Claims

1. A computing device for classifying the differentiation state of an *in vitro* population of cells, the device comprising a memory that comprises:

a first reference dataset that comprises a representation of gene expression levels for one or more genes that are differentially expressed between cells at a first differentiation state and cells at a second differentiation state; and

a second reference dataset that comprises a representation of gene expression levels for one or more genes that are differentially expressed between cells at the second differentiation state and cells at a third differentiation state.

2. The computing device of claim 1, further comprising a processor that implements instructions stored in the memory to perform a method comprising:

(a) receiving as input a test dataset that comprises expression levels for genes that are expressed in one or more test cells comprised in an *in vitro* population of cells, wherein the expression levels in the test dataset comprise expression levels for (i) one or more of the genes for which a representation of expression levels are included in the first reference dataset, and (ii) one or more of the genes for which a representation of expression levels are included in the second reference dataset;

(b) calculating, using the test dataset and the first reference dataset, a first similarity score indicating whether the differentiation state of the test cells is more similar to the first differentiation state or to the second differentiation state;

(c) calculating, using the test dataset and the second reference dataset, a second similarity score indicating whether the differentiation state of the test cells is more similar to the second differentiation state or to the third differentiation state; and

(d) classifying the differentiation state of the one or more test cells based on one or both of the first similarity score and the second similarity score.

3. The computing device of claim 1 or claim 2, wherein the memory further comprises a control dataset that comprises a representation of gene expression levels for one or more genes that are expressed in cells at one or more control differentiation states, which control differentiation state may be the same as or different than one of the first, second, or third differentiation states.

4. The computing device of claim 3, wherein:

the test dataset comprises gene expression levels for one or more of the genes for which a representation of expression levels are included in the control dataset;

the instructions comprise calculating a degree of correlation between the representation of gene expression levels for one or more genes in the control dataset and gene expression levels for the one or more genes in the test dataset to calculate a correlation score; and

the classifying the differentiation state of the one or more test cells is based on the correlation score and the one or both of the first similarity score and the second similarity score.

5. The computing device of claim 4, wherein the correlation score is calculated prior to calculating the first similarity score and the second similarity score, and the method is terminated if the correlation score for the test cells does not meet a predefined cutoff value.

6. The computing device of any one of claims 3-5, wherein the control dataset comprises gene expression levels that are normalized by counts per million mapped reads (CPM) and filtered to include only gene expression levels that exceed a threshold CPM value.

7. The computing device of any one of claims 3-6, wherein the control dataset comprises a centroid of gene expression levels of the one or more genes in the control dataset.

8. The computing device of claim 7, wherein the correlation score is calculated by normalizing the gene expression levels of the one or more genes in the test dataset and calculating a correlation of the gene expression levels of the one or more genes in the test dataset to the centroid.

9. The computing device of claim 8, wherein the control dataset comprises coefficient of variation (CV) values of gene expression levels of the one or more genes in the control dataset, and the correlation to the centroid is weighted by the inverse of the CV values.

10. The computing device of any one of claims 1-9, wherein the population of cells is from a culture of cells differentiated from pluripotent cells that are subjected to suitable differentiation conditions.

11. The computing device of any one of claims 1-10, wherein the first differentiation state is earlier in a stem cell differentiation pathway than the second differentiation state.

12. The computing device of any one of claims 1-11 wherein the second differentiation state is earlier in a stem cell differentiation pathway than the third differentiation state.

13. The computing device of any one of claims 1-10, wherein the first differentiation state is in a cell differentiation pathway that is parallel to a cell differentiation pathway of the second differentiation state.

14. The computing device of any one of claims 1-13, wherein the population of cells are selected from the group consisting of stem-cell derived cardiac muscle cells, stem-cell derived skeletal muscle cells, stem-cell derived kidney tubule cells, stem-cell derived red blood cell cells, stem-cell derived smooth muscle cells, stem-cell derived lung cells, stem-cell derived thyroid cells, stem-cell derived pancreatic cells, stem-cell derived epidermal cells, stem-cell derived pigment cells, and stem-cell derived neuronal cells.

15. The computing device of any one of claims 1-14, wherein the population of cells are stem-cell derived neuronal cells.

16. The computing device of any one of claims 1-15, wherein the second differentiation state is the differentiation state of a determined dopaminergic neuronal cell or a hematopoietic progenitor cell.

17. The computing device of any one of claims 1-16, wherein the first reference dataset comprises a representation of gene expression levels for one or more genes, or for at least 20 genes, selected from Table E1.

18. The computing device of any one of claims 1-17, wherein the second reference dataset comprises a representation of gene expression levels for one or more genes, or for at least 20 genes, selected from Table E2.

