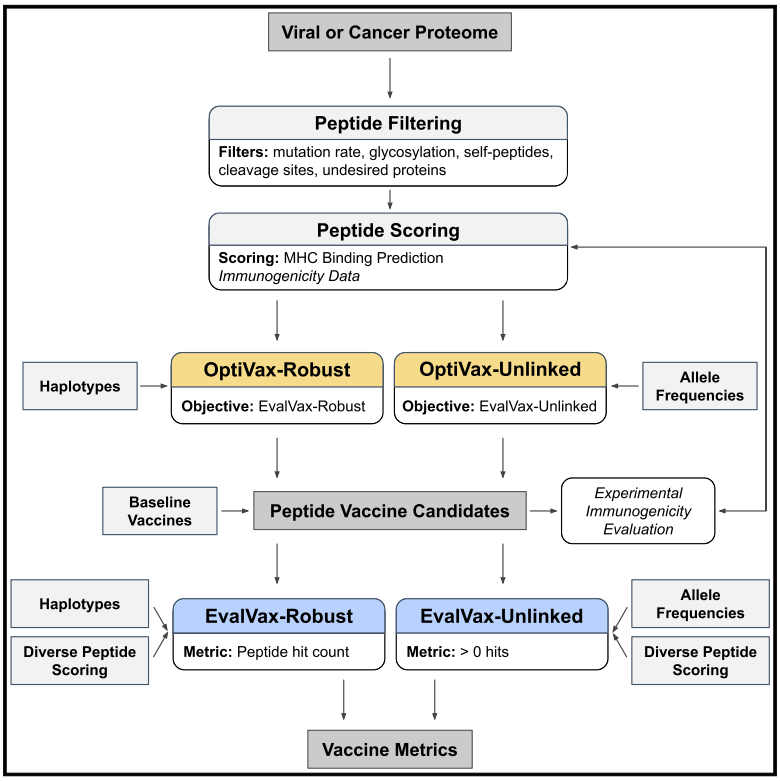
Since January 2020 Elsevier has created a COVID-19 resource centre with free information in English and Mandarin on the novel coronavirus COVID-19. The COVID-19 resource centre is hosted on Elsevier Connect, the company's public news and information website.

Cell Systems

Article

Computationally Optimized SARS-CoV-2 MHC Class I and II Vaccine Formulations Predicted to Target Human Haplotype Distributions

Graphical Abstract



Highlights

- HLA haplotype frequencies are used to predict the population coverage of vaccine designs
- Methods are provided to evaluate and optimize peptide-vaccine formulations
- Vaccine designs are optimized for the HLA display of multiple peptides
- Peptides are scored for selection on multiple criteria including observed mutation rate

Authors

Ge Liu, Brandon Carter, Trenton Bricken, Siddhartha Jain, Mathias Viard, Mary Carrington, David K. Gifford

Correspondence

gifford@mit.edu

In Brief

HLA haplotype frequencies are used to evaluate and design SARS-CoV-2 peptide vaccines. The methods presented optimize the presentation likelihood of a diverse set of vaccine peptides to maximize vaccine immunogenicity. The proposed SARS-CoV-2 MHC class I vaccine formulations provide 93.21% predicted population coverage with at least five vaccine peptide-HLA hits with all vaccine peptides perfectly conserved across 4,690 geographically sampled SARS-CoV-2 genomes. The proposed MHC class II vaccine formulations provide 97.21% predicted coverage with at least five vaccine peptide-HLA hits.

Liu et al., 2020, Cell Systems *11*, 131–144
August 26, 2020
https://doi.org/10.1016/j.cels.2020.06.009

CellPress

Cell Systems

CellPress

Article

Computationally Optimized SARS-CoV-2 MHC Class I and II Vaccine Formulations Predicted to Target Human Haplotype Distributions

Ge Liu,[1,2,7] Brandon Carter,[1,2,7] Trenton Bricken,[4] Siddhartha Jain,[1] Mathias Viard,[5,6] Mary Carrington,[5,6] and David K. Gifford[1,2,3,8,*]

[1]MIT Computer Science and Artificial Intelligence Laboratory, Cambridge, MA, USA
[2]MIT Electrical Engineering and Computer Science, Cambridge, MA, USA
[3]MIT Biological Engineering, Cambridge, MA, USA
[4]Duke University, Durham, NC, USA
[5]Basic Science Program, Frederick National Laboratory for Cancer Research, Frederick, MD, USA
[6]Ragon Institute of Massachusetts General Hospital, MIT and Harvard University, Cambridge, MA, USA
[7]These authors contributed equally
[8]Lead Contact
*Correspondence: gifford@mit.edu
https://doi.org/10.1016/j.cels.2020.06.009

SUMMARY

We present a combinatorial machine learning method to evaluate and optimize peptide vaccine formulations for SARS-CoV-2. Our approach optimizes the presentation likelihood of a diverse set of vaccine peptides conditioned on a target human-population HLA haplotype distribution and expected epitope drift. Our proposed SARS-CoV-2 MHC class I vaccine formulations provide 93.21% predicted population coverage with at least five vaccine peptide-HLA average hits per person ($\geq$ 1 peptide: 99.91%) with all vaccine peptides perfectly conserved across 4,690 geographically sampled SARS-CoV-2 genomes. Our proposed MHC class II vaccine formulations provide 97.21% predicted coverage with at least five vaccine peptide-HLA average hits per person with all peptides having an observed mutation probability of $\leq$ 0.001. We provide an open-source implementation of our design methods (OptiVax), vaccine evaluation tool (EvalVax), as well as the data used in our design efforts here: https://github.com/gifford-lab/optivax.

INTRODUCTION

An effective vaccine for SARS-CoV-2 is urgently needed. For a peptide to be effective in a vaccine to induce cellular immunity, it must first bind within the groove of a major histocompatibility complex (MHC) class I or class II molecule. Second, it must be immunogenic; that is, it must activate T cells when it is bound by MHC proteins and displayed. Immunogenicity is therefore dependent on the sequence of the peptide displayed, the protein sequences of an individual's MHC genes, and the affinity between the two. A challenge for the design of peptide vaccines is the diversity of human MHC gene alleles that each have specific preferences for the peptide sequences they display. The human leukocyte antigen (HLA) loci, located within the MHC, encode the HLA class I and class II molecules; an individual's HLA type describes the alleles that he or she carries at each of the three classical class I loci (HLA-A, HLA-B, and HLA-C) and three class II loci (HLA-DR, HLA-DQ, and HLA-DP).

