprising observing the individual's chromatin status by determining the value of a biomarker metric as described herein. Because the biomarker metrics described herein are associated with the age of the individual, determining the chromatin state of a sample as described herein comprises observing biomarkers associated with the age of the individual. Typically, the skin and blood cells are human fibroblasts, keratinocytes, 10 buccal cells, endothelial cells, lymphoblastoid cells, and/or cells obtained from blood skin, dermis, epidermis or saliva. Embodiments of this method further comprise using the observations to estimate the age of the individual (e.g. using a regression analysis or the like, or by comparing the biomarkers to biomarker values obtained for control samples with known age).

15 [0080] Thus, the present disclosure provides methods for identifying agents that can treat a disorder such as ageing-related disorders, and methods of treating a subject who has been diagnosed as having an ageing-related disorder. An illustrative method will be described by reference to **Figure 3**.

[0081] At optional step 30, a sample comprising cells that have an ageing-related disorder is obtained (diseased sample). The sample may be obtained from a subject 20 who has an ageing-related disorder. Optionally, one or more samples comprising healthy cells may also be obtained (normal sample). At optional step 32, the cells in the diseased sample are exposed to one or more test agents. At optional step 34, light imaging data is obtained from a sample of the cells exposed to each test agent, 25 and optionally also from a diseased sample not exposed to any test agent and/or the normal sample. At step 36, a transcriptomic profile is determined using a method as described herein, such as e.g. by reference to Figure 1, for each sample. At step 38, the one or more test agents are compared based on the determined transcriptomic profiles For example, test agents that bring the transcriptomic profile for a diseased 30 sample treated with the test agent closer to the transcriptomic profile for a normal sample (compared to the initial distance between the transcriptomic profile obtained for a disease sample not treated with the agent) may be considered to be effective in treating the ageing related disorder. At optional step 40, the subject from whom the

sample of cells has been obtained may be treated with a test agent identified at step 38 as effective in treating the ageing related disorder.

Systems

[0082] **Figure 4** illustrates schematically an exemplary system according to the disclosure. The system comprises a computing device 1, which comprises a processor 101 and computer readable memory 102. In the embodiment shown, the computing device 1 also comprises a user interface 103, which is illustrated as a screen but may include any other means of conveying information to a user such as e.g. through audible or visual signals. In the illustrated embodiment, the computing device 1 is operably connected, such as e.g. through a network 6, to a cell culture system comprising a cell culture housing 2, one or more sensors 3, and one or more effectors 4. The cell culture housing may be an incubator or any other kind of housing suitable for live cell culture in a culture dish. The computing device may be a smartphone, tablet, personal computer or other computing device. The computing device is configured to implement a method of determining one or more transcriptomic characteristics and/or a method of providing a tool, as described herein. In alternative embodiments, the computing device 1 is configured to communicate with a remote computing device (not shown), which is itself configured to implement a method of determining one or more transcriptomic profiles and/or a method of providing a tool as described herein. In such cases, the remote computing device may also be configured to send the result of the method of determining one or more transcriptomic profiles and/or a method of providing a tool, to the computing device. Communication between the computing device 1 and the remote computing device may be through a wired or wireless connection, and may occur over a local or public network such as e.g. over the public internet. Each of the sensor(s) 3 and optional effector(s) 4 may be in wired connection with the computing device 1, or may be able to communicate through a wireless connection, such as e.g. through WiFi, as illustrated. The connection between the computing device 1 and the effector(s) 4 and sensor(s) may be direct or indirect (such as e.g. through a remote computer). In alternative embodiments, the computing device 1 is configured to implement a method of determining a transcriptomic characteristic or monitoring or controlling a cell culture, as described herein, using images received from a data store or remote computing device (such as e.g. a computing device associated with the cell culture

