GENETIC SIGNATURES FOR PREDICTING VACCINE RESPONSE AND USES THEREOF

Technical field

The present invention relates generally to genomics and more specifically to methods of predicting vaccine response using indel gene signatures.

Background

Vaccination is one of the most powerful tools for humans to fight infectious diseases. An individual's immune response to vaccination is typically quantified by measuring the antigen-specific antibody levels induced by the vaccine, but there can be considerable variation in these antibody responses among individuals, which suggests the influence of host-specific factors in determining vaccine efficacy. However, there is a lack of methods to identify host-specific factors that can reliably predict antibody response to a broad range of vaccines.

It would be desirable to overcome or ameliorate at least one of the above-described problems, or at least to provide a useful alternative.

Summary

Disclosed herein is a method for predicting the responsiveness of a subject to a vaccine, the method comprising: (a) detecting the presence or absence of, and the copy number of, an insertion/deletion (indel) mutation at one or more genetic loci in Table 1 in a sample obtained from the subject; and (b) calculating, based on the presence or absence of, and the copy number of, an indel mutation at the one or more genetic loci, a value predictive of the responsiveness of the subject to the vaccine.

Disclosed herein is a method of treating or preventing disease in a subject, the method comprising: (a) detecting the presence or absence of, and the copy number of, an indel mutation at one or more genetic loci in Table 1 using a sample obtained from the subject; (b) calculating, based on the presence or absence of, and the copy number of, an indel

mutation at the one or more genetic loci, a value predictive of the responsiveness of the subject to a vaccine against the infection; and (c) providing an effective amount of the vaccine to the subject based on the predicted responsiveness of the subject to the vaccine to treat or prevent the disease.

Disclosed herein is a method of selecting a subject for treatment with a vaccine, the method comprising: (a) detecting the presence or absence of, and the copy number of, an indel mutation at one or more genetic loci in Table 1 using a sample obtained from the subject; (b) calculating, based on the presence or absence of, and the copy number of, an indel mutation at the one or more genetic loci, a value predictive of the responsiveness of the subject to a vaccine; and (c) selecting a subject for treatment with the vaccine based on the predicted responsiveness of the subject to the vaccine.

Disclosed herein is a kit, comprising one or more oligonucleotides for detecting the presence or absence of, and the copy number of, an indel mutation at a genetic locus in Table 1.

Disclosed herein is a composition comprising: (a) a sample obtained from a subject; and (b) one or more oligonucleotides for detecting the presence or absence of, and the copy number of, an indel mutation at a genetic locus in Table 1 in the sample.

Disclosed herein is a computer-implemented method for predicting the responsiveness of a subject to a vaccine, the method comprising: (a) receiving genotype data comprising the presence or absence of, and the copy number of, an insertion/deletion (indel) mutation at one or more genetic loci in Table 1 from a sample from the subject; and (b) calculating a value predictive of the responsiveness of the subject to the vaccine using a machine learning model trained to generate the predictive value based on the presence or absence of, and the copy number of, an indel mutation at one or more genetic loci in Table 1.

Disclosed herein is a system for predicting the responsiveness of a subject to a vaccine, the system comprising: a processor coupled to computer readable memory, the memory comprising instructions that when executed by the processor causes the processor to: (i) receive genotype data comprising the presence or absence of, and the copy number of, an insertion/deletion (indel) mutation at one or more genetic loci in Table 1 from a sample from the subject; and (ii) calculate, based on the presence or absence of, and the copy number of,

an indel mutation at the one or more genetic loci, a value predictive of the responsiveness of the subject to the vaccine using a machine learning model.

Disclosed herein is a system for predicting the responsiveness of a subject to a vaccine, the system comprising: a) a processor coupled to computer readable memory; b) an array for high-throughput detection of the presence or absence of, and the copy number of, an indel mutation at one or more genetic loci in Table 1 in a sample obtained from the subject, wherein the array is configured to generate a plurality of detectable signals indicative of the presence or absence of, and the copy number of, an indel mutation at the one or more genetic loci; and c) a computer algorithm stored in the computer readable memory and comprising instructions executable by the processor to take as input the plurality of detectable signals generated by the array and calculate a value predictive of the responsiveness of the subject to the vaccine.

Brief description of the drawings

Embodiments of the present invention will now be described, by way of non-limiting example, with reference to the drawings in which:

FIG. 1 is a graphical illustration of a method according to an embodiment of the present disclosure. A predictive model for individual antibody response to vaccination is trained with a COVID-19 mRNA vaccine cohort, using age, gender, and 12 genetic variants (indels) as variables, and built with LASSO regression. The model can be used to predict the neutralising antibody level for an individual. The predicted neutralising antibody level can be compared to the "Low" level approved by U.S. FDA or a "Protective" level suggested to WHO.

FIG. 2 shows the significant association of neutralising antibody level on Day 21 after vaccination with that on Day 180, age, and gender. (A) Diagram of the study design. 328 healthy individuals received two doses of BNT162b2 vaccines on Day 0 and 21. Blood was drawn 0, 21, 90, and 180 days after their first dose of vaccines. Whole exome sequencing was carried out using whole blood. (B) Line graph of neutralising antibody level on Day 0, 21, 90, and 180 for 328 people in the cohort. The protective neutralising antibody level (suggested by Feng Zhu et. al, 2022) is indicated by a red line with "P". (C) Scatter plot of

neutralising antibody level on Day 21 and 180. The linear regression line is shown in green, with p value less than $2x10^{-16}$. (D) Association between age and nAb or anti-spike protein IgG levels on day 21 (linear regression, p value < 2.2×10^{-16} for both, adjusted $R^2 = 0.28$ and 0.24 for nAb and anti-spike protein IgG levels, respectively). The lines are linear regression lines. (E) Association between gender and nAb or anti-spike protein IgG levels on day 21 (linear regression, p value = 6.0×10^{-9} and 7.3×10^{-8}). (F) Association between Chinese ethnicity with nAb or anti-spike protein IgG levels on day 21 (linear regression, p value = 3.2×10^{-8} and 5.2×10^{-6} for nAb and anti-spike protein IgG levels, respectively). (G) Left: Distribution of nAb levels on day 21 in a multiple linear regression plane: nAb levels on day 21 ~ age + gender, in a 3D plot (adjusted $R^2 = 0.30$). Right: Distribution of anti-spike protein IgG levels on day 21 in a multiple linear regression plane: anti-spike protein IgG levels on day 21~ age + gender, in a 3D plot (adjusted $R^2 = 0.26$). nAb: neutralising antibody, S IgG: anti-spike protein IgG.

FIG. 3 shows the significant association of the 12 indels with neutralising antibody levels. Box plot for the distribution of individuals of each genotype for each indel. P values are derived from the association test with the additive model. nAb level on D21: neutralising antibody level on Day 21.

FIG. 4 shows the correlation of indels with cellular phenotypes. (A) Bar graph of alternative allele frequency of Indel 2, 3, 5–8 in different populations. (B) Box plot of fold change of memory B cells between Day 0 and 21, or Day 0 and 90 in people with different genotypes of Indel 8 and 11, respectively. P values* are derived from 10,000 permutations. (C) Box plot of pseudovirus neutralisation of COVID-19 omicron and delta strain for people with different genotypes of Indel 8 and 11, respectively. P values* are derived from 10,000 permutations. (D) Box plot of T follicular helper cells ($CXCR5^+CD4^+$ cells) in 1 million PBMC for people with different genotypes of Indel 8. P value is derived from the association test of additive model.

FIG. 5 shows a model to predict neutralising antibody level. (A) Diagram of random sampling for training, validation, and test set. (B) Scatter plot of observed and predicted values of the test set for three linear regression models, nAb level on D21 ~ Age + gender + 0, 5, or 12 indels. nAb on D21: neutralising antibody level on Day 21. MSE.test: mean-squared error (MSE) of the test set. (C) The plot of log(λ) and MSE from LASSO regression

(D) Violin plot of MSE for 100 re-samplings of linear and LASSO regression models. (E) Scatter plot of observed and predicted values of the test set for the best LASSO regression model (smallest MSE). The red lines indicate the "Low" level of neutralising antibody level as approved by U.S. Food and Drug Administration (FDA). Error rate of predicting people with "Low" neutralising antibody level in the test set is 18.46%, indicated in red.

FIG. 6 shows indel distribution and association test. (A) The distribution of 1528 indels in different genomic locations, CDS: coding regions, UTR: untranslated region. (B) Additive model for neutralising antibody level on Day 21 ~ $\beta_0 + \beta_1$*Age + β_2*Gender + β_3*genotype (0: REF/REF, 1: REF/ALT, 2: ALT/ALT).

FIG. 7 shows the significant association of the 8 indels with antibody waning rate. Box plot for the distribution of individuals of each genotype for each indel. nAb waning rate: neutralising antibody waning rate.

FIG. 8 shows a model to predict neutralising antibody waning rate. (A) Diagram of random sampling for training, validation, and test set. (B) The plot of log(λ) and MSE from LASSO regression (C) Scatter plot of observed and predicted values of the test set for the best LASSO regression model (smallest MSE). MSE of the test set: 5.92.

FIG. 9 shows associations between antibody responses and infection outcomes. (A) Associations between infection outcomes within 1 year of vaccination and nAb levels on days 21, 90, and 180 (logistic regression, p value = 0.0013, 0.0038, and 0.0030 for days 21, 90, and 180, respectively). The boxplots showed the median and interquartile range of the data. (B) Associations between infection outcomes within 1 year of vaccination and anti-spike protein levels on days 21, 90, and 180 (logistic regression, p value = 6.7×10^{-5}, 0.013, and 0.69 for days 21, 90, and 180, respectively). nAb: neutralising antibody, S IgG: anti-SARS-CoV-2 spike protein IgG.

FIG. 10 shows a model to predict vaccine-induced antibody responses. (A) Comparison of nAb level distributions on days 0, 21, 90, and 180 in the Sinopharm inactivated virus and BNT162b2 mRNA vaccine cohorts. (B) Violin plot representation of the data in (A). (C) Random Forest model predictions of Low/High antibody responses to COVID-19 vaccines. The model was trained on the training set from the mRNA vaccine cohort. Predictive

performance was assessed by predicting the antibody responses of the test and Sinopharm sets. The receiver operating characteristic (ROC) curves are presented for the predictive model with the best performance in the training, test, and Sinopharm sets (area under the curve (AUC) is 0.80, 0.81, and 0.66 for the training, test, and Sinopharm sets, respectively).

Detailed description

Systems vaccinology predicts vaccine efficacy using systematic approaches to identify predictive signatures, which may be a group of genes or genetic loci strongly associated with particular vaccination outcomes. The inventors have identified stable and predictive DNA signatures based on small insertion or deletion mutations (indels) in the genome. Without being bound by theory, indels alter the number of DNA nucleotides in the genome and are thus likely to have a larger functional impact on cellular processes and in turn a larger contribution to individual variations in vaccine response. Signatures that include indels may be broadly predictive of responsiveness to a range of vaccines, and may be used to develop subject-specific vaccination regimens.

Accordingly, this disclosure provides methods of predicting responsiveness to vaccines using indel signatures and in particular indel mutations selected from Table 1. Methods herein generally comprise detecting the presence or absence of, and the copy number of, an indel mutation at one or more genetic loci in Table 1 in a sample obtained from a subject, and calculating, based on the presence or absence of, and the copy number of, an indel mutation at the one or more genetic loci, a value predictive of the responsiveness of the subject to the vaccine. Also provided are systems and machine learning models for implementing the method, and methods of treatment and prophylaxis based on the predicted response.

General definitions

An "indel" as used herein refers to a mutation class that includes nucleotide insertions, deletions, and a combination thereof. An indel at a genetic locus results in a net gain or loss of nucleotides. As indels can be a combination of an insertion and a deletion, the altered nucleic acid sequence may be larger than the length difference (e.g., a deletion of 5 nucleotides combined with an insertion of 3 nucleotides leads to an altered length of 2, but

the sequence at the locus may have changed as well). In coding regions of the genome, unless the length of an indel is a multiple of 3, the indel will produce a frameshift mutation. Indels contemplated herein may be of any length, but are generally 50 nucleotides or shorter, and preferably 10 nucleotides or shorter. For instance, the indels may have a length of between 1 and 10 nucleotides (i.e., the nucleic acid sequence is 1 to 10 nucleotides longer or shorter than the normal known length of the sequence at a locus). An indel can be located in a gene or coding region of the genome (such as in an exon, intron or untranslated region (UTR)), and may also be found in non-coding regions such as in gene regulatory regions, microsatellites, sequences encoding non-coding RNA, etc.

