

CLAIMS

1. A method comprising:
using at least one computer hardware processor to perform:
 - 5 obtaining expression data for a biological sample, the biological sample previously obtained from a subject, the expression data including first expression data associated with a first set of genes associated with a first cell type;
determining a first cell composition percentage for the first cell type using the expression data and one or more non-linear regression models including a first non-
10 linear regression model, wherein the first cell composition percentage indicates an estimated percentage of cells of the first cell type in the biological sample, wherein determining the first cell composition percentage for the first cell type comprises:
processing the first expression data with the first non-linear regression
model to determine the first cell composition percentage for the first cell type;
15 and
outputting the first cell composition percentage.
2. The method of claim 1, wherein the subject has, is suspected of having, or is at risk of
20 having cancer.
3. The method of claim 1 or any other preceding claim, wherein the expression data is
RNA expression data.
4. The method of claim 3 or any other preceding claim, wherein processing the first
25 expression data with the first non-linear regression model comprises:
providing the first expression data as input to the first non-linear regression model to
obtain a corresponding output representing an estimated percentage of RNA from the first
cell type; and
determining, based on the estimated percentage of RNA from the first cell type, the
30 first cell composition percentage for the first cell type.
5. The method of claim 4 or any other preceding claim, wherein:
the expression data includes second expression data associated with the first set of
genes associated with the first cell type; and

the first non-linear regression model comprises:

a first sub-model configured to generate, using the first expression data as input, a first value for the estimated percentage of RNA from the first cell type; and
a second sub-model configured to generate, using the second expression data
5 and the first value for the estimated percentage of RNA from the first cell type as input, a second value for the estimated percentage of RNA from the first cell type.

6. The method of claim 1 or any other preceding claim,
wherein the expression data includes second expression data associated with a second
10 set of genes associated with a second cell type different from the first cell type,
wherein the one or more non-linear regression models comprise a second non-linear regression model, and
wherein the method further comprises determining a second cell composition
percentage for the second cell type at least in part by processing the second expression data
15 with the second non-linear regression model to determine the second cell composition percentage for the second cell type.

7. The method of claim 1 or any other preceding claim, wherein the first cell type is selected from the group consisting of B cells, CD4+ T cells, CD8+ T cells, endothelial cells,
20 fibroblasts, lymphocytes, macrophages, monocytes, NK cells, neutrophils, and T cells.

8. The method of claim 1 or any other preceding claim, wherein the first expression data includes expression data for at least 10 genes selected from the group of genes for the first cell type in Table 2.

25

9. The method of claim 1 or any other preceding claim,
wherein the expression data includes expression data associated with a plurality of gene sets associated with a respective plurality of cell types, the plurality of gene sets including the first gene set and the plurality of cell types including the first cell type,
30 wherein the one or more non-linear regression models comprise a plurality of non-linear regression models, and
wherein the method further comprises determining a plurality of cell composition percentages for the plurality of cell types using the expression data associated with the plurality of gene sets, the plurality of cell composition percentages including the first cell

composition percentage, wherein determining the plurality of cell composition percentages comprises:

5 for each cell type of the plurality of cell types, determining a respective cell composition percentage for the cell type at least in part by processing expression data associated with a set of genes associated with the cell type using a respective non-linear regression model of the plurality of non-linear regression models to determine the cell composition percentage for the cell type.

10 10. The method of claim 9 or any other preceding claim, wherein genes in the plurality of gene sets include at least 25 genes selected from the group of genes in Table 2, and wherein determining the plurality of cell composition percentages comprises processing expression data for the at least 25 genes.

15 11. The method of claim 9 or any other preceding claim, wherein genes in the plurality of gene sets include at least 35 genes selected from the group of genes in Table 2, and wherein determining the plurality of cell composition percentages comprises processing expression data for the at least 35 genes.

20 12. The method of claim 9 or any other preceding claim, wherein genes in the plurality of gene sets include at least 50 genes selected from the group of genes in Table 2, and wherein determining the plurality of cell composition percentages comprises processing expression data for the at least 50 genes.

25 13. The method of claim 9 or any other preceding claim, wherein genes in the plurality of gene sets include at least 75 genes selected from the group of genes in Table 2, and wherein determining the plurality of cell composition percentages comprises processing expression data for the at least 75 genes.