system). Thus, the computing device 1 may not be directly connected to the cell culture system. In such embodiments, the computing device 1 may provide results of the methods as described herein to a remote computing device or data store. When the results are provided to a remote computing device directly or indirectly associated with the cell culture system, the results may be used by the remote computing device to implement a control action, for example by determining a control action to be implemented based on the results and/or to control the one or more effectors 4 to implement the control action. Similarly, when the computing device 1 is connected to the cell culture system, the computing device 1 may be configured to determine a control action based on the results of a method of monitoring a cell population in a cell culture, and to control one or more effectors 4 to implement the control action. The one or more sensors 3 comprise at least one sensor configured to acquire live images of one or more cell population(s) in the cell culture housing (such as e.g. a phase contrast microscope or bright-field microscope). The sensors 3 may further comprise at least one sensor configured to acquire transcriptomics data from the cell population(s) in the cell culture housing (such as e.g. a MERfish system, or sequence data acquisition means). The one or more effectors 4 may be configured to control one or more process parameters of the cell culture process being performed in the cell culture housing 2. The measurements from the sensors 3 are communicated to the computing device 1, which may store the data permanently or temporarily in memory 102. The computing device memory 102 may store a trained machine learning model as described herein. The processor 101 may execute instructions to predict one or more transcriptomic profiles using the trained machine learning model, and/or to provide a tool for predicting one or more transcriptomic profiles as described herein, using the data from the one or more sensors 3, as described herein (such as e.g. by reference to Figure 1 or Figure 2).

[0083] The following is presented by way of example and is not to be construed as a limitation to the scope of the claims.

Examples

[0084] Omic measurements provide a holistic and unbiased view of entire classes of cellular measurements, but these technologies are almost entirely destructive. In contrast, brightfield imaging offers non-perturbational observations that have historically been considered to have low information content in the absence of

accompanying labeling methods. In these examples, the inventors report an end-to-end pipeline that accurately infers perturbed bulk transcriptomes from single-cell unlabelled images. They show the generalizability of the method by measuring transcriptomes of exact replicate conditions, unseen perturbation dosages; and
 5 unseen perturbations. Finally, they show proof-of-concept “photomics” experiments that showcase the power of multi-modal learning.

Example 1 – Building an end-to-end cross-domain mapping framework.

Methods

[0085] *Cell culture.* BJ fibroblasts (ATCC ID CRL-2522, LOT#70027151) and
 10 U2OS (ATCC ID HTB-96) were grown in phenol red free growth media (Fluorobrite (Gibco #A1896701), 10% FBS (Omega Scientific International LLC #FB-01 Lot#229025), 1%NEAA (Gibco #11140050), 1% Glutamax #35050061) under standard conditions (37C, 5%CO₂) in multi-well glass bottom imaging plates (Cellvis #P121.5HN) coated with Poly-D-Lysine (Gibco #A3890401). The cells were exposed
 15 to a curated list of chemical perturbations as detailed in Table 1 below. The perturbations were chosen based on the diversity of molecular effects associated with these, the diversity and presence of observed phenotypes, and practical considerations such as ease of use and availability. All chemical dosages were established based on existing literature to ensure doses were within an appropriate
 20 working range. Cells were exposed to chemicals for 6 hrs prior to imaging.

Chemical	Dose 1	Dose 2
Methyl viologen	100uM	10uM
D-sorbitol	100mM	10mM
2-DG	10mM	1mM
tunicamycin	1ug/ml	100ng/ml
Rapamycin	1uM	0.1uM
Bafilomycin A	100nM	10nM
Actinomycin D	5uM	0.5uM
cycloheximide	100uM	10uM
NaCl	100mM	20mM
5'AZA	2mM	200uM
Control - PBS	100ul/ml	100ul/ml

Control - DMSO	3.125ul/ml	2.75ul/ml
----------------	------------	-----------

Table 1. List of perturbations used.

[0086] *Imaging.* Cells were imaged on a brightfield microscope using a 40x air objective. Kohler correction was applied, on epi-fluorescence mode, on a Brightfield channel. For each imaged location within each well, a z-stack of 7 images with 1µm spacing was obtained. Each well was imaged at 10x 10 locations with 1400µm spacing. The image corresponding to the focus plane of the stack was used for all computations. Therefore, the pipeline is compatible with both 2D (as used here) or 3D images. 2D images were found to be advantageous because they require less memory for processing than 3D images, thereby enabling the use of deeper models given memory constraints, as well as analysis of time lapse images (i.e. freeing memory for analysis of images with greater time resolution). Additionally, machine learning models designed to process 2D images can be pretrained on large image collections (e.g. pretrained for object recognition on large collections of images such as ImageNet), resulting in an improved performance of the model when fine tuned for the virtual staining task described below. The inventors found this to result in improved performance outweighing gains in performance that could be achieved from considering 3D information.

