SUMMARY

[0004a] In one aspect, the present disclosure refers to a method comprising:

(a) processing amino acid information of a plurality of candidate peptide sequences using a machine learning HLA peptide presentation prediction model to generate a plurality of presentation predictions, wherein each candidate peptide sequence of the plurality of candidate peptide sequences is encoded by a genome or exome of a subject, or a pathogen or a virus in the subject, wherein the plurality of presentation predictions comprises an HLA presentation prediction for each of the plurality of candidate peptide sequences, wherein each HLA presentation prediction is indicative of a likelihood that one or more proteins encoded by a class II HLA allele of a cell of the subject can present a given candidate peptide sequence of the plurality of candidate peptide sequences,

wherein the machine learning HLA peptide presentation prediction model is trained using training data comprising sequence information of sequences of training peptides identified by mass spectrometry to be presented by an HLA protein expressed in training cells, wherein the training cells express an HLA allele with an affinity tag; and

(b) identifying, based at least on the plurality of presentation predictions, a peptide sequence of the plurality of peptide sequences as being presented by at least one of the one or more proteins encoded by a class II HLA allele of a cell of the subject;

wherein the machine learning HLA peptide presentation prediction model has a positive predictive value (PPV) of at least 0.07, wherein the PPV is calculated as a fraction of peptide sequences in the training data identified as being presented by a class II HLA allele of a cell in peptide sequences predicted to be presented by the class II HLA allele of the cell.

[0004b] In another aspect, the present disclosure refers to a method comprising:

(a) processing amino acid information of a plurality of peptide sequences encoded by a genome or exome of a subject using a machine learning HLA peptide binding prediction model to generate a plurality of binding predictions, wherein the plurality of binding predictions comprises an HLA binding prediction for each of the plurality of candidate peptide sequences, each binding prediction indicative of a likelihood that one or more proteins encoded by a class II HLA allele of a cell of the subject binds to a given candidate peptide sequence of the plurality of candidate peptide sequences,

wherein the machine learning HLA peptide binding prediction model is trained using training data comprising sequence information of sequences of peptides identified to bind to an HLA class II protein or a derivative thereof expressed in training cells, wherein the training cells express an HLA allele with an affinity tag; and

(b) identifying, based at least on the plurality of binding predictions, a peptide sequence of the plurality of peptide sequences that has a probability greater than a threshold binding prediction probability value of binding to at least one of the one or more proteins encoded by a class II HLA allele of a cell of the subject;

wherein the machine learning HLA peptide binding prediction model has a positive predictive value (PPV) of at least 0.1, wherein the PPV is calculated as a fraction of peptide sequences in the training data identified as being presented by a class II HLA allele of a cell in peptide sequences predicted to be presented by the class II allele of the cell.

[0005] The methods and compositions described herein find uses in a wide range of applications. For example, the methods and compositions described herein can be used to identify immunogenic antigen

CLAIMS

1. A method comprising:

(a) processing amino acid information of a plurality of candidate peptide sequences using a machine learning HLA peptide presentation prediction model to generate a plurality of presentation predictions, wherein each candidate peptide sequence of the plurality of candidate peptide sequences is encoded by a genome or exome of a subject, or a pathogen or a virus in the subject, wherein the plurality of presentation predictions comprises an HLA presentation prediction for each of the plurality of candidate peptide sequences, wherein each HLA presentation prediction is indicative of a likelihood that one or more proteins encoded by a class II HLA allele of a cell of the subject can present a given candidate peptide sequence of the plurality of candidate peptide sequences,

wherein the machine learning HLA peptide presentation prediction model is trained using training data comprising sequence information of sequences of training peptides identified by mass spectrometry to be presented by an HLA protein expressed in training cells, wherein the training cells express an HLA allele with an affinity tag; and

(b) identifying, based at least on the plurality of presentation predictions, a peptide sequence of the plurality of peptide sequences as being presented by at least one of the one or more proteins encoded by a class II HLA allele of a cell of the subject;

wherein the machine learning HLA peptide presentation prediction model has a positive predictive value (PPV) of at least 0.07, wherein the PPV is calculated as a fraction of peptide sequences in the training data identified as being presented by a class II HLA allele of a cell in peptide sequences predicted to be presented by the class II HLA allele of the cell.

