

WHAT IS CLAIMED IS:

1. A method for identifying targets to distinguish cell types, comprising:
 - (a) receiving a target counts data structure, wherein the target counts data structure comprises expression profiles of a plurality of cells, and wherein the expression profiles of the plurality of cells comprises a number of each target of a plurality of targets for each cell of the plurality of cells;
 - (b) hierarchically clustering expression profiles of the plurality of cells based on the target counts data structure and distances between the expression profiles of the plurality of cells to generate a dendrogram representing the expression profiles of plurality of cells, wherein the dendrogram comprises a plurality of nodes, wherein the plurality of nodes comprise a root node, a plurality of leaf nodes, and a plurality of non-root, non-leaf nodes, wherein each leaf node of the plurality of leaf nodes represents an expression profile of a different cell of the plurality of cells, and wherein the root node represents expression profiles of the plurality of cells;
 - (c) while traversing through each node of the plurality of nodes of the dendrogram from the root node of the dendrogram to the plurality of leaf nodes of the dendrogram:
 - (1) determining whether a splitting of the node into child nodes of the node is valid or invalid; and
 - (2) if the splitting of the node into the child nodes of the node is invalid, adding the node to a merging cluster set;
 - (d) iteratively, for each first node in the merging cluster set,
 - if a distance between the first node in the merging cluster set and a second node in the merging cluster set that is closest to the first node is within a merging distance threshold, merging the first node with the second node to generate a merged node comprising expression profiles represented by the first node and the second node; and
 - (e) for each node in the merging cluster set, identifying targets for distinguishing cell types based on expression profiles of the plurality of targets of cells represented by the node.

2. The method of claim 1, wherein the target counts data structure comprises a target counts matrix.

3. The method of claim 2, wherein each row or each column of the target counts matrix comprises a number of each target of a plurality of targets for a different individual cell of the plurality of cells.

4. The method of claim 1,
wherein each of the plurality of leaf nodes and the plurality of non-root, non-leaf nodes is associated with a parent node, and
wherein each of the root node and the plurality of non-root, non-leaf nodes is associated with a left child node and a right child node and represents expression profiles represented by the left child node and the right child node of the node.

5. The method of any one of claims 1-4, comprising, prior to receiving the target counts data structure in (a):

(f) barcoding the plurality of targets in the 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 have different cell labels, and wherein barcoded targets created from targets of one cell of the plurality of cells have different molecular labels;

(g) obtaining sequencing data of the plurality of barcoded targets; and

(h) for each of the plurality of cells:

(1) counting the number of molecular labels with distinct sequences associated with each target of the plurality of targets in the sequencing data for the cell; and

(2) estimating the number of each target of the plurality of targets for the cell based on the number of molecular labels with distinct sequences associated with the target in the sequencing data counted in (h)(1).

6. The method of claim 5, wherein receiving the target counts data structure comprises:

generating a target counts data structure from the number of each target of the plurality of targets for the cell estimated in (h)(2), wherein the expression profile of the cell

of the plurality of cells comprises the number of each target of the plurality of targets for the cell estimated in (h)(2).

7. The method of any one of claims 1-6, comprising, prior to hierarchically clustering the expression profiles of the plurality of cells based on the target counts data structure and the distances between the expression profiles of the plurality of cells to generate the dendrogram representing the expression profiles of the plurality of cells in (b):

(i) determining a distance data structure of elements of the target counts data structure, wherein the distance data structure comprises distances between the expression profiles of the plurality of cells.

8. The method of claim 7, wherein the distance data structure comprises a distance matrix.

9. The method of claim 8, wherein each diagonal element of the distance matrix has a value of zero.

10. The method of any one of claims 7-9, wherein hierarchically clustering the expression profiles of the plurality of cells based on the target counts data structure and the distances between the expression profiles of the plurality of cells to generate the dendrogram representing the expression profiles of the plurality of cells in (b) comprises: hierarchically clustering the expression profiles of the plurality of cells based on the target counts data structure and the distance data structure.

11. The method of any one of claims 7-10, wherein the distances between the expression profiles of the plurality of cells comprise pairwise correlation distances between the expression profiles of the plurality of cells.