To create effective vaccines, it is necessary to consider the HLA allelic frequency in the target population as well as linkage disequilibrium between HLA genes to discover a set of peptides that is likely to be robustly displayed. Human populations that originate from different geographies have differing frequencies of HLA alleles, and these populations exhibit linkage disequilibrium between HLA loci that result in population-specific haplotype frequencies. However, previous computational peptide vaccine design and evaluation methods do not utilize the distribution of HLA haplotypes in a population, and thus, cannot accurately assess the coverage provided by a vaccine. Present population-based methods, such as iVax (Moise et al., 2015) and SARS-CoV-2-specific efforts (Fast et al., 2020), do not take into account haplotypes and rare HLA allelic combinations. The immune epitope database (IEDB) population coverage tool (Bui et al., 2006) estimates peptide-HLA binding coverage and the distribution of peptides displayed for a given population but assumes independence between different loci, and thus, does not consider linkage disequilibrium.

Here, we utilize human HLA haplotype frequencies of three major populations, those self-reporting as having White, Black, or Asian ancestry, to compute population coverage of SARS-CoV-2 peptides with high predicted HLA binding affinity for inclusion in MHC class I or II vaccine formulations. We examined 4,690 geographically sampled SARS-CoV-2 genomes to exclude peptides with undesired mutation rates. Recent

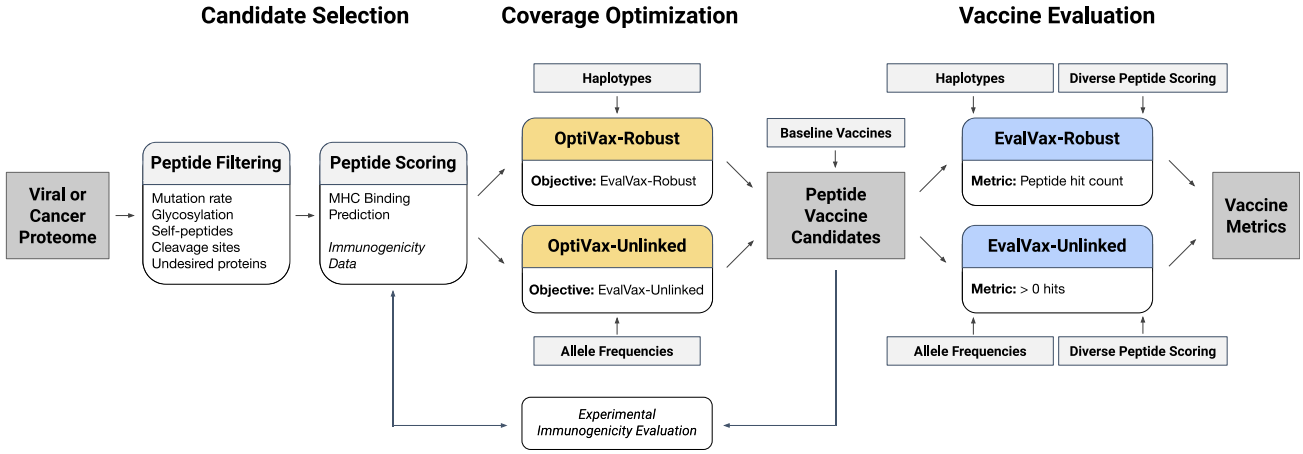


Figure 1. The OptiVax and EvalVax Machine Learning System for Combinatorial Vaccine Optimization and Evaluation

These methods can be used to design new peptide vaccines, evaluate existing vaccines, or augment existing vaccine designs. Peptides are scored by machine learning and immunogenicity data for population coverage optimization and evaluation.

advances in machine learning have produced models that can predict the presentation of peptides by hundreds of allelic variants of both class I and class II MHC molecules (Zeng and Gifford, 2019; Jurtz et al., 2017; O'Donnell et al., 2018; Jensen et al., 2018; Peters et al., 2020). These models were evaluated on their ability to accurately predict data that are not observed during their training on hundreds of HLA alleles. Given the fact that different models may be more or less accurate for different sequence families and can make idiosyncratic errors, we used an ensemble of models for vaccine design. We evaluated completed designs using eleven models to provide a conservative evaluation of vaccine peptide presentation.

Using conservative metrics of peptide-HLA binding, we find that our optimization methods provide both a higher likelihood of peptide display as well as a larger number of associated peptides than other published SARS-CoV-2 peptide vaccine designs with fewer than 150 peptides. Our proposed SARS-CoV-2 MHC class I vaccine formulations provide 93.21% predicted population coverage with at least five vaccine peptide-HLA average hits per person ($\geq$ 1 peptide: 99.91%), with all vaccine peptides perfectly conserved across 4,690 geographically sampled SARS-CoV-2 genomes. Our proposed MHC class II vaccine formulations provide 97.21% predicted coverage with at least five vaccine peptide-HLA average hits per person with all peptides having an observed mutation probability of $\leq$ 0.001. We also show that OptiVax can be used to augment S protein vaccine designs to increase their population coverage.

RESULTS

Our approach to vaccine design uses combinatorial optimization to select peptides to achieve specific population-level objectives. We provide two methods for peptide vaccine evaluation: EvalVax-Unlinked, which considers HLA allele frequencies and assumes independence between HLA loci, and EvalVax-Robust, which considers haplotype frequencies and computes population coverage at minimum levels of high-scoring peptide-HLA combinations per individual. We employ these evaluation methods as objective functions for peptide vaccine formulation by combinatorial optimization in OptiVax-Unlinked and OptiVax-Robust. In our framework, vaccine design proceeds by (1) starting with an initial proteome, filtering out peptides with undesired properties, (2) scoring which peptides will be presented, and thus are potentially immunogenic, and (3) selecting an optimized set of candidate peptides, given the frequency of HLA haplotypes or HLA alleles in a target population. Our filtering step eliminates peptides that are expected to be glycosylated, peptides that are expected to drift in sequence, and thus cause vaccine escape, peptides that are cleaved, and peptides that are identical to peptides in the human proteome. Vaccine peptides can be drawn from the entire proteome or from specific proteins of interest. An overview of our system is shown in Figure 1.