19. The computing device of any one of claims 1-18, wherein at least one of the first, second, and third differentiation states is characterized using an *in vitro* assay or an *in vivo* assay.

20. The computing device of claim 19, wherein the *in vivo* assay comprises determining whether reference cells are capable of surviving, engrafting, and/or innervating tissue or ameliorating or reversing symptoms of a neurodegenerative disease when administered to an animal or human subject.

21. The computing device of any one of claims 1-20, wherein the memory further comprises one or more additional reference datasets, wherein each of the additional reference datasets comprises a representation of gene expression levels for one or more genes that are differentially expressed between cells at the second differentiation state and cells at an additional differentiation state, wherein:

the processor implements instructions to calculate, using the additional reference datasets, one or more additional similarity scores indicating whether the differentiation state of the test cells is more similar to the second differentiation state or to one of the one or more additional differentiation states, and

the classifying the differentiation state of the one or more test cells is based on the first similarity score, the second similarity score, and the one or more additional similarity scores.

22. The computing device of any one of claims 1-21, wherein the representations of gene expression levels in the first reference dataset and/or the second reference dataset are obtained using machine learning.

23. The computing device of claim 22, wherein the machine learning comprises principal component analysis.

24. The computing device of any one of claims 1-23, wherein the representations of gene expression levels in the first reference dataset and/or the second reference dataset comprise normalized gene expression levels.

25. The computing device of any one of claims 1-24, wherein the differentiation state of the one or more test cells is classified as being the second differentiation state if the

one or both of the first and second similarity scores indicate that the differentiation state of the one or more test cells is more similar to the second differentiation state.

26. A method for selecting a population of cells having a desired differentiation state, the method comprising:

(a) calculating a first similarity score using a test dataset and a first reference dataset, wherein:

the first reference dataset comprises a representation of gene expression levels for one or more genes that are differentially expressed between cells at a first differentiation state and cells at a second differentiation state,

the test dataset comprises expression levels for genes that are expressed in one or more test cells comprised in an *in vitro* population of cells, wherein the expression levels in the test dataset comprise expression levels for one or more of the genes for which a representation of expression levels are included in the first reference dataset, and

the first similarity score indicates whether the differentiation state of the test cells is more similar to the first differentiation state or to the second differentiation state;

(b) calculating a second similarity score using the test dataset and a second reference dataset, wherein:

the second reference dataset comprises a representation of gene expression levels for one or more genes that are differentially expressed between cells at the second differentiation state and cells at a third differentiation state,

the expression levels in the test dataset comprise expression levels for one or more of the genes for which a representation of expression levels are included in the second reference dataset, and

the second similarity score indicates whether the differentiation state of the test cells is more similar to the second differentiation state or to the third differentiation state; and

(c) classifying the differentiation state of the one or more test cells based on one or both of the first similarity score and the second similarity score.

27. The method of claim 26, wherein:

the test dataset comprises gene expression levels for one or more genes for which a representation of expression levels are included in a control dataset that comprises a representation of gene expression levels for one or more genes that are expressed in cells at a control differentiation state, which control differentiation state may be the same as or different than one of the first, second, or third differentiation states;

the method further comprises calculating a degree of correlation between the representation of gene expression levels for one or more genes in the control dataset and gene expression levels for the one or more genes in the test dataset to calculate a correlation score; and

the classifying the differentiation state of the one or more test cells is based on the correlation score and the one or both of the first similarity score and the second similarity score.

28. The method of claim 27, wherein the correlation score is calculated prior to calculating the first similarity score and the second similarity score and the method is terminated if the correlation score for the test cells does not meet a predefined cutoff value.

29. The method of claim 27 or claim 28, wherein the control dataset comprises gene expression levels that are normalized by counts per million mapped reads (CPM) and filtered to include only gene expression levels that exceed a threshold CPM value.

30. The method of any one of claims 27-29, wherein the control dataset comprises a centroid of gene expression levels of the one or more genes in the control dataset.

31. The method of claim 30, wherein the correlation score is calculated by normalizing the gene expression levels of the one or more genes in the test dataset and calculating a correlation of the gene expression levels of the one or more genes in the test dataset to the centroid.

32. The method of claim 38, wherein the control dataset comprises coefficient of variation (CV) values of gene expression levels of the one or more genes in the control dataset, and the correlation to the centroid is weighted by the inverse of the CV values.

33. The method of any one of claims 26-32, wherein the first differentiation state is earlier in a stem cell differentiation pathway than the second differentiation state.

34. The method of any one of claims 26-33, wherein the second differentiation state is earlier in a stem cell differentiation pathway than the third differentiation state.