2. A method comprising:

(a) processing amino acid information of a plurality of peptide sequences encoded by a genome or exome of a subject using a machine learning HLA peptide binding prediction model to generate a plurality of binding predictions, wherein the plurality of binding predictions comprises an HLA binding prediction for each of the plurality of candidate peptide sequences, each binding prediction indicative of a likelihood that one or more proteins encoded by a class II HLA allele of a cell of the subject binds to a given candidate peptide sequence of the plurality of candidate peptide sequences,

wherein the machine learning HLA peptide binding prediction model is trained using training data comprising sequence information of sequences of peptides identified to

bind to an HLA class II protein or a derivative thereof expressed in training cells, wherein the training cells express an HLA allele with an affinity tag; and

(b) identifying, based at least on the plurality of binding predictions, a peptide sequence of the plurality of peptide sequences that has a probability greater than a threshold binding prediction probability value of binding to at least one of the one or more proteins encoded by a class II HLA allele of a cell of the subject;

wherein the machine learning HLA peptide binding prediction model has a positive predictive value (PPV) of at least 0.1, wherein the PPV is calculated as a fraction of peptide sequences in the training data identified as being presented by a class II HLA allele of a cell in peptide sequences predicted to be presented by the class II allele of the cell.

3. The method of claim 2, wherein the machine learning HLA peptide presentation prediction model is trained using training data comprising sequence information of sequences of training peptides identified by mass spectrometry to be presented by an HLA protein expressed in the training cells.

4. The method of claim 1 or 2, wherein the method comprises ranking, based on the presentation predictions, at least two peptides identified as being presented by at least one of the one or more proteins encoded by the class II HLA allele of the cell of the subject.

5. The method of claim 1 or 2, wherein the method comprises selecting one or more peptides of the two or more ranked peptides.

6. The method claim 1 or 2, wherein the method comprises selecting one or more peptides of the plurality that were identified as being presented by at least one of the one or more proteins encoded by the class II HLA allele of the cell of the subject.

7. The method of claim 1 or 2, wherein the method comprises selecting one or more peptides of two or more peptides ranked based on the presentation predictions.

8. The method of claim 1, wherein the machine learning HLA peptide presentation prediction model has a positive predictive value (PPV) of at least 0.07 when amino acid information of a plurality of test peptide sequences are processed to generate a plurality of test presentation predictions, each test presentation prediction indicative of a likelihood that the one or more proteins encoded by the class II HLA allele of the cell of the subject can present a given test peptide sequence of the plurality of test peptide sequences, wherein the plurality of test peptide sequences comprises at least 500 test peptide sequences comprising (i) at least one hit peptide sequence identified by mass spectrometry to be presented by an HLA protein expressed in cells and (ii) at least 499 decoy peptide sequences contained within a protein encoded by a genome of an organism, wherein the organism and the subject are the same species, wherein the plurality of test peptide sequences comprises a ratio of

1:499 of the at least one hit peptide sequence to the at least 499 decoy peptide sequences and a top percentage of the plurality of test peptide sequences are predicted to be presented by the HLA protein expressed in cells by the machine learning HLA peptide presentation prediction model.

9. The method of claim 2, wherein the machine learning HLA peptide presentation prediction model has a positive predictive value (PPV) of at least 0.1 when amino acid information of a plurality of test peptide sequences are processed to generate a plurality of test binding predictions, each test binding prediction indicative of a likelihood that the one or more proteins encoded by the class II HLA allele of the cell of the subject binds to a given test peptide sequence of the plurality of test peptide sequences, wherein the plurality of test peptide sequences comprises at least 20 test peptide sequences comprising (i) at least one hit peptide sequence identified by mass spectrometry to be presented by an HLA protein expressed in cells and (ii) at least 19 decoy peptide sequences contained within a protein comprising at least one peptide sequence identified by mass spectrometry to be presented by an HLA protein expressed in cells, such as a single HLA protein expressed in cells (e.g., mono-allelic cells), wherein the plurality of test peptide sequences comprises a ratio of 1:19 of the at least one hit peptide sequence to the at least 19 decoy peptide sequences and a top percentage of the plurality of test peptide sequences are predicted to bind to the HLA protein expressed in cells by the machine learning HLA peptide presentation prediction model.

10. The method of claims 8 or 9, wherein no amino acid sequence overlap exist among the at least one hit peptide sequence and the decoy peptide sequences.

