

Single Cell Screening Method

Cell therapy, the treatment of patients with severe diseases using cells expanded and potentially modified in the laboratory, is a very promising technology that is delivering
5 excellent results in oncology (both for leukemias and solid tumors) (Ottaviano et al. 2022) and regenerative medicine (for instance in the recovery of burn victims) (Cossu et al. 2018). Cells can be safely derived from patients or donors, manipulated to expand, differentiate and potentially correct genetic defects or enhance their therapeutic capabilities, and then administered back to the patient. Unfortunately, though, the
10 production of these advanced therapy medicinal products (ATMPs) is still very challenging. One of the major limitations is the lack of access to therapeutically useful amounts of cells.

The promise of stem cells for regenerative medicine consists in the unprecedented
15 possibility to derive any type of cell from a human pluripotent stem cell such as an induced pluripotent stem cell (iPSC) or a human embryonic stem cell (hESCs). These pluripotent stem cells have the potential to generate *in vitro* any human tissue, since they are part – or very similar to - the embryonic structures that give rise to all primordial structures in our body. The main challenge, however, is directing these stem cells in
20 robust and scalable ways to develop into specific cell types useful for therapy. The expansion and/or differentiation of clinically useful cell products from donor-derived stem cells requires costly and time-consuming testing and optimisation of cell culture conditions that are tailored to the specific application and cell type of interest. Cell culture conditions must be formulated by carefully considering the developmental cues in the human
25 embryo, many of which are imperfectly understood. In any case, these may not be effective in the absence of the physiological niche which provides additional stimuli that are difficult to model in cell culture, e.g. specific cell-cell interactions, extracellular matrix composition and tissue stiffness. There is no *a priori* method to determine which of these conditions will yield a highly homogeneous, correctly differentiated cell population, with
30 high reproducibility and in a process that is scalable and has the potential to meet regulatory requirements.

Cell reprogramming, the process of converting one cell type to another cell type through manipulation of gene regulatory networks that define cell identity, may have great promise for regenerative medicine but has yet to be applied broadly. Although it is now well known
35 that it is possible to switch the phenotype of one cell type to another by manipulating genetic circuits, the elements required for cell fate conversion are difficult to identify and in

most instances unknown. The identification of factors to directly reprogram the identity of cell types is currently limited by, amongst other things, the cost of exhaustive experimental testing of plausible sets of factors, an approach that is inefficient and unscalable.

- 5 One of the most important examples of cell reprogramming is the conversion of a somatic cell to a pluripotent state using exogenous transcription factors (TFs). For example, Yamanaka (Yamanaka S. Cell 2006; 126:663-76; PMID:16904174) demonstrated that a small set of TFs, namely OCT4, KLF4, SOX2 and MYC (OKSM), when introduced into the cell can reprogram a human fibroblast to an induced pluripotent stem cell (iPSC).
- 10 Subsequently, other groups have demonstrated that it is possible to convert fibroblasts to hepatocytes, to cardiomyocytes and to several other cell types (Huang P, et al.. Nature 2011; 475:386-9; Sekiya S, Suzuki A., Nature 2011; 475:390-3; Kogiso T, et al., Hepatol Int 2013; 7:937-44; Ieda M, et al., Cell 2010; 142:375-86; Song K, et al., Nature 2012; 485:599-604; Qian L, et al., Nature 2012; 485:593-8; Takahashi K, Yamanaka S., Nat Rev
- 15 Mol Cell Biol 2016; 17:183- 93; Sadahiro T, et al., Circ. Res. 2015; 116:1378-91. Tsunemoto RK, et al., EMBO J 2015; 34:1445-55; Heinaniemi M, et al., Nat Methods 2013; 10:577-83; Lang AH, et al., PLoS Comput Biol 2014; 10:e1003734; Del Sol ICA, Stem Cells 2013; 31:2127-35; Davis FP, Eddy SR. PLoS One 2013; 8:1-8; D'Alessio et al., Stem Cell Reports 2015; 5:763-75).
- 20 In addition to reprogramming cells to obtain a pluripotent cell (iPSC), a number of related reprogramming strategies have been demonstrated, such as: (i) 'forward programming' where a less differentiated cell such as an iPSC is reprogrammed into a specific cell type (Dalby et al, Stem Cell Reports 11:1462-1478) (ii) 'direct conversion' where one
- 25 differentiated cell type is switched into another without going through a pluripotent or progenitor state (Zhou et al, Nature 455:627-632); as well as combinations of the foregoing such as (iii) 'indirect (or primed) reprogramming' where ectopic OKSM is expressed transiently to effect partial reprogramming to a less differentiated (but not fully pluripotent) intermediate, followed by forward programming to obtain a more differentiated cell type that differs from the starting cell type (Efe et al, Nat Cell Biol 13:215-222).
- 30 Note that '*cell reprogramming*' differs qualitatively from traditional '*cell differentiation*', both in biological mechanism and by the practical method by which it is achieved. Cell differentiation occurs by the progressive restriction of developmental potential until a terminal state is reached - famously illustrated by Waddington who imagined a ball rolling from the top of a mountain towards a lower energy resting place in the valley below (*The*

Strategy of the Genes, pub. Allen & Unwin, London, 1957) - whereas in cell reprogramming this is not necessarily true (and in the case of iPSCs it is precisely the opposite). In cell differentiation methods, (stem) cells in culture are exposed to developmental cues such as morphogens, growth factors, hormones, small molecules etc. which are extracellular effector molecules added to cell culture media and typically function by binding protein elements of the cell signalling apparatus such as cell surface receptors, and cytoplasmic signalling mechanisms such as those mediated by kinase cascades etc. In the case of cell reprogramming, the cell fate switch is directed by ectopic expression of transcription factors (TFs) which translocate to the cell nucleus and directly control gene expression, typically by binding gene regulatory elements in enhancer and promoter DNA and stimulating (or inhibiting) the transcription of mRNA. These two mechanistically different strategies for manipulating cell fate -- the first of which is mediated by controlling the cellular environment (i.e. directed differentiation) and the second of which is mediated by rewiring genetic circuits that determine cell fate (cell reprogramming) -- have also been termed the "outside-in" and "inside-out" approaches, respectively (Shakiba et al, Cell Syst. 12:561-592).

The TFs that orchestrate cell differentiation or reprogramming may themselves regulate one or more TFs amongst other genes, inducing and/or stabilising a Gene Regulatory Network (GRN) that specifies cell fate. One of the best understood GRNs is the pluripotency gene regulatory network (PGRN) that is induced and stabilised by adding the Yamanaka factors OKSM to a (differentiated) cell. At the core of the PGRN is a triumvirate of pluripotency-inducing TFs, namely OCT4, SOX2 and NANOG. These TFs regulate a larger network of interconnected secondary genes. The core TFs do so in a cooperative manner, and form autoregulatory and feed-forward gene circuits that result in network stability. It is known that NANOG and LIN28 can replace KLF4 and MYC in the Yamanaka cocktail, indicating a specific GRN can be induced by alternative means. Indeed, it is known that reprogramming can also be affected by non-coding RNAs that regulate gene expression, such as LincRoR and Let7 that affect key nodes in the PGRN, therefore in the present application non-coding RNAs that regulate gene expression are included in the definition of TFs.

GRNs are essentially composed of genomic components such as genes and their cis-regulatory modules at the nodes, and of regulatory state components, i.e. TFs (including non-coding RNAs) which provide regulatory input into these modules. The connections between different regulatory gene products (i.e. TFs) and the cognate cis-regulatory elements controlling other network nodes, specify a network circuitry that can explain

developmental processes such as the maintenance of pluripotency in the blastocyst, or e.g. the differentiation of a pluripotent stem cell into a pancreatic beta cell.

Few if any GRNs are as well defined as the PGRN, and also the TFs that effect specific cell reprogramming transitions are largely unknown. Recently, a number of computational methods for predicting TFs useful in promoting reprogramming from one cell type to another have been developed (reviewed by Kamaraj et al (Cell Cycle 15: 3343–3354). These methods typically consider differential gene expression between starting and target cell types and use knowledge of GRN architecture to rank TFs in order of influence in establishing the core GRN of the target cell. The availability of such methods to evaluate TFs *a priori* for their ability to influence cell fate is an important advantage of the “inside-out” approach, in contrast to the “outside-in” approach where currently there is no broadly applicable strategy to predict the instructive components of cell culture media that direct cell differentiation.

Although computational methods such as those referenced above have been used to predict OKSM can induce the PGRN, and also TFs capable of inducing a relatively small number of other cell fate transitions, the algorithms (and the input data they require) are far from perfect and frequently result in TF shortlists with little or no overlap. Consequently, the literature shows cell reprogramming is difficult to achieve, generally of low efficiency and/or reproducibility when at all successful. Experimental determination of useful transcription factor cocktails by methods used in the prior art is laborious and suffers from two related, major limitations.

Firstly, the conventional method of determining reprogramming cocktails is to screen candidate TFs in a large pool (or a set of smaller pools) and the minimal core TFs then identified through a process of eliminating each TF in turn, as is exemplified in Takahashi and Yamanaka’s (Cell 126:663-676) seminal experiment in which pools of 24 TFs were screened in a cell-based assay to identify OKSM in reprogramming fibroblasts to iPSCs. Yamanaka used mouse fibroblasts in cell culture, observing mass effects on a large cell sample.

In a further demonstration of this stratagem, Zhou et al (Nature 455:627-633) introduced a pool of 9 TFs into mouse pancreas *in vivo*, and observed pancreatic exocrine cells were reprogrammed into insulin-expressing beta cells. In this case, Zhou et al were able to deduce, again by a process of elimination, that the effective reprogramming cocktail comprises PDX1, NGN3 and MAFA. In both examples, these pooled screening experiments were unable to discover functional combinations of TFs beyond the minimal

core cocktail, since the presence of the core TFs in a pool will mask the effect of eliminating other productive TF cocktails from the pool. Moreover, in neither experiment was it possible to follow the effects of individual TFs at the cellular level.

This limitation could be overcome if a set of TFs could be screened systematically to evaluate all combinations in separate experiments. However, the number of experiments
5 needed is relatively large since there are 10,626 different combinations of four TFs in Yamanaka's set of 24 candidate TFs.

In addition, the number of cells required to perform such a large screen would also be large, since the experiment should be performed in replicates and the efficiency of reprogramming e.g. human cells with OKSM was only 0.01% - 0.02% (Takahashi et al,
10 Cell 131:861–872). There is therefore a need for scalable screening methods capable of systematically evaluating individual combinations of TFs, since "large scale screening of combinatorial TF cocktails remains challenging" (Li and Hon, Frontiers in Bioengineering and Biotechnology (2021) vol 9 article 748942).

Secondly, the pooled screening approach does not allow the discovery of optimal protocols in which the sequence of TF addition is important for certain types of reprogramming. Recent advances in single cell -omics technologies (e.g. RNA-Seq, ATAC-Seq) have allowed us to study cells during reprogramming at unprecedented resolution. These techniques have revealed cell reprogramming is a dynamic process, in
20 which cells transition through a continuum of cell states characterised by different patterns of TF expression that accompany GRN remodelling. For example, optimal reprogramming of pancreatic progenitor cells to insulin expressing beta cells using TFs PDX1, NGN3 and MAFA is only achieved by differential expression dynamics (Saxena et al, Nature communications 7 Article number 11247). There is therefore a need for screening
25 methods capable of interrogating combinations of TFs expressed in sequences, in order to discover optimal cell reprogramming methods.

In our 2004 patent application WO2004/031369, we described a method for determining the factors required to differentiate pluripotent cells into more committed cell types utilising a splitting and pooling culture approach, together with labelling of multicell units in
30 the form of beads, wherein the labels identify the conditions to which the beads have been exposed.

In the present invention, we provide an improvement of our 2004 process, in which cells are optimally converted from one cell type to another. We combine single-cell analysis of

the effects of reagents including TFs on cellular conversion between cell types with flexibility in addition of reagents to cells in order to better understand the genetic pathways leading to cell type conversion.

5 Summary of the Invention

In a first aspect, the invention provides a method for determining the conditions required for conversion of a first cell type to a second cell type, by analysing one or more markers in a single cell which has been exposed to combinations of reagents.

Markers of gene expression which relate to cellular differentiation can be used to follow the differentiation pathway of a cell. Many markers associated with gene expression are transcription factors (TFs), but there are many markers which are not, for example CD antigens, glycoproteins such as synaptophysin (Smith et al., Clin Neuropathol. 1993 Nov-Dec;12(6):335-42), Beta-Tubulin (Korzhevskii, et al., *Neurosci Behav Physi* **42**, 215-222 (2012)), Desmin (Capetanaki et al., Cell Struct Funct. 1997 Feb;22(1):103-16) and the like. Thus, the marker analysed in the cells may be a TF or may be a marker other than a TF. Patterns of expression of differentiation markers may be resolved to identify GRNs involved in cell type conversion.

In one embodiment, the invention provides a method comprising the steps of:

- a. exposing a single cell of a first cell type to a first reagent or combination of reagents;
- b. optionally removing the first reagent and exposing the cell to a different reagent or combination of reagents;
- c. optionally, repeating the foregoing step (b); and
- d. identifying one or more cells which displays one or more markers of a second cell type and determining the identity and sequence of reagents to which the cell has been exposed.

In embodiments, the cells are labelled to identify exposure to each reagent or combination of reagents. Labelling of cells is advantageously by means of genetic barcoding, such as retroviral infection and/or genome editing.

The split/pool techniques described in WO2004/031369 are adaptable to single cell screening of the effects of the exposure of cells to multiple individual reagents and

combinations of reagents. Cell units in WO2004/031369 are envisaged therein as beads comprising multiple cells which can be sorted together; in the present invention, single cells are used and individually sorted.

In embodiments, the invention provides a method comprising the steps of:

- 5 a. Subdividing a plurality of single cells into a first set of groups of single cells, separately exposing the groups of single cells to different reagents or combinations of reagents and labelling the single cells to indicate exposure to said reagents or combinations of reagents;
- 10 b. pooling the groups of single cells, and subdividing the pool of cells to form a second set of groups of single cells;
- c. separately exposing the second set of groups of single cells to different reagents or combinations of reagents;
- d. optionally iteratively repeating steps (a) to (c) as required;
- 15 e. identifying cell type markers in the cells and deconvoluting the labels to identify the identity, sequence and timing of reagents to which the cells have been exposed;
- f. optionally, using the data from step (e) to inform the choice of reagents used in step (a), and repeating steps (a) to (e) as required;
- 20 g. identifying a cell of the second cell type from the cell type markers, and deconvoluting the labels to thereby identify the identity and sequence of reagents required to reprogram a cell of the first cell type into a cell of the second cell type.

The present invention employs the approach described in WO2004031369 (see Fig. 1). In that approach, known as CombiCult®, a serial splitting and pooling of cells is used to
25 expose multiple cells to many different reagent combinations. Sequential method steps in a procedure according to the foregoing aspects of the invention comprise either the addition of single reagents or the addition of combinations of reagents.

In some embodiments, the sequential method steps in a procedure only comprise the addition of individual reagents.

30 The reagents or combinations of reagents used in the first iteration of step (a) in the methods set forth above can be selected from reagents known in the art to promote cell reprogramming. In the prior art, this is the only available approach, and the literature

comprises a number of publications which will guide the skilled person to the selection of apparently suitable reagents.

In an alternative embodiment in accordance with the invention, the reagents or combinations of reagents used in the first iteration of step (c) are selected from reagents
5 known in the art to promote cell type conversion.

With reference to the method set forth above in which data from the selection procedure is analysed, the reagents or combinations of reagents used in second and further iterations of steps (a) and (c) are selected by analysis of the data produced by step (e).

Thus, in one embodiment, the reagents used in step (a) and/or step (c) can be chosen on
10 the basis of computational analysis of known cell conversion events. The data used for analysis can be based on previous performance of the method of the invention, or derived from the prior art. The computational selection is combined with the experimental selection provided by the split/pool selection methodology to permit unprecedented precision in determining both the identity of the factors required for cell conversion, and
15 the ideal timing of their administration.