Once candidate peptides are tested, any that are not immunogenic in the context of the restricting HLA allotype can be eliminated from the candidate peptide pool. Draft vaccine designs containing nonimmunogenic peptides can be revised to eliminate them, and the reduced vaccine design can be re-evaluated with EvalVax to see if the design still meets performance criteria. If not, the vaccine design process can be repeated with the revised candidate pool. Immunogenicity data can be incorporated into the peptide scoring process that is used for both vaccine design and evaluation, as shown in italics in Figure 1.

Datasets

A Proteome Is Converted into Candidate Vaccine Peptides

Given a target proteome as input, we identify all potential T cell epitopes for inclusion in a vaccine. We extract peptides of length 8–10 inclusive for consideration of MHC class I binding (Rist et al., 2013) and peptides of length 13–25 inclusive for class II binding (Chicz et al., 1992) by using sliding windows of each size over the entire proteome. While peptides presented by MHC class I molecules can occasionally be longer than 10 residues (Trolle et al., 2016), we conservatively limited our search to length 8–10, since MHC class I presented peptides are predominately 8–10 residues in length (Rist et al., 2013).

Using this sliding-window approach, we created peptide sets from the SARS-CoV-2 (COVID-19) and SARS-CoV (human SARS coronavirus) proteomes. SARS-CoV-2 was processed to discover relevant peptides for a vaccine, and SARS-CoV was processed to reveal common peptides between the two viruses during evaluation. The SARS-CoV-2 proteome comprises four structural proteins (E, M, N, and S) and at least six additional open reading frames (ORFs) encoding nonstructural proteins, including the SARS-CoV-2 protease (Finkel et al., 2020; Zhang et al., 2020a). We obtained the SARS-CoV-2 viral proteome from GISAID (Elbe and Buckland-Merrett, 2017) sequence entry Wuhan/IPBCAMS-WH-01/2019, the first documented case. We used Nextstrain (Hadfield et al., 2018) to identify ORFs and translate the sequence. Our sliding windows on SARS-CoV-2 resulted in 29,403 candidate peptides for MHC class I and 125,593 candidate peptides for MHC class II. We obtained the SARS-CoV proteome from UniProt: UP000000354 (Consortium, 2019). For SARS-CoV, our procedure created 29,661 and 126,711 unique peptides for MHC class I and class II, respectively.

HLA Population Frequency Computation

When we compute the probability of vaccine coverage over a population, we use complementary methods that assume either independence or linkage between allele frequencies in genomically proximal HLA loci. In EvalVax-Unlinked, we assume independence and use HLA allelic frequencies for 2,392 class I alleles and 280 class II alleles across 15 geographic regions from the dbMHC database (Helmberg et al., 2004) obtained from the IEDB population coverage tool (Bui et al., 2006). For each geographic region, we normalize the frequencies within each locus. If the sum of the raw frequencies exceeds one, we normalize them to one; and if the sum of the raw frequencies is less than one, the missing frequency is made up by a placeholder allele that receives no binding. In EvalVax-Robust, we assume linkage and use observed haplotype frequencies of HLA-A, HLA-B, and HLA-C loci for class I computations, or observed haplotype frequencies of HLA-DP, HLA-DQ, and HLA-DR for class II computations. We observed a total of 2,138 distinct haplotypes for the HLA class I locus that included 230 different HLA-A, HLA-B, and HLA-C HLA alleles. We observed a total of 1,711 distinct haplotypes for the HLA class II loci that included 280 different HLA-DP, HLA-DQ, and HLA-DR HLA alleles. We have independent haplotype frequency measurements for three populations self-reporting as having White (European), Black (African), or Asian ancestry.

HLA class I and class II haplotype frequencies were inferred using high-resolution typing of individuals from distinct racial backgrounds. We estimated HLA class I haplotypes from HLA-A, -B, and -C genotypes of 2,886 individuals of Black ancestry (46 distinct HLA-A alleles, 70 distinct HLA-B alleles, and 40 distinct HLA-C alleles), 2,327 individuals of White ancestry (38 distinct HLA-A alleles, 64 distinct HLA-B alleles, and 34 distinct HLA-C alleles), and 1,653 individuals of Asian ancestry (25 distinct HLA-A alleles, 51 distinct HLA-B alleles, and 25 distinct HLA-C alleles). HLA class II haplotypes were estimated based on DR, DQ, and DP genotypes of 2,474 individuals of Black ancestry (10 distinct HLA-DPA1 alleles, 45 distinct HLA-DPB1 alleles, 14 distinct HLA-DQA1 alleles, 21 distinct HLA-DQB1 alleles, and 38 distinct HLA-DRB1 alleles), 1,857 individuals of White ancestry (7 distinct HLA-DPA1 alleles, 29 distinct HLA-DPB1 alleles, 18 distinct HLA-DQA1 alleles, 21 distinct HLA-DQB1 alleles, and 41 distinct HLA-DRB1 alleles), and 1,675 individuals of Asian ancestry (7 distinct HLA-DPA1 alleles, 28 distinct HLA-DPB1 alleles, 16 distinct HLA-DQA1 alleles, 16 distinct HLA-DQB1 alleles, and 36 distinct HLA-DRB1 alleles). For each racial background, HLA class I and class II haplotypes were inferred using Hapferret (hapferret, 2020), an implementation of the Expectation-Maximization algorithm (Excoffier and Slatkin, 1995). A total of 1,200, 779, and 440 class I and 920, 537, and 502 class II haplotype frequencies were derived in Black, White, and Asian populations, respectively.

Peptide Scoring

Computational Models for Candidate Peptide Selection

For a peptide vaccine to be effective, its constituent peptides need to be displayed, and thus, a computational vaccine design must be built upon a solid predictive foundation of what peptides will be displayed by each HLA allele. Incorrect predictions could lead to failure of a preclinical or clinical trial at great human cost. To this end, we were concerned about the precision (true positives divided by all positives) of our predictions such that we maximize the chance that a peptide predicted to be displayed would in fact be displayed. We were less concerned with our ability to recall all of the peptides that would be displayed, as long as we had a set of suitable size that would be displayed. We reduced the risk of false positives by employing multiple computational methods to predict peptide-HLA binding. For MHC class I vaccine design, we used an ensemble of methods, and for evaluation of MHC class I and class II vaccines we used all of the methods separately. See STAR Methods for details.