The term "allele" refers to sequence variants of the DNA sequence at a genetic locus. A diploid subject such as a human subject has two alleles for most genetic loci (one on each chromosome) and may be homozygous or heterozygous for allelic forms at a genetic locus.

A "reference allele" or "REF" herein refers to an allelic sequence derived from a consensus or widely accepted genomic sequence representing the most common or standard form of the DNA sequence at a particular genetic locus. For humans, a reference allelic sequence may be derived from a consensus human genome such as the GRCh38/hg38 reference genome. The reference allele serves as a baseline or standard for identifying alternative alleles, such as alleles of indel mutations, present in different individuals or populations.

An "alternative allele" or "ALT" herein refers to an allelic sequence that differs from the reference allele sequence at a genetic locus, and in particular to an allelic sequence containing an indel mutation.

As used herein, the term "genotype", in the context of a given genetic locus, refers to the alleles and the copy number of each allele present at that locus. A human subject can have three genotypes at any given genetic locus herein: reference allele/reference allele (REF/REF), reference allele/alternative allele (REF/ALT), or alternative/alternative allele (ALT/ALT).

As used herein a "genetic signature" or "indel signature" denotes a collection of one or more indels, the combination of which is associated with a phenotype (such as neutralising

antibody level after one or more doses of a vaccine). A phenotype may have more than one genetic signature.

The term "non-genetic factor" in this disclosure refers to a non-heritable variable that may be associated with a phenotype. Non-genetic factors may include a range of demographic and clinical variables specific to a subject, non-limiting examples of which include subject gender, age, ethnicity, clinical history, history of infection with related pathogens, levels of specific biomarkers, etc. Non-genetic factors may be detected or quantitated, for example, through consultation with a subject, through a review of a subject's clinical or medication history, or through clinical testing of a sample from the subject.

The term "respiratory pathogen" herein refers to a pathogen capable of infecting the respiratory tract of a subject, in particular a mammalian or human subject. The pathogen may be, for example, a virus, bacterium or fungus.

As used herein, an "individual", "subject" or "patient" is a vertebrate. In certain embodiments, the vertebrate is a mammal. Mammals include, but are not limited to, primates (including human and non-human primates) and rodents (e.g., mice and rats). In preferred embodiments, a subject is a human.

A "sample" herein refers to a biological sample, typically derived from a biological fluid, cell, tissue, organ, or organism, containing DNA.

The "responsiveness" or "response" of a subject to a vaccine refers to the extent of the subject's immunological reaction following the administration of the vaccine. This can include both the magnitude of the immunological reaction (e.g., the level of neutralising antibodies produced) and the duration of the immunological reaction (e.g., the time taken for neutralising antibody levels to drop below a threshold). Thus, predicting vaccine responsiveness in a subject may include predicting whether a subject is likely or unlikely to respond to vaccination; the magnitude of a subject's immunological response to one or more doses of a vaccine; and/or the duration of a subject's immunological response to one or more doses of a vaccine.

As used herein, "antibody waning rate" refers to the rate at which neutralising antibody levels (nAbs) decline following one or more doses of a vaccine. Typically, nAb levels begin to rise a few days post-vaccination, eventually reaching a peak after weeks or months before starting to decline. The decline may not be linear, e.g., antibody levels may decline faster initially (e.g., over a few weeks), before slowing down or stabilising at a lower level over the following months. Antibody waning rate can be calculated based on nAb titres at two or more time points after one or more vaccine doses (e.g., a few days apart, a weeks apart, or a few months apart), with the first time point typically around the time of the nAb peak or after the peak. Computational methods may be used to determine the antibody waning rate.

The terms "treat", "treating", and "treatment" and "prevent", "preventing", and "prevention" as used herein, refer to eliciting a desired biological response, such as a therapeutic and prophylactic effect, respectively. A therapeutic effect may refer to (1) preventing or delaying the appearance of one or more symptoms of the disease; (2) inhibiting the development of the disease or one or more symptoms of the disease; (3) relieving the disease, i.e., causing regression of the disease or at least one or more symptoms of the disease; and/or (4) causing a decrease in the severity of one or more symptoms of the disease. A prophylactic effect may refer to (1) a complete or partial inhibition of a disease; or (2) a delay of disease development or progression.

The methods as disclosed herein may comprise the administration of an "effective amount" of an agent (e.g., a vaccine) to a subject. As used herein, an "effective amount", when referring to preventing a disease or infection, is any non-toxic amount of an agent (e.g., a vaccine) that, when administered to a subject prone to a disease or infection or prone to affliction with an infection-associated disorder, induces in the subject an immune response that protects the subject from becoming infected by a pathogen or afflicted with the disease or disorder. "Protecting" the subject means either reducing the likelihood of the subject becoming infected with a pathogen, or lessening the likelihood of the disease or disorder's onset in the subject, by at least two-fold. Most preferably, an effective dose induces in the subject an immune response that completely prevents the subject from becoming infected by a pathogen or prevents the onset of the disease or disorder in the subject entirely.

As used herein, "machine learning" refers to algorithms that give a computer the ability to learn and perform tasks (such as making predictions) without relying on explicit or rules-

based programming. Machine learning systems or models improve their performance on a specific task over time by identifying and learning from data inputs and adjusting their algorithms accordingly. Key components of machine learning include data pre-processing, feature selection, model training, model validation, and model deployment. Machine learning algorithms include, but are not limited to, classifier algorithms (such as support vector machines, k-nearest neighbour, logistic regression, decision trees); regression algorithms (e.g., linear regression, LASSO regression); Naïve Bayes classifiers; random forests; clustering algorithms (e.g., k-means clustering, hierarchical clustering); and artificial neural networks (including multilayer perceptrons and deep learning networks).

Indel signatures for predicting vaccine response

Methods herein comprise detecting the presence or absence of, and the copy number of, an indel mutation at one or more genetic loci in a sample from a subject and calculating a value indicative of vaccine responsiveness for the subject. The method may further comprise comparing this value to a reference value (e.g., a reference value established by a vaccine manufacturer or a regulatory body) to generate a prediction of whether the subject is likely to be responsive or non-responsive to the vaccine. The method may rely on a computer algorithm (e.g., a machine learning algorithm) to calculate the value and/or generate a prediction of whether the subject is likely to be responsive or non-responsive to the vaccine based on the similarity or difference of the value to a reference value.

Table 1. Exemplary list of genetic loci and indel mutations predictive of responsiveness to vaccine (inserted/deleted nucleotides underlined).

Genetic Locus	**Chromo-some**	**Start position in hg38**	**Reference allele (REF)**	**Alternative allele (ALT)**	**Mutation**
1	Chr1	31433042	A	A<u>CAG</u>	Insertion
2	Chr2	96115238	CTCCTCCTCT (SEQ ID NO: 1)	C	Deletion
3	Chr12	8946399	G	G<u>AAC</u>	Insertion
4	Chr20	257794	CTGGTCT	C	Deletion
5	Chr10	13142923	T	T<u>C</u>	Insertion

6	Chr16	89860945	T	TC	Insertion
7	Chr7	131506795	TC	T	Deletion
8	Chr12	107320533	T	TC	Insertion
9	Chr6	26501243	AG	A	Deletion
10	Chr11	6786184	CT	C	Deletion
11	Chr10	63208190	ACTAAAC	A	Deletion
12	Chr20	49036827	C	CTTTG	Insertion
13	Chr17	7850977	TCCTCTG	T	Deletion
14	Chr14	55026541	AC	A	Deletion
15	Chr7	12629398	C	CCA	Insertion
16	Chr15	59123735	CAT	C	Deletion
17	Chr3	172444590	GGTTTCTGTAA (SEQ ID NO: 2)	G	Deletion
18	Chr15	33670239	AT	A	Deletion
19	Chr8	27010286	A	AC	Insertion
20	Chr17	78494380	GC	G	Deletion
21	Chr15	76311269	CA	C	Deletion

An indel mutation may be directly detected and/or quantitated or the DNA sequence at a genetic locus may be copied and/or amplified to allow detection of the indel. Detection can be performed by any means known to one skilled in the art, including, for example, restriction fragment length polymorphism (RFLP) analysis using restriction enzymes, single-strand conformational polymorphism (SSCP) analysis, heteroduplex analysis, chemical cleavage analysis, microarray analysis using hybridisation probes, dual ligation and hybridisation assays, allele-specific amplification, mass spectrometry methods, isothermal amplification methods, or sequencing methods including Sanger sequencing, high-throughput sequencing such as nanopore sequencing and single-molecule real-time (SMRT) sequencing, and next-generation sequencing technologies applied to a whole genome or exome. Combinations of these methods may also be used.

In some embodiments, indel detection or genotyping involves the use of a physical array of oligonucleotides (e.g., primers or probes), subject samples, or both. Common array formats include both liquid and solid phase arrays. For example, assays employing liquid phase

arrays, e.g., for hybridisation of nucleic acids, can be performed in multi-well or microtitre plates. Microtitre plates with 96, 384 or 1536 wells are widely available, and even higher numbers of wells, e.g., 3456 and 9600 can be used. In general, the choice of microtitre plates is determined by the methods and equipment, e.g., robotic handling and loading systems, used for sample preparation and analysis. Exemplary systems include, e.g., xMAP® technology from Luminex (Austin, TX), the SECTOR® Imager with MULTI-ARRAY® and MULTI-SPOT® technologies from Meso Scale Discovery (Gaithersburg, MD), the ORCA™ system from Beckman-Coulter, Inc. (Fullerton, CA) and the ZYMATE™ systems from Zymark Corporation (Hopkinton, MA).

A variety of solid phase arrays can favourably be employed to genotype indels in the context of the disclosed methods. Exemplary formats include membrane or filter arrays (e.g., nitrocellulose, nylon), pin arrays, and bead arrays (e.g., in a liquid "slurry"). Typically, oligonucleotide probes complementary and hybridisable to a target allele sequence are immobilised, for example by direct or indirect cross-linking, to the solid support. Essentially any solid support capable of withstanding the reagents and conditions necessary for performing the particular assay can be utilised. For example, functionalised glass, silicon, silicon dioxide, modified silicon, any of a variety of polymers, such as polytetrafluoroethylene, polyvinylidenedifluoride, polystyrene, polycarbonate, or combinations thereof can all serve as the substrate for a solid phase array.

Table 2 provides possible combinations of genetic loci for use in the methods herein.

Table 2. Possible combinations of loci

1	A method of predicting the responsiveness of a subject to a vaccine, the method comprising: (a) detecting the presence or absence of, and the copy number of, an insertion/deletion (indel) mutation at one or more genetic loci in Table 1 using a sample obtained from the subject; and (b) calculating, based on the presence or absence of, and the copy number of, an indel mutation at the one or more genetic loci, a value predictive of the responsiveness of the subject to the vaccine.
2	The method of 1, wherein the panel of loci comprises 1 locus.
3	The method of 1, wherein the panel of loci comprises 2 loci.
4	The method of 1, wherein the panel of loci comprises 3 loci.

5	The method of 1, wherein the panel of loci comprises 4 loci.
6	The method of 1, wherein the panel of loci comprises 5 loci.
7	The method of 1, wherein the panel of loci comprises 6 loci.
8	The method of 1, wherein the panel of loci comprises 7 loci.
9	The method of 1, wherein the panel of loci comprises 8 loci.
10	The method of 1, wherein the panel of loci comprises 9 loci.
11	The method of 1, wherein the panel of loci comprises 10 loci.
12	The method of 1, wherein the panel of loci comprises 11 loci.
13	The method of 1, wherein the panel of loci comprises 12 loci.
14	The method of 1, wherein the panel of loci comprises 13 loci.
15	The method of 1, wherein the panel of loci comprises 14 loci.
16	The method of 1, wherein the panel of loci comprises 15 loci.
17	The method of 1, wherein the panel of loci comprises 16 loci.
18	The method of 1, wherein the panel of loci comprises 17 loci.
19	The method of 1, wherein the panel of loci comprises 18 loci.
20	The method of 1, wherein the panel of loci comprises 19 loci.
21	The method of 1, wherein the panel of loci comprises 20 loci.
22	The method of 1, wherein the panel of loci comprises 21 loci.
23	The method of any one of 1 to 22, wherein the panel of genes comprises locus 1 in Table 1.
24	The method of any one of 1 to 23, wherein the panel of genes comprises locus 2 in Table 1.
25	The method of any one of 1 to 24, wherein the panel of genes comprises locus 3 in Table 1.
26	The method of any one of 1 to 25, wherein the panel of genes comprises locus 4 in Table 1.
27	The method of any one of 1 to 26, wherein the panel of genes comprises locus 5 in Table 1.
28	The method of any one of 1 to 27, wherein the panel of genes comprises locus 6 in Table 1.
29	The method of any one of 1 to 28, wherein the panel of genes comprises locus 7 in Table 1.