[0087] *RNA-seq.* Cells were collected in 12-well plates and lysed using Buffer RLT with beta-mercaptoethanol following the manufacturer's instructions. Samples were flash frozen and stored in -80C for extraction. RNA was extracted using the RNeasy Micro Kit (Qiagen). Extracted RNA quality was checked using RNA ScreenTape on the TapeStation (Agilent Technologies) and quantified via RNA Broad Range Qubit assay (Life Technologies). RNA-Seq libraries were generated using Illumina Stranded mRNA Prep, Ligation Kit. The purified libraries were quantified using DNA High Sensitivity Qubit assay (Life Technologies), and size distribution was checked using TapeStation D1000 ScreenTape (Agilent Technologies). Pooled samples were qPCR using KAPA Illumina Quantification kit (Roche) and sequenced on 2 lanes of the NovaSeq6000 S4, XP 200 cycle kit (Illumina). The run was sequenced pair-end 100 cycles, with 96 multiplexed samples per run. Sequences were demultiplexed using bclconvert (Illumina) and aligned using STAR-salmon via Nextflow (github.com/nf-core/rnaseq). The Nextflow pipeline (github.com/nf-core/rnaseq) was used to obtain transcript per million (TPM) data for each sample.

[0088] *RNA-seq data analysis.* RNA-seq analysis was performed using the TPM files output by the Nextflow pipeline. The data was filtered to only keep genes that had at least 300TPMs cumulatively across all conditions. TPMs were log-transformed using numpy's 'log1p' function. The 'StandardScaler' function from scikit-learn was
5 used to normalise the data prior to PCA transformation. PCA was performed on the normalised log transformed data using scikit-learn with default parameters. Only the first 10 principal components were kept.

[0089] *Machine learning.* Individual cells were identified within each field of view using a virtual staining model to produce a nucleus segmentation mask. The virtual
10 staining model was a model (in this case a V-net, although other deep neural network architectures such as a U-net are envisaged) trained to predict stained images from brightfield images. In this particular case the stained images were stained to visualise a particular cellular component (which could be any organelle, or subcellular structure, but in this case was the nuclei). In this particular example, the model was a
15 model pretrained for image recognition using ImageNet data. The model was trained in two steps, using training data comprising pairs of training brightfield images and corresponding stained images. In a first step the model was trained to predict stained cell images corresponding to training brightfield images using a first type of loss (e.g. a mean squared error (MSE), mean absolute error (MAE), or other measure of error
20 that is computed between pixel values) between the predicted and training stained cell images. In a second step the model trained in the first step was used as a starting point for retraining, where the model was trained to predict stained cell images corresponding to training brightfield images using a second type of loss. The second type of loss was computed based on semantic segmentations of the training and
25 predicted stained cell images. The second type of loss was based on texture-based and shape-based attributes of cellular structures (e.g. nuclei) depicted in the images. This was used to obtain a fine-tuned machine learning model that is capable of accurately reconstructing the shape and textures of the structures within cells depicted in the stained cell images. The shape-based loss is a loss (e.g. a cross-
30 entropy loss) associated with the difference between an image segmentation of a predicted image and the corresponding training stained image. Thus, by applying segmentation to ground truth and predicted stained images (e.g. using nuclei stains) and computing a loss that compares these segmentations, the model is trained to accurately capture the shape of the segmented features (e.g. nuclei). The texture-

based loss is a loss associated with the difference in texture between predicted and ground truth stained images, such as e.g. a structural similarity index measure (SSIM) or a perceptual loss that is computed between sets of features generated by a pre-trained feature extractor for the predicted and ground truth stained images.