11. The method of any one of claims 1-10, wherein the machine learning HLA peptide presentation prediction model has a positive predictive value (PPV) of at least 0.08, 0.09, 0.1, 0.11, 0.12, 0.13, 0.14, 0.15, 0.16, 0.17, 0.18, 0.19, 0.2, 0.21, 0.22, 0.23, 0.24, 0.25, 0.26, 0.27, 0.28, 0.29, 0.3, 0.31, 0.32, 0.33, 0.34, 0.35, 0.36, 0.37, 0.38, 0.39, 0.4, 0.41, 0.42, 0.43, 0.44, 0.45, 0.46, 0.47, 0.48, 0.49, 0.5, 0.51, 0.52, 0.53, 0.54, 0.55, 0.56, 0.57, 0.58, 0.59, 0.6, 0.61, 0.62, 0.63, 0.64, 0.65, 0.66, 0.67, 0.68, 0.69, 0.7, 0.71, 0.72, 0.73, 0.74, 0.75, 0.76, 0.77, 0.78, 0.79, 0.8, 0.81, 0.82, 0.83, 0.84, 0.85, 0.86, 0.87, 0.88, 0.89, 0.9, 0.91, 0.92, 0.93, 0.94, 0.95, 0.96, 0.97, 0.98 or 0.99.

12. The method of any one of claims 8-11, wherein the at least one hit peptide sequence comprises at least 5, 10, 20, 50 or 100 hit peptide sequences.

13. The method of any one of claims 8 and 10-12, wherein the at least 499 decoy peptide sequences comprises at least 2500, 5000, 10000, 25000, 50000 or 100000 decoy peptide sequences.

14. The method of any one of claims 8 and 10-13, wherein the at least 500 test peptide sequences comprises at least 2500, 5000, 10000, 25000, 50000 or 100000 test peptide sequences.

15. The method of any one of claims 8-14, wherein the top percentage of the plurality of test peptide sequences is a top 0.2%, 0.3%, 0.4%, 0.5%, 0.6%, 0.7%, 0.8%, 0.9%, 1.00% or 2.00%.

16. The method of any one of claims 9-12, wherein the at least 19 decoy peptide sequences comprises at least 500, 1000, 2000, 5000 or 10000 decoy peptide sequences.

17. The method of any one of claims 9-12 and 16, wherein the at least 20 test peptide sequences comprises at least wherein the at least 500 test peptide sequences comprises at least 500, 1000, 2000, 5000 or 10000 test peptide sequences test peptide sequences.

18. The method of any one of the preceding claims, wherein the PPV is more than the PPV of column 2 of Table 11 for the protein encoded by the corresponding HLA allele of Table 11.

19. The method of any one of the preceding claims, wherein the PPV is more than the PPV of column 2 of Table 16 for the protein encoded by the HLA class II allele.

20. The method of any one of the preceding claims, wherein the subject is a single subject.

21. The method of any one of the preceding claims, wherein the training cells are cells expressing a single protein encoded by the class II HLA allele of the cell of the subject.

22. The method of any one of claims 1-21, wherein the training cells are monoallelic HLA cells.

23. The method of any one of the preceding claims, wherein each peptide sequence of the plurality of peptide sequences is associated with a cancer.

24. The method of any one of the preceding claims, wherein at least one peptide sequence of the plurality of peptide sequences is overexpressed by a cancer cell of the subject.

25. The method of any one of the preceding claims, wherein each peptide sequence of the plurality of peptide sequences is overexpressed by a cancer cell of the subject.

26. The method of any one of the preceding claims, wherein at least one peptide sequence of the plurality of peptide sequences is a cancer cell-specific peptide.

27. The method of any one of the preceding claims, wherein each peptide sequence of the plurality of peptide sequences is a cancer cell-specific peptide.

28. The method of any one of the preceding claims, wherein each peptide sequence of the plurality of peptide sequences is expressed by a cancer cell of the subject.

29. The method of any one of the preceding claims, wherein at least one peptide sequence of the plurality of peptide sequences is not encoded by a non-cancer cell of the subject.

30. The method of any one of the preceding claims, wherein each peptide sequence of the plurality of peptide sequences is not encoded by a non-cancer cell of the subject.

31. The method of any one of the preceding claims, wherein at least one peptide sequence of the plurality of peptide sequences is not expressed by a non-cancer cell of the subject.

32. The method of any one of the preceding claims, wherein each peptide sequence of the plurality of peptide sequences is not expressed by a non-cancer cell of the subject.

33. The method of any one of the preceding claims, wherein the method comprises obtaining the plurality of peptide sequences of the subject.

34. The method of any one of the preceding claims, wherein the method comprises obtaining a plurality of polynucleotide sequences of the subject.

35. The method of any one of the preceding claims, wherein the method comprises obtaining a plurality of polynucleotide sequences of the subject that encodes the plurality of peptide sequences encoded by a genome or exome of a subject.