30	The method of any one of 1 to 29, wherein the panel of genes comprises locus 8 in Table 1.
31	The method of any one of 1 to 30, wherein the panel of genes comprises locus 9 in Table 1.
32	The method of any one of 1 to 31, wherein the panel of genes comprises locus 10 in Table 1.
33	The method of any one of 1 to 32, wherein the panel of genes comprises locus 11 in Table 1.
34	The method of any one of 1 to 33, wherein the panel of genes comprises locus 12 in Table 1.
35	The method of any one of 1 to 34, wherein the panel of genes comprises locus 13 in Table 1.
36	The method of any one of 1 to 35, wherein the panel of genes comprises locus 14 in Table 1.
37	The method of any one of 1 to 36, wherein the panel of genes comprises locus 15 in Table 1.
38	The method of any one of 1 to 37, wherein the panel of genes comprises locus 16 in Table 1.
39	The method of any one of 1 to 38, wherein the panel of genes comprises locus 17 in Table 1.
40	The method of any one of 1 to 39, wherein the panel of genes comprises locus 18 in Table 1.
41	The method of any one of 1 to 40, wherein the panel of genes comprises locus 19 in Table 1.
42	The method of any one of 1 to 41, wherein the panel of genes comprises locus 20 in Table 1.
43	The method of any one of 1 to 42, wherein the panel of genes comprises locus 21 in Table 1.

In some embodiments, the method comprises detecting the genotype at or proximate to a locus selected from the group of Chr1:31433042, Chr2:96115238, Chr12:8946399, Chr20:257794, Chr10:13142923, Chr16:89860945, Chr7:131506795, Chr12:107320533, Chr6:26501243, Chr11:6786184, Chr10:63208190, Chr20:49036827, Chr17:7850977, Chr14:55026541, Chr7:12629398, Chr15:59123735, Chr3:172444590, Chr15:33670239,

Chr8:27010286, Chr17:78494380 and Chr15:76311269. In one embodiment, step (a) of the method comprises detecting the genotype at each of locus 1 to 12 in Table 1. In one embodiment, step (a) of the method comprises detecting the genotype at each of locus 1 to 21 in Table 1.

In some embodiments, the method comprises detecting the genotype at a position located from 1, from 2, from 3, from 4, from 5, from 6, from 7, from 8, from 9, from 10, from 11, from 12, from 13, from 14, from 15, from 16, from 17, from 18, from 19, or from 20 nucleotides of a locus selected from the group of Chr1:31433042, Chr2:96115238, Chr12:8946399, Chr20:257794, Chr10:13142923, Chr16:89860945, Chr7:131506795, Chr12:107320533, Chr6:26501243, Chr11:6786184, Chr10:63208190, Chr20:49036827, Chr17:7850977, Chr14:55026541, Chr7:12629398, Chr15:59123735, Chr3:172444590, Chr15:33670239, Chr8:27010286, Chr17:78494380, and Chr15:76311269.

In some embodiments, step (a) of the method comprises detecting the presence or absence of, and the copy number of, an indel mutation at locus 1, 5 and/or 21 in Table 1. In one embodiment, the method comprises detecting the presence or absence of, and the copy number of, an indel mutation at locus 1 in Table 1. The presence of an ALT/ALT genotype (i.e., the presence of two alleles with an indel mutation) at locus 1 may indicate that a subject is likely to be responsive to a vaccine. In one embodiment, the method comprises detecting the presence or absence of, and the copy number of, an indel mutation at locus 5 in Table 1. The presence of a REF/ALT or ALT/ALT genotype (i.e., the presence of one or more alleles with an indel mutation) at locus 5 may indicate that a subject is likely to be non-responsive to a vaccine. In one embodiment, the method comprises detecting the presence or absence of, and the copy number of, an indel mutation at locus 21 in Table 1. The presence of a REF/ALT or ALT/ALT genotype (i.e., the presence of one or more alleles with an indel mutation) at locus 21 may indicate that a subject is likely to be non-responsive to a vaccine.

In some embodiments, step (a) of the method comprises detecting the presence or absence of, and the copy number of, an indel mutation at each of locus 1 to 12 in Table 1. In some embodiments, step (a) of the method comprises detecting the presence or absence of, and the copy number of, an indel mutation at each of locus 1 to 21 in Table 1.

In some embodiments, step (a) comprises detecting the presence or absence of, and the copy number of, one or more alleles in Table 1. The method may comprise detecting one or more reference alleles (REF) in Table 1, one or more alternative alleles (ALT) containing an indel mutation in Table 1, or a combination thereof. In one embodiment, the method comprises detecting all of the alternative alleles comprising an indel mutation in Table 1. The method may also comprise detecting one or more REF alleles in Table 1 to determine the copy number of the ALT allele and/or the genotype at each genetic locus.

Table 3 provides possible combinations of alleles for use in the methods herein.

Table 3. Possible combinations of alleles

1	A method of predicting the responsiveness of a subject to a vaccine, the method comprising: (a) detecting the presence or absence of, and the copy number of, one or more reference and/or alternative alleles in Table 1 using a sample obtained from the subject; and (b) calculating, based on the presence or absence of, and the copy number of, an allele, a value predictive of the responsiveness of the subject to the vaccine.
2	The method of 1, wherein the panel of alleles comprises 1 allele.
3	The method of 1, wherein the panel of alleles comprises 2 alleles.
4	The method of 1, wherein the panel of alleles comprises 3 alleles.
5	The method of 1, wherein the panel of alleles comprises 4 alleles.
6	The method of 1, wherein the panel of alleles comprises 5 alleles.
7	The method of 1, wherein the panel of alleles comprises 6 alleles.
8	The method of 1, wherein the panel of alleles comprises 7 alleles.
9	The method of 1, wherein the panel of alleles comprises 8 alleles.
10	The method of 1, wherein the panel of alleles comprises 9 alleles.
11	The method of 1, wherein the panel of alleles comprises 10 alleles.
12	The method of 1, wherein the panel of alleles comprises 11 alleles.
13	The method of 1, wherein the panel of alleles comprises 12 alleles.
14	The method of 1, wherein the panel of alleles comprises 13 alleles.
15	The method of 1, wherein the panel of alleles comprises 14 alleles.
16	The method of 1, wherein the panel of alleles comprises 15 alleles.
17	The method of 1, wherein the panel of alleles comprises 16 alleles.

18	The method of 1, wherein the panel of alleles comprises 17 alleles.
19	The method of 1, wherein the panel of alleles comprises 18 alleles.
20	The method of 1, wherein the panel of alleles comprises 19 alleles.
21	The method of 1, wherein the panel of alleles comprises 20 alleles.
22	The method of 1, wherein the panel of alleles comprises 21 alleles.
23	The method of 1, wherein the panel of alleles comprises 22 alleles.
24	The method of 1, wherein the panel of alleles comprises 23 alleles.
25	The method of 1, wherein the panel of alleles comprises 24 alleles.
26	The method of 1, wherein the panel of alleles comprises 25 alleles.
27	The method of 1, wherein the panel of alleles comprises 26 alleles.
28	The method of 1, wherein the panel of alleles comprises 27 alleles.
29	The method of 1, wherein the panel of alleles comprises 28 alleles.
30	The method of 1, wherein the panel of alleles comprises 29 alleles.
31	The method of 1, wherein the panel of alleles comprises 30 alleles.
32	The method of 1, wherein the panel of alleles comprises 31 alleles.
33	The method of 1, wherein the panel of alleles comprises 32 alleles.
34	The method of 1, wherein the panel of alleles comprises 33 alleles.
35	The method of 1, wherein the panel of alleles comprises 34 alleles.
36	The method of 1, wherein the panel of alleles comprises 35 alleles.
37	The method of 1, wherein the panel of alleles comprises 36 alleles.
38	The method of 1, wherein the panel of alleles comprises 37 alleles.
39	The method of 1, wherein the panel of alleles comprises 38 alleles.
40	The method of 1, wherein the panel of alleles comprises 39 alleles.
41	The method of 1, wherein the panel of alleles comprises 40 alleles.
42.	The method of 1, wherein the panel of alleles comprises 41 alleles.
43	The method of 1, wherein the panel of alleles comprises 42 alleles.
44	The method of any one of 1 to 43, wherein the panel of alleles comprises reference and/or alternative allele of locus 1 in Table 1.
45	The method of any one of 1 to 44, wherein the panel of alleles comprises reference and/or alternative allele of locus 2 in Table 1.
46	The method of any one of 1 to 45, wherein the panel of alleles comprises reference and/or alternative allele of locus 3 in Table 1.

47	The method of any one of 1 to 46, wherein the panel of alleles comprises reference and/or alternative allele of locus 4 in Table 1.
48	The method of any one of 1 to 47, wherein the panel of alleles comprises reference and/or alternative allele of locus 5 in Table 1.
49	The method of any one of 1 to 48, wherein the panel of alleles comprises reference and/or alternative allele of locus 6 in Table 1.
50	The method of any one of 1 to 49, wherein the panel of alleles comprises reference and/or alternative allele of locus 7 in Table 1.
51	The method of any one of 1 to 50, wherein the panel of alleles comprises reference and/or alternative allele of locus 8 in Table 1.
52	The method of any one of 1 to 51, wherein the panel of alleles comprises reference and/or alternative allele of locus 9 in Table 1.
53	The method of any one of 1 to 52, wherein the panel of alleles comprises reference and/or alternative allele of locus 10 in Table 1.
54	The method of any one of 1 to 53, wherein the panel of alleles comprises reference and/or alternative allele of locus 11 in Table 1.
55	The method of any one of 1 to 54, wherein the panel of alleles comprises reference and/or alternative allele of locus 12 in Table 1.
56	The method of any one of 1 to 55, wherein the panel of alleles comprises reference and/or alternative allele of locus 13 in Table 1.
57	The method of any one of 1 to 56, wherein the panel of alleles comprises reference and/or alternative allele of locus 14 in Table 1.
58	The method of any one of 1 to 57, wherein the panel of alleles comprises reference and/or alternative allele of locus 15 in Table 1.
59	The method of any one of 1 to 58, wherein the panel of alleles comprises reference and/or alternative allele of locus 16 in Table 1.
60	The method of any one of 1 to 59, wherein the panel of alleles comprises reference and/or alternative allele of locus 17 in Table 1.
61	The method of any one of 1 to 60, wherein the panel of alleles comprises reference and/or alternative allele of locus 18 in Table 1.
62	The method of any one of 1 to 61, wherein the panel of alleles comprises reference and/or alternative allele of locus 19 in Table 1.

63	The method of any one of 1 to 62, wherein the panel of alleles comprises reference and/or alternative allele of locus 20 in Table 1.
64	The method of any one of 1 to 63, wherein the panel of alleles comprises reference and/or alternative allele of locus 21 in Table 1.

In one embodiment, the method comprises detecting the presence or absence of, and the copy number of, reference or alternative allele of locus 1 and/or 21 in Table 1. In one embodiment, the method comprises detecting the presence or absence of, and the copy number of, all of reference and/or alternative alleles 1 to 12 in Table 1. In one embodiment, the method comprises detecting the presence or absence of, and the copy number of, all of reference and/or alternative alleles 1 to 21 in Table 1.

In some embodiments the methods herein further comprise the use of at least one non-genetic factor to calculate the value. In certain embodiments, the non-genetic factor is selected from the age of the subject, the gender of the subject and the ethnicity of the subject. In one embodiment, the methods herein comprise the use of the age and gender of the subject to calculate the value predictive of vaccine responsiveness. In one embodiment, the methods herein comprise the use of the age, gender and ethnicity of the subject to calculate the value predictive of vaccine responsiveness.

Machine learning models and systems

Methods herein may rely on the use of machine learning models to generate a prediction of vaccine responsiveness in a subject based on a gene signature. Any machine learning algorithm can be used to generate the machine learning model, including but not limited to classification algorithms, regression algorithms, clustering algorithms, association algorithms, ensemble learning algorithms, artificial neural networks, deep learning algorithms, and combinations thereof.