5 **[0090]** Single cell patches centred on each identified nucleus were obtained by defining 224x224 pixel bounding boxes centred on the centroid of each segmented nucleus. The information in each single cell patch was embedded into a 512D representation using a ResNet-18 architecture tasked with predicting the cell line that
10 each patch, the embedding was extracted from the last layer, after adaptive pooling. The embedding dimension was selected based on available architectures, but larger embedding dimensions are expected to also be suitable and potentially advantageous.

15 **[0091]** To perform the multi-modal inference, for each well treated with one perturbation condition, the embeddings from all segmented cells were averaged to obtain an “average cell embedding”. These averaged vectors were used to fit a full linear model, a multi-task LASSO linear model, a random forest model, a fully connected neural network, and a neural network regularised using the lockout method. The linear models and the random forest model were implemented using
20 standard functions in scikit-learn. All code was implemented in Python, using the scikit-learn (scikit-learn.org), pytorch (pytorch.org), matplotlib (matplotlib.org) and numpy (numpy.org/) libraries. All models were trained by cross-validation (specifically, 5-fold or 4-fold cross-validation), resulting in an ensemble of models (a model trained on each fold of the cross-validation process). The prediction results shown in these
25 examples relate to the average across the ensemble of models obtained by 5-fold cross-validation.

30 **[0092]** *Functional enrichment analysis.* Enrichment of Gene Ontology (GO) terms in the genes that most contribute to each of the principal components was performed using gProfiler (Kolber et al. 2023), accessed through a scanpy wrapper function. This
uses a Fisher’s one-tailed test (cumulative hypergeometric probability) to obtain the p-value of measuring the intersection between a gene list and an ontology term. The gene list was selected as the top genes (top x genes, x being e.g. 10, 20, 30 40 or 50 genes) that contribute to PC loading, separately for the genes with positive and

negative weights to the loadings, using as background the set of expressed genes in the dataset (the background being used as a the population from which genes are sampled in the hypergeometric model).

Results

5 [0093] The problem of mapping cell images to transcriptomes was approached using a multi-step framework, illustrated on **Figure 5**. Briefly, each brightfield image was segmented to identify individual cells. Next, the images were embedded into a latent space. This latent space was used to predict the transcriptome. However, the present inventors recognised that it is unreasonable to try to predict the ~10,000
10 genes that are expressed in any one cell type. Therefore, they postulated that it is also important to reduce the dimensionality of the transcriptome space by embedding it in a lower-dimensional representation. This was performed using principal component analysis (PCA). Finally, the image embedding and the transcriptome embedding were aligned and trained such that the image embedding can predict the
15 latent transcriptome embedding. Following embedding, the whole transcriptome can be inferred by inverting the components of the PCA.

[0094] The brightfield images obtained contain 20-50 cells per field-of-view, and are largely empty. Thus, the inventors segmented the brightfield images into information-rich cell patches. Each patch was identified by virtually staining each
20 brightfield image using a V-net trained to predict nuclear segments. For both BJ fibroblasts and U2OS cells, the DICE score for this nuclear segmentation model was >0.9 – this score is enough to identify the locations that contain a nucleus in each image. The centroid of each segmented nucleus was used to define the centroid of a 224x224 pixel bounding box containing a cell patch. In total, 12,007 images were
25 segmented into 322,879 cell patches.

[0095] Each cell patch was embedded into a 512D space using a ResNet -18 image classifier trained to predict cell type (classification accuracy: 99.6%). In particular, an embedding for each bounding box image was obtained using the last layer of a Res-Net 18 classifier trained on cell-type, as shown on **Figure 7**.

30 [0096] A variety of models were then tested to regress the transcriptome PCAs on the embeddings from the image embedding model, including linear (e.g. simple linear regression model, LASSO, multi-task LASSO, elastic net) and non-linear (e.g.

random forest, neural networks) regression methods as well as regularized (e.g. LASSO, multi-task LASSO, Ridge) and non-regularized methods (e.g. simple linear regression, random forest, fully connected neural network). The data on **Figure 9** shows that non-linear regression methods outperformed linear regression methods (compare the regression methods and the neural network), and that regularization methods can improve performance even when using linear methods (compare linear regression and regularised linear regression approaches including LASSO, Ridge and Elastic net). Further, the data shows that L2-regularization outperforms L1-regularization (compare ridge and Lasso regression), suggesting that all 512-D of the embedding contain relevant information.