30 14. The method of claim 9 or any other preceding claim, wherein genes in the plurality of gene sets include at least 100 genes selected from the group of genes in Table 2, and wherein determining the plurality of cell composition percentages comprises processing expression data for the at least 100 genes.

15. The method of claim 1 or any other preceding claim, wherein the one or more non-linear regression models comprise one or more random forest regression models.
16. The method of claim 1 or any other preceding claim, wherein the one or more non-linear regression models comprise one or more neural network regression models.
17. The method of claim 1 or any other preceding claim, wherein the one or more non-linear regression models comprise one or more support vector machine regression models.
18. The method of claim 1 or any other preceding claim, wherein the first non-linear regression model was trained at least in part by:
obtaining simulated expression data; and
training the first non-linear regression model using the simulated expression data.
19. The method of claim 18 or any other preceding claim, further comprising:
obtaining the simulated expression data; and
training the first non-linear regression model using the simulated expression data.
20. The method of claim 18 or any other preceding claim, wherein obtaining the simulated expression data comprises generating the simulated expression data, wherein generating the simulated expression data comprises:
obtaining a set of RNA expression data from one or more biological samples, the set of RNA expression data comprising microenvironment cell expression data and malignant cell expression data;
generating simulated microenvironment cell expression data using the microenvironment cell expression data;
generating simulated malignant cell expression data using the malignant cell expression data; and
combining the simulated microenvironment cell expression data and the simulated malignant cell expression data to produce at least a part of the simulated expression data.
21. The method of claim 1 or any other preceding claim, further comprising:
determining a malignancy expression profile using an expression profile for the first cell type and the first cell composition percentage for the first cell type.

22. The method of claim 1 or any other preceding claim, wherein the first non-linear regression model has been trained by:

obtaining training data comprising simulated RNA expression data, the simulated
5 RNA expression data including first RNA expression data for the first set of genes associated with the first cell type;

training the first non-linear regression model to estimate a percentage of RNA from the first cell type, the training comprising:

generating, using the first non-linear regression model and the first RNA
10 expression data, an estimated percentage of RNA from the first cell type; and
updating parameters of the first non-linear regression model using the estimated percentage of RNA from the first cell type.

23. A system, comprising:

15 at least one hardware processor; and

at least one non-transitory computer-readable storage medium storing processor-executable instructions that, when executed by the at least one hardware processor, cause the at least one hardware processor to perform:

obtaining expression data for a biological sample, the biological sample
20 previously obtained from a subject, the expression data including first expression data associated with a first set of genes associated with a first cell type;

determining a first cell composition percentage for the first cell type using the expression data and one or more non-linear regression models including a first non-linear regression mode, wherein the first cell composition percentage indicates an
25 estimated percentage of cells of the first cell type in the biological sample,

wherein determining the first cell composition percentage for the first cell type comprises:

processing the first expression data with the first non-linear regression model to determine the first cell composition percentage for the first cell type;
30 and

outputting the first cell composition percentage.

24. At least one non-transitory computer-readable storage medium storing processor-executable instructions that, when executed by at least one hardware processor, cause the at least one hardware processor to perform:

obtaining expression data for a biological sample, the biological sample previously
5 obtained from a subject, the expression data including first expression data associated with a first set of genes associated with a first cell type;

determining a first cell composition percentage for the first cell type using the
expression data and one or more non-linear regression models including a first non-linear
regression model, wherein the first cell composition percentage indicates an estimated
10 percentage of cells of the first cell type in the biological sample,

wherein determining the first cell composition percentage for the first cell type
comprises:

processing the first expression data with the first non-linear regression model
to determine the first cell composition percentage for the first cell type; and
15 outputting the first cell composition percentage.

25. A method, comprising:

using at least one computer hardware processor to perform:

obtaining RNA expression data for a biological sample, the biological sample
20 previously obtained from a subject having, suspected of having, or at risk of having
cancer,

wherein the RNA expression data includes first RNA expression data
associated with a first set of genes associated with a first cell type,

wherein the first RNA expression data includes expression data for at
least 10 genes selected from the group of genes for the first cell type in Table
25 2,

wherein the first cell type is selected from the group consisting of B
cells, CD4+ T cells, CD8+ T cells, endothelial cells, fibroblasts, lymphocytes,
macrophages, monocytes, NK cells, neutrophils, and T cells; and

30 determining a first cell composition percentage for the first cell type, using the first
RNA expression data, the first cell composition percentage indicating an estimated
percentage of cells of the first cell type in the biological sample, wherein determining
the first cell composition percentage for the first cell type comprises:

providing the first RNA expression data as input to a first non-linear regression model to obtain a corresponding output representing an estimated percentage of RNA from the first cell type; and

5 determining, based on the estimated percentage of RNA from the first cell type, the first cell composition percentage for the first cell type.