In one embodiment, the machine learning model relies on a classifier algorithm. For example, the machine learning model may rely on a decision tree classifier, logistic regression classifier, nearest neighbour classifier, neural network classifier, Gaussian mixture model (GMM), Support Vector Machine (SVM) classifier, nearest centroid classifier, linear regression classifier, linear discriminant analysis (LDA) classifier,

quadratic discriminant analysis (QDA) classifier, LogitBoost classifier, rotation forest classifier, random forest classifier, extreme gradient boosting (XG Boost) classifier, linear mixed effects model classifier and variations and combinations thereof. In one embodiment, the machine learning model is a random forest classifier model that uses a random forest classifier.

In one embodiment, the machine learning model is trained to generate a value predictive of responsiveness to a vaccine based on the presence or absence of, and the copy number of, an indel mutation at one or more genetic loci in Table 1.

In one embodiment, the machine learning model is trained using a dataset comprising genotype data and vaccine response data from a cohort of subjects who have received one or more doses of the vaccine. The genotype data comprises the presence or absence of, and the copy number of, an indel mutation at one or more genetic loci in Table 1. The vaccine response data may comprise the level of neutralising antibodies and/or the antibody waning rate in the cohort of subjects after the subjects have received one or more doses of the vaccine.

Disclosed herein is a computer-implemented method for predicting the responsiveness of a subject to a vaccine, the method comprising: (a) receiving genotype data comprising the presence or absence of, and the copy number of, an insertion/deletion (indel) mutation at one or more genetic loci in Table 1 from a sample from the subject; and (b) calculating a value predictive of the responsiveness of the subject to the vaccine using a machine learning model trained to generate the predictive value based on the presence or absence of, and the copy number of, an indel mutation at one or more genetic loci in Table 1. The machine learning model may be as described herein.

Disclosed here is a computer-implemented method for determining a genetic signature for predicting the responsiveness of a subject to a vaccine, the method comprising:

a) receiving whole exome sequencing records of a plurality of subjects who are responsive to a vaccine at different responsiveness levels;
b) aligning whole exome sequencing records to a human reference genome;

c) identifying indels that are significantly associated with the responsiveness level, wherein each indel with significant association is treated as a candidate variable for a predictive model;

d) dividing the whole exome sequencing records into a training and a validation dataset;

e) performing model training using the identified indels and one or more non-genetic factors from the training and validation dataset as variables, to identify a model that is predictive of the responsiveness of a subject to a vaccine.

Disclosed herein is a system for predicting the responsiveness of a subject to a vaccine, the system comprising: a processor coupled to computer readable memory, the memory comprising instructions that when executed by the processor causes the processor to: (i) receive genotype data comprising the presence or absence of, and the copy number of, an insertion/deletion (indel) mutation at one or more genetic loci in Table 1 from a sample from the subject; and (ii) calculate, based on the presence or absence of, and the copy number of, an indel mutation at the one or more genetic loci, a value predictive of the responsiveness of the subject to the vaccine using a machine learning model.

Disclosed herein is a system for predicting the responsiveness of a subject to a vaccine, the system comprising: (a) a processor coupled to computer readable memory; (b) an array for high-throughput detection of the presence or absence of, and the copy number of, an indel mutation at one or more genetic loci in Table 1 using a sample obtained from the subject, wherein the array is configured to generate a plurality of detectable signals indicative of the presence or absence of, and the copy number of, an indel mutation at the one or more genetic loci; and (c) a computer algorithm stored in the computer readable memory and executable by the processor to take as input the plurality of detectable signals generated by the array, and calculate a value predictive of the responsiveness of the subject to the vaccine.

Vaccines and pathogens

In some embodiments, vaccines herein are directed against a viral, bacterial and/or fungal pathogen. Examples of bacterial pathogens include, but are not limited to, *Clostridium tetani*, *Salmonella typhi*, *Vibrio cholera*, *Neisseria meningitides*, *Streptococcus pneumonia*, *Bacillus anthracis*, and *Haemophilus influenza*. Examples of fungal pathogens include, but are not limited to, *Candida sp.*, *Cryptococcus sp.* and *Aspergillus sp.* Examples of viral

pathogens include, but are not limited to, *Retroviridae* (e.g., human immunodeficiency viruses); *Picornaviridae* (e.g., polio viruses, hepatitis A virus; enteroviruses, human coxsackie viruses, rhinoviruses, echoviruses); *Calciviridae* (e.g., Norwalk virus and noroviruses that cause gastroenteritis); *Togaviridae* (e.g., rubella viruses); *Flaviridae* (e.g., dengue viruses, encephalitis viruses, yellow fever viruses); *Coronaviridae* (e.g., coronaviruses, including SARS-CoV-2 virus); *Rhabdoviridae* (e.g., vesicular stomatitis viruses, rabies viruses); *Filoviridae* (e.g., ebola viruses); *Paramyxoviridae* (e.g., parainfluenza viruses, mumps virus, measles virus, respiratory syncytial virus); *Orthomyxoviridae* (e.g., influenza viruses); *Bungaviridae* (e.g., Hantaan viruses, bunga viruses, phleboviruses and Nairo viruses); *Arena viridae* (hemorrhagic fever viruses); *Reoviridae* (e.g., reoviruses, orbiviurses and rotaviruses); *Bimaviridae*; *Hepadnaviridae* (Hepatitis B virus); *Parvoviridae* (parvoviruses); *Papovaviridae* (papilloma viruses, polyoma viruses); *Adenoviridae* (most adenoviruses); *Herpesviridae* (herpes simplex viruses, varicella zoster virus, cytomegalovirus (CMV), herpes viruses); *Poxyiridae* (variola viruses, vaccinia viruses, pox viruses); *Iridoviridae* (e.g., African swine fever virus); and hepatitis C virus. In some embodiments, the vaccine is directed against one or more viral pathogens.

In some embodiments, the vaccine is directed against one or more respiratory pathogens, including but not limited to influenza virus A (including the H1, H3, H5 and H7 subtypes), influenza virus B, parainfluenza virus types I–IV, respiratory syncytial virus A and B, adenovirus group B, C, D and E, rhinoviruses, enteroviruses, bocavirus, measles virus picornavirus, human metapneumovirus, coronaviruses (including betacoronaviruses such as 229E, OC43, NL63, HKU1, SARS-CoV-1, SARS-CoV-2 and MERS-CoV), *Acinetobacter baumannii*, *Aspergillus fumigatus*, *Bordetella pertussis*, *Chlamydia pneumoniae*, *Cryptococcus spp.*, *Haemophilus influenzae*, *Klebsiella pneumoniae*, *Legionella pneumophila*, *Moraxella catarrhalis*, *Mycoplasma pneumoniae*, *Pneumocystis jirovecii*, *Pseudomonas aeruginosa*, *Rickettsia spp.*, *Staphylococcus spp.*, and *Streptococcus spp.*

In some embodiments, the respiratory pathogen is a viral pathogen, such as a coronavirus, an influenza virus, a parainfluenza virus, or respiratory syncytial virus. In one embodiment, the vaccine is directed against a betacoronavirus. In one embodiment, the vaccine is directed against an influenza virus. In one embodiment, the vaccine is directed against an

parainfluenza virus. In one embodiment, the vaccine is directed against respiratory syncytial virus. In one embodiment, the vaccine is directed against the SARS-CoV-2 virus.

There is no particular limitation on the types of vaccines the response to which can be predicted using the methods of this disclosure. Thus, methods herein may be used to predict response to live attenuated vaccines, inactivated vaccines, subunit vaccines, recombinant vaccines, conjugate vaccines, toxoid vaccines, polypeptide vaccines, virus-like particle (VLP)-based vaccines, mRNA vaccines, DNA vaccines, viral vector-based vaccines, and other relevant vaccine modalities.

Vaccine response

Vaccine responsiveness may be determined with respect to a humoral reaction (e.g., the neutralising antibody titre) or a cellular reaction (e.g., the number and/or proportion of particular immune cells). Responsiveness may be determined with reference to clinical data or guidelines (e.g., guidelines established by a vaccine manufacturer or a regulatory body) for particular vaccines. For instance, the level of neutralising antibodies in a subject may be compared to reference antibody levels (e.g., levels established by a vaccine manufacturer or a regulatory body) to determine whether the subject is protected from a pathogen. Vaccine responsiveness may be determined at specific time points following administration of one or more doses of a vaccine or following completion of a vaccination schedule (e.g., 21 days or 6 months after a vaccine dose). Vaccine responsiveness can also be determined for a duration of time following vaccination to evaluate the durability of the immune response. Vaccine responsiveness can vary depending on the subject (e.g., depending on the age, gender, medical condition, or genetic background of the subject), the type of vaccine, and the specific immune parameters being evaluated. In general, a subject's responsiveness to a vaccine declines over time.

Vaccine responsiveness may be measured using methods commonly known in the art, such as assays quantifying levels of neutralising antibodies (e.g., ELISA, pathogen neutralisation assays, flow cytometry), assays measuring cell-mediated immunity (e.g, cytokine, secretome or surface marker profiling of T or B cells using ELIspot assays or flow cytometry, T cell differentiation assays), immune cell profiling (e.g., using flow cytometry) or challenge studies (i.e., deliberate exposure to the pathogen, e.g., in a controlled setting). The skilled

person can identify appropriate response parameters and values indicative of responsiveness or non-responsiveness to a vaccine based on clinical data and established guidelines for the vaccine and general knowledge of immunology or vaccinology.

In some embodiments, the value predictive of responsiveness is a predicted level of neutralising antibodies in the subject after one or more doses of the vaccine. The predicted neutralising antibody level may be used to determine whether the vaccine is to be administered to the subject, or to determine an effective vaccine dose or immunisation schedule for the subject. The neutralising antibodies may be of IgG, IgA and/or IgM classes. The method may predict the level of neutralising antibodies about 14 days after, about 21 days after, about 1 month after, about 2 months after, about 3 months after, about 4 months after, about 5 months after, about 6 months after, about 7 months after, about 8 months after, about 9 months after, about 10 months after, about 11 months after, about 1 year after, about 2 years after, about 3 years after, about 4 years after, or about five years after a vaccine dose. In one embodiment, the method predicts the level of neutralising antibodies in the subject 21 days after one or more doses of the vaccine.

In some embodiments, the value predictive of responsiveness is a predicted rate of change of neutralising antibody levels in the subject after one or more doses of the vaccine. The rate of change may be determined from neutralising antibody levels at two or more time points after a vaccine dose. The time points may be days, weeks, months or years apart. The rate of change may be a rate of increase in neutralising antibody levels, or a rate of decrease in neutralising antibody levels.

In some embodiments, the value predictive of responsiveness is a predicted antibody waning rate in the subject after one or more doses of the vaccine. The predicted antibody waning rate may be used to determine whether the vaccine is to be administered to the subject, or to determine an effective vaccine dose or immunisation schedule (e.g., when a booster dose should be administered) for the subject. The neutralising antibodies may be of IgG, IgA and/or IgM classes. In general, the method predicts the antibody waning rate after neutralising antibody levels have peaked. The method may predict the antibody waning rate over a period of about 14 days after, about 21 days after, about 1 month after, about 2 months after, about 3 months after, about 4 months after, about 5 months after, about 6 months after, about 7 months after, about 8 months after, about 9 months after, about 10 months after,

about 11 months after, about 1 year after, about 2 years after, about 3 years after, about 4 years after, or about five years after one or more vaccine doses. The method may predict the antibody waning rate over a period of about 14 days after, about 21 days after, about 1 month after, about 2 months after, about 3 months after, about 4 months after, about 5 months after, about 6 months after, about 7 months after, about 8 months after, about 9 months after, about 10 months after, about 11 months after, about 1 year after, about 2 years after, about 3 years after, about 4 years after, or about five years after neutralising antibody levels have peaked. In one embodiment, the method predicts the antibody waning rate in the subject over a period of about 90 days after neutralising antibody levels have peaked.

Samples

A sample herein may be blood, blood fractions, sputum, tissue biopsy, tissue explant, mucosal swab, mucosal smear, buccal swab, buccal smear, connective tissue, muscle, nervous tissue, epithelia, bone, cartilage, gastrointestinal tissue, reproductive tissue, vascular tissue, or any other tissue or cell preparation, or fraction or derivative thereof or isolated therefrom. The sample may be used directly as obtained from the subject or following a pretreatment to modify the character of the sample. For example, such pretreatment may include preparing plasma or cells from blood, diluting viscous fluids and so forth. Methods of pretreatment may also involve, but are not limited to, filtration, precipitation, dilution, distillation, mixing, centrifugation, freezing, lyophilisation, concentration, amplification, nucleic acid fragmentation, inactivation of interfering components, the addition of reagents, lysis, etc. If such methods of pretreatment are employed with respect to the sample, such pretreatment methods are typically such that the nucleic acid(s) of interest remain in the test sample, sometimes at a concentration proportional to that in an untreated test sample (e.g., namely, a sample that is not subjected to any such pretreatment). A sample expressly encompasses fractions or processed portions thereof. In some embodiments the sample is a blood sample or a swab. In one embodiment, the sample is a blood sample. In one embodiment, the sample is a swab, e.g., a buccal swab.