In one embodiment, the data produced in step (e) are processed by an algorithm which analyses the effect of the identity and timing of administration of each reagent on the reprogramming of the first cell type.

For example, the reagents or combinations of reagents used in the first iteration of step
20 (a) can be selected using data produced by step(e) in a previous performance of the method.

Labelling of cells can be performed in a variety of ways, for example by modifying the nucleic acid of each cell, which allows its exposure to reagents to be tracked precisely. Nucleic acid can be modified, for example, by retrovirus integration or genome editing
25 techniques.

In embodiments, genome editing comprises CRISPR.

Re-use of reagents at different times during the split-pool sequences allows the effect of a reagent at different times during the development of a cell to be assessed. For example, a pool may be exposed to a combination of four reagents in a first culture, and then split.
30 Different splits of the pool may be exposed to different combinations of reagents, which may or may not comprise any of the reagents used in the first culture. After pooling and

resplitting, further different splits may be exposed to yet further combinations of reagents which may comprise reagents from the first culture or the second culture, as well as novel reagents. Thus, cells in different splits are exposed to different reagents, as well as to the same reagents at different times.

- 5 In some embodiments, the reagents are added individually to cells in sequential method steps. For example, sequential method steps in a procedure may comprise either the addition of single reagents or the addition of combinations of reagents; alternatively sequential method steps in a procedure only comprise the addition of individual reagents. Thus, a procedure may combine method steps which add groups of reagents with steps
10 which add individual reagents, or may only add individual reagents.

A "procedure", as used in this context, represents the performance of an entire method as represented by steps (a) to (f) above.

- Although repetitive cycles of splitting and pooling may be used highly efficiently, in a similar manner to combinatorial chemistry protocols, given the necessary processing
15 power protocols involving at least two sequential splitting steps (without re-pooling) may be used. The disadvantage of such protocols is that they can quickly generate a very large number of separate samples, which have been handled differently. The advantage, however, is that each sample does not require laborious deconvolution, since the cells therein have only been exposed to one set of conditions. Accordingly, given suitable
20 sample handling facilities, a splitting approach can yield rapid results.

Another advantage of the present invention is that reagents may be added sequentially and individually, or in defined groups, in a large number of combinations, as explained above. This prevents the masking of the effect of a reagent by dominant reagents which may be present in pooled combinations of reagents used in conventional experiments.

- 25 In a preferred embodiment, the reagents are individually added sequentially to cells, such that the effect of each reagent can be assessed in isolation.

The method of the invention allows thousands or millions of reagents and combinations of reagents to be tested, in a multiplexed high-throughput assay, to determine the conditions necessary to achieve the desired result with respect to any cellular process.

- 30 In contrast to methods in which multiple reagents are added together, and then individually removed to identify the effective reagents, the method of the invention allows

testing of sequential addition of reagents and therefore the determination of the action of each reagents, when added in a plurality of sequential combinations, as well as the testing of multiple combinations amongst groups of reagents. This permits the elucidation of the separate steps in a pathway of conversion from one cell type to a second cell type,
5 rather than merely the overall conversion.

Advantageously, the reagents used cause a change in the cellular process; these reagents are selected for the production of cells in which gene expression is analysed. Gene expression may conveniently be analysed using any comparative expression monitoring technology including PCR-based techniques such as RT-PCR, Serial Analysis
10 of Gene Expression (SAGE), RNA sequencing (RNA seq) including single-cell RNA seq, in-situ hybridisation including single-cell hybridisation, or array technology such as is widely available from suppliers such as Affymetrix. In another aspect, the invention provides a method for producing a nucleic acid which encodes a gene product which influences a cellular process, comprising identifying a gene as above, and producing at
15 least the coding region of said gene by nucleic acid synthesis or biological replication.

In a further aspect, there is provided a method for inducing a cellular process in a cell, comprising the steps of: a) identifying one or more genes which are differentially expressed in association with a cellular process in accordance with the invention; and b) modulating the expression of said one or more genes in the cell. The expression of the
20 genes in the cell can be modulated by, for example, transfecting or otherwise transferring the gene into the cell such that it is overexpressed in a transient or permanent manner. Alternatively, the expression of the endogenous gene may be altered, such as by targeted enhancer insertion or the administration of exogenous agents which cause an increase (e.g. gratuitous inducers) or decrease (e.g. antisense, RNAi, transcription factors) in the
25 expression of the gene. Moreover, the product of the gene may itself be administered to or introduced into the cell to achieve an increase in its activity. Furthermore, agents which increase or decrease the activity of the gene product (e.g. competitive and non-competitive inhibitors, drugs, pharmaceuticals) can be administered to the cell. In a still further aspect, the invention provides a method for identifying the state of a cellular
30 process in a cell, comprising the steps of: a) identifying one or more genes which are differentially expressed in association with a cellular process as set forth above; and b) detecting the modulation of expression of said one or more genes in a cell, thereby determining the state of the cellular process in said cell. Advantageously, the genes employed in this analysis encode cellular markers, which may be detected for instance by

immunoassay. Alternatively, the gene products may be enzymes that can be assayed for activity with fluorometric, colorimetric, radiometric, or other methodologies.

The invention further provides a method for regulating a cellular process, comprising the steps of: a) determining the effect of one or more reagents on a cell, in accordance with
5 the foregoing aspect of the invention; b) exposing a cell to reagents which effect a change in the cellular process; and c) isolating the desired cell.

Accordingly, the invention provides for a method for producing a differentiated cell from a second, different differentiated cell type, in accordance with the foregoing aspect of the invention.

10 There is also provided a method for identifying a reagent which is capable of inducing a cellular process, comprising the steps of: a) determining the effect of one or more reagents on a cell, in accordance with the foregoing aspect of the invention; and b) identifying those reagent(s) which induce the desired cellular process in the cells. reagents identified in accordance with the invention may be expressed or synthesised by
15 conventional chemical, biochemical or other techniques, and used in methods for regulating particular cellular processes in cells for example as described herein.

In a further aspect, the invention relates to methods for developing and using a database of cell gene expression data correlated to exposure to reagents, which can provide improved reagent selection in step (a) of the method of first aspect of the invention. Using
20 machine learning and inputting experimental data from the method of the present invention, the database can be improved, further optimising reagent selection and therefore the performance of the method of the invention.

Brief Description of the Figures

Figure 1 is a schematic of the operation of the CombiCult® split/pool selection method.

25 Figure 2 is a schematic of genetic barcode labelling of cells in the method of the present invention.

Figure 3 is a schematic of the 27 well experimental design. The experiment tracks cell differentiation using fluorescent protein tag sequences without splitting/pooling (1 well out of 27). The protocol is derived from Feng et al. Stem Cell Reports 2014.

Figure 4 shows UMAP plots of cells isolated from control, GFP and Luciferase wells. The GFP wells are wells exposed to the final reagent with a GFP marker, and have thus been exposed to all sets of conditions. The Luciferase wells have not been exposed to any conditions. From these wells, BFP, RFP and GFP markers are screened for.

- 5 Figure 5 shows a similar UMAP plot as in Figure 4, except that tag presence is scored on as present or absent, rather than using a threshold level.

Figure 6 shows a similar UMAP plot as in Figure 4, in which RFP and BFP are co-detected, such that only the presence of both markers is scored.

- 10 Figures 7, 8 and 9 illustrate gene expression analysis and clustering of gene expression inn UMAP plots from the 27 well experiment.

Figure 10 illustrates the experimental design used in Examples 2 and 3.

Figure 11 is a UMAP plot of the results from the experiment of Example 2.

Figure 12 is a clustering analysis of the genes identified as expressed in the cells identified in Example 2.

- 15 Figure 13 is a UMAP plot of the results from the experiment of Example 3.

Figure 14 is a representation of the experimental design of Example 4.

Figures 15 and 16 are UMAP plots of the cells isolated from Example 4, and the presence of markers in those cells.

Detailed Description of the Invention

- 20 Unless defined otherwise, all technical and scientific terms used herein have the same meaning as commonly understood by those of ordinary skill in the art, such as in the arts of peptide chemistry, cell culture and phage display, nucleic acid chemistry and biochemistry. Standard techniques are used for molecular biology, genetic and biochemical methods (see Sambrook et al., Molecular Cloning: A Laboratory Manual, 3rd
25 ed., 2001, Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY; Ausubel et al., Short Protocols in Molecular Biology (1999) 4th ed., John Wiley & Sons, Inc.; Patil and Sivaram, A Complete Guide to Gene Cloning: From Basic to Advanced, Springer: ISBN978-3-030-96853-3, 28 April 2023; Jaroszewicz et al., Phage display and other

peptide display technologies, FEMS Microbiology Reviews, fuab052, 46, 2022, 1–25, as well as resources on Addgene.org), which are incorporated herein by reference.

Culture Conditions As used herein, the term "culture conditions" refers to the environment which cells are placed in or are exposed to in order to promote growth or differentiation of said cells. Thus, the term refers to the medium, temperature, atmospheric conditions, substrate, stirring conditions and the like which may affect the growth and/or differentiation of cells. More particularly, the term refers to specific reagents which may be incorporated into culture media and which may influence the growth and/or differentiation of cells. In a preferred embodiment, the culture conditions are the reagent or combination of reagents to which a cell is exposed.

Cell A cell, as referred to herein, is defined as the smallest structural unit of an organism that is capable of independent functioning, or a single-celled organism, consisting of one or more nuclei, cytoplasm, and various organelles, all surrounded by a semipermeable cell membrane or cell wall. The cell may be prokaryotic, eukaryotic or archaebacterial. For example, the cell may be a eukaryotic cell. Mammalian cells are preferred, especially human cells. Cells may be natural or modified, such as by genetic manipulation or passaging in culture, to achieve desired properties. A **stem cell** is defined in more detail below, and is a totipotent, pluripotent or multipotent cell capable of giving rise to more than one differentiated cell type. Stem cells may be differentiated *in vitro* to give rise to differentiated cells, which may themselves be multipotent, or may be terminally differentiated. Cells differentiated *in vitro* are cells which have been created artificially by exposing stem cells to one or more agents which promote cell differentiation.

Cellular process A cellular process is any characteristic, function, process, event, cause or effect, intracellular or extracellular, which occurs or is observed or which can be attributed to a cell. Examples of cellular processes include, but are not limited to, viability, senescence, death, pluripotency, morphology, signalling, binding, recognition, molecule production or destruction (degradation), mutation, protein folding, transcription, translation, catalysis, synaptic transmission, vesicular transport, organelle function, cell cycle, metabolism, proliferation, division, differentiation, phenotype, genotype, gene expression, or the control of these processes.

Single Cell A single cell, which may be present in a culture comprising a plurality of single cells, present as a pool of individual cells. Pools of cells may be sorted, subdivided and handled, and ultimately separated into single cells using any available technique. Examples include manual picking under a microscope, wherein individual cells are

manually picked using a micropipette; flow cytometry and cell sorting (FACS), wherein cells are suspended in a fluid and fluorescence is excited, for example using a laser beam. Detectors analyse the cells by size, shape, and fluorescence, and then droplets containing single cells are sorted and collected. In magnetic-activated cell sorting

- 5 (MACS), cells are labelled with magnetic beads that bind to specific cell surface markers. The cell suspension is then passed through a magnetic column that captures the labelled cells, which can then be eluted as single cells. Cells can also be sorted on microfluidic devices, by limiting dilution, and by automated robotic systems which physically separate cells.

- 10 **Totipotent** A totipotent cell is a cell with the potential to differentiate into any type of somatic or germ cell found in the organism. Thus, any desired cell may be derived, by some means, from a totipotent cell.

Pluripotent A pluripotent cell is a cell which may differentiate into more than one, but not all, cell types.

- 15 **Label** A label or **tag**, as used herein, is a means to identify a cell unit and/or determine a culture condition, or a sequence of culture conditions, to which the cell unit has been exposed. Preferably, a label is a genetic barcode, as described further herein.

- Exposure to culture conditions or reagents** A cell is exposed to culture conditions or reagents when it is placed in contact with a medium, or grown under conditions which
20 affect one or more cellular process(es) such as the growth, differentiation, or metabolic state of the cell. Thus, if the culture conditions comprise culturing the cell in a medium in the presence of a reagent, the cell is placed in the medium with a reagent present for a sufficient period of time for it to have an effect. Likewise, if the conditions are temperature conditions, the cells are cultured at the desired temperature.

- 25 **Pooling** The pooling of one or more cells or groups of cells involves the admixture of the cells or groups to create a single group or pool which comprises cells of more than one background, that is, that have been exposed to more than one different sets of culture conditions. A pool may be subdivided further into groups, either randomly or non-
30 themselves be pooled by combination, for example after exposure to different sets of culture conditions.

Proliferation Cell growth and cell proliferation are used interchangeably herein to denote multiplication of cell numbers without differentiation into different cell types or lineages. In other words, the terms denote increase of viable cell numbers. Preferably proliferation is not accompanied by appreciable changes in phenotype or genotype.

- 5 **Cell Conversion** As referred to herein, the conversion of a cell from a first cell type to a second cell type encompasses all means of cell type conversion, including differentiation, transdifferentiation, reprogramming and other definitions of cell type plasticity and modification.

- 10 **Differentiation** Cell differentiation is the development, from a cell type, of a different cell type. For example, a bipotent, pluripotent or totipotent cell may differentiate into a neural cell. Differentiation may be accompanied by proliferation, or may be independent thereof. The term 'differentiation' generally refers to the acquisition of a phenotype of a mature cell type from a less developmentally defined cell type, e.g. a neuron, or a lymphocyte, but does not preclude transdifferentiation, whereby one mature cell type may convert to
15 another mature cell type e.g. a neuron to a lymphocyte.

Differentiation state The differentiation state of a cell is the level to which a cell has differentiated along a particular pathway or lineage.

- State of a cellular process** The state of a cellular process refers to whether a cellular process is occurring or not and in complex cellular processes can denote a particular step
20 or stage in that cellular process. For example, a cellular differentiation pathway in a cell may be inactive or may have been induced and may comprise a number of discrete steps or components such as signalling events characterised by the presence of a characteristic set of enzymes or intermediates.

- Gene** A gene is a nucleic acid which encodes a gene product, be it a polypeptide or an
25 RNA gene product. As used herein, a gene includes at least the coding sequence which encodes the gene product; it may, optionally, include one or more regulatory regions necessary for the transcription and/or translation of the coding sequence.

- Gene Product** A gene product is typically a protein encoded by a gene in the conventional manner. However, the term also encompasses non-polypeptide gene
30 products, such as ribonucleic acids, which are encoded by the gene.

Nucleic acid synthesis Nucleic acids may be synthesised according to any available technique. Preferably, nucleic acid synthesis is automated. Moreover, nucleic acids may be produced by biological replication, such as by cloning and replication in bacterial or eukaryotic cells, according to procedures known in the art.

- 5 **Differential Expression** Genes which are expressed at different levels in response to cell culture conditions can be identified by gene expression analysis, such as on a gene array, by sequencing such as RNA-seq, or by any methods known in the art. Genes which are differentially expressed display a greater or lesser quantity of mRNA or gene product in the cell under the test conditions than under alternative conditions, relative to overall gene
10 expression levels.

Transfection Genes may be transfected into cells by any appropriate means. The term is used herein to signify conventional transfection, for example using calcium phosphate, but also to include other techniques for transferring nucleic acids into a cell, including transformation, viral transduction, electroporation and the like.

- 15 **Modulation** The term modulation is used to signify an increase and/or decrease in the parameter being modulated. Thus, modulation of gene expression includes both increasing gene expression and decreasing gene expression.

Stem Cells Stem cells are described in detail in *Stem Cells: Scientific Progress and Future Research Directions*. Department of Health and Human Services. June 2001.

- 20 <http://www.nih.gov/news/stemcell/scireport.htm>. The contents of the report are herein incorporated by reference.