12. The method of any one of claims 7-11, comprising, prior to determining the distance data structure of elements of the target counts data structure in (i), log-transforming the target counts data structure into a log-transformed target counts data structure,

wherein determining the distance data structure of elements of the target counts data structure comprises determining the distance data structure of the log-transformed target counts data structure, and

wherein hierarchically clustering the expression profiles of the plurality of cells based on the target counts data structure and the distances between the expression profiles of the plurality of cells in (b) comprises: hierarchically clustering the expression profiles of the

plurality of cells based on the log-transformed target counts data structure and the distance data structure to generate the dendrogram.

13. The method of claim 12, wherein log-transforming the target counts data structure into the log-transformed target counts data structure comprises increasing the value of each element of the target counts data structure by an increment.

14. The method of claim 13, wherein the increment is one.

15. The method of any one of claims 1-14, wherein hierarchically clustering the expression profiles of the plurality of cells based on the target counts data structure and the distances between the expression profiles of the plurality of cells in (b) comprises:

assigning each expression profile of the plurality of cells to a different leaf node; and

iteratively combining a first node and a second node of the plurality of nodes to generate a parent node of the first node and the second node if the second node is the closest node of the plurality of nodes to the first node.

16. The method of claim 15, wherein the distance between the first node and the second node is the maximum distance between any cell with an expression profile represented by the first node and any cell with an expression profile represented by the second node.

17. The method of any one of claims 15-16, wherein at least one of an intra-node correlation of the first node and an intra-node correlation of the second node is higher than an inter-node correlation of the first node and the second node.

18. The method of any one of claims 15-16, wherein an indication of an intra-node correlation of the first node and an intra-node correlation of the second node is higher than an inter-node correlation of the first node and the second node.

19. The method of claim 18, wherein the indication of the intra-node correlation of the first node and the intra-node correlation of the second node is based on at least one of:

an intra-node maximum correlation of the first node and the second node,

an intra-node average correlation of the first node and the second node,

an intra-node median correlation of the first node and the second node,

an intra-node minimum correlation of the first node and the second node, and

any combinations thereof.

20. The method of any one of claims 17-19, wherein the intra-node correlation of the first node is based on at least one of:

- an intra-node maximum correlation of the first node,
- an intra-node average correlation of the first node,
- an intra-node median correlation of the first node,
- an intra-node minimum correlation of the first node, and
- any combinations thereof.

21. The method of any one of claims 17-20, wherein the intra-node correlation of the second node is based on at least one of:

- an intra-node maximum correlation of the second node,
- an intra-node average correlation of the second node,
- an intra-node median correlation of the second node,
- an intra-node minimum correlation of the second node, and
- any combinations thereof.

22. The method of any one of claims 17-21, wherein the inter-node correlation of the first node and the second node is based on at least one of:

- an inter-node maximum correlation of the first node and the second node,
- an inter-node average correlation of the first node and the second node,
- an inter-node median correlation of the first node and the second node,
- an inter-node minimum correlation of the first node and the second node, and
- any combinations thereof.

23. The method of any one of claims 4-16, comprising, at each node when traversing the plurality of nodes of the dendrogram:

- if the splitting is valid, continuing traversing from the node to the left child node and the right child node of the node; and

- if the splitting is invalid, discontinuing traversing from the node to the left child node and the right child node of the node.

24. The method of any one of claims 4-23, wherein determining whether the splitting of the node with the child nodes of the node is valid or invalid comprises:

- determining the splitting to be valid if the distance between the left child node and the right child node is above a splitting threshold, and

otherwise invalid.

25. The method of claims 24, wherein the distance between the left child node and the right child node is determined based on a statistical test performed on each target of the plurality of targets between expression profiles represented by the left child node and the right child node.

26. The method of claim 25, wherein the statistical test comprises a Welch's *t*-test.

27. The method of any one of claims 25-26, wherein the distance between the left child node and the right child node is determined based on the maximum *p*-value of the statistical test performed on the each target of the plurality of targets between each expression profile represented by the left child node and each expression profile represented by the right child node.

28. The method of any one of claims 1-27, comprising, at each node when traversing the plurality of nodes of the dendrogram: (3) adding the node to the merging cluster set if the node represents an expression profile of a single cell.