35. The method of any one of claims 26-32, wherein the first differentiation state is in a cell differentiation pathway that is parallel to a cell differentiation pathway of the second differentiation state.

36. The method of any one of claims 26-35, wherein the population of cells are selected from the group consisting of stem-cell derived cardiac muscle cells, stem-cell derived skeletal muscle cells, stem-cell derived kidney tubule cells, stem-cell derived red blood cell cells, stem-cell derived smooth muscle cells, stem-cell derived lung cells, stem-cell derived thyroid cells, stem-cell derived pancreatic cells, stem-cell derived epidermal cells, stem-cell derived pigment cells, and stem-cell derived neuronal cells.

37. The method of any one of claims 26-36, wherein the population of cells are stem-cell derived neuronal cells.

38. The method of any one of claims 26-37, wherein the second differentiation state is the differentiation state of a determined dopaminergic neuronal cell or a hematopoietic progenitor cell.

39. The method of any one of claims 26-38, wherein the first reference dataset comprises a representation of gene expression levels for one or more genes, or for at least 20 genes, selected from Table E1.

40. The method of any one of claims 26-39, wherein the second reference dataset comprises a representation of gene expression levels for one or more genes, or for at least 20 genes, selected from Table E2.

41. The method of any one of claims 26-40, wherein at least one of the first, second, and third differentiation states is characterized using an *in vitro* assay or an *in vivo* assay.

42. The method of claim 41, wherein the *in vivo* assay comprises determining whether reference cells are capable of surviving, engrafting, and/or innervating tissue or ameliorating or reversing symptoms of a neurodegenerative disease when administered to an animal or human subject.

43. The method of any one of claims 26-42, wherein the method further comprises calculating one or more additional similarity scores using one or more additional reference datasets, wherein:

each of the additional reference datasets comprises a representation of gene expression levels for one or more genes that are differentially expressed between cells at the second differentiation state and cells at an additional differentiation state;

the one or more additional similarity scores indicate whether the differentiation state of the test cells is more similar to the second differentiation state or to one of the one or more additional differentiation states, and

the classifying the differentiation state of the one or more test cells is based on the first similarity score, the second similarity score, and the one or more additional similarity scores.

44. The method of any one of claims 26-43, wherein the representations of gene expression levels in the first reference dataset and/or the second reference dataset are obtained using machine learning.

45. The method of claim 44, wherein the machine learning comprises principal component analysis.

46. The method of any one of claims 26-43, wherein the representations of gene expression levels in the first reference dataset and/or the second reference dataset comprise normalized gene expression levels.

47. The method of any one of claims 26-46, wherein the method further comprises classifying the differentiation state of the one or more test cells as being the second differentiation state if the one or both of the first and second similarity scores indicate that the differentiation state of the one or more test cells is more similar to the second differentiation state.

48. The method of any one of claims 26-47, wherein the method further comprises selecting the *in vitro* population of cells comprising one or more test cells classified as having the second differentiation state as having the desired differentiation state.

49. A method for training a machine learning model classifying the differentiation state of an *in vitro* population of cells, the method comprising:

(a) obtaining, for a plurality of reference populations of cells, gene expression levels for one or more genes that are differentially expressed between cells at a first differentiation state and cells at a second differentiation state and applying the gene expression levels as input to train a first machine learning model to predict if an *in vitro* population of cells comprises one or more test cells having a differentiation state that is more similar to the first differentiation state or to the second differentiation state; and

(b) obtaining, for a plurality of reference populations of cells, gene expression levels for one or more genes that are differentially expressed between cells at the second differentiation state and cells at a third differentiation state and applying the gene expression levels as input to train a second machine learning model to predict if an *in vitro* population of cells comprises one or more test cells having a differentiation state that is more similar to the second differentiation state or to the third differentiation state.

50. A method for training a machine learning model classifying the differentiation state of an *in vitro* population of cells, the method comprising:

(a) selecting one or more genes that are differentially expressed between cells at a first differentiation state and cells at a second differentiation state and applying expression levels of the selected genes for a plurality of reference populations of cells as input to train a first machine learning model to predict if an *in vitro* population of cells comprises one or more test cells having a differentiation state that is more similar to the first differentiation state or to the second differentiation state; and

(b) selecting one or more genes that are differentially expressed between cells at the second differentiation state and cells at a third differentiation state and applying expression levels of the selected genes for a plurality of reference populations of cells as input to train a second machine learning model to predict if an *in vitro* population of cells comprises one or more test cells having a differentiation state that is more similar to the second differentiation state or to the third differentiation state.