

WHAT IS CLAIMED IS:

1. A method for identifying a signal cell label, comprising:
 - (a) barcoding a plurality of targets in a plurality of cells using a plurality of barcodes to create a plurality of barcoded targets, wherein each of the plurality of barcodes comprises a cell label and a molecular label;
 - (b) obtaining sequencing data of the plurality of barcoded targets;
 - (c) determining the number of molecular labels with distinct sequences associated with each of the cell labels of the plurality of barcodes;
 - (d) determining a rank of each of the cell labels of the plurality of barcodes based on the number of molecular labels with distinct sequences associated with each of the cell labels;
 - (e) generating a cumulative sum plot based on the number of molecular labels with distinct sequences associated with each of the cell labels determined in (c) and the rank of each of the cell labels determined in (d);
 - (f) generating a second derivative plot of the cumulative sum plot;
 - (g) determining a minimum of the second derivative plot of the cumulative sum plot, wherein the minimum of the second derivative plot corresponds to a cell label threshold; and
 - (h) identifying each of the cell labels as a signal cell label or a noise cell label based on the number of molecular labels with distinct sequences associated with each of the cell labels determined in (c) and the cell label threshold determined in (g).
2. The method of claim 1, comprising, if a cell label of the plurality of barcodes is identified as a noise cell label in (h), removing sequencing information associated with the identified cell label from the sequencing data obtained in (b).
3. The method of any one of claims 1-2, comprising removing sequencing information associated with molecular labels with distinct sequences associated with a target of the plurality of targets from the sequencing data obtained in (b) if the number of the molecular labels with distinct sequences associated with the target of the plurality of targets is above a molecular label occurrence threshold.
4. The method of any one of claims 1-3, wherein determining the number of molecular labels with distinct sequences associated with each of the cell labels in (c)

comprises removing sequencing information associated with non-unique molecular labels associated with each of the cell labels from the sequencing data.

5. The method of any one of claims 1-4, wherein the cumulative sum plot is a log-log plot.

6. The method of claim 5, wherein the log-log plot is a $\log_{10}$ - $\log_{10}$ plot.

7. The method of any one of claims 1-6, wherein generating the cumulative sum plot based on the number of molecular labels with distinct sequences associated with each of the cell labels determined in (c) and the rank of each of the cell labels determined in (d) comprises:

determining a cumulative sum for each rank of the cell labels, wherein the cumulative sum for the rank comprises a summation of a number of molecular labels with distinct sequences associated with each of the cell labels with a lower rank.

8. The method of claim 7, wherein generating the second derivative plot of the cumulative sum plot comprises determining a difference between a cumulative sum of a first rank of the cell labels and a cumulative sum of a second rank of the cell labels over a difference between the first rank and the second rank.

9. The method of claim 8, wherein the difference between the first rank and the second rank is one.

10. The method of any one of claims 1-9, wherein the minimum is a global minimum.

11. The method of any one of claims 1-10, wherein determining the minimum of the second derivative plot comprises determining a minimum of the second derivative plot above a threshold of a minimum number of molecular labels associated with each of the cell labels.

12. The method of claim 11, wherein the threshold of the minimum number of molecular labels associated with each of the cell labels is a percentile threshold.

13. The method of any one of claims 11-12, wherein the threshold of the minimum number of molecular labels associated with each of the cell labels is determined based on the number of cells in the plurality of cells.

14. The method of any one of claims 1-13, wherein determining the minimum of the second derivative plot comprises determining a minimum of the second derivative plot

below a threshold of a maximum number of molecular labels associated with each of the cell labels.

15. The method of claim 14, wherein the threshold of the maximum number of molecular labels associated with each of the cell labels is a percentile threshold.

16. The method of any one of claims 14-15, wherein the threshold of the maximum number of molecular labels associated with each of the cell labels is determined based on the number of cells in the plurality of cells.

17. The method of any one of claims 1-16, wherein each of the cell labels is identified as the signal cell label if the number of molecular labels with distinct sequences associated with the each of the cell labels determined in (c) is greater than the cell label threshold.

18. The method of any one of claims 1-17, wherein each of the cell labels is identified as a noise cell label if the number of molecular labels with distinct sequences associated with each of the cell labels determined in (c) is not greater than the cell label threshold.