- Stem cells are cells that are capable of differentiating to form at least one and sometimes many specialised or differentiated cell types. The repertoire of the different cells that can be formed from stem cells is thought to be exhaustive; that is to say it includes all the
25 different cell types that make up the organism. Stem cells are present throughout the lifetime of an organism, from the early embryo where they are relatively abundant, to the adult where they are relatively rare. Stem cells present in many tissues of adult animals are important in normal tissue repair and homeostasis. The existence of these cells has raised the possibility that they could provide a means of generating specialised functional
30 cells that can be transplanted into humans and replace dead or non-functioning cells in diseased tissues. The list of diseases for which this may provide therapies includes Parkinson's disease, diabetes, spinal cord injury, stroke, chronic heart disease, end-stage kidney disease, liver failure and cancer.

Different stem cells have differing potential to form various cell types: spermatogonial stem cells are unipotent as they naturally produce only spermatozoa, whereas haematopoietic stem cells are multipotent, and embryonic stem cells are thought to be able to give rise to all cell types and are said to be totipotent or pluripotent. To date three
5 types of mammalian pluripotent stem cell have been isolated. These cells can give rise to cell types that are normally derived from all three germ layers of the embryo (endoderm, mesoderm and ectoderm). The three types of stem cell are: embryonal carcinoma (EC) cells, derived from testicular tumours; embryonic stem (ES) cells, derived from the pre-implantation embryo (normally the blastocyst); and embryonic germ (EG) cells derived
10 from the post-implantation embryo (normally cells of the foetus destined to become part of the gonads).

Differentiated cell types are capable of changing phenotype. This phenomenon, termed transdifferentiation, is the conversion of one differentiated cell type to another, with or without an intervening cell division. Differentiation can sometimes be reversed or altered.
15 In vitro protocols are now available in which cell lines can be induced to transdifferentiate. Moreover, specialised cell types can de-differentiate to yield stem-like cells with the potential to differentiate into further cell types.

Many systems that have been devised for the differentiation of stem cells in vitro are complex multi-stage procedures, in which the precise nature of the various steps, as well
20 as the chronology of the various steps, are important. For instance, Lee et al (2000, Nature Biotechnology, vol. 18, p 675-679) used a five stage protocol to derive dopaminergic neurons from mouse ES cells: 1) undifferentiated ES cells were expanded on gelatin-coated tissue culture surface in ES cell medium in the presence of LIF ; 2) embryoid bodies were generated in suspension cultures for 4 days in ES cell medium; 3)
25 nestin-positive cells were selected from embryoid bodies in ITSFn medium for 8 days after plating on tissue culture surface; 4) nestin-positive cells were expanded for 6 days in N2 medium containing bFGF/laminin; 5) finally the expanded neuronal precursor cells were induced to differentiate by withdrawing bFGF from N2 medium containing laminin. In a second example of serial cell culture, Bonner-Weir et al. [Proc. Natl. Acad. Sci. (2000)
30 97: 7999-8004] derived insulin producing cells from human pancreatic ductal cells by: 1) selecting ductal cells over islet cells by selective adhesion on a solid surface in the presence of serum for 2-4 days; 2) subsequently withdrawing serum and adding keratinocyte growth factor to select for ductal epithelial cells over fibroblasts for 5-10 days; and 3) overlaying the cells with the extracellular matrix preparation 'Matrigel' for 3- 6
35 weeks. In a further example of serial cell culture, Lumelsky et al. [Science (2001) 292:

1389- 1394] derived insulin secreting cells by directed differentiation of mouse embryonic stem (ES) cells by: 1) expanding ES cells in the presence of LIF for 2-3 days; 2) generating embryoid bodies in the absence of LIF over 4 days; 3) selecting nestin-positive cells using ITSFn medium for 6-7 days; 4) expanding pancreatic endocrine precursors in
5 N2 medium containing 827 media supplement and bFGF for 6 days; and 5) inducing differentiation to insulin secreting cells by withdrawing bFGF and adding nicotinamide.

However it is not only the sequence and duration of the various steps or the series of addition of different factors that is important in the determination of cell differentiation. As embryonic development is regulated by the action of gradients of signalling factors that
10 impart positional information, it is to be expected that the concentration of a single signalling factor, and also the relative concentration of two or more factors, will be important in specifying the fate of a cell population in vitro and in vivo. Factor concentrations vary during development and stem cells respond differently to different concentrations of the same molecule. For instance, stem cells isolated from the CNS of
15 late stage embryos respond differently to different concentrations of EGF: low concentrations of EGF result in a signal to proliferate, while higher concentrations of EGF result in proliferation and differentiation to astrocytes. Many of the factors that have been found to influence self-renewal and differentiation of stem cells in vitro are naturally-occurring molecules. This is to be expected, as differentiation is induced and controlled by
20 signalling molecules and receptors that act along signal transduction pathways. However, by the same token, it is likely that many synthetic compounds will have an effect on stem cell differentiation. Such synthetic compounds that have high probability of interacting with cellular targets within signalling and signal transduction pathways (so called drugable targets) are routinely synthesised, for instance for drug screening by pharmaceutical
25 companies. Once known, these compounds can be used to direct the differentiation of stem cells ex vivo, or can be administered in vivo in which case they would act on resident stem cells in the target organ of a patient.

Common variables in tissue culture In developing conditions for the successful culture of a particular cell type, or in order to achieve or modulate a cellular process, it is often
30 important to consider a variety of factors. One important factor is the decision of whether to propagate the cells in suspension or as a monolayer attached to a substrate. Most cells prefer to adhere to a substrate although some, including transformed cells, haematopoietic cells, and cells from ascites, can be propagated in suspension.

Assuming the culture is of adherent cells, an important factor is the choice of adhesion substrate. Most laboratories use disposable plastics as substrates for tissue culture. The plastics that have been used include polystyrene (the most common type), polyethylene, polycarbonate, Perspex, PVC, Teflon, cellophane and cellulose acetate. It is likely that any plastic can be used, but many of these will need to be treated to make them wettable and suitable for cell attachment. Furthermore it is very likely that any suitably prepared solid substrate can be used to provide a support for cells, and the substrates that have been used to date include glass (e.g. alum-borosilicate and soda-lime glasses), rubber, synthetic fibres, polymerised dextrans, metal (e.g. stainless steel and titanium) and others. Some cell types, such as bronchial epithelium, vascular endothelium, skeletal muscle and neurons require the growth substrate to be coated with biological products, usually extracellular matrix materials such as fibronectin, collagen, laminin, polylysine or others. The growth substrate and the method of application (wet or dry coating, or gelling) can have an effect on cellular processes such as the growth and differentiation characteristics of cells, and these must be determined empirically as discussed above. Probably the most obviously important of the variables in cell culture is the choice of culture medium and supplements such as serum. These provide an aqueous compartment for cell growth, complete with nutrients and various factors, some of which have been listed above, others of which are poorly defined. Some of these factors are essential for adhesion, others for conveying information (e.g. hormones, mitogens, cytokines) and others as detoxificants. Commonly used media include RPMI 1640, MEM/Hank's salts, MEM/Earle's salts, F12, DMEM/F12, L15, MCDB 153, and others. The various media can differ widely in their constituents - some of the common differences include sodium bicarbonate concentration, concentration of divalent ions such as Ca and Mg, buffer composition, antibiotics, trace elements, nucleosides, polypeptides, synthetic compounds, drugs, etc. It is well known that different media are selective, meaning they promote the growth of only some cell types. Media supplements- such as serum, pituitary brain and other extracts, are often essential for the growth of cells in culture, and in addition are frequently responsible for determining the phenotype of cells in culture, i.e. they are capable of determining cell survival or directing differentiation. The role of supplements in cell processes such as differentiation is complex and depends on their concentration, the time point at which they are added to the culture, the cell type and medium used. The undefined nature of these supplements, and their potential to affect the cell phenotype, have motivated the development of serum-free media. As with all media, their development has come about largely by trial and error, as has been discussed above. The gas phase of the tissue culture is also important and its composition and volume which should be used can depend on the type of medium used,

the amount of buffering required, whether the culture vessel is open or sealed, and whether a particular cellular process needs to be modulated. Common variables include concentration of carbon dioxide and oxygen. Other conditions important to tissue culture include the choice of culture vessel, amount of headspace, inoculation density,

5 temperature, frequency of media changes, treatment with enzymes, rate and mode of agitation or stirring. Varying the cell culture conditions is therefore a method of achieving a desired cellular process. One aspect of the invention recognises that variation of the cell culture conditions in a serial manner can be a highly effective method for achieving a cellular effect. In various applications, for instance in studies of cell differentiation, it will
10 often be the case that a specific series of different tissue culture conditions are required to effect a cellular process. The different conditions may include additions or withdrawals to/from the media or the change of media at specific time points. Such a set of conditions, examples of which are given below, are commonly developed by trial and error as has been discussed above.

15 **Combinatorial serial culture of cells (Split-pool cell culture)**

A particularly efficient method for sampling a large number of cell culture conditions is referred to as Combinatorial Cell Culture or split-pool cell culture (Fig 1) and in one embodiment involves the serial subdividing and combining of groups of cells in order to sample multiple combinations of cell culture conditions. In one aspect of the invention the
20 method operates by taking an initial starter culture (or different starter cultures) of cells divided into X number of aliquots each containing multiple beads (groups/colonies/carriers/cells) which are grown separately under different culture conditions. Following cell culture for a given time, the cells can be pooled by combining and mixing the beads from the different aliquots. This pool can be split again into X2
25 number of aliquots, each of which is cultured under different conditions for a period of time, and subsequently also pooled. This iterative procedure of splitting, culturing and pooling (or pooling, splitting and culturing; depending on where one enters the cycle) cells allows systematic sampling of many different combinations of cell culture conditions.

The complexity of the experiment, or in other words the number of different combinations
30 of cell culture conditions tested, is equal to the product of the number of different conditions ($X_1 \times X_2 \times \dots \times X_n$) sampled at each round. Note that the step of pooling all the cells prior to a subsequent split can be optional - a step in which a limited number of cells are pooled can have the same effect.

The invention therefore embodies a number of related methods of systematically sampling multiple combinations of cell culture conditions where groups of test units are handled in bulk. Regardless of the precise manner in which a diversity of cell culture conditions is sampled by this means the procedure is efficient because multiple cells can
5 share a single vessel, where they are cultured under identical conditions, and it can be carried out using only a few culture vessels at any one time (the number of culture vessels in use is equal to the number of split samples).

In many respects the principle of this procedure resembles that of split synthesis of large chemical libraries (known as combinatorial chemistry), which samples all possible
10 combinations of linkage between chemical building block groups (see for example: Combinatorial Chemistry, Oxford University Press (2000), Hicham Fenniri (Editor)). Split-pool cell culture can be repeated over any number of rounds, and any number of conditions can be sampled at each round. So long as the number of test units (cells or colonised beads in this example) is greater than or equal to the number of different
15 conditions sampled over all rounds, and assuming that the splitting of cells occurs totally randomly, it is expected that there will be at least one cell unit that has been cultured according to each of the various combinations of culture conditions sampled by the experiment. This procedure can be used to sample growth or differentiation conditions for any cell type, or the efficiency of biomolecule production (e.g. production of erythropoietin
20 or interferon) by any cell type. Because the procedure is iterative, it is ideally suited to testing multistep tissue culture protocols - for instance those described above in connection with stem cell differentiation. The variables which can be sampled using this technique include cell type, duration of cell culture round, temperature, different culture media (including different concentrations of constituents), growth factors, conditioned
25 media, co-culture with various cell types (e.g. feeder cells), animal or plant extracts, drugs, other synthetic chemicals, infection with viruses (incl. transgenic viruses), addition of transgenes, addition of antisense or anti-gene molecules (e.g. RNAi, triple helix), sensory inputs (in the case of organisms), electrical, light, or red-ox stimuli and others.

Split-split cell culture The purpose of performing split-pool processes on cells is to
30 systematically expose these to a pre-defined combination of conditions. The person skilled in the art will conceive of many different means of achieving this outcome. In addition to split-pool processes and variations thereof, it is worthwhile briefly discussing split-split processes. A split-split process involves subdividing a group of cells at least twice, without intervening pooling of cells. If split-split processes are used over a large
35 number of rounds, the number of separate samples that are generated increases

exponentially. In this case it is important to employ some level of automation, for example the use of a robotic platform and sophisticated sample tracking systems. The advantage of split-split steps is that (since cells are not combined) it is possible to segregate lineages of the various cells based on their cell culture history. Consequently split-split steps can
5 be used to deduce if a particular cell culture condition is responsible for any given cellular process and therefore used to deduce the culture history of cells.

Predetermined protocols The splitting and/or pooling of cells may be accomplished totally randomly or may follow a predetermined protocol. Where cells are split and/or pooled randomly, the segregation of a given cell unit into any group is not predetermined
10 or prejudiced in any way. In order to result in a high probability that at least one cell unit has been exposed to each of the possible combinations of cell culture conditions, it is advantageous to employ a larger number of cells than the total number of combinations of cell culture conditions that are being tested. Under certain circumstances it is therefore advantageous to split and/or pool cells according to a predetermined protocol, the overall
15 effect being that adventitious duplications or omissions of combinations are prevented. Predetermined handling of cells can be optionally planned in advance and logged on a spreadsheet or computer programme, and splitting and/or pooling operations executed using automated protocols, for instance robotics. Labelling of cells (see below) can be by any of a number of means, for instance labelling by RFID, optical tagging or spatial
20 encoding. Robotic devices capable of determining the identity of a sample, and therefore partitioning the samples according to a predetermined protocol, have been described (see 'Combinatorial Chemistry, A practical Approach', Oxford University Press (2000), Ed H. Fenniri). Alternatively, standard laboratory liquid handling and/or tissue culture robotics (for example such as manufactured by: Beckman Coulter Inc, Fullerton, CA; The
25 Automation Partnership, Royston, UK) is capable of spatially encoding the identity of multiple samples and of adding, removing or translocating these according to pre-programmed protocols.

Analysis and/or separation of cells Following each round of cell culture, or after a defined number of rounds, the cells can be studied to observe any given cellular process
30 that may have been affected by the tissue culture conditions. The examples below are illustrative and not intended to limit the scope of the invention.

Following each round of cell culture, or after a defined number of rounds, the cells can be assayed to determine whether there are members displaying increased cell proliferation. This can be achieved by a variety of techniques, for instance by visual inspection of the

cells under a microscope, or by quantitating a marker product characteristic of the cell. This may be an endogenous marker such as a particular DNA sequence, or a cell protein which can be detected by a ligand or antibody. Alternatively an exogenous marker, such as green fluorescent protein (GFP), can be introduced into the cells being assayed to

5 provide a specific readout of (living) cells. Live cells can be visualised using a variety of vital stains, or conversely dead cells can be labelled using a variety of methods, for instance using propidium iodide. Furthermore the labelled cells can be separated from unlabelled ones by a variety of techniques, both manual and automated, including affinity purification ('panning'), or by fluorescence activated cell sorting (FACS) or broadly similar

10 techniques (Fig. 17). Depending on the application it may be possible to use standard laboratory equipment, or it may be advantageous to use specialised instrumentation. For instance, certain analysis and sorting instruments (e.g. see Union Biometrica Inc., Somerville MA, USA) have flow cell diameters of up to one millimeter, which allows flow sorting of beads with diameters up to 500 microns. These instruments provide a reading

15 of bead size and optical density as well as two fluorescent emission wavelengths from tags such as GFP, YFP or OS-red. Sorting speeds of 180,000 beads per hour and dispensing into multi-well plates or into a bulk receptor are possible. Following each round of cell culture, or after a defined number of rounds, the cells can be assayed to determine whether there are members displaying a particular genotype or phenotype. Genotype

20 determination can be carried out using well known techniques such as the polymerase chain reaction (PCR), fluorescence in situ hybridisation (FISH), DNA sequencing, and others. Phenotype determination can be carried out by a variety of techniques, for instance by visual inspection of the cells under a microscope, or by detecting a marker product characteristic of the cell. This may be an endogenous marker such as a particular

25 DNA or RNA sequence, or a cell protein which can be detected by a ligand, conversion of an enzyme substrate, or antibody that recognises a particular phenotypic marker (For instance see Appendix E of *Stem Cells: Scientific Progress and Future Research Directions*. Department of Health and Human Services. June 2001; appendices incorporated herein by reference). A genetic marker may also be exogenous, i.e. one that

30 has been introduced into the cell population, for example by transfection or viral transduction. Examples of exogenous markers are the fluorescent proteins (e.g. GFP) or cell surface antigens which are not normally expressed in a particular cell lineage or which are epitope-modified, or from a different species. A transgene or exogenous marker gene with associated transcriptional control elements can be expressed in a manner that

35 reflects a pattern representative of an endogenous gene(s). This can be achieved by associating the gene with a minimal cell type-specific promoter, or by integrating the transgene into a particular locus (e.g. see European patent No. EP 0695351). The

labelled cells can be separated from unlabelled ones by a variety of techniques, both manual and automated, including affinity purification ('panning'), or by fluorescence activated cell sorting (FACS). Nishikawa et al (1998, Development vol 125, p1747-1757) used cell surface markers recognised by antibodies to follow the differentiation of
5 totipotent murine ES cells. Using FACS they were able to identify and purify cells of the haematopoietic lineage at various stages in their differentiation.