All models take a HLA-peptide pair as input and output predicted peptide-HLA binding affinity (IC50) in nanomolar units. For both MHC class I and class II models, we considered peptides to be binders if the predicted HLA binding affinity was $\leq$ 50 nM (Sette et al., 1994), providing a conservative threshold to increase the probability of peptide display. We validated our computational models on a dataset of SARS-CoV-2 peptides evaluated for stability (Prachar et al., 2020). We found that scoring peptides by predicted binding affinity maximized AUROC as compared with alternative scoring methods, and selecting peptides using a 50 nM binding affinity threshold maximized precision in classification of stable binders compared to alternative binding criteria (STAR Methods, Table S1). Our ensemble of NetMHCpan-4.0 and MHCflurry further increased AUROC and precision over individual predictors.

Our computational predictions of peptide display include supporting HLA alleles, thus enabling immunogenicity testing of peptides on HLA-matched individuals. When available, these data can be used to eliminate peptide support by particular HLA alleles when the peptides are found to be nonimmunogenic (Figure 1).

Combinatorial Vaccine Design (OptiVax) and Evaluation (EvalVax) Use Coverage-Based Objectives

Coverage optimization is performed by OptiVax using beam search to efficiently select an optimal subset of peptides that maximizes a desired population coverage objective. Starting from an empty set, it iteratively expands solutions in the beam by adding one peptide at a time and keeps the top k solutions

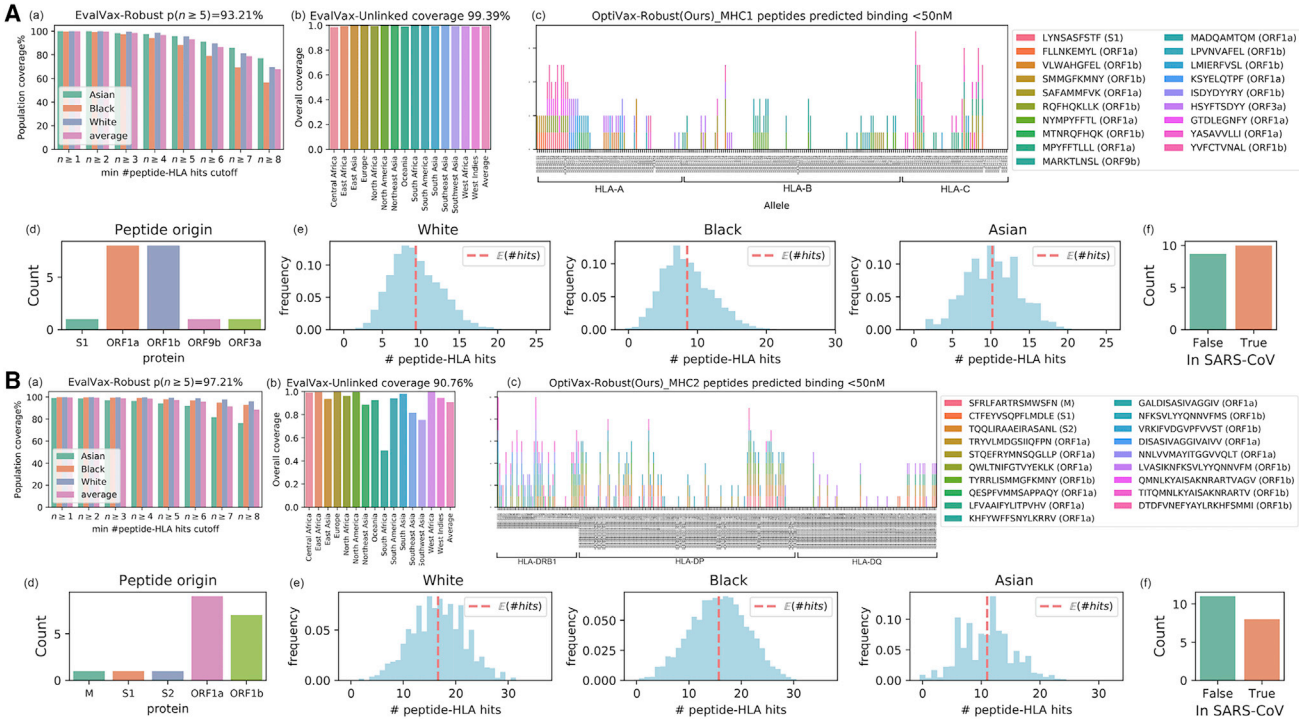


Figure 2. SARS-CoV-2 OptiVax-Robust Selected Peptide Vaccine Sets for (A) MHC Class I and (B) MHC Class II

(a) EvalVax-Robust population coverage at different per-individual number of peptide-HLA hit cutoffs for populations self-reporting as having White, Black, or Asian ancestry and average values.

(b) EvalVax-Unlinked population coverage on 15 geographic regions and averaged population coverage.

(c) Binding of vaccine peptides to each of the available alleles in MHC I and II.

(d) Peptide viral protein origins.

(e) Distribution of the number of per-individual peptide-HLA hits in populations self-reporting as having White, Black, or Asian ancestry.

(f) Vaccine peptide presence in SARS-CoV.

over all possible expansions in the beam. OptiVax-Robust uses the EvalVax-Robust objective function, which aims to find a minimal set of peptides that reaches a desired population coverage at a threshold of n predicted peptide-HLA hits per individual. EvalVax-Robust utilizes HLA haplotype frequencies for MHC class I (HLA-A, -B, -C) and MHC class II (HLA-DP, -DQ, -DR) genes. OptiVax-Unlinked uses the EvalVax-Unlinked objective function that considers HLA allele frequencies at each HLA locus independently and computes the likelihood that at least one peptide from a vaccine set is displayed at any locus. Both methods take into consideration HLA allele frequency, allelic zygosity, and for EvalVax-Robust, linkage disequilibrium (LD) among loci. OptiVax reduces vaccine peptide redundancy by not selecting peptides with closely related sequences for a vaccine formulation. EvalVax can be used independently to evaluate candidate peptide vaccine coverage metrics (Figure 1). We use a beam size of $k = 10$ for MHC class I and $k = 5$ for MHC class II. See STAR Methods for details.

