

CLAIMS

We claim:

1. A method comprising:

receiving single cell sequence data, dextramer sequence data, and single cell T Cell

Receptor (TCR) sequence data;

determining, for each cell represented in the dextramer sequence data, based on the single cell sequence data, a number of genes;

removing, from the dextramer sequence data, data associated with cells having a number of genes outside of a gene threshold range;

determining, for each cell represented in the dextramer sequence data, based on the single cell sequence data, a fraction of mitochondrial gene expression;

removing, from the dextramer sequence data, data associated with cells having a fraction of mitochondrial gene expression that exceeds a gene expression threshold;

determining, based on the dextramer sequence data,

sorted dextramer sequence data wherein the sorted dextramer sequence data comprises sorted test dextramer sequence data and negative control dextramer sequence data and

unsorted dextramer sequence data, wherein the unsorted dextramer sequence data comprises unsorted test dextramer sequence data;

determining, for each cell represented in the dextramer sequence data, based on the negative control dextramer sequence data, a maximum negative control dextramer signal;

determining, for each cell represented in the dextramer sequence data, based on the sorted test dextramer sequence data, a maximum sorted dextramer signal;

determining, for each cell represented in the dextramer sequence data, based on the unsorted test dextramer sequence data, a maximum unsorted dextramer signal;

estimating, based on the maximum negative control dextramer signals, a dextramer binding background noise;

estimating, based on the maximum sorted dextramer signals and the maximum unsorted dextramer signals, a dextramer sorting gate efficiency;

determining, based on the dextramer binding background noise and the dextramer sorting gate efficiency, a measure of background noise;

- subtracting, for each cell represented in the dextramer sequence data, the measure of background noise from a dextramer signal associated with each cell;
- performing, for each cell represented in the dextramer sequence data, cell-wise normalization on the dextramer signals associated with each cell;
- performing, for each cell represented in the dextramer sequence data, pMHC-wise normalization;
- determining, for each cell represented in the dextramer sequence data, based on the single cell TCR sequence data, a presence or an absence of at least one α -chain and at least one β -chain;
- removing, from the normalized dextramer sequence data, based on the presence or the absence of the at least one α -chain and the at least one β -chain, data associated with cells having only an α -chain, only a β -chain, or multiple α - or β -chains;
- and
- identifying data remaining in the normalized dextramer sequence data as associated with reliable TCR-pMHC binding events.
2. The method of claim 1, wherein the gene threshold range is from about 200 genes to about 2,500 genes.
 3. The method of claim 1, wherein the gene expression threshold is about 40 percent of total unique molecular identifier counts.
 4. The method of claim 1, wherein estimating, based on the maximum sorted dextramer signals and the maximum unsorted dextramer signals, the dextramer sorting gate efficiency comprises determining a maximum difference between the maximum sorted dextramer signals and the maximum unsorted dextramer signals.
 5. The method of claim 1, further comprising training a predictive model based on the data associated with reliable TCR-pMHC binding events.
 6. The method of claim 5, further comprising predicting a binding status of a newly presented receptor sequence according to the trained predictive model.
 7. The method of claim 5, further comprising:

- presenting, to the predictive model, subject TCR sequence data;
determining, by the predictive model, based on the subject TCR sequence data, a subject TCR binding pattern; and
determining, based on a repository of antigen locations and the subject TCR binding pattern, a likelihood that a subject associated with the TCR sequence data has traveled to one or more locations.
8. The method of claim 1, further comprising generating, based on the data remaining in the normalized dextramer sequence data associated with reliable TCR-pMHC binding events, a TCR binding pattern for a subject.
9. The method of claim 8, further comprising:
- receiving, at a subsequent point in time, second single cell sequence data, second dextramer sequence data, and second single cell T Cell Receptor (TCR) sequence data for the subject;
determining, based on the second single cell sequence data, second dextramer sequence data, and second single cell T Cell Receptor (TCR) sequence data for the subject, a second TCR binding pattern; and
identifying, based on a comparison of the TCR binding pattern for the subject and the second TCR binding pattern, the subject.
10. A method comprising:
- receiving single cell sequencing data comprising single cell sequence data, dextramer sequence data, and single cell T-Cell Receptor (TCR) sequence data;
filtering, from the dextramer sequence data, based on the single cell sequence data, data associated with low-quality cells;
adjusting, based on a measure of background noise, the dextramer sequence data;
filtering, from the dextramer sequence data, based on the single cell TCR- data, data according to a presence or an absence of an α -chain or a β -chain; and
identifying data remaining in the normalized filtered dextramer sequence data as associated with reliable TCR-pMHC binding events.