An alternative or complementary technique for enriching cells of a particular genotype or phenotype is to genetically select the desired groups. This can be achieved for instance by introducing a selectable marker into the cells, and to assay for viability under selective
10 conditions, for instance see Soria et al (2000, Diabetes vol 49, p1-6) who used such a system to select insulin secreting cells from differentiated ES cells. Li et al (1998, Curr Biol vol 8, p 971-974) identified neural progenitors by integrating the bifunctional selection marker/reporter f3geo (which provides for f3- galactosidase activity and G418 resistance) into the Sox2 locus by homologous recombination in murine ES cells. Since one of the
15 characteristics of neural progenitors is expression of Sox2, and therefore the integrated marker genes, these cells could be selected from non-neuronal lineages by addition of G418 after inducing differentiation using retinoic acid. Cell viability could be determined by inspection under a microscope, or by monitoring f3-gal activity. Unlike phenotype-based selection approaches, which can be limited by the availability of an appropriate ligand or
20 antibody, genetic selection can be applied to any differentially expressed gene.

Determination of the identity or cell culture history of a cell When handling large numbers of cells, their identity and/or cell culture history (for example the chronology and the exact nature of a series of culture conditions that any one group or cell may have been exposed to) can become confused. For instance, the split-pool protocol of cell
25 culture necessarily involves mixing cells in each round, making it difficult to follow individual cells. Determining the cell culture history of a cell in a mixture of cells which have been subjected to multiple culture conditions is sometimes referred to as 'deconvolution' of the cell culture history. One method of doing this is to label cells and it is therefore advantageous to label the cells. Labelling may be performed at the beginning
30 of an experiment, or during each round of an experiment and may involve a unique label (which may or may not be modified in the course of an experiment) or a series of labels which comprise a unique aggregate. Similarly, reading of the label(s) may take place during each round or simply at the end of the experiment.

A method of labelling cells involves sequentially associating unique tags with the cells whenever they are cultured under different conditions, such that subsequent detection and identification of the tags provides for an unambiguous record of the chronology and identity of the cell culture conditions to which the cell unit has been exposed. Tags can be taken up by cells, or attached to the cell surface by adsorption, or a suitable ligand or antibody. For instance, one simple tag that can be introduced to cells is an oligonucleotide of defined length and/or sequence. Oligonucleotides may comprise any class of nucleic acid (e.g. RNA, DNA, PNA, linear, circular or viral) and may contain specific sequences for amplification (e.g. primer sequences for PCR) or labels for detection (e.g. fluorophores or quenchers, or isotopic tags). The detection of these may be direct, for instance by sequencing the oligos or by hybridising them to complementary sequences (e.g. on an array or chip), or indirect as by monitoring an oligonucleotide-encoded gene product, or the interference of the nucleotide with a cellular activity (e.g. antisense inhibition of a particular gene). An advantageous method of amplifying nucleic acids is by rolling circle amplification (RCA; 2002, V. Demidov, *Expert Rev. Mol. Diagn.* 2(6), p.89- 95) where nucleic acid tags can comprise RCA templates, elongation primers, or struts that aid the circularization of minicircle templates).

A preferred labelling technique, advantageously used when the test units are single cells, comprises nucleic acid modification of the cell by insertion of a nucleic acid "barcode", by viral infection (such as by a lentivirus) or gene editing, for instance by CRISPR.

The reading of genetic modification is advantageously carried out by single-cell sequencing. Single-cell sequencing can be used for examining the genomic, transcriptomic, and epigenomic landscapes at the resolution of individual cells. Exemplary technologies include:

Single-Cell RNA Sequencing (scRNA-seq), including full-length transcript methods such as SMART-seq and SMART-seq2, which allow the capture of full-length transcripts, providing information about transcript isoforms; and 3' and 5' end methods, including Drop-seq, 10x Genomics Chromium, and CEL-seq2. These methods capture either the 3' or 5' ends of mRNA, allowing for counting of gene expression levels. See - Picelli, S., Faridani, O. R., Björklund, A. K., Winberg, G., Sagasser, S., & Sandberg, R. (2014). Full-length RNA-seq from single cells using Smart-seq2. *Nature Protocols*, 9(1), 171-181. DOI: 10.1038/nprot.2014.006; Macosko, E. Z., Basu, A., Satija, R., Nemesh, J., Shekhar, K., Goldman, M., Tirosh, I., Bialas, A. R., Kamitaki, N., Martersteck, E. M., Trombetta, J. J., Weitz, D. A., Sanes, J. R., Shalek, A. K., Regev, A., & McCarroll, S. A. (2015). Highly

Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets. *Cell*, 161(5), 1202-1214. DOI: 10.1016/j.cell.2015.05.002; Zheng, G. X. Y., Terry, J. M., Belgrader, P., Ryvkin, P., Bent, Z. W., Wilson, R., Ziraldo, S. B., Wheeler, T. D., McDermott, G. P., Zhu, J., Gregory, M. T., Shuga, J., Montesclaros, L., Underwood, J. G.,
5 Masquelier, D. A., Nishimura, S. Y., Schnall-Levin, M., Wyatt, P. W., Hindson, C. M., ... & Livak, K. J. (2017). Massively parallel digital transcriptional profiling of single cells. *Nature Communications*, 8, 14049. DOI: 10.1038/ncomms14049.

Another approach is Single-Cell DNA Sequencing (scDNA-seq). Examples include Whole Genome Amplification (WGA). Techniques such as MDA (Multiple Displacement
10 Amplification), DOP-PCR (Degenerate Oligonucleotide Primed PCR), and MALBAC (Multiple Annealing and Looping Based Amplification Cycles) are used for amplifying the DNA from a single cell.

See Zong, C., Lu, S., Chapman, A. R., & Xie, X. S. (2012). Genome-wide Detection of Single-Nucleotide and Copy-Number Variations of a Single Human Cell. *Science*,
15 338(6114), 1622-1626. DOI: 10.1126/science.1229164; Navin, N., Kendall, J., Troge, J., Andrews, P., Rodgers, L., McIndoo, J., Cook, K., Stepanisky, A., Levy, D., Esposito, D., Muthuswamy, L., Krasnitz, A., McCombie, W. R., Hicks, J., & Wigler, M. (2011). Tumour evolution inferred by single-cell sequencing. *Nature*, 472(7341), 90-94. DOI: 10.1038/nature09807.

20 If the tag is an epigenetic tag, Single-Cell Epigenomics may be used to detect the tag, including techniques such as single-cell bisulfite sequencing to study DNA methylation at the single-cell level.

See Smallwood, S. A., Lee, H. J., Angermueller, C., Krueger, F., Saadeh, H., Peat, J., Andrews, S. R., Stegle, O., Reik, W., & Kelsey, G. (2014). Single-cell genome-wide
25 bisulfite sequencing for assessing epigenetic heterogeneity. *Nature Methods*, 11(8), 817-820. DOI: 10.1038/nmeth.3035

Certain techniques include a multiplicity of approaches, and can be referred to as single-cell multi-omics. Examples include G&T-seq (Genome and Transcriptome Sequencing), which allows simultaneous sequencing of the genome and transcriptome of a single cell.

30 CITE-seq (Cellular Indexing of Transcriptomes and Epitopes by sequencing) combines scRNA-seq with antibody-derived tags to profile protein and RNA from the same cell.

- See Macaulay, I. C., Haerty, W., Kumar, P., Li, Y. I., Hu, T. X., Teng, M. J., Goolam, M., Saurat, N., Coupland, P., Shirley, L. M., Smith, M., Van der Aa, N., Banerjee, R., Ellis, P. D., Quail, M. A., Swerdlow, H. P., Zernicka-Goetz, M., Livesey, F. J., & Ponting, C. P. (2015). G&T-seq: parallel sequencing of single-cell genomes and transcriptomes. *Nature Methods*, 12(6), 519-522. DOI: 10.1038/nmeth.3370; - Stoeckius, M., Hafemeister, C., Stephenson, W., Houck-Loomis, B., Chattopadhyay, P. K., Swerdlow, H., Satija, R., & Smibert, P. (2017). Simultaneous epitope and transcriptome measurement in single cells. *Nature Methods*, 14(9), 865-868. DOI: 10.1038/nmeth.4380.

- Protein expression can also be studied at the single cell level, for example using Mass Cytometry (CyTOF), which uses heavy metal-labeled antibodies to study protein expression at the single-cell level. See Bendall, S. C., Simonds, E. F., Qiu, P., Amir, E. D., Krutzik, P. O., Finck, R., Bruggner, R. V., Melamed, R., Trejo, A., Ornatsky, O. I., Balderas, R. S., Plevritis, S. K., Sachs, K., Pe'er, D., Tanner, S. D., & Nolan, G. P. (2011). Single-cell mass cytometry of differential immune and drug responses across a human hematopoietic continuum. *Science*, 332(6030), 687-696. DOI: 10.1126/science.1198704.

Cell reprogramming involves changing the identity or function of a cell. This can involve transforming a cell from one type to another or reverting mature cells back to a more pluripotent or stem cell-like state, before re-differentiating into a more committed state.

- In the context of the present invention, reprogramming is the conversion of a differentiated, committed cell type into another differentiated, committed cell type. Reprogramming excludes mere conversion of a pluripotent cell into a differentiated cell type.

- Several techniques for reprogramming are known to the skilled person; in one embodiment, reprogramming as referred to herein may encompass any one or more of these techniques.

Induced Pluripotent Stem Cells (iPSCs):

Discovered by Shinya Yamanaka and his colleagues in 2006, this method involves introducing a specific set of transcription factors (often referred to as Yamanaka factors: Oct4, Sox2, Klf4, and c-Myc) into mature cells.

- These factors induce the mature cells to revert to a pluripotent state, similar to embryonic stem cells. The pluripotent cells can then be further differentiated into a desired cell type.

Yamanaka used retroviral vectors to deliver TFs, but many other methods are available, including non-integrating viral vectors, mRNA delivery, plasmids and direct addition of TF proteins, as well as the use of non-coding RNA.

Various types of ncRNAs, including microRNAs (miRNAs), long non-coding RNAs (lncRNAs), and others, can be involved in regulating transcription factors. These ncRNAs can be used to replace TFs in the assays of then invention.

microRNAs (miRNAs) are short (~22 nucleotides) ncRNAs that generally bind to the 3' untranslated region (UTR) of target mRNAs, leading to their degradation or translation repression. Through this mechanism, miRNAs can negatively regulate the levels of specific transcription factors (Bartel DP. (2009) MicroRNAs: target recognition and regulatory functions. **Cell**. 136(2):215-33).

Long non-coding RNAs (lncRNAs) are longer than 200 nucleotides and have diverse roles. Some lncRNAs can interact with transcription factors directly, modulating their activity, and can thus serve to modulate TF activity in an assay according to the invention. See Rinn JL, Chang HY. (2012) Genome regulation by long noncoding RNAs. *Annual Review of Biochemistry*. 81:145-66.

Non-coding RNAs have been shown to play roles in conjunction or as an alternative to the classic Yamanaka factors:

1. miR-302/367 Cluster: This cluster of microRNAs has been shown to enhance reprogramming efficiency when used in combination with the Yamanaka factors (Oct4, Sox2, Klf4, and c-Myc). See Anokye-Danso F, et al. (2011) Highly efficient miRNA-mediated reprogramming of mouse and human somatic cells to pluripotency. *Cell Stem Cell*. 8(4):376-88.

2. ****lncRNA-RoR****: This lncRNA has been implicated in the regulation of reprogramming and maintenance of pluripotency. It was shown that lncRNA-RoR can modulate the core transcriptional network of pluripotency, including influencing the levels of the Yamanaka factors. See Loewer S, et al. (2013) Large intergenic non-coding RNA-RoR modulates reprogramming of human induced pluripotent stem cells. **Nature Genetics**. 45(12):1504-9.

Direct Lineage Conversion (Transdifferentiation):

This method bypasses the pluripotent state. Instead, cells are directly converted from one mature cell type to another using a combination of lineage-specific transcription factors.

For example, fibroblasts can be directly converted to neurons, cardiomyocytes, or other cell types, depending on the factors used. Many techniques are currently available for
5 transdifferentiation:

1. Viral Vector-mediated Transcription Factor Introduction:

Using viral vectors, lineage-specific transcription factors are delivered to cells. For example:

10 Converting fibroblasts to neurons using factors like Ascl1, Brn2, and Myt1l (Vierbuchen, et al., (2010). Nature, 463(7284), 1035-1041).

Converting fibroblasts to cardiomyocytes using factors like Gata4, Mef2c, and Tbx5 (Ieda, et al., (2010) Cell, 142(3), 375-386).

2. RNA-based Transcription Factor Delivery:

15 This method employs synthetic modified mRNA to transiently express the necessary transcription factors. RNA-based methods can alleviate concerns associated with potential DNA integrations into the host genome. Warren, et al., (2010) Cell Stem Cell, 7(5), 618-630.

3. Small Molecules and Chemicals:

20 While transcription factors are most commonly used for direct lineage conversion, some studies have successfully employed combinations of small molecules to either enhance the efficiency of the process or, in rare cases, achieve transdifferentiation without the need for genetic factors.

For example:

25 Hou, et al., (2013). Science, 341(6146), 651-654, successfully reprogrammed mouse somatic cells into pluripotent stem cells using only small molecules, highlighting the potential to generate iPSCs without introducing exogenous genes.

Shi, et al., (2008). Cell stem cell, 2(6), 525-528 showed how a combination of small molecules can significantly enhance the efficiency of iPSC generation from mouse fibroblasts.

5 Fu, et al., (2015). Cell Research, 25(9), 1013-1024 used a set of small molecules to directly convert mouse fibroblasts into cardiomyocyte-like cells.

4. CRISPR/Cas9-based Activation of Endogenous Genes:

10 The CRISPR/Cas9 system, originally known for its gene-editing capabilities, can be adapted for gene activation by using a deactivated Cas9 (dCas9) fused to transcriptional activators. This can be used to activate endogenous lineage-specific genes, promoting transdifferentiation. Gilbert, et al., (2013) Cell, 154(2), 442-451.

5. MicroRNAs (miRNAs):

Certain miRNAs have been found to play roles in cell fate decisions and can be used alongside transcription factors or alone to drive direct lineage conversion. Jayawardena, et al., (2012) Circulation research, 110(11), 1465-1473.

15 6. Epigenetic Modulators:

Direct lineage conversion involves significant changes in the epigenetic landscape of a cell. Incorporating molecules that modulate epigenetic marks can increase the efficiency of transdifferentiation. Polo, Jet al., (2012) Cell, 151(7), 1617-1632

7. Extracellular Cues and Microenvironment:

20 While transcription factor-mediated conversion is the primary driver, the efficiency and success of the conversion process can be influenced by the microenvironment, including extracellular matrix components, neighboring cells, and culture conditions. Engler, A. J., et al., (2006). Cell, 126(4), 677-689.

25 In the context of the present invention, reprogramming is preferably transdifferentiatoin, by which cells are converted directly from one differentiated state to another, without passing through a pluripotent state.

A “cell type”, as referred to herein, is a differentiated cell belonging to a specified lineage. A cell type is not a totipotent cell. In embodiments, a cell type is not a pluripotent cell

capable of differentiating into more than one lineage. A cell lineage may be any of a variety of cells types which can develop during normal embryogenesis. Examples include:

Hematopoietic Lineage: This lineage gives rise to all the different blood cells.