[0097] The data on **Figure 6** show that the end-to-end pipeline described above is able to predict about 70% of the variance associated with transcriptomes.

[0098] The data further shows that the image embeddings obtained from the ResNet model (512D embeddings) contain non-linear information predictive of transcriptomes, as evidenced by the observation that non-linear models outperform linear models. Indeed, this indicates that the embeddings contain non-linear information that can be leveraged to make better prediction, when a nonlinear photomics model is used.

[0099] The inventors further showed that the machine learning reconstructed transcriptomes (obtained by back-transforming the predictions from the model by reverse PCA) are qualitatively similar to real transcriptomes, as the mean and variance of many genes were consistent with expectations (data not shown).

[0100] **Figure 8** summarises the obtained for the best performing model (fully connected neural network). The table shows, for each of the PCs that were predicted, the top significantly enriched Gene Ontology (GO) term (separately for positive and negative genes) amongst the genes that contribute most to the PC loading, the percentage of variance explained by the PC, and the prediction performance obtained for the PC. The data shows that the prediction performance was extremely good for the first, third and 8th principal components (PC), and acceptable for the second and fourth PCs. The data further shows that PC1 reflects biological differences in extracellular matrix production. These are likely to result in morphological features that are identifiable in the images and picked up by the image embeddings. This at

least in part explains the very good prediction performance for this PC. PC3, a strong predictor of cycloheximide treatment, is associated with transcription factors that open RNA Pol 1 promoters, which suggests that the cell is up-regulating ribosomes. This likely corresponds to changes in nucleolar morphology that are identifiable in the images, underlining the strong prediction performance for this PC.

Example 2 – Validation by replication, interpolation and extrapolation

Methods

[0101] The methods described in Example 1 are being applied to additional data including replicates of the experiments above (U2OS cells exposed to the 10 chemicals in Table 1, 4 controls, 2 doses, total 24 samples), interpolated data (U2OS cells exposed to the 10 chemicals in Table 1, 4 controls, 4 new doses, total 44 samples), and extrapolated data (U2OS cells exposed to the new chemical treatments not listed in Table 1).

[0102] The replicate data is being used to show that the models trained on the data in Example 1 are able to make predictions for new data not collected at the same time. The interpolated data is being used to show that the models trained on the data in Example 1 are able to make predictions for new data obtained by applying similar or identical perturbations at different concentrations. The extrapolated data is being used to show that the models trained on the data in Example 1 are able to make predictions for new data including data about previously unseen perturbations. Thus, Each of these represent progressively more stringent tests of the generalisability of the models described herein. The approaches described herein are expected to be robust and able to provide usable predictions in each of these cases.

Example 3 – Discussion

[0103] In these examples, the inventors demonstrated that brightfield images contain sufficient information to quantitatively approximate transcriptomes. Certain classes of approximations are easier than others – for example, inference of transcriptomic state is easiest across cell types. Many cell types grown in culture can be readily distinguished by the human eye, and thus it is expected that inferring the mean transcriptomic differences between them is readily possible. A much more challenging task is prediction within-cell-type of quantitative changes due to arbitrary

genetic or chemical perturbations where morphological changes are not readily visible by the human eye. The examples above show that the methods described herein are able to do this.

[0104] We have systematically observed that we are able to predict certain genetic programmes considerably better than others. There is only a rough correlation between the variance explained by each PC and our ability to infer it. A likely explanation is that the PC space used is heavily influenced by the choice of chemicals in these experiments, and certain chemicals, like cycloheximide, elicit a particular and specific response distinct from the others, and which therefore explains less variation. However, some of these perturbations elicit a clear change in features that are observable in the imaging data, leading to a very good ability to predict PCs associated with these perturbations (despite the variance explained by these PCs being low).

[0105] For example, certain perturbations such as cycloheximide affect e.g. ribosomes, which lead to changes in the nucleolus. The nucleolus is a relatively prominent structure in the cells, which is easily observable in a brightfield setting. As a result, even if the “ribosome perturbations” explained a small amount of the transcriptional variance in the principal component space, principal components associated with these particular perturbed genetic programmes will be easier to predict than others (because they are linked to prominent changes in observable features such as nucleolar size and shape).