Nucleic acid amplification

The method herein may involve isolation of nucleic acids from a sample for indel genotyping. Nucleic acid can be obtained or isolated from a sample using any method of

nucleic acid isolation known to one skilled in the art. Examples of nucleic acid isolation methods can be found in general laboratory manuals, such as Sambrook and Russel, *Molecular Cloning: A Laboratory Manual*, 3rd Ed., Cold Spring Harbor Laboratory Press, Cold Spring Harbor, N.Y. (2001). The step of isolating nucleic acids is not restricted to any particular degree of purity of the isolated nucleic acids.

An amplification step may be performed before proceeding with the detection method. For example, exons or introns of a gene can be amplified and then directly sequenced or detected with labelled probes. Dye primer sequencing can be used to increase the accuracy of detecting heterozygous samples. Amplification may be performed with allele-specific primers, e.g., primers that recognise particular reference or alternative alleles. By using labelled primers or amplification-responsive indicators, a detectable signal may be generated during amplification that would identify the presence of and/or copy number of specific alleles.

In some embodiments, step (a) comprises amplifying the one or more genetic loci to detect an indel mutation. Amplification may be carried out non-isothermally (e.g., using polymerase chain reaction) or isothermally. Examples of isothermal amplification methods include but are not limited to loop-mediated isothermal amplification (LAMP), recombinase polymerisation amplification (RDA), helicase-dependent amplification (HDA), rolling circle amplification (RCA), and nickase-mediated strand displacement amplification (NMDA). In one embodiment, the one or more genetic loci is amplified using loop-mediated isothermal amplification (LAMP).

In some embodiments, locus amplification is performed using allele-specific primers. The primers may comprise a detectable label. In some embodiments, the primers are for loop-mediated isothermal amplification (LAMP). In one embodiment, the primers are for loop-mediated isothermal amplification (LAMP) of an allele in Table 1.

In some embodiments, locus amplification is performed in the presence of an indicator which is capable of undergoing a change when a nucleic acid is amplified. The change may be, for example, a colour change or a change in fluorescence or chemoluminescence. In some embodiments, the indicator is a metallochromic dye, a fluorescent dye, a pH-sensitive colorimetric dye or a nucleic acid intercalating dye.

In one embodiment, the indicator is a pH-sensitive dye, e.g., a pH-sensitive colorimetric or fluorescent dye. When the dye is used in an amplification reaction that alters the pH of the reaction mix, the spectral or fluorescent properties of the dye changes (e.g., the dye changes colour), which provides confirmation that amplification has occurred. Examples of pH-sensitive dyes include colorimetric dyes such as phenol red, cresol red, m-cresol purple, bromocresol purple, neutral red, phenolphthalein, naphtholphthaein, and thymol blue; and fluorescent dyes such as 2',7'-Bis-(2-carboxyethyl)-5-(and-6)-carboxyfluorescein or a carboxyl seminaphthorhodafluor (e.g. SNARF-1).

In one embodiment, the indicator is a metallochromic indicator. When a metallochromic indicator is used in an amplification reaction that alters the availability of one or more metal ions in the reaction mix, the spectral or fluorescent properties of the dye changse (e.g., the dye changes colour), which provides confirmation that amplification has occurred. Non-limiting examples of metallochromic indicators include 4-(2-pyridylazo) resorcinol (PAR) and hydroxynaphthol blue. If PAR is used, the amplification reaction mix may additionally comprise manganese ions such that the PAR in the master mix is complexed with Mn ions to form a PAR-Mn complex.

In some embodiments, the method herein is high throughput. As used herein, a "high-throughput method" or "high throughput detection" refers to a process that uses a combination of modern robotics, data processing and control software, liquid handling devices and/or sensitive detectors to efficiently process a large number of samples (e.g., dozens, hundreds, thousands, tens of thousands) in biochemical or genetic analysis, either in parallel or in sequence, within a reasonably short period of time (e.g., minutes, hours, days). The process may be amenable to automation, such as robotic simultaneous handling of 96 samples, 384 samples, 3546 samples or more. The samples are often in small volumes, such as no more than 1 ml, 500 µl, 200 µl, 100 µl, 50 µl or less. Through this process one can rapidly amplify nucleic acids for analysis, identify genetic variants (e.g., indels) at a large number of loci and/or calculate scores for a large number of samples.

Methods of treatment and prevention

Disclosed herein is a method of treating or preventing a disease in a subject, the method comprising: (a) detecting the presence or absence of, and the copy number of, an indel mutation at one or more genetic loci in Table 1 using a sample obtained from the subject; (b) calculating, based on the presence or absence of, and the copy number of, an indel mutation at the one or more genetic loci, a value predictive of the responsiveness of the subject to a vaccine against the infection; and (c) providing an effective amount of the vaccine to the subject based on the predicted responsiveness of the subject to the vaccine to treat or prevent the disease.

In one embodiment, the disease is an infection. In some embodiments, the infection is caused by a viral, bacterial or fungal pathogen, such as a viral, bacterial or fungal pathogen as disclosed herein. In one embodiment, the infection is caused by a viral pathogen. In some embodiments, the infection is caused by a respiratory pathogen. In some embodiments, the respiratory pathogen is a viral pathogen, such as a coronavirus, an influenza virus, a parainfluenza virus, or respiratory syncytial virus. In one embodiment, the viral pathogen is a betacoronavirus. In one embodiment, the viral pathogen is an influenza virus. In one embodiment, the viral pathogen is an parainfluenza virus. In one embodiment, the viral pathogen is respiratory syncytial virus. In one embodiment, the infection is caused by the SARS-CoV-2 virus.

In some embodiments, the value predictive of responsiveness is a predicted level of neutralising antibodies in the subject after one or more doses of the vaccine, and step (c) comprises providing an effective amount of the vaccine based on the predicted level of neutralising antibodies.

In some embodiments, the value predictive of responsiveness is a predicted antibody waning rate after one or more doses of the vaccine, and step (c) comprises providing an effective amount of the vaccine based on the predicted antibody waning rate.

The "effective amount" of a vaccine will vary from subject to subject depending on factors such as the genetic background, age and general condition of the subject, the severity of the condition being treated, the particular agent or vaccine being administered, the mode of administration, the predicted responsiveness of the subject to a vaccine and so forth. Thus, it is not possible to specify an exact "effective amount". However, for any given case, an

appropriate "effective amount" may be determined by one of ordinary skill in the art using only routine experimentation or established clinical guidance. For instance, where a subject is predicted to respond poorly to a vaccine (e.g., the neutralising antibody levels are not predicted to be protective), the dose of the vaccine may be increased, or the immunisation schedule may be modified to include additional (booster) doses to extend the duration of the protective response. Where the neutralising antibody levels of a subject are predicted to decrease below a threshold after a period of time after a vaccine dose, an additional (booster) dose may be scheduled after that period of time to maintain an effective level of protection.

Also disclosed herein is a method of selecting a subject for treatment with a vaccine, the method comprising: (a) detecting the presence or absence of, and the copy number of, an indel mutation at one or more genetic loci in Table 1 using a sample obtained from the subject; (b) calculating, based on the presence or absence of, and the copy number of, an indel mutation at the one or more genetic loci, a value predictive of the responsiveness of the subject to a vaccine; and (c) selecting a subject for treatment with the vaccine based on the predicted responsiveness of the subject to the vaccine.

In some embodiments, the value predictive of responsiveness is a predicted level of neutralising antibodies in the subject after one or more doses of the vaccine, and step (c) comprises selecting a subject for treatment with the vaccine based on the predicted level of neutralising antibodies. The method may further comprise providing an effective amount of the vaccine to the subject based on the predicted responsiveness of the subject to the vaccine.

In some embodiments, the value predictive of responsiveness is a predicted antibody waning rate after one or more doses of the vaccine, and step (c) comprises selecting a subject for treatment with the vaccine based on the predicted antibody waning rate. The method may further comprise providing an effective amount of the vaccine to the subject based on the predicted responsiveness of the subject to the vaccine.

Kits

Disclosed herein is a kit, the kit comprising one or more oligonucleotides for detecting the presence or absence of, and the copy number of, an indel mutation at a genetic locus in Table 1. The oligonucleotide may be a labelled nucleic acid probe that hybridises to a target

sequence at a genetic locus. Alternatively, the oligonucleotide may be a primer for amplifying a target sequence at a genetic locus for further detection.

In some embodiments, the one or more oligonucleotides are primers for amplifying a genetic locus. Amplification may be via methods known in the art, including but not limited to non-isothermal methods (e.g., PCR) and isothermal methods (e.g., loop-mediated isothermal amplification (LAMP)).

In some embodiments, the one or more oligonucleotides are primers for amplifying a genetic locus in Table 1. In some embodiments, the primers are allele-specific primers for detecting an allele in Table 1, e.g., a reference or alternative allele in Table 1. The primers may comprise a detectable label.

In some embodiments, the primers are for loop-mediated isothermal amplification (LAMP). In one embodiment, the primers are for loop-mediated isothermal amplification (LAMP) of an allele in Table 1.

Detectable labels include, for example, chromogens, fluorophores, near-infrared dyes, chemiluminescent molecules, bioluminescent molecules, colloidal metal particles (such as gold or silver particles), lanthanide ions (e.g., Eu^{3+}), semiconductor nanocrystals (e.g., quantum dots), radioisotopes, epitopes, enzymes, coloured beads (such as glass or plastic beads). Detectable labels also include barcodes such as molecular barcodes and fluorescent barcodes. Alternatively or additionally, the probe may contain a sequence for further amplification and detection, e.g., by sequencing.

Kits herein may comprise reagents for generating a detectable signal from the label. For example, where the label is an enzyme, the kit may contain reagents and substrates for the enzyme to generate a detectable signal, such as a substrate that produces a colour change or a change in fluorescence or chemiluminescence when acted on by the enzyme.

In one embodiment, the one or more oligonucleotides in the kit comprises a nucleic acid sequence selected from SEQ ID NO: 3–30.

Table 4. Primers for amplifying indels in Table 1 for Sanger sequencing.

Target indel	Primer name	Sequence
1	Indel1_forward	AACATCAAGGCTCATCACAGA (SEQ ID NO: 3)
1	Indel1_reverse	GTCCTTGATACACACCGCCA (SEQ ID NO: 4)
2	Indel2_forward	CCAATGACCACAGCCAGCA (SEQ ID NO: 5)
2	Indel2_reverse	GGTCATCCTGGCTGTGTTGA (SEQ ID NO: 6)
3	Indel3_forward	CAGTGGTTTCAGCCTCTTCA (SEQ ID NO: 7)
3	Indel3_reverse	AACCAAAGTCCAGACTTGCAG (SEQ ID NO: 8)
4	Indel4_forward	CCAGGCTCCAGTCCAGTAGA (SEQ ID NO: 9)
4	Indel4_reverse	CTTCCTGGATTGATGGCCTG (SEQ ID NO: 10)
5	Indel5_forward	TCTGGATGAAGATAGGCTAGGA (SEQ ID NO: 11)
5	Indel5_reverse	GAACTTCCAGCCATTCACCA (SEQ ID NO: 12)
6	Indel6_forward	CTGTGGCTGGTGAGTGAGG (SEQ ID NO: 13)
6	Indel6_reverse	CCTAACTGCAGCCTCACGAA (SEQ ID NO: 14)
7	Indel7_forward	CCTTGCCCTCCACTCACTGT (SEQ ID NO: 15)
7	Indel7_reverse	GGCATGGCAGGACTGGATAC (SEQ ID NO: 16)
8	Indel8_forward	TCACCTAACCTGGCCAACTC (SEQ ID NO: 17)
8	Indel8_reverse	GGAAACAGACAGGCAGCAAC (SEQ ID NO: 18)
9	Indel9_forward	GCCTCACAGTTCCCTTCTCC (SEQ ID NO: 19)
9	Indel9_reverse	CTCACCAGCCACCTTCAGAT (SEQ ID NO: 20)
10	Indel10_forward	AGTTCCTTCCACAGCACCAG (SEQ ID NO: 21)
10	Indel10_reverse	CTCTACTCAGCCTTGCCTTCC (SEQ ID NO: 22)
11	Indel11_forward	CACCAGCATTTACTGGCATC (SEQ ID NO: 23)
11	Indel11_reverse	TACAGCCCATTCCAGTCCAC (SEQ ID NO: 24)
12	Indel12_forward	CCATGCTTTGATCAGGTGGT (SEQ ID NO: 25)
12	Indel12_reverse	CACCATGCCAGACCCTAGAC (SEQ ID NO: 26)
21	Indel21_forward1	CAGAGAGTGGGCCAGAAGAA (SEQ ID NO: 27)
21	Indel21_reverse1	TAGTATGGACCAGGCATTGTG (SEQ ID NO: 28)
21	Indel21_forward2	CCAAACCTCTGACTTCTGCC (SEQ ID NO: 29)
21	Indel21_reverse2	ACAGTCTGAGCTCGGTCATC (SEQ ID NO: 30)

As used herein, "and/or" refers to and encompasses any and all possible combinations of one or more of the associated listed items, as well as the lack of combinations when interpreted in the alternative (or).