26. The method of claim 25, wherein:

the RNA expression data includes second RNA expression data associated with the first set of genes associated with the first cell type; and

10 the first non-linear regression model comprises:

a first sub-model configured to generate, using the first RNA expression data as input, a first value for the estimated percentage of RNA from the first cell type; and

a second sub-model configured to generate, using the second expression data and the first value for the estimated percentage of RNA from the first cell type as input, a second value for the estimated percentage of RNA from the first cell type.

27. The method of claim 25 or any other preceding claim

wherein the RNA expression data includes second RNA expression data associated with a second set of genes associated with a second cell type,

20 wherein the second RNA expression data includes expression data for at least 10 genes selected from the group of genes for the second cell type in Table 2,

wherein the second cell type is selected from the group consisting of B cells, CD4+ T cells, CD8+ T cells, endothelial cells, fibroblasts, lymphocytes, macrophages, monocytes, NK cells, neutrophils, and T cells; and

25 wherein determining a second cell composition percentage for the second cell type comprises:

processing the second RNA expression data with a second non-linear regression model to determine the second cell composition percentage for the second cell type.

30

28. The method of claim 25 or any other preceding claim,

wherein the RNA expression data includes RNA expression data associated with a plurality of gene sets associated with a respective plurality of cell types, the plurality of gene sets including a first gene set and the plurality of cell types including the first cell type;

wherein the method further comprises determining a plurality of cell composition percentages for the plurality of cell types using the RNA expression data associated with the plurality of gene sets, the plurality of cell composition percentages including the first cell composition percentage, wherein determining the plurality of cell composition percentages
5 comprises:

for each cell type of the plurality of cell types, determining a respective cell composition percentage for the cell type at least in part by processing RNA expression data associated with a set of genes associated with the cell type using a respective non-linear regression model to determine the cell composition percentage for the cell type.

10

29. The method of claim 25 or any other preceding claim, wherein the first non-linear regression model comprises a random forest regression model.

30. The method of claim 25 or any other preceding claim, wherein the first non-linear
15 regression model comprises a neural network regression model.

31. The method of claim 25 or any other preceding claim, wherein the first non-linear regression model comprises a support vector machine regression model.

20 32. The method of claim 25 or any other preceding claim, wherein the first non-linear regression model was trained at least in part by generating training data comprising simulated RNA expression data, wherein generating the training data comprises:

obtaining a set of RNA expression data from one or more biological samples, the set of RNA expression data comprising microenvironment cell RNA expression data and
25 malignant cell RNA expression data;

generating simulated microenvironment cell RNA expression data using the microenvironment cell RNA expression data;

generating simulated malignant cell RNA expression data using the malignant cell RNA expression data; and

30 combining the simulated microenvironment cell RNA expression data and the simulated malignant cell RNA expression data to produce at least a part of the simulated RNA expression data.

33. The method of claim 25 or any other preceding claim, further comprising:

determining a malignancy expression profile using an RNA expression profile for the first cell type and the first cell composition percentage for the first cell type.

34. The method of claim 25, or any other preceding claim, wherein the first RNA
5 expression data includes expression data for at least 25 genes selected from the group of genes in Table 2.

35. The method of claim 25 or any other preceding claim, wherein the first RNA
10 expression data includes expression data for at least 50 genes selected from the group of genes in Table 2.

36. The method of claim 25 or any other preceding claim, wherein the first RNA
expression data includes expression data for at least 100 genes selected from the group of
15 genes in Table 2.

37. The method of claim 25 or any other preceding claim, wherein the first non-linear
regression model has been trained by:

obtaining training data comprising simulated RNA expression data, the simulated
RNA expression data including second RNA expression data for the first set of genes
20 associated with the first cell type;

training the first non-linear regression model to estimate a percentage of RNA from
the first cell type, the training comprising:

generating, using the first non-linear regression model and the second RNA
expression data, an estimated percentage of RNA from the first cell type; and
25 updating parameters of the first non-linear regression model using the
estimated percentage of RNA from the first cell type.