[0106] The examples above are not based on paired information between cell morphology and single-cell RNA-seq. However, the methods described are equally applicable to such paired information. Although we expect we could significantly improve our model by expanding the perturbation space, expanding the number of measurements and further tuning our models, to fully explore the limits of transcriptome prediction from brightfield imaging will certainly require paired information. Such information would theoretically become accessible using techniques such as MERFISH that perform in situ transcriptome quantification. However, these techniques frequently are unable to estimate more than 500-1000 genes, so further anchoring methods, such as (Regev method for spatial omics anchoring – see e.g. Biancalani et al. 2021) will be necessary to map the full transcriptome.

The full power of multimodal analyses will come from the ability to exploit a fundamentally new class of experiments that we term “photomics”, which ought to allow real-time or near-real time estimation of single-cell transcriptomes, freeing up laser channels for optogenetic control of cells. By enabling real-time estimation of cellular states, known and unknown, multimodal mapping of microscopy images to transcriptomes can allow implementation of adaptive algorithms to better understand how layered interventions can steer cell fate.

References

- [0107] Jin et al. 2023. Robotic data acquisition with deep learning enables cell image-based prediction of transcriptomic phenotypes. *Proc. Nat. Acad. Sci.* December 28, 2022. 120 (1) e2210283120
- [0108] Wakui et al. 2023. Predicting reprogramming-related gene expression from cell morphology in human induced pluripotent stem cells. *Mol Biol Cell.* 2023 May 1;34(5):ar45. Epub 2023 Mar 22. PMID: 36947171
- [0109] Valdes et al., 2021. Lockout: Sparse Regularization of Neural Networks. arXiv:2107.07160v1. Submitted on 15 Jul 2021
- [0110] Liis Kolberg, Uku Raudvere, Ivan Kuzmin, Priit Adler, Jaak Vilo, Hedi Peterson: g:Profiler—interoperable web service for functional enrichment analysis and gene identifier mapping (2023 update) *Nucleic Acids Research*, May 2023.
- [0111] Minsheng Hao, Jing Gong, Xin Zeng, Chiming Liu, Yucheng Guo, Xingyi Cheng, Taifeng Wang, Jianzhu Ma, Le Song, Xuegong Zhang. Large Scale Foundation Model on Single-cell Transcriptomics. *bioRxiv* 2023.05.29.542705; doi: doi.org/10.1101/2023.05.29.542705
- [0112] Haotian Cui, Chloe Wang, Hassaan Maan, Bo Wang. scGPT: Towards Building a Foundation Model for Single-Cell Multi-omics Using Generative AI. *bioRxiv* 2023.04.30.538439; doi: doi.org/10.1101/2023.04.30.538439
- [0113] Biancalani, T., Scalia, G., Buffoni, L. et al. Deep learning and alignment of spatially resolved single-cell transcriptomes with Tangram. *Nat Methods* 18, 1352–1362 (2021).

- [0114] Rivenson Y, Wang H, Wei Z, de Haan K, Zhang Y, Wu Y, Günaydin H, Zuckerman JE, Chong T, Sisk AE, Westbrook LM, Wallace WD, Ozcan A. Virtual histological staining of unlabelled tissue-autofluorescence images via deep learning. Nat Biomed Eng. 2019 Jun;3(6):466-477.
- 5 [0115] Guo SM, Yeh LH, Folkesson J, Ivanov IE, Krishnan AP, Keefe MG, Hashemi E, Shin D, Chhun BB, Cho NH, Leonetti MD, Han MH, Nowakowski TJ, Mehta SB.
- [0116] Revealing architectural order with quantitative label-free imaging and deep learning. Elife. 2020 Jul 27;9:e55502.
- 10 [0117] Zhou Wang, A. C. Bovik, H. R. Sheikh and E. P. Simoncelli, "Image quality assessment: from error visibility to structural similarity," in IEEE Transactions on Image Processing, vol. 13, no. 4, pp. 600-612, April 2004
- [0118] All references cited herein are incorporated herein by reference in their entirety and for all purposes to the same extent as if each individual publication or
- 15 patent or patent application was specifically and individually indicated to be incorporated by reference in its entirety.

