

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. 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

(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; 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), 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 is greater than the cell label threshold.

5. The method of claim 4, comprising determining the number of molecular labels
with distinct sequences associated with each of the cell labels.

6. The method of claim 5, 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.

7. The method of claim 4, wherein the minimum is a global minimum.

8. The method of claim 4 or 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.

9. The method of claim 8, 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.

10. The method of any one of claims 4-9, 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.

11. The method of claim 10, 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.

12. The method of any one of claims 4-11, 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).

13. The method of any one of claims 4-12, further comprising: (e) re-identifying at least one of the plurality of cell labels that is identified as a noise cell label in (d) as a signal cell label, and/or re-identifying at least one of the plurality of cell labels that is identified as a signal cell label in (d) as a noise cell label.

14. The method of claim 13, wherein re-identifying at least one of the plurality of cell labels that is identified as a signal cell label in (d) as a noise cell label comprises:

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

re-identifying at least one of the plurality of cell labels that is identified as a signal cell label in (d) as a noise cell label based on, for each of the plurality of cell labels, (1) the number of molecular labels with distinct sequences associated with the plurality of second targets and (2) the cell label threshold.

15. The method of any one of claims 13 or 14, wherein re-identifying at least one of the plurality of cell labels that is identified as a noise cell label in (d) as a signal cell label comprises:

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

re-identifying at least one of the plurality of cell labels that is identified as a noise cell label in (d) as a signal cell label, for each of the plurality of cell labels, based on (1) the number of molecular labels with distinct sequences associated with the plurality of third targets, and (2) the cell label threshold.

16. The method of any one of claims 4-15, 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.

17. 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.

18. The method of claim 17, 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.

19. The method of claim 17, 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.