11. The method of claim 10, wherein filtering, from the dextramer sequence data, based on the single cell sequence data, data associated with low-quality cells comprises:

determining, for each cell represented in the dextramer sequence data, based on the single cell sequence data, a number of genes;
removing, from the dextramer sequence data, data associated with cells having a number of genes outside of a gene threshold range;
determining, for each cell represented in the dextramer sequence data, based on the single cell sequence data, a fraction of mitochondrial gene expression; and
removing, from the dextramer sequence data, data associated with cells having a fraction of mitochondrial gene expression that exceeds a gene expression threshold.
12. The method of claim 11, wherein the gene threshold range is from about 200 genes to about 2,500 genes.
13. The method of claim 11, wherein the gene expression threshold is about 40 percent of total unique molecular identifier counts.
14. The method of claim 10, further comprising determining, based on the dextramer sequence data, sorted dextramer sequence data wherein the sorted dextramer sequence data comprises sorted test dextramer sequence data and negative control dextramer sequence data and unsorted dextramer sequence data, wherein the unsorted dextramer sequence data comprises unsorted test dextramer sequence data.
15. The method of claim 14, further comprising:

determining, for each cell represented in the dextramer sequence data, based on the negative control dextramer sequence data, a maximum negative control dextramer signal;
determining, for each cell represented in the dextramer sequence data, based on the sorted test dextramer sequence data, a maximum sorted dextramer signal; and
determining, for each cell represented in the dextramer sequence data, based on the unsorted test dextramer sequence data, a maximum unsorted dextramer signal.

16. The method of claim 15, wherein adjusting, based on the measure of background noise, the dextramer sequence data comprises:
- estimating, based on the maximum negative control dextramer signals, a dextramer binding background noise;
 - estimating, based on the maximum sorted dextramer signals and the maximum unsorted dextramer signals, a dextramer sorting gate efficiency;
 - determining, based on the dextramer binding background noise and the dextramer sorting gate efficiency measure of background noise; and
 - subtracting, for each cell represented in the dextramer sequence data, the measure of background noise from a dextramer signal associated with each cell.
17. The method of claim 16, wherein estimating, based on the maximum sorted dextramer signals and the maximum unsorted dextramer signals, the dextramer sorting gate efficiency comprises determining a maximum difference between the maximum sorted dextramer signals and the maximum unsorted dextramer signals.
18. The method of claim 10, further comprising normalizing the dextramer sequence data.
19. The method of claim 18, wherein normalizing the dextramer sequence data comprises:
- performing, for each cell represented in the dextramer sequence data, cell-wise and normalization on the dextramer signals associated with each cell; and
 - performing, for each cell represented in the dextramer sequence data, pMHC-wise normalization.
20. The method of claim 10, wherein filtering, from the dextramer sequence data, based on the single cell TCR- data, data according to the presence or the absence of the α -chain or the β -chain comprises:
- determining, for each cell represented in the dextramer sequence data, based on the single cell TCR sequence data, a presence or an absence of at least one α -chain and at least one β -chain; and

removing, from the normalized dextramer sequence data, based on the presence or the absence of the at least one α -chain and the at least one β -chain, data associated with cells having only an α -chain, only a β -chain, or multiple α - or β -chains.

21. The method of claim 10, further comprising training a predictive model based on the data remaining in the normalized filtered dextramer sequence data.

22. The method of claim 21, further comprising predicting a binding status of a newly presented receptor sequence according to the trained predictive model.

23. The method of claim 22, further comprising:

presenting, to the predictive model, subject TCR sequence data;
determining, by the predictive model, based on the subject TCR sequence data, a subject TCR binding pattern; and
determining, based on a repository of antigen locations and the subject TCR binding pattern, a likelihood that a subject associated with the TCR sequence data has traveled to one or more locations.

24. The method of claim 10, further comprising generating, based on the data remaining in the normalized dextramer sequence data associated with reliable TCR-pMHC binding events, a TCR binding pattern for a subject.

25. The method of claim 24, further comprising:

receiving, at a subsequent point in time, second single cell sequence data, second dextramer sequence data, and second single cell T Cell Receptor (TCR) sequence data for the subject;
determining, based on the second single cell sequence data, second dextramer sequence data, and second single cell T Cell Receptor (TCR) sequence data for the subject, a second TCR binding pattern; and
identifying, based on a comparison of the TCR binding pattern for the subject and the second TCR binding pattern, the subject.

26. A method comprising:

performing TCR-pMHC binding specificity data normalization on dextramer sequence data to identify a plurality of TCR-pMHC binding events;
determining, based on the normalized dextramer sequence data, a training dataset comprising a plurality of TCR sequences wherein each TCR sequence is associated with a binding affinity;
determining, based on the plurality of TCR sequences, a plurality of features for a predictive model;
training, based on a first portion of the training dataset, the predictive model according to the plurality of features;
testing, based on a second portion of the training dataset, the predictive model; and
outputting, based on the testing, the predictive model.