CLAIMS

1. A computer-implemented method of determining a transcriptomic profile of a sample comprising one or more cells, the method comprising:
 - obtaining one or more single cell images of the sample, wherein the images
 - 5 have been obtained using a live cell imaging technology;
 - predicting a transcriptomic profile for one or more cells of the sample using a machine learning model configured to take as input one or more single cell images and produce as output a predicted transcriptomic profile for the one or more cells in the single cell images,
 - 10 wherein the machine learning model has been trained using training data comprising for each of a plurality of samples, a plurality of training single cell images and corresponding measured transcriptomic profiles, wherein a measured transcriptomic profile comprises a set of values each value being indicative of the level of expression of a respective transcript, and a predicted transcriptomic profile
 - 15 comprises a plurality of values that are each indicative of the expression level of a plurality of transcripts.
2. The computer-implemented method according to claim 1, wherein the machine learning model comprises:
 - 20 a first machine learning model configured to take as input a single cell image and produce as output an image embedding; and
 - a second machine learning model configured to take as input an image embedding and produce as output a predicted transcriptomic profile.
- 25 3. The computer-implemented method according to claim 2, wherein the first and second machine learning models have been trained independently.
4. The computer-implemented method according to claim 2 or claim 3, wherein the first machine learning model comprises a machine learning model that has been
- 30 trained in a supervised manner using training data comprising a plurality of single cell images and associated labels, wherein the labels are not derived from transcriptomic data.

5. The computer-implemented method according to any of claims 2 to 4, wherein the first machine learning model comprises a machine learning models that has been trained as a part of a machine learning model configured to classify single cell images between a plurality of classes, the plurality of classes associated with respective cell
5 types, cell cycle phases, staining patterns, morphologies, and/or experimental conditions, or a plurality of said machine learning models, the outputs of which are concatenated.

6. The computer-implemented method according to any preceding claim, wherein the
10 predicted transcriptomic profile and the measured transcriptomic profile are single cell transcriptomic profiles or bulk transcriptomic profiles.

7. The computer-implemented method according to any preceding claim, wherein the machine learning model is configured to determine an image embedding for each of a
15 plurality of single cell images, and predict a transcriptomic profile for the plurality of cells in the plurality of single cell images using the plurality of image embeddings, optionally wherein the machine learning model is configured to obtain a summarised image embedding from the plurality of image embeddings, and predict a transcriptomic profile for the plurality of cells in the plurality of single cell images
20 using the summarised image embedding.

8. The computer-implemented method according to any preceding claim, wherein obtaining the one or more single cell images of the sample comprises obtaining one or more images of the sample, each image comprising one or more cells, and
25 obtaining one or more smaller segments of the one or more images each comprising a single cell.

9. The computer-implemented method according to claim 8, wherein obtaining one or more smaller segments of the one or more images each comprising a single cell
30 comprises analysing the one or more images using an image analysis algorithm configured to identify regions of an image that contain a whole cell or a part thereof that is unique to a single cell.

10. The computer-implemented method according to claim 9, wherein obtaining one or more smaller segments of the one or more images each comprising a single cell comprises performing cell segmentation or nuclear segmentation to identify regions of an image that contain individual cells or nuclei, and/or wherein the image analysis
5 algorithm is a machine learning model that has been trained to predict a stained image from a label free image using training comprising pairs of label free images and corresponding stained images, wherein the stained images comprise signal associated with a label specific to a particular subcellular structure, and/or wherein obtaining one or more smaller segments of the one or more images comprises
10 obtaining a smaller segment of the one or more images for each of at least one of one or more identified regions of the image, optionally wherein each smaller segment has a predetermined size and is centred on the centre of mass of an identified region.

11. The computer-implemented method according to any preceding claim, wherein
15 the live cell imaging technology is label free or uses one or more labels that are not transcriptomic labels, and/or wherein the one or more single cell images are 2D images.