As used in this application, the singular form "a," "an," and "the" include plural references unless the context clearly dictates otherwise. For example, the term "an agent" includes a plurality of agents, including mixtures thereof.

Throughout this specification and the claims which follow, unless the context requires otherwise, the word "comprise", and variations such as "comprises" and "comprising", will be understood to imply the inclusion of a stated integer or step or group of integers or steps but not the exclusion of any other integer or step or group of integers or steps.

Throughout this specification and the claims which follow, unless the context requires otherwise, the phrase "consisting essentially of", and variations such as "consists essentially of" will be understood to indicate that the recited element(s) is/are essential i.e. necessary elements of the invention. The phrase allows for the presence of other non-recited elements which do not materially affect the characteristics of the invention but excludes additional unspecified elements which would affect the basic and novel characteristics of the method defined.

The reference in this specification to any prior publication (or information derived from it), or to any matter which is known, is not, and should not be taken as an acknowledgment or admission or any form of suggestion that that prior publication (or information derived from it) or known matter forms part of the common general knowledge in the field of endeavour to which this specification relates.

Those skilled in the art will appreciate that the invention described herein is susceptible to variations and modifications other than those specifically described. It is to be understood that the invention includes all such variations and modifications, which fall within the spirit and scope. The invention also includes all of the steps, features, compositions and compounds referred to or indicated in this specification, individually or collectively, and any and all combinations of any two or more of said steps or features.

Unless otherwise defined, all technical and scientific terms used herein have the same meanings as commonly understood by one of ordinary skill in the art to which this invention belongs.

Certain embodiments of the invention will now be described with reference to the following examples which are intended for the purpose of illustration only and are not intended to limit the scope of the generality hereinbefore described.

EXAMPLES

Methods

Surrogate virus neutralisation assay (SVNT assay using cPASS, GenScript), memory B cell ELISpot assay, and pseudovirus neutralisation assay

Neutralising antibody levels were measured in surrogate virus neutralisation assay using cPASS kit (GenScript). Spike protein flow cytometry-based assay were used to measure anti-SARS-CoV-2 spike protein IgG levels. Memory B cell counts were measured by memory B cell ELISpot assay. Pseudovirus neutralisation assay was used to measure pseudovirus neutralising activity against Delta or Omicron strains.

Association of infection outcomes

Infections were recorded for 201 participants in the cohort for one year post-vaccination. Logistic regression was used to infer the association of infection outcomes with nAb levels on day 21 or anti-spike protein IgG levels on day 21.

Whole exome sequencing (WES)

Genomic DNA was extracted from whole blood samples using abGenix™ whole blood genomic extraction kit (#800815, AITbiotech). Libraries for whole exome sequencing were prepared using genomic DNA and hybridization capture kit (Roche-Nimblegen SeqCap), and later sequenced with pair-end 150bp reads on Illumina instruments.

Whole exome sequencing (WES) data processing

Each sample was sequenced, yielding approximately 45 million reads with an average coverage of 56. The reads were mapped to the human genome build GRCh38.p13 using

Burrows-Wheeler Aligner (version 0.7.17). The germline short variant discovery pipeline of Genome Analysis Toolkit (version 4.2.6.1) was used to call indels. The called indels have undergone stringent filtering. Exact tests of Hardy-Weinberg Equilibrium were performed to filter variants with excess heterozygosity. In addition, VariantRecalibrator, which builds a recalibration model to score variant quality, was used to filter out indels with low quality. After that, hard filters were applied to select indels with high quality: FS=0, SOR < 0.8, MQ ≥ 60, ExcessHet < 15, and exclude multi-allelic indels. Additionally, for any genotype of an indel, if total individuals (n) < 40, the indel was also excluded. In the end, 983 indels pass the filters.

Sanger sequencing

The same genomic DNAs used for WES were used here for Sanger sequencing. In general, primers were designed to amplify the region from 200 bp upstream of the indel to 800 bp downstream of the indel.

40 ng of genomic DNA was amplified using Phusion high-fidelity DNA polymerase (NEB, M0530L) and indel-specific primers for 30 cycles in Bio-rad theromocycler C1000. KAPA HiFi PCR kit (Roche, 07958935001) was used to amplify high GC-content indel sequences (e.g. Indel 21). After confirming PCR products to be a single band in 1% TAE gel using gel electrophoresis, PCR products were purified using KAPA HyperPure beads (Roche, 08963843001). Purified PCR products were sent for Sanger sequencing. 15 individuals from the BNT162b2 mRNA vaccine cohort were sequenced. Additionally, all 27 individuals in the Sinopharm cohort were sequenced.

Association test of indels with antibody responses to vaccines

For the 983 indels that passed the filters, 25 indels are on the X chromosome. For the 958 autosomal indels, genotypes of each indel were coded as 0 for REF/REF, 1 for REF/ALT, and 2 for ALT/ALT. The additive model was used to test the association of an indel with nAb levels or anti-spike protein IgG levels on day 21 using the additive model:

- nAb levels on day 21 ~ $\beta_0 + \beta_1$*age + β_2*sex + β_3*Chinese ethnicity + β_4*genotype of Indel N (0: REF/REF, 1: REF/ALT, 2: ALT/ALT), or
- anti-spike protein IgG levels on day 21 ~ $\beta_0 + \beta_1$*age + β_2*sex + β_3*Chinese ethnicity + β_4*genotype of Indel N (0: REF/REF, 1: REF/ALT, 2: ALT/ALT).

Age, sex, and Chinese ethnicity were used as covariates. Because the additive model assumes normal distribution, permutation tests were further conducted with 10,000 permutations to obtain p value without normality assumption. To account for multiple testing burden, false discovery rate (FDR) was calculated using the Benjamini-Hochberg procedure and permutation p values. FDR of 10% was used as the cutoff.

Model development using machine learning

nAb levels on day 21 were categorised into low or high antibody responses using the FDA-approved cutoff of 60% inhibition. The quartiles of the age distribution in the mRNA vaccine cohort was used to categorise age into 4 groups: < 34, 34–51.5, 51.5–66, and > 66 years old. To develop a model to predict vaccine-induced antibody response, the Low/High categorised antibody response was employed as the response of the model. Different combinations of these host factors were tested as predictors: age, sex, Chinese ethnicity, and indels.

328 individuals were randomly separated into training and test sets (295 and 33 individuals, respectively). Classification Random Forest models were trained on a training dataset using R package ranger. After testing different combinations of these predictors, a model with the best performance was obtained, using age group, Chinese ethnicity, and Indel 1 as predictors: Vaccine-induced antibody response (Low/High) ~ age group + Chinese ethnicity + Indel 1, with these hyperparameters: num.trees = 200, mtry = 2, min.node.size = 4, replace = T, sample.fraction = 0.25, verbose = F, respect.unordered.factors= "order". R package ROCR was used to plot ROC curve and calculate AUC.

EXAMPLE 1: Indels at certain loci and non-genetic factors are associated with vaccine responsiveness to COVID-19 mRNA vaccines

The method was developed by investigating vaccine response to a COVID-19 mRNA vaccine. 328 individuals were vaccinated with the BNT1626 mRNA vaccines, with first dose on Day 0 and second dose on Day 21. Blood was collected on Day 0, 21, 90, and 180 after the first dose of the vaccine (FIG. 2A). All individuals tested negative for antibodies against the SARS-CoV-2 N protein using the commercial Roche N serology assay, suggesting none of the individuals were infected with SARS-CoV-2 viruses prior to vaccination. The level of vaccine-induced neutralising antibody (nAb) was measured using surrogate virus neutralisation test (cPASS, % inhibition) for the Wuhan strain. Anti-spike protein antibody

(S protein IgG) levels were measured via a flow cytometry-based assay. Levels of nAbs and S protein IgG allowed characterisation of the dynamic antibody responses that followed vaccination.

The overview of the nAb level of 328 people on Day 0, 21, 90, and 180 after the first dose of the vaccine showed substantial variation in nAb levels on Day 21 and Day 180 (FIG. 2B).Considerable variation in S protein IgG levels was also observed on days 21, 90, and 180, with the greatest variance in S protein IgG levels occurring on day 21. The majority of people have nAb levels of Day 21 and 180 below the protective level defined by Feng Zhu et. al (Zhu et al., Lancet Microbe, 5(2): e105, 2022). Interestingly, the nAb level on Day 21 is significantly associated with that on Day 180 with p value $< 2\times10^{-16}$ (FIG. 2C), suggesting people who respond poorly to the first dose of the vaccine are likely to be people who end up with lower level of nAb level six months after vaccination. Hence, the short-term antibody response to the vaccine (nAb level on Day 21) could be predictive of the long-term antibody response to the vaccine (nAb level on Day 180).

Given age and gender are known to affect vaccine response, their associations with nAb levels on day 21 were explored. Using linear regression to test the association, age or gender is found to be significantly associated with the nAb on Day 21, with p value $< 2\times10^{-16}$ for age and p value = 5.964×10^{-9} for gender (FIG. 2D, E). Ethnicity is another factor known to influence vaccine responses. The cohort of 328 individuals consisted of 69.82% Chinese, 10.06% Malay, 8.54% Indian, 7.93% Filipino, and 3.65% individuals from other ethnic backgrounds. To delve deeper, the population was stratified into two groups: Chinese and non-Chinese. Individuals of Chinese ethnicity exhibited significantly lower nAb levels on day 21, compared to their non-Chinese counterparts (linear regression, p value = 3.2×10^{-8}, FIG. 2F). The association between Chinese ethnicity and lower nAb levels on day 21 remained significant after correcting for age and gender (linear regression, Chinese ethnicity p value = 0.0017). Similar associations were also found for S protein IgG levels of day 21 with age, gender, or Chinese ethnicity (linear regression, p value $< 2.2 \times 10^{-8}$ for age, p value = 7.3×10^{-8} and 5.2×10^{-6} for gender and Chinese ethnicity, respectively).

To quantify the contribution of age, gender and ethnicity to the variance in antibody responses to the vaccine, multiple linear regression was performed for nAb levels on day 21 using age and gender as variables. The adjusted R^2 of 0.30 indicated that 30% of the variance

could be explained by age and gender (FIG. 2G). When Chinese ethnicity was included in the model, the adjusted R^2 increased to 0.32, suggesting all three factors combined explained 32% of the variance in nAb levels on day 21. Similar analyses for the S protein IgG levels on day 21 yielded adjusted R^2 values of 0.26 or 0.27 for two factors (age + gender) or three factors combined (age + gender + Chinese). These findings underline the existence of additional factors influencing antibody responses to the vaccine.

It is hypothesised that the individuals carry genetic variants which affect how they respond to the mRNA vaccine. To identify genetic variants, whole exome sequencing (WES) was performed from whole blood samples from the individuals. Each sample was sequenced for ~ 45 million reads with ~56 coverage. WES data was aligned to the human genome build hg38 and processed by GATK pipelines for mapping germline small genetic variants in 328 individuals, including indels and SNPs. Indels were chosen and filtered based on the following criteria. Indels that have multiple alternative genotypes were first filtered out. The remaining indels were further filtered to ensure that each genotype comprised more than 16 individuals (5% of 328). Additionally, the individuals were divided into two groups based on the median of the nAb level on day 21. The allele frequency of the reference allele for each group was calculated together with the log2 foldchange (log2FC) between the two groups for each indel, and indels with log2FC less than 0.3 were filtered out. In the end, 1528 indels passed all filters.