38. A system, comprising:
at least one hardware processor; and
30 at least one non-transitory computer-readable storage medium storing processor-
executable instructions that, when executed by the at least one hardware processor, cause the
at least one hardware processor to perform:

obtaining RNA expression data for a biological sample, the biological sample previously obtained from a subject having, suspected of having, or at risk of having cancer,

wherein the RNA expression data includes first RNA expression data associated with a first set of genes associated with a first cell type,

wherein the first RNA expression data includes expression data for at least 10 genes selected from the group of genes for the first cell type in Table 2,

wherein the first cell type is selected from the group consisting of B cells, CD4+ T cells, CD8+ T cells, endothelial cells, fibroblasts, lymphocytes, macrophages, monocytes, NK cells, neutrophils, and T cells; and

determining a first cell composition percentage for the first cell type, using the first RNA expression data, the first cell composition percentage indicating an estimated percentage of cells of the first cell type in the biological sample, wherein determining the first cell composition percentage for the first cell type comprises:

providing the first RNA expression data as input to a first non-linear regression model to obtain a corresponding output representing an estimated percentage of RNA from the first cell type; and

determining, based on the estimated percentage of RNA from the first cell type, the first cell composition percentage for the first cell type.

39. At least one non-transitory computer-readable storage medium storing processor-executable instructions that, when executed by at least one hardware processor, cause the at least one hardware processor to perform:

obtaining RNA expression data for a biological sample, the biological sample previously obtained from a subject having, suspected of having, or at risk of having cancer,

wherein the RNA expression data includes first RNA expression data associated with a first set of genes associated with a first cell type,

wherein the first RNA expression data includes expression data for at least 10 genes selected from the group of genes for the first cell type in Table 2,

wherein the first cell type is selected from the group consisting of B cells, CD4+ T cells, CD8+ T cells, endothelial cells, fibroblasts, lymphocytes, macrophages, monocytes, NK cells, neutrophils, and T cells; and

determining a first cell composition percentage for the first cell type, using the first RNA expression data, the first cell composition percentage indicating an estimated percentage of cells of the first cell type in the biological sample, wherein determining the first cell composition percentage for the first cell type comprises:

- 5 providing the first RNA expression data as input to a first non-linear regression model to obtain a corresponding output representing an estimated percentage of RNA from the first cell type; and
- determining, based on the estimated percentage of RNA from the first cell type, the first cell composition percentage for the first cell type.

10

40. A method, comprising:

 using at least one computer hardware processor to perform:

- obtaining training data comprising simulated RNA expression data, the simulated RNA expression data including first RNA expression data for first genes
- 15 associated with a first cell type and second RNA expression data for second genes associated with a second cell type different from the first cell type; and
- training a plurality of non-linear regression models to estimate percentages of RNA from one or more respective cell types, the plurality of non-linear regression models comprising a first non-linear regression model for estimating percentage of
- 20 RNA from the first cell type and a second non-linear regression model for estimating percentage of RNA from the second cell type, wherein training the plurality of non-linear regression models comprises training the first non-linear regression model at least in part by:

- generating, using the first non-linear regression model and the first
- 25 RNA expression data, an estimated percentage of RNA from the first cell type; and
- and

- updating parameters of the first non-linear regression model using the estimated percentage of RNA from the first cell type; and
- outputting the trained plurality of non-linear regression models including the
- 30 first non-linear regression model and the second non-linear regression model.

41. The method of claim 40 or any other preceding claim, wherein obtaining the training data comprises generating at least a part of the simulated RNA expression data by:

obtaining a set of RNA expression data from one or more biological samples, the set of RNA expression data comprising microenvironment cell RNA expression data and malignant cell RNA expression data;

5 obtaining simulated microenvironment cell RNA expression data based on the microenvironment cell RNA expression data;

obtaining simulated malignant cell RNA expression data based on the malignant cell RNA expression data; and

10 combining the simulated microenvironment cell RNA expression data and the simulated malignant cell RNA expression data to produce the at least the part of the simulated RNA expression data.