12. The computer-implemented method according to any preceding claim, wherein
20 the live cell imaging technology is selected from: optical microscopy and live fluorescence imaging, or wherein the live cell imaging technology is brightfield microscopy, phase contrast microscopy, differential interference contrast microscopy, quantitative label-free imaging with phase and polarization (QLIPP), labelled fluorescence imaging or autofluorescence imaging.

25 13. The computer-implemented method according to any preceding claim, further comprising determining a value for each of the transcripts of the plurality of transcripts from the predicted transcriptomic profile.

14. The computer-implemented method according to any preceding claim, wherein
30 the machine learning model has been trained to predict training predicted transcriptomic profiles derived from measured transcriptomic profiles corresponding to the plurality of training single cell images, wherein the training predicted transcriptomic profiles have been derived from the measured transcriptomic profiles using an embedding method, optionally wherein the method comprises applying a

reverse of the embedding method to a predicted transcriptomic profile provided as output of the machine learning model, thereby obtaining set of values each value being indicative of the level of expression of a respective transcript present in the measured transcriptomic profiles.

- 5 15. The computer-implemented method according to claim 14, wherein the training predicted transcriptomic profiles have been derived from the measured transcriptomic profiles using an embedding method, and wherein the method comprises applying a reverse embedding to a predicted transcriptomic profile provided as output of the machine learning model.
- 10 16. The computer-implemented method according to claim 15, wherein the embedding method is principal component analysis, and the reverse embedding method is reverse PCA.
17. The computer-implemented method according to claim 16, wherein a training predicted transcriptomic profile comprises the values of the first n principal
- 15 components for a measured transcriptomic profile, wherein n is between 5 and 15, between 8 and 12, or 10.
18. The computer-implemented method according to any preceding claim, wherein the machine learning model has been trained using training data comprising training single cell images and corresponding measured transcriptomics profiles for each of a
- 20 plurality of samples exposed to different perturbations, optionally wherein a perturbation is a physicochemical condition that applies a stress to the cells.
19. The computer-implemented method according to any of claims 2 to 17, wherein the first machine learning model comprises a deep neural network, optionally a convolutional neural network, and/or wherein the second machine learning model
- 25 comprises a regression model, optionally a non-linear regression model and/or a regularised regression model.
20. The computer-implemented method according to any of claims 2 to 18, wherein the second machine learning model is a non-linear and/or regularised regression model, such as a Ridge regression model, a LASSO regression model, an elastic net
- 30 model, a neural network or a regularised neural network.

21. A computer-implemented method of providing a tool for determining a transcriptomic profile of a sample comprising one or more cells; the method comprising:

5 obtaining training data comprising for each of a plurality of samples, a plurality of training single cell images and corresponding measured transcriptomic profiles, wherein a measured transcriptomic profile comprises a set of values each value being indicative of the level of expression of a respective transcript, and

10 training a machine learning model to take as input one or more single cell images and produce as output a predicted transcriptomic profile for the one or more cells in the single cell images, wherein a predicted transcriptomic profile comprises a plurality of values that are each indicative of the expression level of a plurality of transcripts.

22. A method of monitoring a cell population in cell culture, the method comprising: optionally culturing a cell population and/or obtaining one or more images of the cell
15 population during the cell culture process; and determining a transcriptomic profile of the cell population or one or more cells thereof, using a method according to any of claims 1 to 20.

23. A method of characterising and/or comparing one or more perturbations applied to one or more cells, the method comprising:

20 culturing one or more cells in the presence of the one or more perturbations and/or obtaining one or more images of the one or more cells in culture in the presence of the one or more perturbations;

determining a transcriptomic profile of the one or more cells using a method according to any of claims 1 to 20; and

25 optionally comparing the one or more transcriptomic profile(s) determined for cells cultured in the presence of a first and second perturbation.

24. A system comprising: at least one processor; and at least one non-transitory computer readable medium storing instructions that, when executed by the at least one processor, cause the at least one processor to perform the method of any of
30 claims 1 to 23.

25. A non-transitory computer medium or media comprising instructions that, when executed by at least one processor, cause the at least one processor to perform the method of any of claims 1 to 23.