It was reasoned that indels located in different genomic regions would have different effect sizes on the antibody response. For example, indels in the coding regions change the amino acid sequence of the protein and could have a larger effect on antibody response. The 1528 indels were therefore grouped based on their locations in coding regions (CDS), untranslated regions (UTRs), introns, and non-coding regions. While 1136 indels were located in introns, 52 and 138 indels fell in CDS and UTRs, respectively (FIG. 6A). The 52 indels in CDS were first examined. Three genotypes of reference (REF) and alternative (ALT) alleles, REF/REF, REF/ALT, and ALT/ALT were coded 0, 1, and 2. The association of each indel with the nAb level on day 21 was tested by fitting an additive model, nAb level on Day 21 ~ Age + gender + genotype (0, 1, or 2), using age and gender as covariates (FIG. 6B). To account for multiple testing burden, false discovery rate (FDR) was calculated using the Benjamini-Hochberg procedure. The same process was carried out for indels in UTRs, introns, and noncoding regions. Because the additive model assumes normal distribution, permutation

tests were further conducted with 10,000 permutations to acquire p value without normality assumption. In general, the p values from permutation tests are comparable to the p values from the additive model (Table 5), suggesting normality assumption is reasonable. The results show four indels in CDS, one indel in noncoding regions, and three indels in introns, (Indels 1–8, Table 1) are significantly associated with nAb level on Day 21 with FDR < 0.1 (Table 5). If the FDR cut-off was relaxed to be 0.39, two indels in UTRs, one indel in CDS, and one indel in noncoding regions (Indels 9–12, Table 1) are also significantly associated with nAb level on Day 21 after the first dose of the vaccine. Additionally, it was discovered that an indel at chr15:76311269 (Indel 21), that was significantly associated with S protein IgG levels on day 21.

To predict the effect of these indels on antibody response, the genotypes and nAb level for these 12 indels were plotted (FIG. 3). Interestingly, while the alternative genotypes of some indels were associated with lower nAb level, those of other indels were associated with higher nAb level. It indicates that these indels affect genes of different functions and the variable level of neutralising antibody is a combinatorial effect resulting from multiple genetic variants. Indels 1–4 and 11 in CDS cause in-frame insertion or deletion, which change 1–3 amino acids in the proteins encoded by the nearest genes, *SERINC2*, *ADRA2B*, *M6PR*, *DEFB132*, and *JMJD1C*. On the other hand, Indel 5 and Indel 12 are located in noncoding regions, within 5 kb and 2 kb regions of the *OPTN* and *ARFGEF2* genes. Indels 6–8 are in the introns of *SPIRE2*, *PODXL*, and *BTBD11* genes. While it would be reasonable to assume Indels 5–8 and 12 affect the expression of their nearest genes, these indels also likely to alter the expression of other neighbouring genes. Indels 9–10 fall in the UTRs of *BTN2A1* and *OR2AG1*, so they may affect the translation of these two genes.

The dbGap database and gnomAD were used to examine the distribution of these indels in global populations. It was found that indels from East Asian populations only represent 0.4–2.2% of the total number of indels in those databases. Potentially due to being under-represented, only six of the twelve indels were reported in those databases. Indel 6 has allele frequency of 0.93 in East Asians, which is much higher than in other populations (0.34–0.43) (FIG. 4A). Indel 3 also has relatively higher frequency of 0.36 in East Asians when compared to other populations. Notably, the frequency of Indel 3 is much lower in other populations,

ranging from 0.03–0.11. Unlike Indels 3 and 6, Indels 2, 5, 7 and 8 have comparable frequencies in most populations.

The potential cell types that might be affected by these indels were investigated. Single-cell RNA-seq data (scRNA-seq) was obtained for multiple tissues or immune cells from The Human Protein Atlas for genes nearest to these indels. Many of the genes are expressed in multiple tissues while some have specific expression in one or two tissues. Gene ontology analysis of these 12 genes showed no enrichment in specific pathways. It was found that *M6PR*, *OPTN*, *BTN2A1* and *JMJD1C* are genes that are expressed in immune cells. Among them, *JMJD1C* is expressed in neutrophils, and naïve and memory B cells (FIG. 4B). *JMJD1C* was recently reported to restrain plasma cell differentiation by demethylating STAT3 in a mouse model. As Indel 11 is a 6 bp deletion in the CDS of *JMJD1C* and is associated with lower neutralising antibody level, Indel 11 may function in enhancing the activity of *JMJD1C* to suppress plasma cell differentiation and result in lower level of neutralizing antibody. Notably, scRNA-seq data in this database mostly captures steady-state expression. Some of those genes might only express upon stimulation, such as *DEFB132*.

These indels might affect the level of neutralising antibody by altering relevant cellular functions. Thus, for each indel, the difference in the level of memory B cells (mBCs) or pseudovirus neutralisation activity between people with different genotypes was tested. When the foldchange of mBC count between Day 0 and a later time point was calculated, it was found that people with ALT/ALT genotype of Indel 8 had more mBCs on Day 21 than people with REF/ALT genotype of Indel 8 (Wilcoxon rank sum test, p value: 0.00591) (FIG. 4C). On the other hand, people with REF/ALT genotype of Indel 11 had fewer mBCs on Day 90 than people with REF/REF genotype of Indel 11 (Wilcoxon rank sum test, p value: 0.04895). Given that n in the mBC analysis is smaller than 50 for each group, 10,000 permutations were performed by shuffling the foldchange and p values with permutations remain to be significant, 0.0025 for Indel 8 and 0.0251 for Indel 11. As the alternative genotype of Indels 8 and 11 are associated with higher and lower nAb level respectively, the results of changes in mBC are consistent: people with ALT genotype of Indel 8 had more mBCs and people with ALT genotype of Indel 11 had fewer mBCs, comparing to the REF genotype.

A similar phenomenon was observed for pseudovirus neutralisation activity on Day 90 (FIG. 4D). People with ALT/ALT genotype of Indel 8 show higher neutralising activity for pseudovirus of omicron strain than people with REF/ALT genotype (Wilcoxon rank sum test with 10,000 permutations, p value: 0.0054). People with REF/ALT genotype of Indel 11 exhibit lower neutralising activity for pseudovirus of delta strain than people with REF/REF genotype of Indel 11 (p value with permutations: 0.0068). The results aligned with the previous neutralising antibody and mBC results.

Besides B cells and antibodies, the association of each indel with cell count of different T cell types was also tested after *in vitro* stimulation with SARS-CoV-2 Spike protein peptides. Interestingly, it was found that the alternative genotype of Indel 8 is associated with increased cell count of T follicular helper (Tfh) cells ($CXCR5^+$ $CD4^+$ cells) (FIG. 4E). Tfh cells help B cells produce antibody. Taken together, these results revealed that Indel 8 is associated with higher nAb level on Day 21, higher mBCs on Day 21, higher pseudovirus neutralising activity on Day 90, and higher cell count of Tfh cells. On the contrary, the alternative genotype of Indel 11 is associated with lower neutralising antibody level on Day 21, fewer mBCs on Day 90, and lower pseudovirus neutralising activity on Day 90. This suggests that Indel 8 and 11 might affect the expression of genes that have direct functions in B cell and T cell pathways.

EXAMPLE 2: A LASSO regression model predicts the neutralising antibody level to COVID-19 mRNA vaccine

A model was built to predict neutralising antibody (nAb) levels for individuals following administration of COVID-10 mRNA vaccine. Age and gender contributed about 30% variance to the nAb level on Day 21, thus the predictive power of the multiple linear regression model would likely be improved by adding indels as variables.

The 328 individuals were first randomly separated into training set of 263 people and test set of 65 people (FIG. 5A). The training set was further randomly divided into final training set of 237 people and validation set of 26 people. Additionally, 100 random resamplings were performed to obtain different training and test sets to understand how sampling affects the model and performance. Each model was built based on 100 resamplings with K=10 cross validation.

To test if addition of indels improved the model, multiple linear regression was first used to model nAb level on Day 21 with age, gender, and 0, 5, or 12 indels, nAb level on Day 21 ~ age + gender + 0, 5, or 12 indels. For each of the three models with different number of indels, the model with the resampling which gives the least mean-squared error (MSE) of the test set, MSE.test, was chosen. The predicted values based on the test set are plotted against observed values for the three models (FIG. 5B). The results showed the dots moving towards the diagonal line as more indels were included. To quantify the difference between predicted and observed values, MSE.tests were calculated to be 240.74, 197.41, and 169.40 for the best models of 0, 5, and 12 indels. The model with 12 indels had the lowest MSE.test score, suggesting the model with 12 indels was the best of the three models. The adjusted R-squared also increased from 0.3 to 0.44 when 12 indels were used in the model.

While the model improved with the addition of more indels as variables, the variables need to be carefully selected to avoid overfitting. To perform variable selection and improve the linear regression model, LASSO regression was applied using the 14 variables (age, gender, and 12 indels) with the training and test sets, as shown in FIG. 5A. Log(λ) was plotted with MSE for the models with different number of variables, and it was found that the model with the lowest λ and MSE included all 14 variables (FIG. 5C). Thus, the inclusion of all 14 variables in the model gave the best predictive value. It was desired to know how the MSE.test score changes with resampling. From 100 random resamplings, it was found the MSE.test score ranged from 169.24–464.90 (median: 307.12) for the LASSO regression model and ranged from 169.40–465.08 (median: 307.76) for the linear regression model (FIG. 5D). Thus, the LASSO regression model was slightly improved over the linear regression model, and this model was chosen for subsequent validation.

It is desirable to be able to predict the responsiveness to a vaccine for a given individual. Using the LASSO model with the least MSE.test, the predicted values are plotted with the observed values for the test set (Fig. 5E). Using the "Low" level of nAb approved by U.S. FDA as a cut-off, the error rate of predicting people whose nAb levels are lower than the "Low" level (Low responders) is 18.46%. A total of 53 out of 65 people in the test set were predicted correctly for being or not being a low responder.

Table 5. 12 indels that are associated with the neutralising antibody level on Day 21.

The REF genotype sequence for Indel 2 corresponds to SEQ ID NO: 1 in Table 1.

Indel	Chr	Position	REF genotype	ALT genotype	Difference (REF vs ALT)	Genomic location	Nearest gene	REF/REF freq (%)	ClinVar	P value (Additive model)	P value with permutation	FDR
1	chr1	31403842	A	ACAG	3bp insertion	CDS	SERINC2	12.26	-	0.0003	0.0001	0.0052
2	chr2	96115238	CTCCTCCTCT	C	9bp deletion	CDS	ADRA2B	33.84	+	0.0010	0.0006	0.0156
3	chr12	8946386	G	GAAC	3bp insertion	CDS	MGP R	51.62	-	0.0038	0.0038	0.0624
4	chr20	257194	CTCGTCT	C	6bp deletion	CDS	DEFB132	9.78	-	0.0051	0.0048	0.0624
5	chr10	131423833	T	TC	1bp insertion	noncoding	CPTN	28.72	-	0.0003	0.0001	0.0292
6	chr16	88660945	T	TC	1bp insertion	intron	SPIRE2	9.45	-	0.0002	0.0001	0.0566
7	chr7	131936795	TC	T	1bp deletion	intron	PODXL	29.12	+	0.0003	0.0001	0.0568
8	chr12	107328523	T	TC	1bp insertion	intron	B1BD11	8.23	-	0.0023	0.0021	0.0752
9	chr8	26501243	AG	A	1bp deletion	UTR	B1NZA1	8.71	-	0.0013	0.0023	0.2208
10	chr11	67286184	CT	C	1bp deletion	UTR	OR2AG1	60.37	-	0.0032	0.0040	0.2208
11	chr10	63208190	ACTAAAC	A	6bp deletion	CDS	JMJD1C	61.66	+	0.0344	0.0314	0.8286
12	chr20	49636827	C	CTTTG	4bp deletion	noncoding	ARFGEF2	12.80	-	0.0041	0.0038	0.3828

EXAMPLE 3: Indels at certain loci are associated with antibody waning rate after two doses of COVID-19 mRNA vaccines

Using the same cohort, the neutralising antibody level is converted to antibody titre based on the equation provided in Zhu Feng et. al, 2022. Antibody waning rate is calculated by how much antibody titre decreases per day using the data from Day 90 and 180. The same pipeline is applied for identifying indels associated with antibody waning rate. 8 indels are found to be significantly associated with antibody waning rate (FIG. 7).