27. The method of claim 26, wherein performing TCR-pMHC binding specificity data normalization on the dextramer sequence data to identify the plurality of TCR-pMHC binding events comprises one or more of the methods of claims 1-4 or 7-17.

28. The method of claim 26, wherein determining, based on the normalized dextramer sequence data, the training dataset comprising the plurality of TCR sequences wherein each TCR sequence is associated with a binding affinity comprises:

determining, for each TCR sequence of the plurality of TCR sequences, a paired $\alpha\beta$ chain CDR3 amino acid sequence, a V gene segment sequence, and a J gene segment sequence; and
encoding, for each TCR sequence of the plurality of TCR sequences, the paired $\alpha\beta$ chain CDR3 amino acid sequence, the V gene segment sequence, and the J gene segment sequence into a one-dimensional input vector.

29. The method of claim 28, wherein encoding, for each TCR sequence of the plurality of TCR sequences, the paired $\alpha\beta$ chain CDR3 amino acid sequence comprises transforming each alphabetical representation of an amino acid into a numerical representation of the amino acid.

30. The method of claim 28, wherein encoding, for each TCR sequence of the plurality of TCR sequences, the V gene segment sequence and the J gene segment sequence comprises one hot encoding to generate a categorical and discrete representation of gene names in numerical space.
31. The method of claim 28, further comprising clustering the one-dimensional input vectors into one or more clusters.
32. The method of claim 31, wherein clustering the one-dimensional input vectors into one or more clusters comprising applying a KNN clustering algorithm to the one-dimensional input vectors.
33. The method of claim 31, wherein the one or more clusters are indicative binding strength.
34. The method of claim 26, wherein the predictive model comprises a weighted binary classifier or a Convolutional Neural Network (CNN).
35. The method of claim 26, wherein training, based on the first portion of the training dataset, the predictive model according to the plurality of features comprises training a Neural Network by embedding the one-hot encoded V and J genes of each chain of the TCR sequence via learned embeddings, and concatenating these embeddings together with the output of a Convolutional Neural Network for each CDR3, which is fed the embedded CDR3, forming a 1D numerical vector representing the TCR, followed by passing each numeric TCR sequence through a final fully connected layer.
36. The method of claim 26, wherein training, based on a first portion of the training dataset, the predictive model according to the plurality of features comprises applying a class-weighted cost function.
37. The method of claim 26, further comprising:

presenting, to the trained predictive model, an unknown TCR sequence; and
predicting, by the trained predictive model, a binding affinity.

38. The method of claim 26, further comprising:

presenting, to the predictive model, subject TCR sequence data;
determining, by the predictive model, based on the subject TCR sequence data, a
subject TCR binding pattern; and
determining, based on a repository of antigen locations and the subject TCR binding
pattern, a likelihood that a subject associated with the TCR sequence data has
traveled to one or more locations.

39. The method of claim 26, further comprising generating, based on the data remaining
in the normalized dextramer sequence data associated with reliable TCR-pMHC
binding events, a TCR binding pattern for a subject.

40. The method of claim 39, further comprising:

receiving, at a subsequent point in time, second single cell sequence data, second
dextramer sequence data, and second single cell T Cell Receptor (TCR)
sequence data for the subject;
determining, based on the second single cell sequence data, second dextramer
sequence data, and second single cell T Cell Receptor (TCR) sequence data
for the subject, a second TCR binding pattern; and
identifying, based on a comparison of the TCR binding pattern for the subject and the
second TCR binding pattern, the subject.

41. A method comprising:

presenting, to a trained predictive model, an unknown TCR sequence, wherein the
trained predictive model is trained based on a training data set derived
according to the methods of claims 1-4 or 7-17; and
predicting, by the trained predictive model, a binding affinity.

42. A method comprising:

presenting, to a trained predictive model, subject TCR sequence data, wherein the trained predictive model is trained based on a training data set derived according to the methods of claims 1-4 or 7-17

determining, by the predictive model, based on the subject TCR sequence data, a subject TCR binding pattern; and

determining, based on a repository of antigen locations and the subject TCR binding pattern, a likelihood that a subject associated with the TCR sequence data has traveled to one or more locations.