OptiVax-Robust Optimization Results on MHC Class I and II

MHC Class I Results

We selected an optimized set of peptides from all SARS-CoV-2 proteins using the EvalVax-Robust objective function. We limited our candidates to peptides with length 8–10 and excluded peptides that have been observed with any mutation or are predicted to have non-zero probability of glycosylation. For computation of the objective function, we used the mean predicted IC50 values from our NetMHCpan-4.0 and MHCflurry ensemble to obtain reliable binding affinity predictions for evaluation and optimization. After all of our filtering steps, we had 378 candidate peptides. With OptiVax-Robust optimization, we designed a vaccine with 19 peptides that achieves 99.39% EvalVax-Unlinked coverage and 99.91% EvalVax-Robust coverage over three ethnic groups (Asian, Black, and White) with at least one peptide-HLA hit per individual. This set of peptides also provides 93.21% coverage with at least 5 peptide-HLA hits and 67.75% coverage with at least 8 peptide-HLA hits (Figure 2; Table 1). The population-level distribution of the number of peptide-HLA hits in White, Black, and Asian populations is shown in Figure 2, where the expected numbers of peptide-HLA hits are 9.358, 8.515, and 10.206, respectively.

MHC Class II Results

We limited our candidates to peptides with lengths of 13–25 and excluded peptides that have been observed with mutation probability greater than 0.001 or are predicted to have a non-zero glycosylation probability. We used the predicted binding affinity from NetMHCIIpan-4.0 for optimization and evaluation. After all of our

Table 1. Comparison of Baselines, S-protein Peptides, and OptiVax Designed Peptide Vaccines (Using All SARS-CoV-2 Proteins or SMN Proteins Only) on Various Population Coverage Evaluation Metrics and Vaccine Quality Metrics (Percentage of Peptides with Mutation rate > 0.001 or with Non-zero Probability of being Glycosylated)

Peptide Set	Vaccine Size	EvalVax-Unlinked	EvalVax-Robust $p(n \geq 1)$	EvalVax-Robust $p(n \geq 5)$	EvalVax-Robust $p(n \geq 8)$	Exp. # Peptide-HLA Hits/ Vaccine Size	Exp. # Peptide-HLA Hits (White)	Exp. # Peptide-HLA Hits (Black)	Exp. # Peptide-HLA Hits (Asian)	Peptides Glycosylated	Peptides Mutation Rate > 0.001	On Cleavage Site	Protein Origins	In SARS-CoV
MHC Class I Peptide Vaccine Evaluation														
OptiVax Augmented Nonredundant S-Protein	126 + 16	100.00%	100.00%	99.97%	99.27%	20.50%	27.20	27.68	32.44	0.00%	0.00%	0.00%	M, N, ORF1a, ORF1b, ORF3a, S1, S2	30.28%
S-Protein	3795	99.96%	100.00%	99.17%	98.29%	0.91%	30.84	32.14	41.13	15.57%	29.99%	0.63%	S1, S2	29.30%
OptiVax-Unlinked	19	99.79%	99.99%	89.15%	49.59%	40.72%	7.34	6.90	8.97	0.00%	0.00%	0.00%	ORF1a, ORF1b, ORF3a, S1	42.11%
Nonredundant S-protein	126	99.84%	99.93%	97.37%	91.69%	16.82%	19.20	19.99	24.38	0.00%	0.00%	0.00%	S1, S2	27.78%
OptiVax-Robust	19	99.39%	99.91%	93.21%	67.75%	49.26%	9.36	8.52	10.21	0.00%	0.00%	0.00%	ORF1a, ORF1b, ORF3a, ORF9b, S1	52.63%
OptiVax-Robust – size 15	15	99.07%	99.89%	86.69%	54.36%	54.47%	8.17	7.20	9.14	0.00%	0.00%	0.00%	ORF1a, ORF1b, ORF9b, S1	53.33%
Nonredundant S1-subunit	68	99.18%	99.76%	86.53%	56.36%	12.23%	8.31	8.84	7.80	0.00%	0.00%	0.00%	S1	8.82%
(Srivastava et al., 2020)	37	95.86%	99.75%	52.94%	16.00%	13.51%	5.37	4.99	4.64	8.11%	37.84%	0.00%	E, M, N, ORF10, ORF1a, ORF1b, ORF3a, ORF6, ORF7a, ORF7b, ORF8, S1	45.95%
OptiVax-Robust – S/M/N only	26	97.49%	98.15%	67.37%	26.24%	22.31%	5.31	5.64	6.45	0.00%	0.00%	0.00%	M, N, S1, S2	57.69%
(Herst et al., 2020)	52	90.89%	95.82%	56.52%	19.99%	9.88%	5.20	4.44	5.77	7.69%	34.62%	0.00%	N	55.77%
(Herst et al., 2020) – top 16	16	80.41%	93.46%	9.47%	0.03%	15.73%	2.75	2.60	2.20	12.50%	12.50%	0.00%	N	68.75%
Random subset of binders	19	81.04%	90.33%	25.02%	4.58%	16.74%	3.01	2.83	3.70	0.00%	29.89%	0.00%	N/A	40.37%
(Baruah and Bose, 2020)	5	71.91%	90.10%	0.55%	0.00%	33.60%	1.93	1.44	1.67	0.00%	40.00%	0.00%	S1, S2	40.00%
(Fast et al., 2020)	13	78.66%	85.29%	58.51%	30.56%	44.25%	5.59	4.98	6.69	7.69%	30.77%	0.00%	E, M, N, ORF1a, S1, S2	23.08%
(Poran et al., 2020)	10	69.12%	85.13%	3.21%	0.01%	19.23%	1.68	1.72	2.37	0.00%	30.00%	0.00%	ORF1a, ORF1b, ORF3a, ORF8, S1	20.00%

(*Continued on next page*)