19. The method of any one of claims 1-18, comprising:

(i) for one or more of the plurality of targets:

(1) counting the number of molecular labels with distinct sequences associated with the target in the sequencing data; and

(2) estimating the number of the target based on the number of molecular labels with distinct sequences associated with the target in the sequencing data counted in (1).

20. A method for determining a signal cell label, comprising:

(a) obtaining sequencing data of a plurality of barcoded targets, wherein the plurality of barcoded targets is created from a plurality of targets in a plurality of cells that are barcoded using a plurality of barcodes, and wherein each of the plurality of barcodes comprises a cell label and a molecular label;

(b) determining a rank of each of the cell labels of the plurality of barcodes based on the number of molecular labels with distinct sequences associated with each of the cell labels of the plurality of barcodes;

(c) determining a cell label threshold based on the number of molecular labels with distinct sequences associated with each of the cell labels and the rank of each of the cell labels of the plurality of barcodes determined in (b); and

(d) identifying each of the cell labels as a signal cell label or a noise cell label based on the number of molecular labels with distinct sequences associated with each of the cell labels and the cell label threshold determined in (c).

21. The method of claim 20, comprising determining the number of molecular labels with distinct sequences associated with each of the cell labels.

22. The method of claim 21, wherein determining the number of molecular labels with distinct sequences associated with each of the cell labels comprises removing sequencing information associated with non-unique molecular labels associated with each of the cell labels from the sequencing data.

23. The method of any one of claims 20-22, wherein determining the cell label threshold based on the number of molecular labels with distinct sequences associated with each of the cell labels of the plurality of barcodes and the rank of each of the cell labels of the plurality of barcodes determined in (b) comprises:

determining a cumulative sum for each rank of the cell labels, wherein the cumulative sum for the rank comprises a summation of a number of molecular labels with distinct sequences associated with each of the cell labels with a lower rank; and

determining a rank n of the cell labels with the largest change in a cumulative sum for the rank n and a cumulative sum for the next rank $n+1$, wherein the rank n of the cell labels with the largest change in the cumulative sum and the cumulative sum for the next rank $n+1$ corresponds to the cell label threshold.

24. The method of any one of claims 20-22, wherein determining the cell label threshold based on the number of molecular labels with distinct sequences associated with each of the cell labels of the plurality of barcodes comprises:

determining the cell label with the largest change in a cumulative sum for the cell label with a rank n and a cumulative sum for the cell label with the next rank $n+1$, wherein a number of molecular labels with distinct sequences associated with the cell label corresponds to the cell label threshold.

25. The method of any one of claims 20-22, wherein determining the cell label threshold based on the number of molecular labels with distinct sequences associated with each of the cell labels of the plurality of barcodes and the rank of each of the cell labels of the plurality of barcodes determined in (b) comprises:

generating a cumulative sum plot based on the number of molecular labels with distinct sequences associated with each of the cell labels and the rank of each of the cell labels determined in (b);

generating a second derivative plot of the cumulative sum plot; and

determining a minimum of the second derivative plot of the cumulative sum plot, wherein the minimum of the second derivative plot corresponds to the cell label threshold.

26. The method of claim 25, wherein generating the cumulative sum plot based on the number of molecular labels with distinct sequences associated with each of the cell labels and the rank of each of the cell labels determined in (b) comprises:

determining a cumulative sum for each rank of the cell labels, wherein the cumulative sum for the rank comprises a summation of a number of molecular labels with distinct sequences associated with each of the cell labels with a lower rank.

27. The method of any one of claims 25-26, wherein generating the second derivative plot of the cumulative sum plot comprises determining a difference between a cumulative sum of a first rank of the cell labels and a cumulative sum of a second rank of the cell labels over a difference between the first rank and the second rank.

28. The method of claim 27, wherein the difference between the first rank and the second rank is one.

29. The method of any one of claims 20-28, comprising, if a cell label of the plurality of barcodes is identified as a noise cell label in (d), removing sequencing information associated with the identified cell label from the sequencing data obtained in (a).

30. The method of any one of claims 20-29, comprising removing sequencing information associated with molecular labels with distinct sequences associated with a target of the plurality of targets from the sequencing data obtained in (a) if the number of the molecular labels with distinct sequences associated with the target of the plurality of targets is above a molecular label occurrence threshold.