42. The method of claim 41 or any other preceding claim, further comprising:
adding noise to the simulated RNA expression data prior to training the plurality of non-linear regression models.

15

43. The method of claim 42 or any other preceding claim, wherein the noise comprises at least one of Poisson noise or Gaussian noise.

44. The method of claim 41 or any other preceding claim, wherein generating the
20 simulated microenvironment cell RNA expression data comprises:
generating, for a first microenvironment cell type, a first RNA expression profile using a first portion of the microenvironment cell RNA expression data.

45. The method of claim 44 or any other preceding claim, wherein the first portion of the
25 microenvironment cell RNA expression data comprises RNA expression data from multiple subtypes of the first microenvironment cell type.

46. The method of claim 45 or any other preceding claim, wherein generating the first
RNA expression profile comprises resampling the first portion of the microenvironment cell
30 RNA expression data using the multiple subtypes of the first microenvironment cell type.

47. The method of claim 44 or any other preceding claim, wherein the first portion of the microenvironment cell RNA expression data comprises RNA expression data from multiple samples.

48. The method of claim 47 or any other preceding claim, wherein generating the first RNA expression profile comprises resampling a first portion of the microenvironment cell RNA expression data, taking a number of samples included in the multiple samples as input.

5

49. The method of claim 44 or any other preceding claim, wherein generating the simulated microenvironment cell RNA expression data further comprises:

generating, for a second microenvironment cell type, a second RNA expression profile using a second portion of the microenvironment cell RNA expression data; and

10 combining the first RNA expression profile and the second RNA expression profile to generate at least some of the simulated microenvironment cell RNA expression data.

50. The method of claim 49 or any other preceding claim, wherein combining the first RNA expression profile and the second RNA expression profile to generate the at least some of the simulated microenvironment cell RNA expression data comprises:

15

determining a weighted sum of the first RNA expression profile and the second RNA expression profile.

51. The method of claim 41 or any other preceding claim, wherein the malignant cell RNA expression data comprises RNA expression data from multiple malignant cell samples.

20

52. The method of claim 51 or any other preceding claim, wherein generating the simulated malignant cell RNA expression data comprises combining the RNA expression data from the multiple malignant cell samples.

25

53. The method of claim 41 or any other preceding claim, wherein generating the simulated malignant cell RNA expression data comprises adding noise to the simulated malignant cell RNA expression data.

30 54. The method of claim 50 or any other preceding claim, wherein coefficients of the weighted sum are determined using an output of a previously trained non-linear regression model.

55. The method of claim 40 or any other preceding claim, wherein the first RNA expression data includes expression data for at least 10 genes selected from the group of genes for the first cell type in Table 2.

5 56. The method of claim 40 or any other preceding claim, wherein the second RNA expression data includes expression data for at least 10 genes selected from the group of genes for the second cell type in Table 2.

57. The method of claim 40 or any other preceding claim, wherein the first cell type and
10 the second cell type are selected from the group consisting of B cells, CD4+ T cells, CD8+ T cells, endothelial cells, fibroblasts, lymphocytes, macrophages, monocytes, NK cells, neutrophils, and T cells.

58. The method of claim 40 or any other preceding claim, wherein the simulated RNA
15 expression data comprises second RNA expression data for the first genes associated with the first cell type.

59. The method of claim 58 or any other preceding claim, wherein the first non-linear regression model of the one or more non-linear regression models comprises:
20 a first sub-model configured to generate, using the first RNA expression data as input, a first value for the estimated percentage of RNA from the first cell type; and
a second sub-model, configured to generate, using the second RNA expression data and the first value for the estimated percentage of RNA from the first cell type as input, a second value for the estimated percentage of RNA from the first cell type.

25 60. The method of claim 59 or any other preceding claim, wherein the second sub-model is further configured to generate the second value for the estimated percentage of RNA from the first cell type using an estimated percentage of RNA from each of multiple cell types other than the first cell type as input.