Table 1. *Continued*

Peptide Set	Vaccine Size	EvalVax-Unlinked	EvalVax-Robust $p(n \geq 1)$	EvalVax-Robust $p(n \geq 5)$	EvalVax-Robust $p(n \geq 8)$	Exp. # Peptide-HLA Hits/ Vaccine Size	Exp. # Peptide-HLA Hits (White)	Exp. # Peptide-HLA Hits (Black)	Exp. # Peptide-HLA Hits (Asian)	Peptides Glycosylated	Peptides Mutation Rate > 0.001	On Cleavage Site	Protein Origins	In SARS-CoV
(Vashi et al., 2020)	51	68.63%	80.80%	1.52%	0.00%	3.12%	1.90	1.70	1.17	11.76%	43.14%	5.88%	S1, S2	5.88%
(Abdelmageed et al., 2020)	10	66.91%	78.49%	23.49%	2.72%	28.34%	2.93	2.50	3.07	10.00%	10.00%	0.00%	E	80.00%
(Lee and Koohy, 2020)	13	64.96%	75.75%	39.82%	37.09%	34.15%	4.77	3.69	4.86	0.00%	7.69%	0.00%	E, N, ORF1a, ORF1b, S2	53.85%
(Akhand et al., 2020)	31	49.46%	71.24%	0.08%	0.00%	3.47%	1.09	1.11	1.02	3.23%	35.48%	0.00%	E, M, N, S1	41.94%
(Singh et al., 2020)	7	53.91%	66.59%	1.38%	0.00%	19.87%	1.34	1.30	1.53	0.00%	28.57%	0.00%	E, M, N, S1, S2	71.43%
(Bhattacharya et al., 2020)	13	44.56%	61.09%	0.00%	0.00%	5.67%	0.79	0.69	0.73	23.08%	46.15%	7.69%	S1, S2	23.08%
(Ahmed et al., 2020)	16	45.25%	52.30%	35.61%	4.15%	15.57%	2.56	2.18	2.73	12.50%	25.00%	0.00%	N, S2	100.00%
(Saha and Prasad, 2020)	5	29.90%	41.77%	0.00%	0.00%	8.86%	0.56	0.36	0.41	0.00%	20.00%	0.00%	S1	20.00%
(Gupta et al., 2020)	7	30.23%	38.91%	21.08%	1.41%	23.92%	1.32	0.55	3.15	0.00%	42.86%	0.00%	S1, S2	14.29%
(Khan et al., 2020)	3	27.14%	34.98%	0.00%	0.00%	17.33%	0.76	0.56	0.24	0.00%	66.67%	0.00%	S1, S2	0.00%
(Mitra et al., 2020)	9	13.97%	23.86%	0.00%	0.00%	2.83%	0.15	0.08	0.54	22.22%	11.11%	0.00%	S1, S2	11.11%
MHC Class II Peptide Vaccine Evaluation														
OptiVax-Unlinked	19	91.67%	99.67%	95.94%	83.30%	64.45%	14.37	12.71	9.66	0.00%	0.00%	0.00%	M, ORF1a, ORF1b, S2	52.63%
OptiVax-Robust	19	90.76%	99.67%	97.21%	88.48%	76.04%	16.64	15.71	11.00	0.00%	0.00%	0.00%	M, ORF1a, ORF1b, S1, S2	42.11%
OptiVax Augmented Nonredundant S-protein	102 + 26	91.65%	99.67%	98.73%	97.27%	26.81%	43.79	36.06	23.12	0.00%	0.00%	0.00%	M, ORF1a, ORF1b, S1, S2	29.69%
(Ramaiah and Arumugaswami, 2020)	134	87.28%	98.88%	90.20%	83.97%	25.18%	45.04	38.25	17.93	20.15%	44.78%	0.00%	E, M, N, S1, S2	30.60%
S-protein	16315	89.80%	98.76%	95.99%	95.73%	2.22%	492.82	385.60	208.34	30.01%	57.50%	1.43%	S1, S2	16.06%

(*Continued on next page*)

Table 1. *Continued*

Peptide Set	Vaccine Size	EvalVax-Unlinked	EvalVax-Robust $p(n \geq 1)$	EvalVax-Robust $p(n \geq 5)$	EvalVax-Robust $p(n \geq 8)$	Exp. # Peptide-HLA Hits/ Vaccine Size	Exp. # Peptide-HLA Hits (White)	Exp. # Peptide-HLA Hits (Black)	Exp. # Peptide-HLA Hits (Asian)	Peptides Glycosylated	Peptides Mutation Rate > 0.001	On Cleavage Site	Protein Origins	In SARS-CoV
OptiVax-Robust – S/M/N only	22	86.34%	98.57%	85.37%	62.49%	42.51%	11.31	9.69	7.05	0.00%	0.00%	0.00%	M, N, S1, S2	36.36%
Nonredundant S-protein	102	84.91%	98.56%	82.72%	77.19%	16.61%	23.54	17.04	10.23	0.00%	0.00%	0.00%	S1, S2	28.43%
Nonredundant S1-subunit	53	77.14%	95.81%	63.43%	41.82%	16.33%	13.07	8.74	4.16	0.00%	0.00%	0.00%	S1	3.77%
Random subset of binders	19	72.41%	93.61%	58.67%	32.40%	31.59%	7.72	6.49	3.79	0.00%	63.79%	0.00%	N/A	23.55%
(Fast et al., 2020)	13	67.29%	86.99%	15.24%	3.69%	19.69%	3.65	2.26	1.77	30.77%	38.46%	0.00%	E, M, N, ORF1a, S1, S2	0.00%
(Banerjee et al., 2020)	9	56.73%	83.51%	12.49%	0.66%	26.65%	3.16	2.35	1.68	22.22%	44.44%	0.00%	S1, S2	55.56%
(Tahir ul Qamar et al., 2020)	11	39.44%	72.75%	0.27%	0.00%	11.62%	1.84	1.46	0.53	0.00%	72.73%	0.00%	E, M, N, ORF10, ORF6, ORF7a, ORF8	36.36%
(Poran et al., 2020)	10	42.30%	69.37%	0.00%	0.00%	9.83%	1.47	0.91	0.57	20.00%	90.00%	0.00%	ORF1a, ORF1b, ORF3a, S2	20.00%
(Akhand et al., 2020)	31	43.90%	60.45%	9.22%	1.01%	6.08%	2.53	2.54	0.59	3.23%	48.39%	0.00%	E, M, N, S1	29.03%
(Singh et al., 2020)	7	41.48%	56.29%	0.96%	0.00%	14.02%	1.44	1.11	0.39	0.00%	28.57%	0.00%	E, M, N, S1, S2	71.43%
(Ahmed et al., 2020)	5	27.69%	54.96%	0.00%	0.00%	13.08%	0.74	0.72	0.51	0.00%	20.00%	0.00%	N, S2	100.00%
(Mitra et al., 2020)	5	25.46%	47.92%	0.04%	0.00%	13.14%	0.90	0.58	0.49	60.00%	20.00%	0.00%	S1, S2	0.00%
(Vashi et al., 2020)	20	20.78%	35.12%	0.04%	0.00%	3.36%	0.96	0.62	0.44	15.00%	35.00%	5.00%	S1, S2	0.00%
(Abdelmageed et al., 2020)	10	19.15%	28.40%	0.96%	0.00%	4.79%	0.92	0.27	0.24	60.00%	70.00%	0.00%	E	30.00%
(Baruah and Bose, 2020)	3	0.00%	0.00%	0.00%	0.00%	0.00%	0.00	0.00	0.00	66.67%	100.00%	0.00%	S1	0.00%