29. The method of any one of claims 4-28, comprising, at each node when traversing the plurality of nodes of the dendrogram: assigning a node label to the node.

30. The method of claim 29,

wherein, if the node represents an expression profile of a single cell, the node label of the node comprises a single cell designation,

otherwise, if the node is the left child node of the parent node, the node label of the node comprises the node label of the parent node and a left designation, and

otherwise, the node label of the node comprises the node label of the parent node and a right designation.

31. The method of any one of claims 1-30, wherein for each node in the merging cluster set, identifying the targets for distinguishing the cell types based on the expression profiles of the plurality of targets of the cells represented by the node comprises:

determining a difference, between expression profiles represented by the node and expression profiles represented by another node in the merging cluster set, in the numbers of molecular labels with distinct sequences associated with the targets for distinguishing the cell types is greater than a significance threshold.

32. The method of any one of claims 1-31, comprising, prior to merging the first node with the second node to generate the merged node in (d):

merging each third node in the merging cluster set representing an expression profile of a single cell with a fourth node in the merging cluster set if a distance between the third node and the fourth node is within a node distance threshold.

33. The method of any one of claims 1-32, comprising classifying the plurality of cells based on the nodes in the merging cluster set that represent expression profiles of the cells.

34. The method of any one of claims 1-33, comprising designing a whole transcriptome assay based on the targets for distinguishing cell types identified.

35. The method of any one of claims 1-33, comprising designing a targeted transcriptome assay based on the targets for distinguishing cell types identified.

36. A method for identifying targets to distinguish cell types, comprising:

(a) receiving expression profiles of a plurality of cells, wherein the expression profiles comprise a number of each target of a plurality of targets for each cell of the plurality of cells;

(b) clustering the expression profiles of the plurality of cells to generate a plurality of clusters of expression profiles based on the distances between the expression profiles of the plurality of cells, wherein each cluster has one or more associations with one or both of (1) a parent cluster and (2) two or more child clusters, wherein the parent cluster represents expression profiles of one or more cells of the plurality of cells represented by the cluster, and wherein the cluster represents expression profiles represented by the two or more child clusters;

(c) for each cluster with two or more child clusters, if associations between the cluster with the two or more child clusters are invalid, adding the cluster to a merging cluster set;

(d) iteratively, for each first cluster in the merging cluster set, if a distance between the first cluster in the merging cluster set and a second cluster in the merging cluster set that is closest to the first cluster is within a merging distance threshold, merging the first cluster and the second cluster to generate a merged cluster, wherein the merged cluster comprises expression profiles of the first cluster and the second cluster; and

(e) for each cluster in the merging cluster set, identifying targets for distinguishing cell types based on expression profiles of the plurality of targets of cells represented by the cluster.

37. The method of claim 36, wherein receiving expression profiles of the plurality of cells comprises receiving a target counts data structure.

38. The method of claim 37, wherein the target counts data structure comprises a target counts matrix.

39. The method of claim 38, wherein each row or each column of the target counts matrix comprises an expression profile of a different individual cell of the plurality of cells.

40. The method of any one of claims 36-39, wherein clustering the expression profiles of the plurality of cells into the plurality of clusters of expression profiles based on the distances between the expression profiles of the plurality of cells comprises:

hierarchically clustering the expression profiles of the plurality of cells to generate a dendrogram representing the expression profiles of the plurality of cells based on the distances between the expression profiles of the plurality of cells, wherein the dendrogram comprises the plurality of clusters, wherein the plurality of clusters comprise a root cluster, a plurality of leaf clusters, and a plurality of non-root, non-leaf clusters.

41. The method of claim 40,

wherein each of the plurality of leaf clusters and the plurality of non-root, non-leaf clusters has an association with a parent cluster,

wherein each of the root cluster and the plurality of non-root, non-leaf clusters has associations with a left child cluster and a right child cluster and represents expression profiles represented by the left child cluster and the right child cluster of the cluster, and

wherein the root cluster represents the expression profiles of the plurality of cells.

42. The method of any one of claims 40-41, wherein for each cluster with two or more child clusters, if associations between the cluster with the two or more child clusters are invalid, adding the cluster to a merging cluster set comprises:

while traversing through each cluster of the dendrogram from the root cluster of the dendrogram to the plurality of leaf clusters of the dendrogram:

- (1) determining whether associations of the cluster with child clusters of the cluster are valid or invalid; and
- (2) if the associations are invalid, adding the cluster to a merging cluster set.