43. A method comprising:

performing TCR-pMHC binding specificity data normalization on dextramer sequence data of a subject to identify a plurality of TCR-pMHC binding events;

generating, based on the plurality of TCR-pMHC binding events, a TCR binding pattern for the subject.

performing, at a subsequent time, TCR-pMHC binding specificity data normalization on second dextramer sequence data;

determining, based on the second dextramer sequence data, a second TCR binding pattern; and

identifying, based on a comparison of the TCR binding pattern for the subject and the second TCR binding pattern, the second dextramer sequence data as associated with the subject.

44. A method comprising:

receiving single cell sequence data, dextramer sequence data, and single cell T Cell Receptor (TCR) sequence data;

determining, for each cell represented in the dextramer sequence data, based on the single cell sequence data, a number of genes;

removing, from the dextramer sequence data, data associated with cells having a number of genes outside of a gene threshold range;

determining, for each cell represented in the dextramer sequence data, based on the single cell sequence data, a fraction of mitochondrial gene expression;

removing, from the dextramer sequence data, data associated with cells having a fraction of mitochondrial gene expression that exceeds a gene expression threshold;

determining, based on the dextramer sequence data, sorted dextramer sequence data wherein the sorted dextramer sequence data comprises sorted test dextramer sequence data and negative control dextramer sequence data;

determining, for each cell represented in the dextramer sequence data, based on the negative control dextramer sequence data, a maximum negative control dextramer signal;

determining, for each cell represented in the dextramer sequence data, based on the sorted test dextramer sequence data, a maximum sorted dextramer signal;

estimating, based on the maximum negative control dextramer signals and the maximum sorted dextramer signals, a dextramer binding background noise;

determining, for each cell represented in the dextramer sequence data, based on the single cell TCR sequence data, a presence or an absence of at least one α -chain and at least one β -chain;

removing, from the dextramer sequence data, based on the presence or the absence of the at least one α -chain and the at least one β -chain, data associated with cells having only an α -chain, only a β -chain, or multiple α - or β -chains;

determining, for each dextramer binding to a given cell represented in the dextramer sequence data, a ratio of dextramer signal within the cell to a sum of all dextramers binding to the cell (a measure of the dextramer binding specificity to the cell);

determining, for each dextramer binding to a given TCR clonotype of each cell represented in the dextramer sequence data, a fraction of T cells within a clone binding to a particular dextramer (a measure of the dextramer binding specificity to the clonotype to which the cell belongs);

determining, for each dextramer binding to a given cell represented in the dextramer sequence data, based on the measure the of the dextramer binding specificity to the cell and the measure of the dextramer binding specificity to the clonotype to which the cell belongs, a corrected dextramer signal associated with each dextramer binding to the cell;

performing, for each cell represented in the dextramer sequence data, cell-wise normalization on the dextramer signals associated with each cell;
 performing, for each cell represented in the dextramer sequence data, pMHC-wise normalization; and
 identifying, based on a threshold, data remaining in the normalized dextramer sequence data as associated with reliable TCR-pMHC binding events.

45. The method of claim 44, wherein determining, for each dextramer binding to a given cell represented in the dextramer sequence data, a ratio of dextramer signal within the cell to a sum of all dextramers binding to the cell comprises:

determining a background noise subtracted dextramer signal E_{ij} , for the i^{th} T cell binding the j^{th} dextramer; and
 determining a fraction of dextramer signal due to binding of the j^{th} dextramer for the i^{th} T cell by evaluating:

$$RC_{ij} = \frac{E_{ij}}{\sum_{j=1}^n E_{ij}} .$$

46. The method of claim 44, wherein determining, for each dextramer binding to a given TCR clonotype of each cell represented in the dextramer sequence data, a fraction of T cells within a clone binding to a particular dextramer comprises:

determining a TCR clonotype k_i , of the i^{th} T cell;
 determining a number of T cells, $T_{k_{ij}}$, belonging to clonotype k_i that bind dextramer j; and
 determining a fraction of T cells that belong to clonotype k_i that bind the j^{th} dextramer by evaluating:

$$RT_{k_{ij}} = \frac{T_{k_{ij}}}{\sum_{j=1}^n T_{k_{ij}}} .$$

47. The method of claim 44, wherein determining, for each dextramer binding to a given cell represented in the dextramer sequence data, based on the measure the of the dextramer binding specificity to the cell and the measure of the dextramer binding

specificity to the clonotype to which the cell belongs, a corrected dextramer signal associated with each dextramer binding to the cell comprises:

determining the corrected dextramer signal for the i^{th} T cell binding the j^{th} dextramer by evaluating:

$$S_{ij} = E_{ij}(RC_{ij})^2 RT_{kj} .$$

48. An apparatus configured to perform any of the preceding methods.

49. A computer readable medium having processor-executable instructions embodiment thereon configured to cause an apparatus to perform any of the preceding methods.