36. The method of any one of the preceding claims, wherein the method comprises obtaining a plurality of polynucleotide sequences of the subject that encodes the plurality of peptide sequences encoded by a genome or exome of a subject by a computer processor.

37. The method of any one of the preceding claims, wherein the method comprises obtaining a plurality of polynucleotide sequences of the subject by genomic or exome sequencing.

38. The method of any one of the preceding claims, wherein the method comprises obtaining a plurality of polynucleotide sequences of the subject by whole genome sequencing or whole exome sequencing.

39. The method of any one of the preceding claims, wherein the method further comprises formulating a composition comprising one or more of the selected sub-set of peptide sequences.

40. The method of any one of the preceding claims, wherein the HLA class II protein comprises an HLA-DR, HLA-DQ, or an HLA-DP protein.

41. The method of any one of the preceding claims, wherein the HLA class II protein is selected from the group consisting of: HLA-DPB1*01:01/HLA-DPA1*01:03, HLA-DPB1*02:01/HLA-DPA1*01:03, HLA-DPB1*03:01/HLA-DPA1*01:03, HLA-DPB1*04:01/HLA-DPA1*01:03, HLA-DPB1*04:02/HLA-DPA1*01:03, HLA-DPB1*06:01/HLA-DPA1*01:03,HLA-DQB1*02:01/HLA-DQA1*05:01, HLA-DQB1*02:02/HLA-DQA1*02:01, HLA-DQB1*06:02/HLA-DQA1*01:02, HLA-DQB1*06:04/HLA-DQA1*01:02, HLA-DRB1*01:01, HLA-DRB1*01:02, HLA-DRB1*03:01, HLA-DRB1*03:02, HLA-DRB1*04:01, HLA-DRB1*04:02, HLA-DRB1*04:03, HLA-DRB1*04:04, HLA-DRB1*04:05, HLA-DRB1*04:07, HLA-DRB1*07:01, HLA-DRB1*08:01, HLA-DRB1*08:02, HLA-DRB1*08:03, HLA-DRB1*08:04, HLA-DRB1*09:01, HLA-DRB1*10:01, HLA-DRB1*11:01, HLA-DRB1*11:02, HLA-DRB1*11:04, HLA-DRB1*12:01, HLA-DRB1*12:02, HLA-DRB1*13:01, HLA-DRB1*13:02, HLA-DRB1*13:03, HLA-DRB1*14:01, HLA-DRB1*15:01, HLA-DRB1*15:02,

HLA-DRB1*15:03, HLA-DRB1*16:01, HLA-DRB3*01:01, HLA-DRB3*02:02, HLA-DRB3*03:01, HLA-DRB4*01:01 and HLA-DRB5*01:01; an HLA-DR protein selected from the group consisting of: DPA*01:03/DPB*04:01, DRB1*01:01, DRB1*01:02, DRB1*03:01, DRB1*04:01, DRB1*04:02, DRB1*04:04, DRB1*04:05, DRB1*07:01, DRB1*08:01, DRB1*08:02, DRB1*08:03, DRB1*09:01, DRB1*11:01, DRB1*11:02, DRB1*11:04, DRB1*12:01, DRB1*13:01, DRB1*13:02, DRB1*13:03, DRB1*14:01, DRB1*15:01, DRB1*15:02, DRB1*15:03, DRB1*16:02, DRB3*01:01, DRB3*02:01, DRB3*02:02, DRB3*03:01, DRB4*01:01, DRB4*01:03 and DRB5*01:01; an HLA-DP protein selected from the group consisting of: DPB1*01:01, DPB1*02:01, DPB1*02:02, DPB1*03:01, DPB1*04:01, DPB1*04:02, DPB1*05:01, DPB1*06:01, DPB1*11:01, DPB1*13:01, DPB1*17:01; or an HLA-DQ protein complex selected from the group consisting of: A1*01:01+B1*05:01, A1*01:02+B1*06:02, A1*01:02+B1*06:04, A1*01:03+B1*06:03, A1*02:01+B1*02:02, A1*02:01+B1*03:03, A1*03:01+B1*03:02, A1*03:03+B1*03:01, A1*05:01+B1*02:01 and A1*05:05+B1*03:01.

42. The method of any one of the preceding claims, wherein the peptides presented by the HLA protein have a length of from 15-40 amino acids.

43. The method of claim 39, wherein the composition further comprises an adjuvant.

44. The method of claim 39, wherein the composition further comprises an immune checkpoint inhibitor.