30 61. A system, comprising:
at least one computer hardware processor; and

at least one non-transitory computer-readable storage medium storing processor executable instructions that, when executed by the at least one computer hardware processor, cause the at least one computer hardware processor to perform:

- 5 obtaining training data comprising simulated RNA expression data, the simulated RNA expression data including first RNA expression data for first genes associated with a first cell type and second RNA expression data for second genes associated with a second cell type different from the first cell type; and
- 10 training a plurality of non-linear regression models to estimate percentages of RNA from one or more respective cell types, the plurality of non-linear regression models comprising a first non-linear regression model for estimating percentage of RNA from the first cell type and a second non-linear regression model for estimating percentage of RNA from the second cell type, wherein training the plurality of non-linear regression models comprises training the first non-linear regression model at least in part by:
 - 15 generating, using the first non-linear regression model and the first RNA expression data, an estimated percentage of RNA from the first cell type; and
 - updating parameters of the first non-linear regression model using the estimated percentage of RNA from the first cell type; and
 - 20 outputting the trained plurality of non-linear regression models including the first non-linear regression model and the second non-linear regression model.

62. At least one non-transitory computer-readable storage medium storing processor executable instructions that, when executed by the at least one computer hardware processor, cause the at least one computer hardware processor to perform:
- 25 obtaining training data comprising simulated RNA expression data, the simulated RNA expression data including first RNA expression data for first genes associated with a first cell type and second RNA expression data for second genes associated with a second cell type different from the first cell type; and
 - 30 training a plurality of non-linear regression models to estimate percentages of RNA from one or more respective cell types, the plurality of non-linear regression models comprising a first non-linear regression model for estimating percentage of RNA from the first cell type and a second non-linear regression model for estimating percentage of RNA

from the second cell type, wherein training the plurality of non-linear regression models comprises training the first non-linear regression model at least in part by:

- generating, using the first non-linear regression model and the first RNA expression data, an estimated percentage of RNA from the first cell type; and
- 5 updating parameters of the first non-linear regression model using the estimated percentage of RNA from the first cell type; and
- outputting the trained plurality of non-linear regression models including the first non-linear regression model and the second non-linear regression model

- 10 63. A method comprising:
- using at least one computer hardware processor to perform:
- obtaining expression data for a biological sample, the biological sample previously obtained from a subject having, suspected of having, or at risk of having cancer;
 - 15 obtaining a plurality of expression profiles for a corresponding plurality of cell types, each of the expression profiles comprising respective expression data from one or more genes associated with a respective cell type from the plurality of cell types; and
 - determining a plurality of cell composition percentages for the plurality of cell
 - 20 types at least in part by optimizing a piecewise continuous error function between the expression data and the plurality of expression profiles.

64. The method of claim 63, wherein the expression data is RNA expression data, and the plurality of expression profiles are RNA expression profiles.

25

65. The method of claim 63, wherein determining the plurality of cell composition percentages for the plurality of cell types comprises determining a weighted sum of error values, the error values determined using the piecewise continuous error function.

- 30 66. The method of claim 65, wherein determining the plurality of cell composition percentages for the plurality of cell types comprises minimizing the weighted sum of error values.

67. The method of claim 63, wherein the one or more genes comprise less than 5000 genes and at least two genes in Table 2.

68. The method of claim 63 or any other preceding claim, further comprising:

5 determining a malignancy expression profile using the plurality of expression profiles for the corresponding plurality of cell types and the plurality of cell composition percentages.

69. A system, comprising:

at least one computer hardware processor; and

10 at least one computer readable storage medium storing processor executable

instructions that, when executed by the at least one computer hardware processor, cause the at least one computer hardware processor to perform:

obtaining expression data for a biological sample, the biological sample previously obtained from a subject having, suspected of having, or at risk of having cancer;

15 obtaining a plurality of expression profiles for a corresponding plurality of cell types, each of the expression profiles comprising respective expression data from one or more genes associated with a respective cell type from the plurality of cell types; and

20 determining a plurality of cell composition percentages for the plurality of cell types at least in part by optimizing a piecewise continuous error function between the expression data and the plurality of expression profiles.

70. At least one computer readable storage medium storing processor executable

25 instructions that, when executed by at least one computer hardware processor, cause the at least one computer hardware processor to perform:

obtaining expression data for a biological sample, the biological sample previously obtained from a subject having, suspected of having, or at risk of having cancer;

30 obtaining a plurality of expression profiles for a corresponding plurality of cell types, each of the expression profiles comprising respective expression data from one or more genes associated with a respective cell type from the plurality of cell types; and

determining a plurality of cell composition percentages for the plurality of cell types at least in part by optimizing a piecewise continuous error function between the expression data and the plurality of expression profiles.