- 5 Hematopoietic stem cells (HSCs) in the bone marrow differentiate into various types of blood cells, such as erythrocytes (red blood cells), various leukocytes (neutrophils, lymphocytes, and monocytes), and platelets.

- 10 Neuronal Lineage: Neural stem cells differentiate into various types of neuronal cells and glial cells. This includes motor neurons, sensory neurons, astrocytes, oligodendrocytes, and microglia.

Epidermal Lineage: The epidermal stem cells in the skin differentiate into various skin cells, including keratinocytes, melanocytes, and cells of the hair follicles.

Muscle Lineage: Myogenic stem cells (or satellite cells) can differentiate into muscle fibres or myocytes.

- 15 Mesenchymal Stem Cell (MSC) Lineage: MSCs are multipotent and can give rise to a variety of cell types, including osteoblasts, chondrocytes, fibroblasts and adipocytes.

Intestinal Lineage: Stem cells in the crypts of the intestines differentiate into various types of intestinal cells, such as absorptive enterocytes, goblet cells, Paneth cells, and enteroendocrine cells.

- 20 Hepatic Lineage: Liver stem cells can differentiate into hepatocytes, which are the main functional cells of the liver, and cholangiocytes, which line the bile ducts.

Cardiac Lineage: Cardiac progenitor cells can differentiate into various cell types of the heart, such as cardiomyocytes, endothelial cells, and smooth muscle cells.

- 25 Germ Cell Lineage: This lineage leads to the formation of ova in females and sperm in males.

Retinal Lineage: Retinal progenitor cells can differentiate into various types of retinal cells, such as photoreceptors, bipolar cells, and ganglion cells.

Reprogramming can involve the conversion of a cell from one lineage to another, or from one cell type within a lineage to another cell type within the same lineage, for example from a fibroblast to an osteoblast.

“**Reagents**” as referred to herein may be any chemical, small molecule, transcription
5 factor, nucleic acid, protein, compound or ion which can interact with a cell. In preferred embodiments, reagents are transcription factors, small molecule or protein compounds, or nucleic acids. In the most preferred embodiment, reagents are transcription factors (TFs) or agents which influence transcription factor activity, such as ncRNAs.

Certain reagents are known in the art to promote reprogramming, as described in
10 references cited herein and in more detail below. Transcription factors are often used to influence reprogramming.

Deconvolution of labels refers to the identification of labels on a cell, and optionally the order in which the labels were acquired by the cell, and thereby determining which reagents or combinations of reagents the cell has been exposed to. Since the cell is
15 labelled whenever it is exposed to a given reagent, the reagent leaves a “mark” on the cell which can be used to map the cell’s history in exposure to reagents as well as the timing of such exposure.

Predicting the most effective conversion-inducing reagents requires a combination of experimental data, computational modelling, and optimization techniques. An algorithm
20 can be used to interrogate the results of exposure of cells to given reagents, and thereby predict which reagents can be employed to induce a desired conversion in a given cell type. The results of the reagent exposure should be presented as a consistent metric to the algorithm; for example, the results can be in the form of gene expression data, cell morphology scores, specific gene activation events, and the like. Output from the
25 algorithm may be used to select initial reagents for the performance of the method of the invention. Where the method is performed multiple times, the output data can be used to inform further instances of the performance of the method.

Machine learning (ML, AI) can be employed to better predict input reagents for desired conversion outcomes. A database may be provided of gene expression data related to
30 reagent, and gene regulatory networks constructed further comprising the machine-learning model configured to output a gene regulatory graph.

Pooling and splitting or subdividing of cell populations is carried out as described in WO2004013369. Iterative repetition of the steps of the process of the present invention can also be performed in a manner analogous to that described in WO2004013369 and above.

- 5 Identification of cell types can be performed by any suitable means, including genetic, morphological, immunochemical or other features.

Gene expression analysis includes the analysis of one or more reporter genes, as well as analysis based on a wide variety of cellular genes including using arrays which analyse expression of the entire genome. In one embodiment, gene expression analysis focusses
10 on one or more genes which are markers for the desired cell type. The method of the invention can be used to determine such genes by analysing gene expression changes in cells which are exposed to reagents which are shown to lead to desired conversion outcomes by analysis of alternative markers of cell differentiation, such as immunochemical markers or cell morphology.

- 15 Genes which influence cell conversion can in turn be purposefully modulated in order to induce conversion; for example, information about a gene being modulated in response to exposure to a reagent can be used to select a gene modulation technique, such as genome editing or the like, to further modulate the gene in place of administration of the reagent.

20 **CombiCult®**

A procedure for splitting and pooling cellular populations to determine the effect of reagents on cell fate has been described in WO2004031369. This combinatorial screening platform allows testing thousands of time-resolved combinations of culture conditions at a time. It is based on combinatorial science, which had already been
25 successfully employed in chemical synthesis of new compounds. Briefly, stem cells are grown in or on beads that can be tagged with condition-specific fluorescent labels. For every step in the differentiation course, beads are pooled and split, and subjected to different conditions, where they are concurrently labelled in a condition-specific fashion.

This process is repeated for several steps, ensuring that every bead is exposed to one
30 specific combination of conditions, with multiple beads going through the same combination. Cells in beads are then screened for markers, using antibodies, and selected through a large particle sorter. Positive beads are isolated and dissociated, and

their tags are read by a FACS, thus detecting all the combinations of tags enriched in the positive beads. A computational deconvolution strategy allows to identify and quantify the most frequent tag combination, which will correspond to a specific set and order of addition of molecules to the cell culture (also termed "protocol"). Highly represented

5 protocols are then validated independently in vitro and compared to gold standard differentiation or expansion procedures. CombiCult® has been successfully deployed in several cases, allowing the discovery of for example novel protocols for expansion of hematopoietic stem cells, differentiation of macrophages, neutrophils, natural killer cells, megakaryocytes, smooth muscle cells, and oligodendrocyte progenitors.

- 10 The decision on which molecules to include in a CombiCult® screening comes from careful reading of literature, meaning that its power to discover new combinations is somehow limited by already suggested hypotheses. By harnessing the power of modern machine learning models combined with the large amount of high-resolution single cell transcriptomics datasets, the invention permits the generation of new data-driven
- 15 hypotheses with an unprecedented breadth.

Moreover, the multiplexing potential of CombiCult® is limited by the use of beads and the selection of one positive cell fate per screen, discarding the rest of the material.

The present invention provides:

1. Bespoke computational analyses to define the input of molecules to be tested,
- 20 including a machine learning algorithm to prioritize molecular drivers of cell fate acquisition and prevent contradictory or opposing pathways to be tested together;
2. A single cell tagging approach to entirely bypass the need for beads and large particle sorting, while simultaneously acquiring single cell definition and gene expression-based monitoring of differentiation results through low cost single cell sequencing (e.g. Oxford
- 25 Nanopore benchtop sequencing);
3. A second computational pipeline to analyse the results of the combinatorial screen by identifying viral barcodes and their representation in the acquisition of cell fates of interest.

- These 3 elements create a virtuous cycle in which the information from one screen can be
- 30 used to better refine the initial predictions, as more and more experimental evidence linking molecules and their transcriptional effects is generated at every round. Eventually,

once a critical mass of evidence is generated from several experiments, a new AI- based approach can be implemented in which, for a given profile, an ordered set of molecular drivers can be efficiently predicted, effectively reducing the need for combinatorial testing.

Computational analysis

- 5 The results of one CombiCult screen can be used to select input reagents for a further screen, or alternatively to determine optimal reagents for a desired transdifferentiation protocol.

The wealth of publicly available single-cell RNA-sequencing (scRNA-seq) and single cell chromatin accessibility data (scATAC-seq) from cells actively differentiating shows a
10 continuum of gene expression changes.

In general, an algorithm for selection of reagents can process the results of screens as follows:

Input:

1. List of cells with their identities: cell_list
- 15 2. List of reagents: reagent_list
3. Data on the effect of each reagent on each cell: effect_data (this could be gene expression profiles, cell phenotype changes, or any measurable metric post-exposure)
4. Desired cell type profile: desired_profile

Output:

- 20 List of selected reagents: selected_reagents

Steps:

1. Initialize an empty list: selected_reagents = []
2. For each cell in cell_list:
 - a. For each reagent in reagent_list:
25 i. Get the effect of this reagent on this cell from effect_data.

- ii. Compute the similarity score between the effect and the desired_profile. (This could be a correlation score, frequency metric, or any relevant measure of similarity.)
 - b. Rank the reagents for the cell based on the similarity scores.
 - c. Store the top-ranked reagent for this cell.
- 5 3. Aggregate the top-ranked reagents across all cells.
- 4. For each reagent in the aggregated list:
 - a. Compute the frequency of its selection across all cells.
 - b. If the frequency is above a certain threshold (indicating its consistent efficacy):
 - i. Add the reagent to selected_reagents.
- 10 5. Return selected_reagents.

By further sorting cells into cell types, it is possible for the algorithm to return selected_reagents associated with specific cell type transdifferentiation events.

- Machine learning may be applied to the algorithm, to further refine the selection of reagents. For example, a machine learning algorithm may be applied in the following manner:
- 15

Input:

1. Training dataset: Contains cell identities, timing and identity of reagent exposure, and observed effects post-exposure.

2. Desired cell type profile: desired_profile

- 20 Output:

Optimal reagent(s) for transdifferentiation: optimal_reagents

Steps:

1. Data Preprocessing:

a. Normalize and standardize the data (e.g., z-score normalization for gene expression levels).

b. Split the dataset into training and validation sets.

2. Feature Selection:

5 a. Select relevant features (e.g., specific gene expression levels, cell markers).

b. If dataset is high-dimensional, consider dimensionality reduction methods like PCA.

3. Model Selection:

a. Choose a suitable ML model. Regression models like Random Forest or Gradient Boosting Machines (GBM) are examples.

10 b. Train the model on the training set.

c. Validate the model on the validation set.

4. Hyperparameter Tuning:

a. Use techniques like grid search or random search to find the optimal hyperparameters for the chosen model.

15 b. Retrain the model using the optimal hyperparameters

5. Prediction:

a. Use the trained model to predict the effect of each reagent on transdifferentiation towards the desired cell type profile.

b. Rank reagents based on their predicted effectiveness.

20 6. Model Interpretation (Optional but recommended):

a. Use techniques like SHAP (SHapley Additive exPlanations) or feature importance scores to interpret which features (e.g., specific timings, cell markers) are most influential in the predictions.

7. Return:

- a. Return the top-ranked reagent(s) as optimal_reagents.

The algorithm of the invention applies gold standard data pre-processing steps and machine learning methods to identify temporal (or pseudo-temporal) trajectories and gene expression changes associated with them. These trajectories will connect stem cells to terminally differentiated cells through a continuum of gene expression changes. This continuum can be treated as a high-resolution sequence of molecular events that are specific to the cell fate acquisition process. Within this continuum, it will be possible to detect a set of reagents that drive gene expression changes. The algorithm of the invention prioritizes and orders these reagents based on 4 main features:

1. Their trend along the differentiation trajectories (differential expression)
2. Their gene regulatory network, inferred through multiple regression and externally validated reagent-target networks and protein-protein interaction networks (PPINs)
3. Their specificity, i.e. preferential activity within a trajectory compared to all the other ones in a similar physiological process (e.g. selecting a T-cell specific set of drivers compared to other lymphoid lineages in haematopoiesis)
4. Their dynamics in terms of switching time and/or transient expression.

In the case of TF reagents, if matching public scATAC data is available, it can supplement TF activity by performing footprinting/TF binding site analysis within the open chromatin regions, thus providing orthogonal evidence for the gene regulatory network, and complementing the other dataset(s).

Once TFs have been identified, target networks from each of them will be screened for annotated interactions. If the interactors of two TFs show a significant degree of overlap but are involved in opposing interactions (e.g., one TF is reported to inhibit these interactors, the other one is reported to activate them), they will be flagged for incompatibility within the same screen, thus providing a more rational way to arrange combinations.

The algorithm of the invention can take advantage of regression-based gene- gene correlation networks and existing PPINs to identify molecules that act as transducers and receptors in a signalling pathway that is upstream of the TFs and is also changing along the trajectory. Once candidate receptors have been identified, their ligands can be inferred and included in the combinatorial screen.

Finally, small molecule-protein interaction networks overlaid on the PPINs, and drug-transcriptome correlation databases whose signature can be interrogated along a trajectory, can also be included in the software for their prioritization and inclusion in the screening.

- 5 Along the generation of new hypotheses, the CombiCult platform can be improved by greatly extending its multiplexing and resolution capabilities. The implementation in WO2004031369 is limited to the identification of one cell differentiation at a time, which means that several other possibilities that may have emerged in culture – such as alternative cell fates – remain unexplored. In the present invention, tagging (“barcoding”)
- 10 single cells using viral transduction avoids this problem. To avoid using beads (whose encapsulation represents a time- consuming and resource-intensive step) and to track all single cells during the screening, barcoding is achieved through a unique, short, expressed DNA sequence that functions as a Unique Condition Tag (UCT) delivered through a lentiviral vector, which stably integrates in cells in culture (**Fig. 2**). As cells go
- 15 through different splits and are treated with different molecules/TFs/growth factors (“drivers”), they will be tagged by lentiviral vectors bearing a barcode unique to the condition. By using viral, integrating tags at low doses – so that cells do not become saturated at every transduction cycle – cells can be uniquely barcoded in every condition, and the UCT stays within each cell, tracking the condition in which they were inserted. In
- 20 the first split, candidate drivers and their corresponding vectors are added to the culture media, which can be refreshed to remove viral vectors after transduction according to the needs of the protocol.

Then, cells are detached, resuspended, pooled together and split in different plates for the second split.

- 25 Pooled and split cells undergo a second round of transduction and driver treatment; media is refreshed, and drivers are re-added until the third pool and split process.

- This is repeated for as many splits as necessary, until the final split is reached and cells are detached, optionally sorted via FACS and prepared for single cell RNA sequencing. The fundamental difference between this approach and the approach of WO2004031369
- 30 is that in the present invention barcoding happens at the single cell level – and the data consist of the transcriptome of every cell, with its corresponding combination of UCTs.

Viral tagging and the delivery of reagents to cells can be combined. Viral delivery and overexpression of TFs is described in Joung et al., Cell 186, 209–229, January 5, 2023;

the same methods can be used to deliver a tag to label a cell, or the reagent transgene itself can be considered a tag and detected by sequencing.

Development of a new analytical pipeline to select protocols

The analysis of single cell RNA coupled with UCT detection resolves the effect of different cell fate choices (both successful and unsuccessful) in a single snapshot. Data from the transcriptome are analysed, annotating cells through their similarity and expression of cell-type specific signatures. By reconstructing UCT "lineage" trees and matching them to signatures, it is possible to chart the sequence of variables that gives rise to cell-fate specific transcriptional profiles. There are three fundamental advantages in using single cell data coupled to UCTs: first, it allows to resolve even small differences in cell fate acquisition; second, if the whole experiment is sequenced and analysed, it allows to identify several cell fates at once; third, it generates data that can be reused and integrated.

Regarding the first two aspects, scRNA-seq has been already deployed in similar ways to track guide RNA constructs in CRISPR screens (Dixit et al. 2016. Cell 167, 1853–1866) or lentivirus-mediated TF transduction (Joung et al. 2023. Cell 186, 209–229). Furthermore, complex studies using sequential genome editing ("scarring") for lineage tracing at the single cell level have been successfully carried out (Spanjaard et al. Nat Biotechnol . 2018 June ; 36(5): 469–473), providing a first analytical framework to connect different events in a single experiment.