31. The method of any one of claims 20-30, wherein the cumulative sum plot is a log-log plot.
32. The method of claim 31, wherein the log-log plot is a $\log_{10}$ - $\log_{10}$ plot.
33. The method of any one of claims 25-32, wherein the minimum is a global minimum.
34. The method of any one of claims 25-33, wherein determining the minimum of the second derivative plot comprises determining a minimum of the second derivative plot above a threshold of a minimum number of molecular labels associated with each of the cell label.
35. The method of claim 34, wherein the threshold of the minimum number of molecular labels associated with each of the cell labels is a percentile threshold.
36. The method of any one of claims 34-35, wherein the threshold of the minimum number of molecular labels associated with each of the cell labels is determined based on the number of cells in the plurality of cells.
37. The method of any one of claims 25-36, wherein determining the minimum of the second derivative plot comprises determining a minimum of the second derivative plot below a threshold of a maximum number of molecular labels associated with each of the cell label.
38. The method of claim 37, wherein the threshold of the maximum number of molecular labels associated with each of the cell labels is a percentile threshold.
39. The method of any one of claims 37-38, wherein the threshold of the maximum number of molecular labels associated with each of the cell labels is determined based on the number of cells in the plurality of cells.
40. The method of any one of claims 20-39, wherein each of the cell labels is identified as a signal cell label if the number of molecular labels with distinct sequences associated with the each of the cell labels determined is greater than the cell label threshold.
41. The method of any one of claims 20-40, wherein each of the cell labels is identified as a noise cell label if the number of molecular labels with distinct sequences associated with each of the cell labels determined is not greater than the cell label threshold.
42. The method of any one of claims 20-41, comprising:
- (e) for one or more of the plurality of targets:

- (1) counting the number of molecular labels with distinct sequences associated with the target in the sequencing data; and
 - (2) estimating the number of the target based on the number of molecular labels with distinct sequences associated with the target in the sequencing data counted in (1).
43. A method for identifying a signal cell label, comprising:
 - (a) obtaining sequencing data of a plurality of targets of cells, wherein each target is associated with a number of molecular labels with distinct sequences associated with each cell label of a plurality of cell labels;
 - (b) determining a cell label threshold based on the number of molecular labels with distinct sequences associated with each of the cell labels; and
 - (c) identifying each of the cell labels as a signal cell label or a noise cell label based on the number of molecular labels with distinct sequences associated with each of the cell labels and the cell label threshold.
44. The method of claim 43, wherein obtaining sequencing data comprises:
 - barcoding the plurality of targets of the cells using a plurality of barcodes to create a plurality of barcoded targets, wherein each of the plurality of barcodes comprises a cell label of the plurality of cell labels and a molecular label; and
 - determining the number of molecular labels with distinct sequences associated with each of the cell labels of the plurality of barcodes.
45. The method of any one of claims 43-44, comprising:
 - for one or more of the plurality of targets:
 - (1) counting the number of molecular labels with distinct sequences associated with the target in the sequencing data; and
 - (2) estimating the number of the target based on the number of molecular labels with distinct sequences associated with the target in the sequencing data counted in (1).
46. The method of any one of claims 43-45, comprising, if a cell label of the plurality of barcodes is identified as a noise cell label, removing sequencing information associated with the identified cell label from the sequencing data.

47. The method of any one of claims 43-46, comprising removing sequencing information associated with molecular labels with distinct sequences associated with a target of the plurality of targets from the sequencing data if the number of the molecular labels with distinct sequences associated with the target of the plurality of targets is above a molecular label occurrence threshold.

48. The method of any one of claims 43-47, wherein determining the number of molecular labels with distinct sequences associated with each of the cell labels in (c) comprises removing sequencing information associated with non-unique molecular labels associated with each of the cell labels from the sequencing data.

49. The method of any one of claims 43-48, wherein determining the cell label threshold comprises:

determining an inflection point of a cumulative sum plot,

wherein the cumulative sum plot is based on the number of molecular labels with distinct sequences associated with each of the plurality of cell labels and a rank of each of the cell labels, and

wherein the inflection point corresponds to the cell label threshold.