S-protein includes all possible S-protein peptides of lengths 8–10 (MHC class I) and 13–25 (MHC class II). Nonredundant peptide sets are a result of OptiVax analysis of nonredundant displayed peptides. The table is sorted by EvalVax-Robust $p(n \geq 1)$. Random subsets are generated 200 times. The binders used for generating random subsets are defined as peptides that are predicted to bind with affinity $\leq$ 50 nM to more than 5 of the alleles.

CellPress

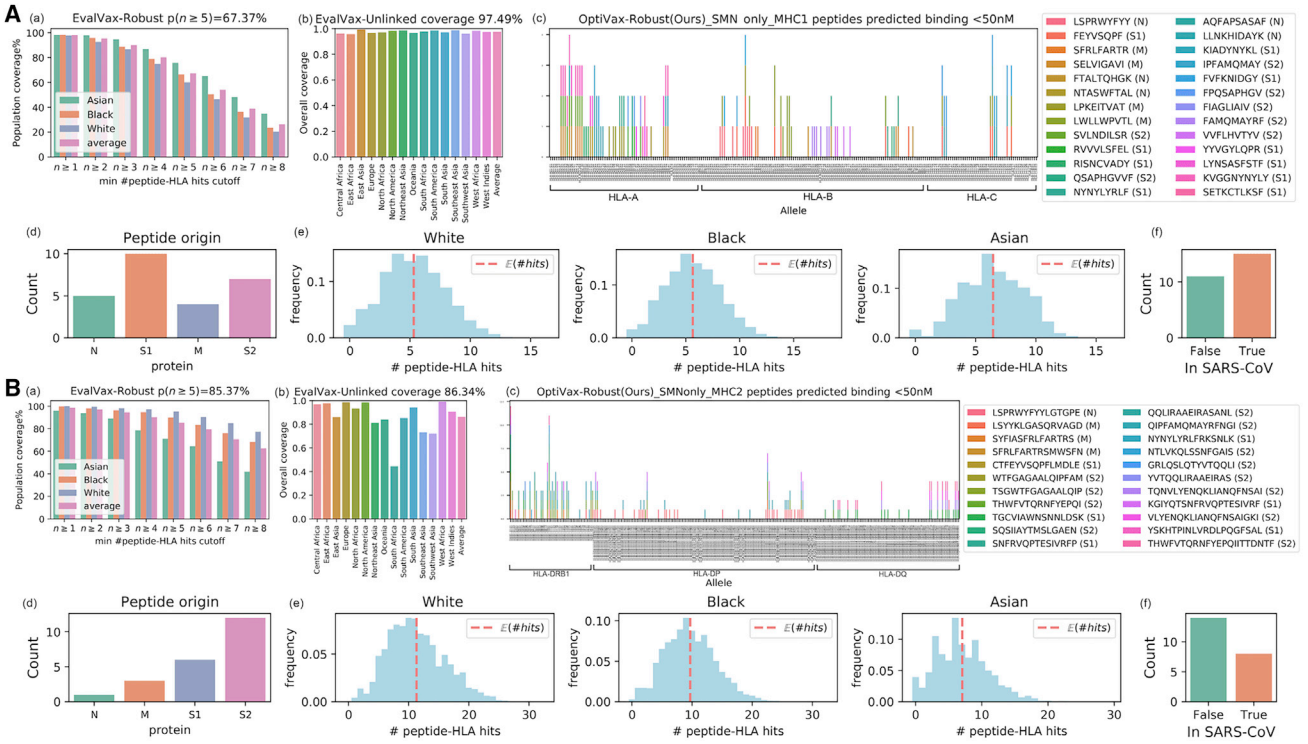


Figure 3. OptiVax-Robust-Designed Peptide Vaccine Using Peptides for (A) MHC class I and (B) MHC class II from the SARS-CoV-2 S, M, and N Proteins Only

(a) EvalVax-Robust population coverage at different minimum number of peptide-HLA hit cutoffs for populations self-reporting as having White, Black, or Asian ancestry and average values.

(b) EvalVax-Unlinked population coverage on 15 geographic regions and averaged population coverage.

(c) Binding of vaccine peptides to each of the available alleles in MHC I and II.

(d) Peptide viral protein origins.

(e) Distribution of the number of per-individual peptide-HLA hits in populations self-reporting as having White, Black, or Asian ancestry.

(f) Vaccine peptide presence in SARS-CoV.

filtering steps we had 7,977 candidate peptides. With OptiVax-Robust optimization, we designed a vaccine with 19 peptides that achieved 90.76% EvalVax-Unlinked coverage and 99.67% EvalVax-Robust coverage over three ethnic groups (Asian, Black, and White) with at least one peptide-HLA hit per individual. This set of peptides also provided 97.21% coverage with at least 5 peptide-HLA hits and 88.48% coverage with at least 8 peptide-HLA hits (Figure 2; Table 1). The population-level distribution of the number of peptide-HLA hits per individual in White, Black, and Asian populations is shown in Figure 2, where the expected peptide-HLA hits are 16.635, 15.708, and 11, respectively.