These pioneering approaches have been used to study transcriptional responses to perturbations at a systematic level, including cell fate acquisition, proving the feasibility in principle of a genetic tag-based tracing of experimental and/or physiological conditions at the single cell level. The innovative and ground-breaking aspect of this proposal lies in the application of CombiCult® principles, which allow to keep track of thousands of combinations of sequential addition of molecules or TFs in culture, thus resolving a much larger experimental space in a single assay. The third advantage allows to greatly enhance predictive power and ties the three elements of this platform together: the generation of datasets where transcriptional changes at the single cell level are matched to specific cell culture conditions provides an excellent opportunity to feed the algorithm again with experimentally validated driver-target interaction data. This virtuous cycle allows the algorithms of the invention to become more and more precise, learning from their own experimental results, and paves the way for a supervised learning framework that, in the long run, can predict with high precision the molecules required for any type of

differentiation. In fact, given a large number of ground truth (i.e., data resulting from a known intervention) it is in principle possible to create a Language Model (LM) that is able to predict the steps needed to generate specific cell types. Different types of LMs have been successfully deployed in studies of viral evolution and antibody evolution (Hie et al.

- 5 2023. Nat biotechnol doi:10.1038/s41587-023-01763-2) and chemical modifications (Kosonocky et al. arXiv 2023 doi:10.48550/arXiv.2305.16330).

The successful application of the sequencing-based platform of the invention will quickly generate many such data points, thus greatly accelerating ATMP engineering and production.

10 Transcription Factors

The most commonly recognised reagents for promoting transdifferentiation are transcription factors. Transdifferentiation can be driven by overexpressing specific transcription factors, as has been observed various transdifferentiation experiments:

1. Neuron:

- 15 - Ascl1, Brn2, and Myt1l can reprogram fibroblasts into induced neurons (iNs).

- NeuroD1 alone or in combination with other factors can also convert fibroblasts into neurons.

2. Cardiomyocytes:

- 20 - Gata4, Mef2c, and Tbx5 can convert fibroblasts into induced cardiomyocyte-like cells (iCMs).

3. Endoderm cells:

- FoxA2 and Hnf1 α can convert fibroblasts into hepatic cell types.

4. Pancreatic beta cells:

- 25 - Pdx1, Ngn3, MafA, and sometimes with other factors, can convert various cell types into insulin-producing beta-like cells.

5. Endothelial cells:

- Fli1, FoxC2, and Etv2 can convert fibroblasts into endothelial-like cells.

6. Myocytes:

- MyoD is a classic example that can convert fibroblasts into skeletal muscle cells.

Further transcription factors useful in transdifferentiation can be derived from the
5 literature.

The foregoing are examples of agents known in the art to promote reprogramming of cells.

Improved reagent database

The method of the invention permits the generation of a database which can improve the
10 starting parameters for an experiment designed to identify reagents suitable for converting cell types. Therefore, the invention provides a method for selecting the starting reagents in a method according to the preceding aspects of the invention, comprising the steps of:

compiling a database of reagents and cell conversion results; and selecting from the
database a set of reagents or combinations of reagents for use in a method for
15 determining the conditions for converting a cell.

The invention also provides a method for improving a database of reagents and cell transdifferentiating results, comprising using the database to select reagents as in the preceding embodiment, and inputting the results of the methods for determining the conditions for reprogramming a cell into the database, thus providing further data points.

20 The algorithms and machine learning approaches described above may also be used to further improve the database.

In another aspect, the method according to the invention provides: a database of reagents (for example transcription factors) useful for reprogramming cells; at least one computer processor; a memory in operative communication with the processor, the memory
25 comprising instructions configuring the processor to:

- (a) generate gene regulatory networks (GRN) from the database of transcription factors;
- (b) determine a candidate transcription factor for a desired reprogramming event

- (c) analyse an impact of the candidate transcription factor in cell reprogramming;
- (d) optionally, iteratively repeat steps (a)-(c); and
- (e) output a set of optimal transcription factors.

The database preferably comprises gene expression data, such as RNA-seq data.

- 5 Determining a candidate transcription factor comprises interrogating a gene regulatory graph to identify the set of optimal transcription factors to differentiate any given cell type from any given starting cell. A machine-learning model can be used to generate a metric calculation as a function of the gene regulatory graph.

- 10 Examples of a metric calculation include a criticality algorithm, which may be configured for time-series RNA-seq data; see Oh et al., Biomed Res Int. 2013;2013:203681. doi: 10.1155/2013/203681. Epub 2013 Mar 24.

- 15 The invention further provides a method for determining optimal transcription factors cell reprogramming, comprising: curating, using a computing device, a gene expression database related to use of reagents in cell culture; generating gene regulatory networks from a plurality of gene expression datasets; determining a candidate optimal reagent;
- 20 analysing an impact of the optimal reagent in cell differentiation; and outputting a set of optimal reagents.

Gene regulatory networks can be predicted computationally, as described in Kamaraj, Cell Cycle 2016, vol. 15, no. 24, 3343–3354.

- 20 Methods include CellNet (et al., Cell 2014; 158:903-15), which uses microarray gene expression data and correlation-based net-work score, the method of D'Alessio et al. (D'Alessio et al., Stem Cell Reports 2015; 5:763–75) whose method uses microarray gene expression data and an information theoretic approach (Jensen- Shannon Divergence) and Mogrify (Rackham et la., Nat Genet 2016; 48(3):331–5;
- 25 PMID:26711105) which uses cap analysis of gene expression (CAGE) data and a network based score.

Regardless of the computational approach selected, the iteration of the experimental steps with computational analysis provides a novel advantage over computational and experimental approaches used alone.

- 30 **Bioinformatic Methods**

To detect viral tags a new reference genome is created using the Parse Bio split-pipe pipeline (v 1.2.0) using the *mkref* mode and adding the following sequences to the Ensembl GRCh38 Release 108 FASTA file:

```
5  >LSSmOrange dna:chromosome
   chromosome:GRCh38:LSSmOrange:1:722:1 REF
   ATGGTGAGCAAGGGCGAGGAGAATAACATGGCCATCATCAAGGAGTTCATGCGCTTCAAG
   GTGCGCATGGAGGGCTCCGTGAACGGCCACGAGTTCGAGATCGAGGGCGAGGGCGAGGGC
   CGCCCCCTACGAGGGCTTTCAGACCGTTAAGCTGAAGGTGACCAAGGGTGGCCCCCTGCCC
10  TTCGCCTGGGACATCTTGTCCCCTCAGTTCACCTACGGCTCCAAGGCCTACGTGAAGCAC
   CCCGCCGACATCCCCGACTACCTCAAGCTGTCCTTCCCCGAGGGCTTCAAGTGGGAGCGC
   GTGATGAACTTCGAGGACGGCGGCGTGGTGACCGTGACTCAGGACTCCTCCCTGCAGGAC
   GGCGAGTTCATCTACAAGGTGAAGCTGCGCGGCACCAACTTCCCCCTCCGACGGCCCCGTA
   ATGCAGAAGAAGACCATGGGCATGGAGGCCCTCTCCGAGCGGATGTACCCCGAGGACGGC
15  GCCCTGAAGGGCGAGGACAAGCTCAGGCTGAAGCTGAAGGACGGCGGCCACTACACCTCC
   GAGGTCAAGACCACCTACAAGGCCAAGAAGCCCGTGCAGTTGCCCGGCGCCTACATCGTC
   GACATCAAGTTGGACATCACCTCCCACAACGAGGACTACACCATCGTGGAACAGTACGAA
   CGCGCCGAGGGCCGCCACTCCACCGGCGGCATGGACGAGCTGTACAAGTAA

20  >dTomato dna:chromosome chromosome:GRCh38:dTomato:1:713:1
   REF
   ATGGTGAGCAAGGGCGAGGAGGTCATCAAAGAGTTCATGCGCTTCAAGGTGCGCATGGAG
   GGCTCCATGAACGGCCACGAGTTCGAGATCGAGGGCGAGGGCGAGGGCCGCCCTACGAG
   GGCACCCAGACCGCCAAGCTGAAGGTGACCAAGGGCGGCCCCCTGCCCTTCGCCTGGGAC
25  ATCCTGTCCCCCAGTTCATGTACGGCTCCAAGGCGTACGTGAAGCACCCCGCCGACATC
   CCCGATTACAAGAAGCTGTCCTTCCCCGAGGGCTTCAAGTGGGAGCGCGTGATGAACTTC
   GAGGACGGCGGTCTGGTGACCGTGACCCAGGACTCCTCCCTGCAGGACGGCACGCTGATC
   TACAAGGTGAAGATGCGCGGCACCAACTTCCCCCGACGGCCCCGTAATGCAGAAGAAG
   ACCATGGGCTGGGAGGCCTCCACCGAGCGCCTGTACCCCGCGACGGCGTGCTGAAGGGC
30  GAGATCCACCAGGCCCTGAAGCTGAAGGACGGCGGCCACTACCTGGTGGAGTTCAAGACC
   ATCTACATGGCCAAGAAGCCCGTGCAACTGCCCGGCTACTACTACGTGGACACCAAGCTG
   GACATCACCTCCCACAACGAGGACTACACCATCGTGGAACAGTACGAGCGCTCCGAGGGC
   CGCCACCACCTGTTCTGTACGGCATGGACGAGCTGTACAAG

35  >EGFP dna:chromosome chromosome:GRCh38:EGFP:1:728:1 REF
   ATGGTGAGCAAGGGCGAGGAGCTGTTACCGGGGTGGTGCCCATCCTGGTCGAGCTGGAC
   GGCGACGTAAACGGCCACAAGTTCAGCGTGTCGGCGAGGGCGAGGGCGATGCCACCTAC
   GGCAAGCTGACCCCTGAAGTTCATCTGCACCACCGGCAAGCTGCCCCTGCCCTGGCCACC
   CTCGTGACCACCCCTGACCTACGGCGTGCAAGTTCAGCCGCTACCCCGACCACATGAAG
40  CAGCAGCACTTCTTCAAGTCCGCCATGCCCCGAAGGCTACGTCCAGGAGCGCACCATCTTC
   TTCAAGGACGACGGCAACTACAAGACCCGCGCCGAGGTGAAGTTCGAGGGCGACACCCCTG
   GTGAACCGCATCGAGCTGAAGGGCATCGACTTCAAGGAGGACGGCAACATCCTGGGGCAC
   AAGCTGGAGTACAACATAACAGCCACAACGTCTATATCATGGCCGACAAGCAGAAGAAC
   GGCATCAAGGTGAAGTTCAAGATCCGCCACAACATCGAGGACGGCAGCGTGACGCTCGCC
45  GACCACTACCAGCAGAACACCCCCATCGGCGACGGCCCCGTGCTGCTGCCCGACAACCAC
   TACCTGAGCACCCAGTCCGCCCTGAGCAAAGACCCCAACGAGAAGCGCGATCACATGGTC
   CTGCTGGAGTTCGTGACCGCCGCGGGGATCACTCTCGGCATGGACGAGCTGTACAAG

>mRFP dna:chromosome chromosome:GRCh38:mRFP:1:689:1 REF
```

ATGGCCAGCTCCGAGGATGTCATCAAAGAGTTTATGAGATTTAAGGTCAAGATGGAGGGA
AGCGTCAACGGACACGAGTTCGAGATTGAGGGAGAAGGAGAAGGCCGGCCTTACGAGGGC
ACACAAACCGCTAAGCTCAAGGTCACAAAAGGAGGACCCCTCCCTTCTCCTGGGATATT
CTGAGCCCTCAGTTCCAGTACGGAAGCAAAGCCTATGTTAAACACCCTGCCGACATCCCT
5 GACTATCTGAAGCTCTCCTTCCCTGAAGGCTTCAAGTGGGAGAGATTTCATGAACTTCGAG
GACGGAGGCGTGTTGACAGTCACACAAGATAGCACCCCTCCAGGACGGAGAGTTTATTTAT
AAGGTGAAACTCAGAGGAACCAACTTCCCTCCGATGGCCCTGTCATGCAAAAAAACA
ATGGGATGGGAAGCCTCCACCGAGAGAATGTATCCTGAGGATGGCGCTCTGAAAGGCGAA
ATTAATGAGACTGAACTCAAAGACGGAGGACACTACGATGCCGAGGTCAAACAACC
10 TACAAGGCCAAGAAACAAGTGCAGCTGCCTGGCGCCTACATGACTGATATTAACTCGAC
ATTATCAGCCATAATGGGGACTACACCATCGTGGAACAATATGAGAGAGCTGAGGGCAGA
CATAGCACAGGCGCTGGA

>BFP dna:chromosome chromosome:GRCh38:BFP:1:722:1 REF
15 ATGGTGTCTAAGGGCGAAGAGCTGATTAAGGAGAACATGCACATGAAGCTGTACATGGAG
GGCACCGTGGACAACCATCACTTCAAGTGCACATCCGAGGGCGAAGGCAAGCCCTACGAG
GGCACCCAGACCATGAGAATCAAGGTGGTTCGAGGGCGGCCCTCTCCCTTCGCCTTCGAC
ATCCTGGCTACTAGCTTCTCTACGGCAGCAAGACCTTCATCAACCACACCCAGGGCATC
CCCAGCTTCTTCAAGCAGTCTTCCCTGAGGGCTTCACATGGGAGAGAGTCAACACATAC
20 GAAGACGGGGGCGTGCTGACCGCTACCCAGGACACCAGCCTCCAGGACGGCTGCCTCATC
TACAACGTCAAGATCAGAGGGGTGAACTTCACATCCAACGGCCCTGTGATGCAGAAGAAA
ACACTCGGCTGGGAGGCCCTTACCGAGACGCTGTACCCCGCTGACGGCGGCCTGGAAGGC
AGAAACGACATGGCCCTGAAGCTCGTGGGCGGGAGCCATCTGATCGCAAACGCCAAGACC
ACATATAGATCCAAGAAACCCGCTAAGAACCTCAAGATGCCTGGCGTCTACTATGTGGAC
25 TACAGACTGGAAAGAATCAAGGAGGCCAACAACGAGACCTACGTCGAGCAGCACGAGGTG
GCAGTGGCCAGATACTGCGACCTCCCTAGCAAACCTGGGGCACAAGCTTAAT

>FLUC dna:chromosome chromosome:GRCh38:FLUC:1:1677:1 REF
30 ATGGAAGACGCCAAAAACATAAAGAAAGGCCCGGCCATTCTATCCGCTAGAGGATGGA
ACCGCTGGAGAGCAACTGCATAAGGCTATGAAGAGATACGCCCTGGTTCTGGAACAATT
GCTTTTACAGATGCACATATCGAGGTGAACATCACGTACGCGGAATACTTCGAAATGTCC
GTTTCGTTGGCAGAAGCTATGAAACGATATGGGCTGAATACAAATCACAGAATCGTCGTA
TGCAGTGAAACTCTCTTCAATTCTTTATGCCGGTGTGGGCGCGTTATTTATCGGAGTT
GCAGTTGCGCCCGCAACGACATTTATAATGAACGTGAATTGCTCAACAGTATGAACATT
35 TCGCAGCCTACCGTAGTGTGTTTCCAAAAAGGGGTTGCAAAAAATTTTGAACGTGCAA
AAAAAATTACCAATAATCCAGAAAATTATTATCATGGATTCTAAAACGGATTACCAGGGA
TTTCAGTCGATGTACACGTTCTGTCACATCTCATCTACCTCCCGGTTTTAATGAATACGAT
TTTGTACCAGAGTCCTTTGATCGTGACAAAACAATTGCACTGATAATGAACTCCTCTGGA
TCTACTGGGTACCTAAGGGTGTGGCCCTTCCGCATAGAACTGCCTGCGTCGATTCTCG
40 CATGCCAGAGATCCTATTTTTGGCAATCAAATCATTCCGGATACTGCGATTTTAAGTGTT
GTTCCATTCCATCACGGTTTTGGAATGTTTACTACACTCGGATATTTGATATGTGGATTT
CGAGTCGTCTTAATGTATAGATTTGAAGAAGAGCTGTTTTTACGATCCCTTCAGGATTAC
AAAATTCAAAGTGC GTTGTAGTACCAACCCTATTTTCATTCTTCGCCAAAAGCACTCTG
ATTGACAAATACGATTTATCTAATTTACACGAAATTGCTTCTGGGGGCGCACCTCTTTTCG
45 AAAGAAGTCGGGGAAGCGGTTGCAAAACGCTTCCATCTTCCAGGGATACGACAAGGATAT
GGGCTCACTGAGACTACATCAGCTATTCTGATTACACCCGAGGGGGATGATAAACCGGGC
GCGGTCGGTAAAGTTGTTCCATTTTTTGAAGCGAAGGTTGTGGATCTGGATACCGGGAAA
ACGCTGGGCGTTAATCAGAGAGGCGAATTATGTGTGAGAGGACCTATGATTATGTCCGGT
TATGTAAACAATCCGGAAGCGACCAACGCCTTGATTGACAAGGATGGATGGCTACATTCT
50 GGAGACATAGCTTACTGGGACGAAGACGAACACTTCTTCATAGTTGACCGCTTGAAGTCT
TTAATTAAATACAAAGGATACCAGGTGGCCCCCGCTGAATTGGAGTCGATATTGTTACAA