50. The method of claim 49, wherein determining the inflection point of the cumulative sum plot comprises:

generating the cumulative sum plot based on the number of molecular labels with distinct sequences associated with each of the plurality of cell labels and the rank of each of the cell labels;

generating a second derivative plot of the cumulative sum plot; and

determining a minimum of the second derivative plot of the cumulative sum plot, wherein the minimum of the second derivative plot corresponds to a cell label threshold.

51. The method of any one of claims 49-50, wherein determining the cell label threshold comprises: determining the rank of each of the plurality of cell labels based on the number of molecular labels with distinct sequences associated with each of the cell labels.

52. The method of any one of claims 43-51, wherein the cumulative sum plot is a log-log plot.

53. The method of claim 52, wherein the log-log plot is a $\log_{10}$ - $\log_{10}$ plot.

54. The method of any one of claims 50-53, wherein generating the cumulative sum plot based on the number of molecular labels with distinct sequences associated with each of the cell labels and the rank of each of the cell labels comprises:

determining a cumulative sum for each rank of the cell labels, wherein the cumulative sum for the rank comprises a summation of a number of molecular labels with distinct sequences associated with each of the cell labels with a lower rank.

55. The method of claim 54, wherein generating the second derivative plot of the cumulative sum plot comprises determining a difference between a cumulative sum of a first rank of the cell labels and a cumulative sum of a second rank of the cell labels over a difference between the first rank and the second rank.

56. The method of claim 55, wherein the difference between the first rank and the second rank is one.

57. The method of any one of claims 43-56, wherein the minimum is a global minimum.

58. The method of any one of claims 43-57, wherein determining the minimum of the second derivative plot comprises determining a minimum of the second derivative plot above a threshold of a minimum number of molecular labels associated with each of the cell labels.

59. The method of claim 58, wherein the threshold of the minimum number of molecular labels associated with each of the cell labels is a percentile threshold.

60. The method of any one of claims 58-59, wherein the threshold of the minimum number of molecular labels associated with each of the cell labels is determined based on the number of the plurality of cells.

61. The method of any one of claims 43-60, wherein determining the minimum of the second derivative plot comprises determining a minimum of the second derivative plot below a threshold of a maximum number of molecular labels associated with each of the cell labels.

62. The method of claim 61, wherein the threshold of the maximum number of molecular labels associated with each of the cell labels is a percentile threshold.

63. The method of any one of claims 61-62, wherein the threshold of the maximum number of molecular labels associated with each of the cell labels is determined based on the number of the plurality of cells.

64. The method of any one of claims 43-63, wherein each of the cell labels is identified as the signal cell label if the number of molecular labels with distinct sequences associated with the each of the cell labels is greater than the cell label threshold.

65. The method of any one of claims 43-64, wherein each of the cell labels is identified as a noise cell label if the number of molecular labels with distinct sequences associated with each of the cell labels is not greater than the cell label threshold.

66. A method for identifying a signal cell label, comprising:

(a) barcoding a plurality of targets in a plurality of cells using a plurality of barcodes to create a plurality of barcoded targets, wherein each of the plurality of barcodes comprises a cell label and a molecular label, wherein barcoded targets created from targets of different cells of the plurality of cells have different cell labels, and wherein barcoded targets created from targets of the same cell of the plurality of cells have different molecular labels;

(b) obtaining sequencing data of the plurality of barcoded targets;

(c) determining a feature vector of each cell label of the plurality of barcoded targets, wherein the feature vector comprise numbers of molecular labels with distinct sequences associated with the each cell label;

(d) determining a cluster for the each cell label of the plurality of barcoded targets based on the feature vector; and

(e) identifying the each cell label of the plurality of barcoded targets as a signal cell label or a noise cell label based on a number of cell labels in the cluster and a cluster size threshold.

67. The method of claim 66, wherein determining the cluster for the each cell label of the plurality of barcoded targets based on the feature vector comprises clustering the each cell label of the plurality of barcoded targets into the cluster based on a distance of the feature vector to the cluster in a feature vector space.

68. The method of claim 66, wherein determining the cluster for the each cell label of the plurality of barcoded targets based on the feature vector comprises:

projecting the feature vector from a feature vector space into a lower dimensional space; and

clustering the each cell label into the cluster based on a distance of the feature vector to the cluster in the lower dimensional space.

