

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;
 - (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), wherein each of the cell labels is identified as a signal cell label if the number of molecular labels with distinct sequences associated with each of the cell labels is greater than the cell label threshold; and
 - (i) re-identifying as a noise cell label one or more cell labels that are identified as a signal cell label in (h), comprising:
 - determining a plurality of second targets of the plurality of barcoded targets, wherein the variability in the number of molecular labels with distinct sequences associated with each of the plurality of second targets is above a variability threshold; and
 - when the number of molecular labels with distinct sequences associated with a cell label of the plurality of second targets that is identified

as a signal cell label in (h) is not greater than a second cell label threshold, re-identifying the cell label as a noise cell label, wherein the second cell label threshold is a minimum of a second derivative of the cumulative sum plot of the number of molecular labels with distinct sequences associated with cell labels of the plurality of second targets in descending order of the number of molecular labels.

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, further comprising (j) re-identifying as a signal cell label one or more cell labels that are identified as a noise cell label in (h), comprising:

determining a set of signal cell labels remaining among the plurality of cell labels identified as a signal cell label in (h) after re-identifying as a noise cell label one or more cell labels that are identified as a signal cell label in (i);

determining a plurality of under-represented targets based on the difference in the number of molecular labels with distinct sequences associated with (A) signal cell labels identified in (h), and (B) the remaining set of signal cell labels; and

re-identifying as a signal cell label one or more cell labels that are identified as a noise cell label in (h), for each of the plurality of cell labels, when the number of molecular labels with distinct sequences associated with the plurality of under-represented targets is greater than a third cell label threshold, wherein the third cell label threshold is a minimum of a second derivative of the cumulative sum plot of the number of molecular labels with distinct sequences associated with cell labels of the plurality of under-represented targets in descending order of the number of molecular labels.

5. 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),

wherein determining the cell label threshold 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 one of:

(i) 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; or

(ii) 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; or

(iii) generating a cumulative sum plot based on the cumulative sum for each rank 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 the cell label threshold;

(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), wherein each of the cell labels is identified as a signal cell label if the number of molecular labels with distinct sequences associated with each of the cell labels is greater than the cell label threshold; and

(e) re-identifying as a noise cell label one or more cell labels that are identified as a signal cell label in (d), comprising:

determining a plurality of second targets of the plurality of barcoded targets, wherein the variability in the number of molecular labels with distinct sequences associated with each of the plurality of second targets is above a variability threshold; and

when the number of molecular labels with distinct sequences associated with a cell label of the plurality of second targets that is identified as a signal cell label in (d) is not greater than a second cell label threshold, re-identifying the cell label as a noise cell label, wherein the second cell label threshold is based on an analysis of a cumulative sum of the number of molecular labels with distinct sequences associated with each of the cell labels of the plurality of second targets in descending order of the number of molecular labels.

6. The method of claim 5, comprising determining the number of molecular labels with distinct sequences associated with each of the cell labels.

7. The method of claim 6, 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.

8. The method of claim 5, wherein the minimum is a global minimum.

9. The method of any one of claims 5-7, 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.

10. The method of claim 9, 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.

11. The method of any one of claims 5-10, 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.

12. The method of claim 11, 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.

13. The method of any one of claims 5-12, 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).

14. The method of any one of claims 5-13, further comprising: (f) re-identifying as a signal cell label one or more cell labels that are identified as a noise cell label in (d), comprising:

determining a set of signal cell labels remaining among the plurality of cell labels identified as a signal cell label in (d) after re-identifying as a noise cell label one or more cell labels that are identified as a signal cell label in (e);

determining a plurality of under-represented targets based on the difference in the number of molecular labels with distinct sequences associated with (A) signal cell labels identified in (d), and (B) the remaining set of signal cell labels; and

re-identifying as a signal cell label one or more cell labels that are identified as a noise cell label in (d), for each of the plurality of cell labels, when the number of molecular labels with distinct sequences associated with the plurality of under-represented targets is greater than a third cell label threshold, wherein the third cell label threshold is based on an analysis of a cumulative sum of the number of molecular labels with distinct sequences associated with each of the cell labels of the plurality of under-represented targets in descending order of the number of molecular labels.

15. The method of any one of claims 5-14, wherein determining the cell label threshold 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

(i) 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; or

(ii) 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.

16. 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 created 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,

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, and

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.

17. The method of claim 16, 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.

18. The method of claim 16, 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.