CACCCCAACATCTTCGACGCGGGCGTGCGAGGTCTTCCCGACGATGACGCCGGTGAACCTT
CCCGCCGCCGTTGTTGTTTTGGAGCACGGAAAGACGATGACGGAAAAAGAGATCGTGATG
TACGTCGCCAGTCAAGTAACAACCGCGAAAAAGTTGCGCGGAGGAGTTGTGTTTGTGGAC
GAAGTACCGAAAAGGTCTTACCGGAAAACTCGACGCAAGAAAAATCAGAGAGATCCTCATA
5 AAGGCCAAGAAGGGCGGAAAGTCCAAATTG

>copGFP dna:chromosome chromosome:GRCh38:copGFP:1:756:1 REF
ATGGAGAGCGACGAGAGCGGCCTGCCCGCCATGGAGATCGAGTGCCGCATCACCGGCACC
CTGAACGGCGTGAGTTTCGAGCTGGTGGGCGGCGAGAGGGCACCCCAAGCAGGGCCGC
10 ATGACCAACAAGATGAAGAGCACCAAAGGCGCCCTGACCTTCAGCCCCCTACCTGCTGAGC
CACGTGATGGGCTACGGCTTCTACCACTTCGGCACCTACCCAGCGGCTACGAGAACCC
TTCCTGCACGCCATCAACAACGGCGGCTACACCAACACCCGCATCGAGAAGTACGAGGAC
GGCGGCGTGCTGCACGTGAGCTTCAGCTACCGCTACGAGGCCGGCCGCTGATCGGCGAC
TTCAAGGTGGTGGGCACCGGCTTCCCCGAGGACAGCGTGATCTTCACCGACAAGATCATC
15 CGCAGCAACGCCACCGTGGAGCACCTGCACCCCATGGGCGATAACGTGCTGGTGGGCAGC
TTCGCCCCGACCTTCAGCCTGCGCGACGGCGGCTACTACAGCTTCGTGGTGGACAGCCAC
ATGCACTTCAAGAGCGCCATCCACCCAGCATCCTGCAGAACGGGGGGCCCCATGTTCCGCC
TTCGCGCGCGTGAGGAGCTGCACAGCAACACCGAGCTGGGCATCGTGGAGTACCAGCAC
GCCTTCAAGACCCCATCGCCTTCGCCAGATCCCGCGCTCAGTCGTCCAATTCTGCCGTG
20 GACGGCACCGCCGGACCCGGCTCCACCGGATCTCGC

>T2A_PURO dna:chromosome chromosome:GRCh38:T2A_PURO:1:654:1
REF
GAGGGCAGAGGAAGTCTTCTAACATGCGGTGACGTGGAGGAGAATCCCGGCCCTATGACC
25 GAGTACAAGCCACGGTGCGCCTCGCCACCCGCGACGACGTCCCCAGGGCCGTACGCACC
CTCGCCGCCGCGTTTCGCCGACTACCCCGCCACGCGCCACACCGTCGATCCGGACCGCCAC
ATCGAGCGGGTCACCGAGCTGCAAGAACTCTTCCTCACGCGCGTCGGGCTCGACATCGGC
AAGGTGTGGGTTCGCGGACGACGGCGCCGCGGTGGCGGTCTGGACCACGCCGGAGAGCGTC
GAAGCGGGGGCGGTGTTTCGCCGAGATCGGCCGCGCATGGCCGAGTTGAGCGGTTCCCGG
30 CTGGCCGCGCAGCAACAGATGGAAGGCCTCCTGGCGCCGCACCGGCCCAAGGAGCCCGCG
TGGTTCCTGGCCACCGTTCGGCGTCTCGCCGACCACCAGGGCAAGGTCTGGGCAGCGCC
GTCGTGCTCCCCGGAGTGGAGGCGGCCGAGCGCGCCGGGGTGCCCGCCTTCCTGGAGACC
TCCGCGCCCCGCAACCTCCCCTTCTACGAGCGGCTCGGCTTCACCGTCACCGCCGACGTC
GAGGTGCCCCAAGGACCGCGCACCTGGTGCATGACCCGCAAGCCCGGTGCCTGA
35

These sequences are referenced in a custom GTF file where they were included as gene,
transcript and exon.

Raw FASTQ files are aligned to this reference using the split-pipe pipeline using the *all*
mode, and the 8 sub-libraries are combined using the *combine* mode, all with default
40 parameters. Resulting normalized h5ad files are downloaded and imported in R v. 4.3.2 as
SingleCellExperiment objects using the *h5ad2sce()* function from the *schard* package
(<https://github.com/cellgeni/schard>).

For the 27 well and cancer-iPSC experiments, cells in which the T2A_PURO feature is
detected with at least 1 log-normalized count were isolated and divided according to the
45 experiment, as they are cells in which we can detect with certainty that a lentiviral vector

has been integrated and is actively being transcribed. For the 3TF experiments, this is not possible and all cells are used.

Log-normalized counts are used as input for variable feature selection using the

5 *modelGeneVar()* function from the *scrn* package
(<https://bioconductor.org/packages/release/bioc/html/scrn.html>). The top 2000 most
variable features are selected as input to Principal Component Analysis (PCA) for
dimensionality reduction. PCA is run using the *runPCA()* function from the *scater* package
(<https://bioconductor.org/packages/release/bioc/html/scater.html>) and 20 dimensions
10 are retained for all subsequent analyses.

Universal Manifold Approximation and Projection (UMAP) 2D embeddings are created
from the first 20 dimensions of the PCA using the *uwot* package ([https://cran.r-](https://cran.r-project.org/web/packages/uwot/index.html)
[project.org/web/packages/uwot/index.html](https://cran.r-project.org/web/packages/uwot/index.html)) with the following parameters: *n_neighbors*
15 *= 30*, *min_dist = 0.4*, *seed = 1*.

A tag is considered as detected if it has more than 0 log-normalized counts.

Clusters are identified building a Shared-Nearest Neighbor (SNN) graph using the
makeSNNGraph() function from the *bluster* package

20 (<https://bioconductor.org/packages/release/bioc/html/bluster.html>) using *k = 20* and a
Jaccard weighting scheme. Then, the *cluster_leiden()* function from the *igraph* package
(<https://cran.r-project.org/web/packages/igraph/index.html>) is used to identify clusters in
the graph, using a resolution of 0.3.

25 Cluster markers are identified using the *wilcoxauc()* function from the *presto* package
(<https://github.com/immunogenomics/presto>), and prioritized according to a score
calculated by multiplying the area under the Receiver-Operating Characteristic curve
(auROC) by the difference between percentage of cells expressing the marker in the
cluster and in the rest of the data.

30

UMAPs and dot plots are visualized using plotting functions from the *ggplot2* package
(<https://cran.r-project.org/web/packages/ggplot2/index.html>), and heatmaps are created

using the *ComplexHeatmap* package

(<https://bioconductor.org/packages/release/bioc/html/ComplexHeatmap.html>).

Example 1

5 Genetic Tagging of Differentiated Single Cells

This experiment demonstrates:

- Genetic tagging of single cells during a CombiCult experiment
- Tracking single cells exposed to multiple reagent conditions
- 10 • Deconvolution of genetic tags by single-cell RNA sequencing (scRNAseq)
- Correlating cell culture condition history with differentiation outcomes

The experiment consists of 3 splits, with 3 conditions per split, resulting in a total of $3 \times 3 \times 3 = 27$ combinations. Induced pluripotent stem cells (iPSCs) serve as the starting cell type. In each split, only one condition involves differentiation medium for iPSC-
15 megakaryocyte (MK) differentiation. This condition is referred to as the 'specifying' condition. The remaining two conditions contain only basal medium and fetal bovine serum (FBS), which do not specify differentiation into MK and are referred to as 'non-specifying' conditions. Therefore, out of the 27 combinations, only 1 combination allows iPSC-MK differentiation.

20 Cells were cultured in the following media:

Specifying condition (with fluorescent proteins as genetic tags):

Days 1 and 4: APEL2 1% anti-anti, 50 ng/ml BMP4, 50 ng/ml VEGF, 50 ng/ml bFGF; Day 7 (expansion): APEL2 1% anti-anti, TPO 25 ng/ml, SCF 25 ng/ml, Flt3L 25 ng/ml, IL-3 10 ng/ml, IL-6 10 ng/ml, Heparin 5 U/ml.

25 Non-specifying condition (with luciferase as genetic tag): DMEM 1% anti-anti 5% FBS

The specifying condition in each split is tagged with a unique lentivirus carrying a different fluorescent protein transgene, while the non-specifying conditions are tagged with a lentivirus carrying a luciferase transgene in all splits. Theoretically, only cells that pass through the specifying condition in all 3 splits differentiate fully, and these cells are tagged
30 with 3 different lentiviruses carrying 3 distinct fluorescent proteins. A schematic of the procedure described is shown in Figure 3. By the end of the experiment, all cells are collected for scRNAseq. It is expected to see gene expression of all 3 fluorescent proteins

only in differentiated cells. While some differentiated cells may show gene expression of fewer than 3 fluorescent proteins due to inefficient tagging or detection thresholds, no differentiated cells are expected to show luciferase gene expression, as cells passing through the non-specifying condition are not expected to differentiate.

5

Detailed Procedure

- **On Day 0**, iPSCs are seeded as single cells into 27 wells of a multi-well culture plate. Cells in 9 out of the 27 wells are transduced with lentivirus carrying the blue fluorescent protein gene (LV-BFP) (Fig. 3a). Cells in the remaining 18 wells are transduced with lentivirus carrying the luciferase gene (LV-LUC) (Fig. 4a). An MOI of 5 is used for all lentivirus transductions.
 - **On Day 1**, viruses are removed, and cells are rinsed twice to remove residual viruses before switching to fresh media. Cells transduced with LV-BFP are switched to the MK differentiation medium (specifying condition), and cells transduced with LV-LUC are switched to basal medium with FBS (non-specifying condition).
 - **On Day 4**, cells are washed and switched to the second split conditions. Three wells from the specifying condition in the first split and six wells from the non-specifying condition are switched to MK differentiation medium (specifying condition). In this specifying condition, cells are also transduced with lentivirus carrying the red fluorescent protein gene (LV-RFP) (Fig. 3b). Cells in the remaining wells are switched to basal medium with FBS (non-specifying condition) and transduced with LV-LUC (Fig. 4b). On Day 5, viruses are removed, and cells are rinsed twice to remove residual viruses before adding fresh media.
 - **On Day 7**, cells are washed and switched to the third split conditions, following the same procedure. Cells in the specifying condition are transduced with lentivirus carrying the green fluorescent protein gene (LV-GFP) (Fig. 4c). Cells in the non-specifying condition are transduced with LV-LUC (Fig. 3c). On Day 8, viruses are removed, and cells are rinsed twice before adding fresh media.
 - **On Day 18**, cells are collected for scRNAseq.
- Cells are collected and fixed so that they can all be processed at the same time. The last split/tag can be identified through Parse Bio's combinatorial barcoding system, where the first of 3 barcodes indicates the original sample.

RNA reads are aligned to a custom reference human genome in which the following cDNA sequences are added:

- EGFP (from GATA1 lentivirus)
- mRFP
- 5 • BFP
- copGFP
- dTomato (from TAL1 lentivirus)
- LSSmOrange (from FLI1 lentivirus)
- LUC
- 10 • Puro_T2A (LV backbone—does not include TF vectors)

Only cells with enough barcode reads and enough RNA reads can be analysed (~78K good quality from 100K cells total).

The results are shown in Figures 4 to 7.

Figure 4 shows BFP tag detection in a variety of cell clusters on a UMAP plot. The plots
15 shown are for control (CTRL) cells, both in adherent (ADH) and in suspension (SUS); "GFP"
cells, which are cells taken from the final GFP-positive wells in the final step of the
procedure. Thus, in Figure 4, the "GFP" map detects BFP expression in cells that have
been exposed to all three tags. The LUC cells are cells that have been exposed to
luciferase only, that is, have not been tagged with any viral tag, and have not been
20 exposed to any specifying conditions.

Figure 4 also shows identical UMAP plots for cells screened for RFP and GFP expression.
Although GFP expression is low, it is postulated that this may be a result of increasing viral
resistance as the cells differentiate. All of the "GFP" cells are derived from the final GFP-
positive well, so have inevitably been exposed to the third stage of specifying conditions.
25 In Figure 5, UMAP plots are shown in which markers are scored as present or absent,
rather than on intensity. Triple-positive markers are visible in the 27W GFP ADH plot in
Figure 6.

Analysis of clusters in Figures 7 and 8 shows the three-marker cells (27W ADH GFP) have a
strong presence in cluster 7. Gene expression associated with cluster 7 is illustrated and
30 further characterised in Figure 9, which suggests the cells are stromal cells. Further
analysis is required. However, the results clearly show that a single cell can be tagged and
followed through differentiation in a plurality of media.

The experiment successfully demonstrates that single cells can be genetically tagged and tracked through multiple reagent conditions, providing insights into how specific sequences of exposure influence cell differentiation.

5 Example 2

Tagging of Cells Exposed to Transcription Factors (Cells on Microcarrier Beads)

This experiment demonstrates:

- Genetic tagging of cells on microcarrier beads during a CombiCult experiment
- Deconvolution of genetic tags by single-cell RNA sequencing (scRNAseq)
- 10 • Tracing combinations of TFs responsible for reprogramming cells into specific lineages through genetic tag deconvolution

This experiment involves the use of transcription factors (TFs) to drive the reprogramming of induced pluripotent stem cells (iPSCs) into megakaryocytes (MKs). The experiment consists of 3 splits, with 3 conditions per split, resulting in a total of 27 combinations. In
15 each split, cells on beads are transduced with a different lentivirus carrying distinct transgenes. Only cells in one condition are transduced with a lentivirus carrying a TF involved in driving iPSC reprogramming into MK. This condition is referred to as the 'TF condition.' Each lentivirus in the TF conditions also carries a transgene for a fluorescent protein, which acts as a genetic tag. Cells in the other two conditions are transduced with
20 lentivirus carrying only the luciferase gene, which serves as a genetic tag for those conditions. Before each split, beads are washed, pooled, mixed, and split into 3 different conditions for the next round of lentivirus transduction. All conditions are cultured in the same medium, which does not promote differentiation into MK in the absence of TFs. Successful differentiation of iPSCs into MKs is driven only by the presence of TFs. By the
25 end of the experiment, cells are collected for scRNAseq. Cell identities and corresponding genetic tags are characterized. Theoretically, only cells that pass through all 3 TF conditions in the 3 splits are transduced by lentiviruses carrying all the necessary TFs for MK reprogramming. As these lentiviruses also carry distinct fluorescent protein tags, only MKs are expected to express all 3 fluorescent proteins. While some MKs may show fewer
30 than 3 fluorescent proteins due to detection thresholds, none are expected to express the luciferase gene.

Cell Subculture

The Human Episomal iPSC Line (Thermo Fisher Scientific, A18945) is maintained in Essential 8 Medium (Thermo Fisher Scientific, A1517001) in Vitronectin-coated vessels (Thermo Fisher Scientific, A14700) and passaged using Accutase (Thermo Fisher Scientific, A1110501). Y-27632 (ROCK inhibitor) (STEMCELL Technologies, 72304) is added to the medium for 1 day after thawing or passaging.