69. The method of claim 68, wherein the lower dimensional space is a two dimensional space.

70. The method of any one of claims 68-69, wherein projecting the feature vector from the feature vector space into the lower dimensional space comprises projecting the feature vector from the feature vector space into the lower dimensional space using a t-distributed stochastic neighbor embedding (tSNE) method.

71. The method of any one of claims 68-70, wherein clustering the each cell label into the cluster based on the distance of the feature vector to the cluster in the lower dimensional space comprises clustering the each cell label into the cluster based on the distance of the feature vector to the cluster in the lower dimensional space using a density-based method.

72. The method of claim 71, wherein the density-based method comprises a density-based spatial clustering of applications with noise (DBSCAN) method.

73. The method of any one of claims 66-72, wherein the cell label is identified as a signal cell label if the number of cell labels in the cluster is below the cluster size threshold.

74. The method of any one of claims 66-73, wherein the cell label is identified as a noise cell label if the number of cell labels in the cluster is not below the cluster size threshold.

75. The method of any one of claims 66-74, comprising determining the cluster size threshold based on the number of cell labels of the plurality of barcoded targets.

76. The method of claim 75, wherein the cluster size threshold is a percentage of the number of cell labels of the plurality of barcoded targets.

77. The method of any one of claims 66-74, comprising determining the cluster size threshold based on the number of cell labels of the plurality of barcoded targets.

78. The method of claim 77, wherein the cluster size threshold is a percentage of the number of cell labels of the plurality of barcoded targets.

79. The method of any one of claims 66-78, comprising determining the cluster size threshold based on numbers of molecular labels with distinct sequences associated with each cell label of the plurality of barcodes.

80. The method of any one of claims 66-79, comprising:

(f) for one or more of the plurality of targets:

(1) counting the number of molecular labels with distinct sequences associated with the target in the sequencing data; and

(2) estimating the number of the target based on the number of molecular labels with distinct sequences associated with the target in the sequencing data counted in (1).

81. A method for identifying a signal cell label, comprising:

(a) obtaining sequencing data of a plurality of barcoded targets, wherein the plurality of barcoded targets is create from a plurality of targets in a plurality of cells that are barcoded using a plurality of barcodes, wherein each of the plurality of barcodes comprises a cell label and a molecular label, wherein barcoded targets created from targets of different cells of the plurality of cells have different cell labels, and wherein barcoded targets created from targets of the same cell of the plurality of cells have different molecular labels;

(b) determining a feature vector of each cell label of the plurality of barcoded targets, wherein the feature vector comprise numbers of molecular labels with distinct sequences associated with the each cell label;

(c) determining a cluster for the each cell label of the plurality of barcoded targets based on the feature vector; and

(d) identifying the each cell label of the plurality of barcoded targets as a signal cell label or a noise cell label based on a number of cell labels in the cluster and a cluster size threshold.

82. The method of claim 80, wherein determining the cluster for the each cell label of the plurality of barcoded targets based on the feature vector comprises clustering the each cell label of the plurality of barcoded targets into the cluster based on a distance of the feature vector to the cluster in a feature vector space.

83. The method of claim 80, wherein determining the cluster for the each cell label of the plurality of barcoded targets based on the feature vector comprises:

projecting the feature vector from a feature vector space into a lower dimensional space; and

clustering the each cell label into the cluster based on a distance of the feature vector to the cluster in the lower dimensional space.

84. The method of claim 83, wherein the lower dimensional space is a two dimensional space.

85. The method of any one of claims 83-84, wherein projecting the feature vector from the feature vector space into the lower dimensional space comprises projecting the feature vector from the feature vector space into the lower dimensional space using a t-distributed stochastic neighbor embedding (tSNE) method.

86. The method of any one of claims 83-85, wherein clustering the each cell label into the cluster based on the distance of the feature vector to the cluster in the lower dimensional space comprises clustering the each cell label into the cluster based on the distance of the feature vector to the cluster in the lower dimensional space using a density-based method.

87. The method of claim 86, wherein the density-based method comprises a density-based spatial clustering of applications with noise (DBSCAN) method.

88. The method of any one of claims 83-87, wherein the cell label is identified as the signal cell label if the number of cell labels in the cluster is below the cluster size threshold.

89. The method of any one of claims 83-88, wherein the cell label is identified as a noise cell label if the number of cell labels in the cluster is not below the cluster size threshold.