43. The method of any one of claims 36-42, comprising, prior to receiving the expression profiles of the plurality of cells in (a):

(f) barcoding the plurality of targets in the 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 have different cell labels, and wherein barcoded targets created from targets of one cell of the plurality of cells have different molecular labels;

- (g) obtaining sequencing data of the plurality of barcoded targets; and
- (h) for each of the plurality of cells:

(1) counting the number of molecular labels with distinct sequences associated with each target of the plurality of targets in the sequencing data for the cell; and

(2) estimating the number of each target of the plurality of targets for the cell based on the number of molecular labels with distinct sequences associated with the target in the sequencing data counted in (h)(1).

44. The method of claim 43, wherein the expression profile of the cell of the plurality of cells comprises the number of each target of the plurality of targets for the cell estimated in (h)(2).

45. The method of any one of claims 36-44, comprising, prior to clustering the expression profiles of the plurality of cells to generate the plurality of clusters of expression profiles based on the distances between the expression profiles of the plurality of cells in (b):

(i) determining a distance data structure of the expression profiles of the plurality of cells.

46. The method of claim 45, wherein the distance data structure comprises a distance matrix of the expression profiles of the plurality of cells.

47. The method of claim 46, wherein each diagonal element of the distance matrix has a value of zero.

48. The method of any one of claims 46-47, wherein clustering the expression profiles of the plurality of cells to generate the plurality of clusters of expression profiles based on the distances between the expression profiles of the plurality of cells in (b) comprises: clustering the expression profiles of the plurality of cells to generate the plurality of clusters of expression profiles based on the distance matrix.

49. The method of any one of claims 45-48, wherein the distances between the expression profiles of the plurality of cells are pairwise correlation distances between the expression profiles of the plurality of cells.

50. The method of any one of claims 45-49, comprising, prior to determining the distance data structure in (i), log-transforming the target counts data structure into a log-transformed target counts data structure,

wherein determining the distance data structure of elements of the target counts data structure comprises determining the distance data structure of the log-transformed target counts data structure, and

wherein clustering the expression profiles of the plurality of cells to generate the plurality of clusters of expression profiles based on the distances between the expression profiles of the plurality of cells in (b) comprises: clustering the expression profiles of the plurality of cells based on the log-transformed target counts data structure and the distance data structure to generate the plurality of clusters.

51. The method of claim 50, wherein log-transforming the target counts data structure into the log-transformed target counts data structure comprises increasing the value of each element of the target counts data structure by an increment.

52. The method of claim 51, wherein the increment is one.

53. The method of any one of claims 36-52, wherein clustering the expression profiles of the plurality of cells based on distances between the expression profiles of the plurality of cells in (b) comprises:

assigning each expression profile of the plurality of cells to a different leaf cluster in the plurality of clusters; and

iteratively combining a first cluster and a second cluster of the plurality of clusters to generate a parent cluster of the first cluster and the second cluster if the second cluster is the closest cluster of the plurality of cluster to the first cluster.

54. The method of claim 53, wherein the distance between the first cluster and the second cluster is the maximum distance between any expression profile represented by the first cluster and any expression profile represented by the second cluster.

55. The method of any one of claims 53-54, wherein an intra-cluster correlation of the first cluster and an intra-cluster correlation of the second cluster are higher than an inter-cluster correlation of the first cluster and the second cluster.

56. The method of any one of claims 15-16, wherein an indication of an intra-cluster correlation of the first cluster and an intra-cluster correlation of the second cluster is higher than an inter-cluster correlation of the first cluster and the second cluster.

57. The method of any one of claims 17-18, wherein the indication of the intra-cluster correlation of the first cluster and the intra-cluster correlation of the second cluster is based on at least one of:

- an intra-cluster maximum correlation of the first cluster and the second cluster,
- an intra-cluster average correlation of the first cluster and the second cluster,
- an intra-cluster median correlation of the first cluster and the second cluster,
- an intra-cluster minimum correlation of the first cluster and the second cluster, and
- any combinations thereof.