EXAMPLE 4: A LASSO regression model predicts the neutralising antibody waning rate after two doses of COVID-19 mRNA vaccine

A similar approach of random sampling is applied to develop a model to predict antibody waning rate (FIG. 8A). As the neutralising antibody level on Day 21 is associated with that on Day 180, the prediction from the first model that predicts vaccine responsiveness is used as one variable together with the 8 indels (Indel 13 – 20) for LASSO regression, antibody waning rate ~ Age + gender + prediction from the first model + 8 indels. Log(λ) was plotted with MSE for the models with different number of variables, and it was found that the model with the lowest λ and MSE included all 11 variables (FIG. 8B). Thus, the inclusion of all 11 variables in the model gave the best predictive value. The performance of the best model with the least MSE is assessed by comparing the predicted value to the observed value. The prediction could be used to estimate the timing for booster shots for an individual.

EXAMPLE 5: A Random Forest model predicts the categorized "Low/High" antibody responses of a subject to COVID-19 vaccines

It was first investigated whether nAb levels on day 21 could serve as a correlate of protection. After analysing the infection records within one year of vaccination for 201 individuals in the cohort (who had received three COVID-19 vaccine doses in one year), it was found that there was a significant association: individuals who had infections had lower nAb levels on day 21 compared to those who had no infections (logistic regression, p value = 0.0013, Fig. 9A). Similar associations were found for nAb levels on day 90 and day 180, with their p

values higher than the p value for day 21 (logistic regression, p value = 0.0038 and 0.0030 for day 90 and day 180, respectively). Additionally, anti-SARS-CoV-2 spike protein IgG levels on day 21 were associated with infection outcomes (logistic regression, p value = 6.7 × 10−5, Fig. 9B). A similar association was found for anti-SARS-CoV-2 spike protein IgG levels on day 90, but not day 180 (logistic regression, p value = 0.013 and 0.69 for day 90 and 180, respectively). These results suggest antibody responses on day 21 could serve as a correlate of protection.

A model to predict vaccine-induced antibody responses was developed using the identified host factors. To facilitate interpretation, nAb levels on day 21 were categorised as either "Low" or "High" using an FDA-approved cutoff of 60% inhibition for low or high antibody responses. The Low/High categorised antibody responses were employed as the response of the model and these host factors as our predictors. To construct the predictive model, 328 individuals were randomly divided into training and test sets (295 and 33 people, respectively). The performance of the predictive models was validated with the test set or the Sinopharm vaccine cohort as an independent test set. The Sinopharm vaccine cohort included 27 participants who received two doses of the Sinopharm COVID-19 vaccine, an inactivated SARS-CoV-2 virus vaccine, on days 0 and 21. Their nAb levels were measured using the same SVNT assays on days 0, 21, 90, and 180.

The kinetics of the vaccine-induced nAb levels over time were compared between the two vaccines (Fig. 12A). The nAb levels induced by the Sinopharm vaccine exhibited slower kinetics compared to those induced by the BNT162b2 mRNA vaccine (Fig. 10B). When calculated the variance in nAb levels at the four time points for the Sinopharm cohort, nAb levels on day 180 exhibited the greatest variance. Thus, for the Sinopharm vaccine cohort, nAb levels on day 180 were categorised into Low/High antibody responses using the same cutoff value.

Random Forest models were trained using the training set to predict vaccine-induced Low/High antibody responses. Initially, ages of the individuals were categorised into four groups: <34, 34–51.5, 51.5–66, and >66 years old. Different combinations of these predictors were explored to enhance the model. Ultimately, a model was obtained with the best performance using age groups, Chinese ethnicity, and Indel 1 as predictors. The model predicted the test set's Low/High antibody responses with an accuracy of 72.7%. More

importantly, it achieved an accuracy of 66.7% for predicting the day 180 Low/High antibody responses of individuals in the Sinopharm vaccine cohort. The area under the receiver operating characteristic curve (AUC) analysis yielded 80.3%, 81.2%, and 66.4% for the training set, test set, and Sinopharm cohort, respectively (Fig. 10C).

It will be appreciated that many further modifications and permutations of various aspects of the described embodiments are possible. Accordingly, the described aspects are intended to embrace all such alterations, modifications, and variations that fall within the spirit and scope of the appended claims.

CLAIMS

1. A method for predicting the responsiveness of a subject to a vaccine, the method comprising: (a) detecting the presence or absence of, and the copy number of, an insertion/deletion (indel) mutation at one or more genetic loci in Table 1 in a sample obtained from the subject; and (b) calculating, based on the presence or absence of, and the copy number of, an indel mutation at the one or more genetic loci, a value predictive of the responsiveness of the subject to the vaccine.

2. The method of claim 1, wherein step (a) comprises detecting the presence or absence of, and the copy number of, an indel mutation at locus 8 and/or 11 in Table 1.

3. The method of claim 1, wherein step (a) comprises detecting the presence or absence of, and the copy number of, an indel mutation at each of locus 1 to 12 in Table 1.

4. The method of claim 1, wherein step (a) comprises detecting the presence or absence of, and the copy number of, an indel mutation at each of locus 1 to 21 in Table 1.

5. The method of any one of claims 1 to 4, wherein step (a) comprises detecting the presence or absence of, and the copy number of, one or more alleles in Table 1.

6. The method of any one of claims 1 to 5, wherein the method further comprises the use of at least one non-genetic factor to calculate the value.

7. The method of claim 6, wherein the non-genetic factor is selected from the age of the subject, the gender of the subject and the ethnicity of the subject.

8. The method of any one of claims 1 to 7, wherein step (b) is performed using a computer algorithm.

9. The method of claim 8, wherein step (b) is performed using a machine learning model trained to generate a value predictive of responsiveness to a vaccine based on the presence or absence of, and the copy number of, an indel mutation at one or more genetic loci in Table 1.

10. The method of claim 9, wherein the machine learning model is a random forest classifier model.

11. The method of claim 9 or 10, wherein the machine learning model is trained using a dataset comprising genotype data and vaccine response data from a cohort of subjects who have received one or more doses of the vaccine, wherein the genotype data comprises the presence or absence of, and the copy number of, an indel mutation at one or more genetic loci in Table 1.

12. The method of claim 11, wherein the vaccine response data comprises the level of neutralising antibodies and/or the antibody waning rate in the cohort of subjects after the subjects have received one or more doses of the vaccine.

13. The method of any one of claims 1 to 12, wherein the vaccine is directed against a viral, bacterial and/or fungal pathogen.

14. The method of claim 13, wherein the vaccine is directed against one or more respiratory pathogens.

15. The method of claim 13 or 14, wherein the vaccine is directed against one or more viral pathogens.

16. The method of claim 15, wherein the viral pathogen is selected from the group consisting of a coronavirus, an influenza virus, a parainfluenza virus, or respiratory syncytial virus.

17. The method of claim 16, wherein the viral pathogen is a betacoronavirus.

18. The method of claim 17, wherein the vaccine is directed against the SARS-CoV-2 virus.

19. The method of any one of claims 1 to 18, wherein the value predictive of responsiveness is a predicted level of neutralising antibodies in the subject after one or more doses of the vaccine.

20. The method of any one of claims 1 to 19, wherein the value predictive of responsiveness is a predicted antibody waning rate after one or more doses of the vaccine.

21. The method of any one of claims 1 to 20, wherein the sample is a blood sample or a swab.

22. The method of any one of claims 1 to 21, wherein step (a) comprises amplifying the one or more genetic loci.

23. The method of claim 22, wherein the one or more genetic loci is amplified using loop-mediated isothermal amplification (LAMP).

24. The method of any one of claims 1 to 23, wherein the method is high throughput.

25. A method of treating or preventing a disease in a subject, the method comprising: (a) detecting the presence or absence of, and the copy number of, an indel mutation at one or more genetic loci in Table 1 using a sample obtained from the subject; (b) calculating, based on the presence or absence of, and the copy number of, an indel mutation at the one or more genetic loci, a value predictive of the responsiveness of the subject to a vaccine against the disease; and (c) providing an effective amount of the vaccine to the subject based on the predicted responsiveness of the subject to the vaccine to treat or prevent the disease.

26. The method of claim 25, wherein the value predictive of responsiveness is a predicted level of neutralising antibodies in the subject after one or more doses of the vaccine, and wherein step (c) comprises providing an effective amount of the vaccine based on the predicted level of neutralising antibodies.

27. The method of claim 25, wherein the value predictive of responsiveness is a predicted antibody waning rate after one or more doses of the vaccine, and wherein step (c) comprises providing an effective amount of the vaccine based on the predicted antibody waning rate.

28. The method of any one of claims 25 to 27, wherein the disease is an infection.

29. The method of claim 28, wherein the infection is caused by a viral, bacterial or fungal pathogen.

30. The method of claim 29, wherein the pathogen is a respiratory pathogen.

31. The method of claim 29 or 30, wherein the pathogen is a viral pathogen.

32. The method of claim 31, wherein the viral pathogen is selected from the group consisting of a coronavirus, an influenza virus, a parainfluenza virus, or respiratory syncytial virus.

33. The method of claim 32, wherein the viral pathogen is a betacoronavirus.

34. The method of claim 33, wherein the infection is caused by the SARS-CoV-2 virus.

35. A method of selecting a subject for treatment with a vaccine, the method comprising: (a) detecting the presence or absence of, and the copy number of, an indel mutation at one or more genetic loci in Table 1 using a sample obtained from the subject; (b) calculating, based on the presence or absence of, and the copy number of, an indel mutation at the one or more genetic loci, a value predictive of the responsiveness of the subject to a vaccine; and (c) selecting a subject for treatment with the vaccine based on the predicted responsiveness of the subject to the vaccine.

36. The method of claim 35, wherein the value predictive of responsiveness is a predicted level of neutralising antibodies in the subject after one or more doses of the vaccine, and wherein step (c) comprises selecting a subject based on the predicted level of neutralising antibodies.

37. The method of claim 35, wherein the value predictive of responsiveness is a predicted antibody waning rate after one or more doses of the vaccine, and wherein step (c) comprises selecting a subject based on the predicted antibody waning rate.

38. The method of any one of claims 35 to 37, further comprising providing an effective amount of the vaccine to the subject based on the predicted responsiveness of the subject to the vaccine.

39. A kit, comprising one or more oligonucleotides for detecting the presence or absence of, and the copy number of, an indel mutation at a genetic locus in Table 1.

40. The kit of claim 39, wherein the one or more oligonucleotides is a primer for amplifying an allele in Table 2.

41. The kit of claim 40, wherein the primer is for loop-mediated isothermal amplification (LAMP) of an allele in Table 2.

42. The kit of claim 40 or 41, wherein the primer comprises or consists of a nucleic acid sequence having at least 80% sequence identity to a nucleic acid sequence in SEQ ID NO: 3–30.

43. A composition comprising: (a) a sample obtained from a subject; and (b) one or more oligonucleotides for detecting the presence or absence of, and the copy number of, an indel mutation at a genetic locus in Table 1 in the sample.

44. A computer-implemented method for predicting the responsiveness of a subject to a vaccine, the method comprising: (a) receiving genotype data comprising the presence or absence of, and the copy number of, an insertion/deletion (indel) mutation at one or more genetic loci in Table 1 from a sample from the subject; and (b) calculating a value predictive of the responsiveness of the subject to the vaccine using a machine learning model trained to generate the predictive value based on the presence or absence of, and the copy number of, an indel mutation at one or more genetic loci in Table 1.

45. A system for predicting the responsiveness of a subject to a vaccine, the system comprising: a processor coupled to computer readable memory, the memory comprising instructions that when executed by the processor causes the processor to: (i) receive genotype data comprising the presence or absence of, and the copy number of, an insertion/deletion (indel) mutation at one or more genetic loci in Table 1 from a sample from the subject; and (ii) calculate, based on the presence or absence of, and the copy number of, an indel mutation at the one or more genetic loci, a value predictive of the responsiveness of the subject to the vaccine using a machine learning model.

46. A system for predicting the responsiveness of a subject to a vaccine, the system comprising:
    a) a processor coupled to computer readable memory;
    b) an array for high-throughput detection of the presence or absence of, and the copy number of, an indel mutation at one or more genetic loci in Table 1 in a sample obtained from the subject, wherein the array is configured to generate a plurality of detectable signals indicative of the presence or absence of, and the copy number of, an indel mutation at the one or more genetic loci; and
    c) a computer algorithm stored in the computer readable memory and comprising instructions executable by the processor to take as input the plurality of detectable signals generated by the array and calculate a value predictive of the responsiveness of the subject to the vaccine.