Differentiation Media

cells were cultured in the following media:

Day 1-2: E8 media

Day 2-4: E6, FGF2 20 ng/ml, BMP4 10 ng/ml, VEGF 50ng/ml

Day 4-11: Cellgenix SCGM, SCF 25 ng/ml, TPO 20 ng/ml

Day 11-17: SFEM, TPO 50 ng/ml, SCF 20 ng/ml, IL-6 10 ng/ml, IL-9 10 ng/ml, ETP 500 ng/ml, Arachidonic acid 5 μ M, Nicotinamide 2.5 mM

Seeding Cells onto Microcarrier Beads

On Day -1, hydrated and sterilized microcarrier beads (Plasticell) are dispensed into 100 mm square Petri dishes (Thermo Fisher Scientific, 103) at 4,000 beads per well. Sub-confluent (~80%) hiPSCs are dissociated using Accutase and seeded onto beads at 300 cells per bead in Essential 8 Medium with Y-27632.

Combinatorial Lentivirus Transduction on Beads

- **On Day 0**, before lentivirus transduction, the number of cells on the beads is counted using a hemacytometer to calculate the volume of each lentivirus at MOI 2. After replacing the medium with fresh Essential 8 medium, one condition is transduced with a lentivirus carrying TF1 and the enhanced green fluorescent protein (EGFP) gene (LV-TF1-EGFP). The other conditions are transduced with lentivirus carrying the luciferase gene (LV-LUC).
- **On Day 1**, viruses are removed, and cells are rinsed twice to eliminate residual viruses. Cells on microcarrier beads are pooled in fresh Essential 8 medium, mixed, and split into 3 conditions. One condition is transduced with lentivirus carrying TF2

and the dTomato gene (LV-TF2-dTomato), while the other conditions are transduced with LV-LUC.

- **On Day 2**, viruses are removed, and cells are rinsed twice to eliminate residual viruses. Cells on microcarrier beads are pooled in fresh Reprogramming Medium 1, mixed, and split into 3 conditions. One condition is transduced with lentivirus carrying TF3 and the LSSmOrange gene (LV-TF3-LSSmOrange), while the other conditions are transduced with LV-LUC.
- **On Day 3**, viruses are removed, and cells are rinsed twice before replacing the medium with fresh Reprogramming Medium 1. From Day 3 to Day 17, medium changes are performed as specified in Tables 1 and 2. A schematic of the experimental procedure is shown in Figure 10.

•

Single-Cell RNA Sequencing and Deconvolution

Cells are collected for scRNAseq on Day 17.

Results

The scRNAseq results reveal three clusters of cells expressing all three genetic tags (EGFP, dTomato, and LSSmOrange; Figures 11 and 12):

1. A cluster of CD41a⁺ CD235a⁻ cells, which are MKs
2. A cluster of CD41a⁺ CD42b⁺ CD235a⁻ cells, representing more mature MKs
3. A cluster of CD41a⁺ CD235a⁺ cells, which are bipotent progenitors

Conclusion

This experiment exemplifies MK reprogramming through combinatorial TF transduction incorporating genetic tags. It demonstrates a method to label cells on microcarrier beads for CombiCult experiments and trace TFs responsible for lineage-specific reprogramming using genetic tag deconvolution via scRNAseq. By screening a broad array of TF pathways and employing an effective labeling strategy, this approach facilitates the discovery of novel reprogramming protocols.

Example 3

Tagging of Cells Exposed to Transcription Factors (Single Cell on Monolayer)

This experiment demonstrates:

- Genetic tagging of single cells during a CombiCult experiment
- Split-pooling of single cells in a CombiCult experiment
- 5 • Deconvolution of genetic tags by single-cell RNA sequencing (scRNAseq)
- Tracing combinations of transcription factors (TFs) responsible for reprogramming cells into specific lineages through genetic tag deconvolution

This experiment uses transcription factors (TFs) to drive the reprogramming of induced pluripotent stem cells (iPSCs) into megakaryocytes (MKs). The experiment consists of 3
10 splits, with 3 conditions per split, resulting in 27 combinations. In each split, cells are transduced with a different lentivirus carrying distinct transgenes. Only cells in one condition are transduced with a lentivirus carrying a TF involved in driving iPSC reprogramming into MK. This condition is referred to as the 'TF condition.' Each lentivirus in the TF conditions also carries a transgene for a fluorescent protein, which acts as a
15 genetic tag. Cells in the other two conditions are transduced with lentivirus carrying only the luciferase gene, which serves as a genetic tag for those conditions. Before each split, single cells are pooled, mixed, and split into 3 different conditions for the next round of lentivirus transduction. All cells are cultured in the same medium, which does not promote differentiation into MK in the absence of TFs. Successful differentiation of iPSCs
20 into MKs is driven only by the presence of TFs. By the end of the experiment, cells are collected for scRNAseq. Cell identities and corresponding genetic tags are characterized. Theoretically, only cells that pass through all 3 TF conditions in the 3 splits are transduced by lentiviruses carrying all the necessary TFs for MK reprogramming. As these lentiviruses also carry distinct fluorescent protein tags, only MKs are expected to express all 3
25 fluorescent proteins. While some MKs may show fewer than 3 fluorescent proteins due to detection thresholds, none are expected to express the luciferase gene.

Detailed Procedure

- **On Day -1**, iPSCs are seeded into 3 wells of a multi-well culture plate. On Day 0,
30 cells in one well are transduced with lentivirus carrying TF1 and enhanced green fluorescent protein (EGFP) gene (LV-TF1-EGFP). Cells in the other two wells are transduced with lentivirus carrying the luciferase gene (LV-LUC) (Fig. 13).

- **On Day 1**, viruses are removed, and cells are rinsed twice to remove residual viruses. All cells are dissociated into single cells, pooled, mixed, and split evenly into 3 wells of a multi-well culture plate. Cells in one well are transduced with lentivirus carrying TF2 and the dTomato gene (LV-TF2-dTomato), while cells in the other two wells are transduced with LV-LUC (Fig. 13).
- **On Day 2**, viruses are removed, and cells are rinsed twice to remove residual viruses. All cells are dissociated into single cells, pooled, mixed, and split evenly into 3 wells of a multi-well culture plate. Cells in one well are transduced with lentivirus carrying TF3 and the LSSmOrange gene (LV-TF3-LSSmOrange), while cells in the other two wells are transduced with LV-LUC (Fig. 13).
- **On Day 3**, viruses are removed, and cells are rinsed twice to remove residual viruses before adding fresh culture media. An MOI of 2 is used for all lentivirus transductions.
- **On Day 17**, cells are collected for scRNAseq.

Results

Results are similar to those for Example 2.

Example 4

Tagging of iPSC and Cancer Cells

This experiment demonstrates:

- Genetic tagging of single cells during a CombiCult experiment
- Deconvolution of genetic tags by single-cell RNA sequencing (scRNAseq)

This experiment involves 3 cell lines of different types. In the first split, only one cell line is tagged with a lentivirus carrying a fluorescent protein gene. The other two cell lines are tagged with a lentivirus carrying the luciferase gene. The cells are then pooled and split into the second split, where only one condition is tagged with a fluorescent protein, and the other two conditions are tagged with luciferase. This process is repeated for the third split. All cells are cultured in the same medium (DMEM, anti-anti 1%, FBS 10%) to maintain survival without differentiation, as the cells already have distinct identities. By the end of the experiment, cells are collected for scRNAseq. Theoretically, only one cell line or type carries all 3 fluorescent protein tags. All cells expressing all 3 fluorescent

proteins are expected to belong to the cell line or type tagged with a fluorescent protein gene in the first split.

Detailed Procedure

- 5 • **On Day -1**, the A549 lung cancer cell line, T47D breast cancer cell line, and iPSC line are seeded onto culture plates. On Day 1, A549 cells are transduced with a lentivirus carrying the green fluorescent protein (GFP) gene (LV-GFP) (Fig. 13). T47D cells and iPSCs are transduced with lentivirus carrying the luciferase gene (LV-LUC) (Fig. 14).
- 10 • **On Day 2**, viruses are removed, and cells are rinsed twice to remove residual viruses. All cells are dissociated into single cells, pooled, mixed, and split evenly into 3 wells of a multi-well culture plate. Cells in one well are transduced with lentivirus carrying the blue fluorescent protein (BFP) gene (LV-BFP), while cells in the other two wells are transduced with LV-LUC (Fig. 14).
- 15 • **On Day 3**, viruses are removed, and cells are rinsed twice to remove residual viruses. All cells are dissociated into single cells, pooled, mixed, and split evenly into 3 wells of a multi-well culture plate. Cells in one well are transduced with lentivirus carrying the dTomato gene (LV-dTomato), while cells in the other two wells are transduced with LV-LUC (Fig. 14).
- 20 • **On Day 4**, viruses are removed, and cells are rinsed twice to remove residual viruses before adding fresh media. An MOI of 10 is used for all transductions.
- **On Day 6**, cells are collected for scRNAseq.

Results

- 25 Figure 15 shows clustering of single cells using markers for iPSCs and cancer cells as detection probes. The distribution of the tested cell populations is clear.
- In Figure 16, genetic tags are screened, and the cells are clustered as in Figure 17. The tags segregate with the cells clustered according to cell type markers as anticipated, showing that individual cells can be successfully tagged genetically and followed through
- 30 a split-pooling procedure. GFP is only visualised in A549 cells, while following split-pooling, the BFP and RFP tags are distributed amongst all cell types.

The experiment successfully demonstrates that single cells from different cell lines can be tagged and tracked through multiple reagent conditions, providing a method to distinguish and follow individual cells in mixed populations.

Across all examples, the experiments effectively demonstrate the ability to:

- 5 • **Tag Single Cells:** Using lentiviruses carrying fluorescent protein genes or luciferase, individual cells are uniquely marked at each reagent condition.
- **Track Through Multiple Conditions:** The sequential introduction of tags allows for the reconstruction of each cell's exposure history to different reagent conditions.
- **Deconvolute Genetic Tags:** scRNAseq enables the identification of genetic tags
10 within single cells, facilitating the analysis of how specific combinations of conditions affect cell fate.
- **Understand Differentiation Pathways:** By correlating genetic tags with differentiation outcomes, we gain insights into the necessary conditions and factors that drive specific lineage commitments.
- 15 These methodologies provide powerful tools for studying complex cellular processes and can be applied to a wide range of research areas in cell biology and regenerative medicine.

Claims

1. A method for determining the conditions required for promoting the conversion of a first cell type to a second cell type, by analysing one or more markers in a single cell which has been exposed to combinations of reagents.
5
2. A method according to claim 1 comprising the steps of:
 - a. exposing a single cell of a first cell type to a first reagent or combination of reagents;
10
 - b. optionally, removing the first reagent and exposing the cell to a different reagent or combination of reagents;
 - c. optionally, repeating the foregoing step (b); and
 - d. identifying one or more cells which displays one or more markers of a second cell type and determining the identity and sequence of reagents to which the cell has been exposed.
15
3. A method according to claim 1 or claim 2, wherein the cells are labelled to identify exposure to each TF or combination of reagents.
- 20 4. A method according to claim 3, which comprises the steps of:
 - a. Subdividing a plurality of single cells into a first set of groups of single cells, separately exposing the groups of single cells to different reagents or combinations of reagents and labelling the single cells to indicate exposure to said reagents or combinations of reagents;
25
 - b. pooling the groups of single cells, and subdividing the pool of cells to form a second set of groups of single cells;
 - c. separately exposing the second set of groups of single cells to different reagents or combinations of reagents;
 - d. optionally iteratively repeating steps (a) to (c) as required;
 - 30 e. identifying cell type markers in the the cells and deconvoluting the labels to identify the identity, sequence and timing of reagents to which the cells have been exposed;
 - f. optionally, using the data from step (e) to inform the choice of reagents used in step (a), and repeating steps (a) to (e) as required;
 - 35 g. identifying a cell of the second cell type from the cell type markers, and deconvoluting the labels to thereby identify the identity and sequence of reagents required to reprogram a cell of the first cell type into a cell of the second cell type.
- 40 5. A method according to claim 4, wherein sequential method steps in a procedure comprise either the addition of single reagents or the addition of combinations of reagents.
- 45 6. A method according to claim 4, wherein sequential method steps in a procedure only comprise the addition of individual reagents.

7. A method according to claim 2 or claims 4-6, wherein the TFs or combinations of TFs used in the first iteration of step (a) are selected from reagents known in the art to promote cell type conversion.
- 5 8. A method according to claim 7, wherein the reagents or combinations of reagents used in the first iteration of step (c) are selected from reagents known in the art to promote cell type conversion.
- 10 9. A method according to claim 2 or claims 4-8, wherein the reagents or combinations of reagents used in second and further iterations of steps (a) and (c) are selected by analysis of the data produced by step (e).
- 15 10. A method according to claim 10, wherein the data produced in step (e) are processed by an algorithm which analyses the effect of the identity and timing of administration of each reagent on the conversion of the first cell type.
- 20 11. A method according to claims 4-10, wherein the reagents or combinations of reagents used in the first iteration of step (a) are selected using data produced by step(e) in a previous performance of the method.
- 25 12. A method according to any one of claims 3-11, wherein the cells are labelled by modifying the nucleic acid of the cell.
- 30 13. A method according to claim 12, wherein the nucleic acid of the cell is modified by genome editing or by viral integration.
- 35 14. A method according to claim 13, wherein the genome editing comprises CRISPR.
- 40 15. A method according to claim 12, wherein data derived from the fate of cells exposed to different reagents are combined to correlate transcriptional changes at the single cell level with specific cell culture conditions, to generate a model which predicts the culture conditions required to generate specific cell types.
- 45 16. A method according to claim 15, wherein the model is used to generate initial reagent choices for steps (a) to (c), the iterative performance of the steps resulting in an improved model.
17. A method according to any preceding claim, wherein conversion of a cell from a first cell type to a second cell type is a procedure selected from cell differentiation, dell reprogramming, cell transdifferentiation and cell retrodifferentiation.
18. A method for identifying a gene which influences a cellular process, comprising the steps of:
 - a) determining the effect of one or more reagents or combinations of reagents on a single cell, in accordance with any one of the preceding claims;
 - b) analysing gene expression in said cells when exposed to said reagents or combinations of reagents; and

c) identifying genes which are differentially expressed when exposed to said reagents or combinations of reagents.

5 19. A method for producing a nucleic acid which encodes a gene product which influences a cellular process, comprising identifying a gene in accordance with claim 18, and producing at least the coding region of said gene by nucleic acid synthesis or biological replication.

10 20. A method for inducing a cellular process, comprising the steps of:

- a) identifying one or more genes which are differentially expressed in association with the cellular process in accordance with claim 18; and
- b) modulating the expression of said one or more genes in the cell.

15 21. A method according to claim 20, wherein modulation of gene expression in the cell comprises transfection of said one or more genes into the cell.

22. A method for selecting the starting reagents in a method according to claim 4, comprising the steps of:

20 compiling a database of reagents and cell conversion results; and

selecting from the database a set of reagents or combinations of reagents for use in a method according to claim 4.

25 23. A method for improving a database of reagents and cell conversion results, comprising using the database to select reagents according to claim 22, and inputting the results of the method according to claim 4 into the database, thus providing further data points.

30 24. A method according to any preceding claim, wherein the reagents are transcription factors (TFs).

25 25. A system comprising a database of TFs useful for converting cells from a first cell type to a second cell type, at least one computer processor and a memory in operative communication with the processor, the memory comprising instructions configuring the processor to:

- 35
- (a) generate gene regulatory networks (GRN) from the database of transcription factors;
 - (b) determine a candidate transcription factor for a desired conversion event;

- (c) analyse an impact of the candidate transcription factor in cell reprogramming;
- (d) optionally, iteratively repeat steps (a)-(c); and
- output a set of optimal transcription factors.

5

26. The system according to claim 25, wherein the database comprises gene expression data.

10

27. A method for determining optimal transcription factors for cell reprogramming, comprising: using a computing device, curating a gene expression database related to use of TFs in cell culture; generating gene regulatory networks from a plurality of gene expression datasets; determining a candidate optimal TF; analysing an impact of the optimal TF in cell differentiation; and outputting a set of optimal TFs.

15