90. The method of any one of claims 83-89, comprising determining the cluster size threshold based on the number of cell labels of the plurality of barcoded targets.

91. The method of claim 90, wherein the cluster size threshold is a percentage of the number of cell labels of the plurality of barcoded targets.

92. The method of any one of claims 83-91, comprising determining the cluster size threshold based on the number of cell labels of the plurality of barcoded targets.

93. The method of claim 92, wherein the cluster size threshold is a percentage of the number of cell labels of the plurality of barcoded targets.

94. The method of any one of claims 83-93, comprising determining the cluster size threshold based on numbers of molecular labels with distinct sequences associated with each cell label of the plurality of barcoded targets.

95. The method of any one of claims 83-94, comprising:

(e) for one or more of the plurality of targets:

(1) counting the number of molecular labels with distinct sequences associated with the target in the sequencing data; and

(2) estimating the number of the target based on the number of molecular labels with distinct sequences associated with the target in the sequencing data counted in (1).

96. A method for identifying a signal cell label, comprising:

(a) obtaining sequencing data of a plurality of first targets of cells, wherein each first target is associated with a number of molecular labels with distinct sequences associated with each cell label of a plurality of cell labels;

(b) identifying each of the cell labels as a signal cell label or a noise cell label based on the number of molecular labels with distinct sequences associated with each of the cell labels and an identification threshold; and

(c) re-identifying at least one of the plurality of cell labels as a signal cell label identified as a noise cell label in (b) or re-identifying at least one of the cell label as a noise cell label identified as a signal cell label in (b).

97. The method of claim 96, wherein the identification threshold comprises a cell label threshold, a cluster size threshold, or any combination thereof.

98. The method of any one of claims 96-97, comprising removing one or more cell labels of the plurality of cell labels each associated with a number of molecular labels with distinct sequences below threshold of a number of molecular labels.

99. The method of any one of claims 96-98, wherein re-identifying at least one of the plurality of cell labels as a signal cell label identified as a noise cell label in (b) comprises:

determining a plurality of second targets of the plurality of first targets each with one or more variability indications, amongst the plurality of first targets, above a variability threshold; and

re-identifying at least one of the plurality of cell labels as a signal cell label identified as a noise cell label in (b) based on, for each of the plurality of cell labels, the number of molecular labels with distinct sequences associated with the plurality of second targets and the identification threshold.

100. The method of claim 99, wherein the one or more variability indications of the second target comprises an average, a maximum, a median, a minimum, a dispersion, or any combinations thereof, of the numbers of molecular labels with distinct sequences associated with the second target and cell labels of the plurality of cell labels in the sequencing data.

101. The method of any one of claims 99-100, wherein the one or more variability indications of the second target comprises a standard deviation, a normalized dispersion, or any combinations thereof, variability indications of a subset of the plurality of second targets.

102. The method of claim 101, wherein the variability threshold is smaller than or equal to the size of the subset of the plurality of second targets.

103. The method of any one of claims 96-102, wherein re-identifying at least one of the plurality of cell labels as a noise cell label identified as a signal cell label in (b) comprises:

determining a plurality of third targets of the plurality of first targets each with an association with cell labels identified as noise cell labels in (c) above an association threshold; and

re-identifying at least one of the cell label as a noise cell label identified as a signal cell label in (b), for each of the plurality of cell labels, based on the number of molecular labels with distinct sequences associated with the plurality of third targets, and the identification threshold.

104. The method of claim 103, wherein determining the plurality of third targets of the plurality of first targets each with an association with cell labels identified as noise cell labels in (c) above the association threshold comprises:

determining a plurality of remaining cells labels identified as signal cell labels after re-identifying at least one of the cell label as a signal cell label identified as a noise cell label in (b);

determining the plurality of third targets based on

for each of the plurality of cell labels, the number of molecular labels with distinct sequences associated with the plurality of targets, and

for each of the plurality of remaining cell labels, the number of molecular labels with distinct sequences associated with the plurality of targets.

105. A computer system for determining the number of targets comprising:
a hardware processor; and

non-transitory memory having instructions stored thereon, which when executed by the hardware processor causes the processor to perform the method of any one of claims 1-104.

106. A computer readable medium comprising code for performing the method for performing the method of any one of claims 1-104.