CLAIMS

1. A computer-implemented method of selecting one or more amino acid sequences for inclusion in a vaccine from a set of predicted immunogenic candidate amino acid sequences, the method comprising:

identifying an immune profile response value for each candidate amino acid sequence in respect of each one of a plurality of sample components of an immune profile, wherein the immune profile response value represents whether the candidate amino acid sequence results in an immune response for the sample component of an immune profile;

retrieving a plurality of immune profiles for a population;

generating a plurality of representative immune profiles for the population, wherein the representative immune profiles overlap with the sample components of an immune profile; and,

selecting the one or more amino acid sequences for inclusion in the vaccine that minimises a likelihood of no immune response for each representative immune profile, based on the immune profile response values,

wherein the step of selecting comprises applying a mathematical optimisation algorithm to minimise a maximum likelihood of no immune response for each of the representative immune profiles, and

the immune profile comprises a set of HLA alleles and the sample components of an immune profile comprise sample HLA alleles, and wherein the variables of the mathematical optimisation algorithm comprise:

(a) a binary indicator variable for each candidate amino acid sequence which indicates whether the candidate amino acid is included in a vaccine;

(b) a continuous variable for each representative immune profile which gives a log likelihood of no immune response;

(c) a continuous variable for each sample component of an immune profile which gives a log likelihood of no response; and,

(d) a continuous variable which gives a maximum log likelihood that any representative immune profile does not respond to the selected one or more amino acid sequences,

wherein the mathematical optimisation algorithm minimises the continuous variable which gives a maximum log likelihood that any representative immune profile does not respond to the selected one or more amino acid sequences.

2. The computer-implemented method of claim 1, wherein the step of generating comprises:
    (i) creating a first distribution over the plurality of immune profiles; and,
    (ii) sampling the first distribution to create the plurality of representative immune profiles.

3. The computer-implemented method of claim 2, wherein the first distribution is a distribution over the plurality of immune profiles for each region of the population.

4. The computer-implemented method of claim 3, wherein the first distribution is a posterior distribution over genotypes in each region based on a prior distribution and observed genotypes from the plurality of immune profiles in each region of the population.

5. The computer-implemented method of claim 4, wherein the first distribution is a symmetric Dirichlet distribution, wherein the method further comprises the step of collecting all genotypes observed at least once across all regions, and wherein the step of sampling comprises sampling a desired number of genotypes from each region based on counts of each genotype in the sample.

6. The computer-implemented method of any of claims 2 to 5, further comprising:

simulating a digital population based on the retrieved plurality of immune profiles for the population, wherein the step of creating a first distribution is based on the simulated population such that the step of sampling is performed on the distribution of the simulated population.

7. The computer-implemented method of claim 6, wherein the step of simulating a digital population comprises:

defining a population size; and,

creating a second distribution over the regions.

8. The computer-implemented method of claim 7, wherein the second distribution is a Dirichlet distribution.

9. The computer-implemented method of any of the preceding claims, wherein the representative immune profiles are generated such the representative immune profiles maximise coverage of combinations of immune profiles in the population.

10. The computer-implemented method of claim 1, wherein the mathematical optimisation algorithm is a mixed integer linear program.

11. The computer-implemented method of any of the preceding claims, further comprising:

assigning a cost to each candidate amino acid sequence,

wherein the step of selecting is constrained based on the cost assigned to each candidate amino acid sequence, such that the selected one or more amino acid sequences have a total cost below a predetermined threshold budget.

12. The computer-implemented method of any of the preceding claims, wherein the step of selecting is constrained based on a maximum amount of amino acid sequences allowed in a vaccine delivery platform.

13. The computer-implemented method of any of the preceding claims, further comprising:

creating a tripartite graph, wherein:

a first set of nodes corresponds to the candidate amino acid sequences;

a second set of nodes corresponds to the sample components of an immune profile; and,

a third set of nodes corresponds to the representative immune profiles for the population,

and wherein:

weights of edges between the first set of nodes and the second set of nodes are the immune profile response values; and,

weights of edges between the second set of nodes and the third set of nodes represent correspondence between the sample components of an immune profile and each representative immune profile.

14. The computer-implemented method of any of the preceding claims, wherein the immune profile response value is a log likelihood value based on amino acid sub-sequences of the candidate amino acid sequence.

15. The computer implemented method of any of the preceding claims, wherein the step of identifying comprises selecting a best likelihood value as the immune profile response value from a likelihood value for each amino-acid subsequence.

16. The computer-implemented method of any of the preceding claims, wherein the one or more candidate amino acid sequences are comprised in one or more proteins of a coronavirus, preferably the SARS-CoV-2 virus.

17. The computer-implemented method of any of the preceding claims, wherein the representative immune profile may comprise one or more selected from a group comprising: a set of HLA alleles; presence of tumor infiltrating lymphocytes; presence of immune checkpoint markers; presence of hypoxia markers; presence of chemokine receptors; and, previous infection by human papillomavirus.

18. The computer-implemented method of any of the preceding claims, wherein the step of selecting the one or more amino acid sequences for inclusion in

the vaccine is further based on a correspondence between the sample components of an immune profile and the representative immune profiles.

19. A system for selecting one or more amino acid sequences for inclusion in a vaccine from a set of predicted immunogenic candidate amino acid sequences, the system comprising at least one processor in communication with at least one memory device, the at least one memory device having stored thereon instructions for causing the at least one processor to perform a method according to any of claims 1 to 18.

20. A computer readable medium having computer executable instructions stored thereon for implementing the method of any of claims 1 to 18.