58. The method of any one of claims 17-19, wherein the intra-cluster correlation of the first cluster is based on at least one of:

- an intra-cluster maximum correlation of the first cluster,
- an intra-cluster average correlation of the first cluster,
- an intra-cluster median correlation of the first cluster,
- an intra-cluster minimum correlation of the first cluster, and
- any combinations thereof.

59. The method of any one of claims 17-20, wherein the intra-cluster correlation of the second cluster is based on at least one of:

- an intra-cluster maximum correlation of the second cluster,
- an intra-cluster average correlation of the second cluster,

an intra-cluster median correlation of the second cluster,
an intra-cluster minimum correlation of the second cluster, and
any combinations thereof.

60. The method of any one of claims 17-21, wherein the inter-cluster correlation of the first cluster and the second cluster is based on at least one of:

an inter-cluster maximum correlation of the first cluster and the second cluster,
an inter-cluster average correlation of the first cluster and the second cluster,
an inter-cluster median correlation of the first cluster and the second cluster,
an inter-cluster minimum correlation of the first cluster and the second cluster, and
any combinations thereof.

61. The method of any one of claims 42-54, comprising, at each cluster when traversing the plurality of clusters of the dendrogram:

if the associations are valid, continuing traversing from the cluster to the left child cluster and the right child cluster of the cluster; and

if the associations are invalid, discontinuing traversing from the cluster to the left child cluster and the right child cluster of the cluster.

62. The method of any one of claims 42-61, wherein determining whether the associations of the cluster with the child clusters of the cluster are valid or invalid comprises:

determining the associations to be valid if the distance between the left child cluster and the right child cluster is above an association threshold, and

otherwise invalid.

63. The method of claims 62, wherein the distance between the left child cluster and the right child cluster is determined based on a statistical test performed on each target of the plurality of targets between expression profiles represented by the left child cluster and the right child cluster.

64. The method of claim 63, wherein the statistical test comprises a Welch's *t*-test.

65. The method of any one of claims 63-64, wherein the distance between the left child cluster and the right child cluster is determined based on the maximum *p*-value of the statistical test performed on the each target of the plurality of targets between expression profile represented by the left child cluster and each expression profile represented by the right child cluster.

66. The method of any one of claims 42-65, comprising, at each cluster when traversing the plurality of clusters of the dendrogram: (3) adding the cluster to the merging cluster set if the cluster represents an expression profile of a single cell.

67. The method of any one of claims 42-66, comprising, at each cluster when traversing the plurality of clusters of the dendrogram: assigning a cluster label to the cluster.

68. The method of claim 67,
wherein, if the cluster represents an expression profile of a single cell, the cluster label of the cluster comprises a single cell designation,
otherwise, if the cluster is the left child cluster of the parent cluster, the cluster label of the cluster comprises the cluster label of the parent cluster and a left designation, and
otherwise, the cluster label of the cluster comprises the cluster label of the parent cluster and a right designation.

69. The method of any one of claims 36-68, wherein for each cluster in the merging cluster set, identifying the targets for distinguishing the cell types based on the expression profiles of the plurality of targets of the cells represented by the cluster comprises:
determining a difference, between expression profiles represented by the cluster and expression profiles represented by another cluster in the merging cluster set, in numbers of molecular labels with distinct sequences associated with the targets for distinguishing the cell types is greater than a significance threshold.

70. The method of any one of claims 36-69, comprising, prior to merging the first cluster with the second cluster to generate the merged cluster in (d):

merging each third cluster in the merging cluster set representing an expression profile of a single cell with a fourth cluster in the merging cluster set if a distance between the third cluster and the fourth cluster is within a cluster distance threshold.

71. The method of any one of claims 36-70, comprising classifying the plurality of cells based on the clusters in the merging cluster set that represent expression profiles of the cells.

72. The method of any one of claims 36-71, comprising designing a whole transcriptome assay based on the targets for distinguishing cell types identified.

73. The method of any one of claims 36-71, comprising designing a targeted transcriptome assay based on the targets for distinguishing cell types identified.

74. 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 cause the processor to perform the method of any one of claims 1-73.
75. A computer readable medium comprising code for performing the method for performing the method of any one of claims 1-73.