Designing Vaccines with S, M, N Proteins Only

We also used OptiVax-Robust to design vaccines for MHC class I and class II based solely upon peptides from the S, M, and N proteins of SARS-CoV-2 and evaluated vaccine performance. Grifoni et al. (2020b) found that peptides from the S, M, and N proteins produced the majority of the $CD4^+$ (86%) and $CD8^+$ (60%) T cell response in 20 convalescent COVID-19 patients. Since Grifoni et al. (2020b) used megapool-based assays, it is not possible to use their data to identify individual peptides that are immunogenic.

As shown in Table 1, our SMN-only MHC class I vaccine with 26 peptides achieves 98.15% coverage over three ethnic groups (Asian, Black, and White) with at least one average peptide-HLA hit per individual. There was an average of at least five peptide hits in 67.37% of the population, and the expected per-individual numbers of hits for White, Black, and Asian populations are 5.313, 5.643, and 6.448, respectively. The OptiVax-Robust MHC class II SMN-only vaccine with 22 peptides achieves 98.57% coverage with an average of at least one peptide-HLA hit per individual. There was an average of at least five peptide hits in 85.37% of the population, and the expected per-individual number of hits in White, Black, and Asian populations are 11.309, 9.693, and 7.053, respectively. The detailed vaccine designs are shown in Figure 3. We observed that it is more difficult to optimize vaccines with S, M, and N proteins only. We suspect this is because there are fewer candidate peptides to cover all of our haplotype combinations.

OptiVax-Unlinked Optimization Results on MHC Class I and II

MHC Class I Results

We limited our candidates to peptides with length 8–10 and zero predicted probability of glycosylation. We also excluded peptides that have been observed with any mutation. We used the mean predicted binding affinity values from our ensemble of

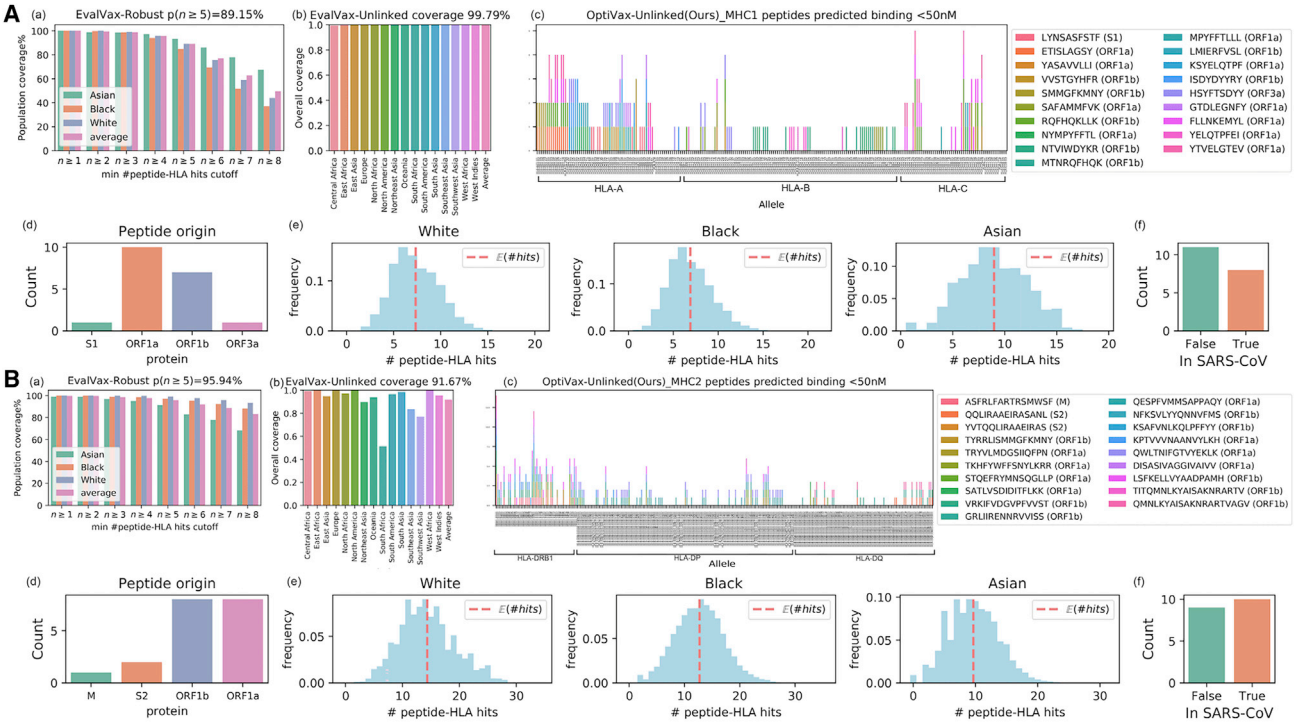


Figure 4. OptiVax-Unlinked Selected SARS-CoV-2 Optimal Peptide Vaccine Sets for (A) MHC Class I and (B) MHC Class II

(a) EvalVax-Robust population coverage at different per-individual numbers of peptide-HLA hits cutoffs for populations self-reporting as having White, Black, or Asian ancestry and average value.

(b) EvalVax-Unlinked population coverage on 15 geographic regions and averaged population coverage.

(c) Binding of vaccine peptides to each of the available alleles in MHC I and II.

(d) Peptide viral protein origins.

(e) Distribution of the number of per-individual peptide-HLA hits in populations self-reporting as having White, Black, or Asian ancestry.

(f) Vaccine peptide presence in SARS-CoV.

NetMHCpan-4.0 and MHCflurry on 2,392 HLA class I alleles to obtain reliable binding affinity predictions for evaluation and optimization. After all of our filtering steps, we had 472 candidate peptides. With OptiVax-Unlinked optimization, we designed a vaccine with 19 peptides that achieved 99.79% EvalVax-Unlinked population coverage (averaged over 15 geographic regions). As shown in Figure 4, the 19 vaccine peptides bind to a diverse range of alleles across the HLA-A, -B, -C loci. Even though less effective than OptiVax-Robust at providing a higher number of expected individual peptide-HLA hits in the population, the OptiVax-Unlinked peptide set still achieves high coverage on EvalVax-Robust metrics (99.99% for $p(n \geq 1)$, 89.15% for $p(n \geq 5)$, and 49.59% for $p(n \geq 8)$). The expected per-individual numbers of peptide-HLA hits for the design are 7.340, 6.899, and 8.971 for White, Black, and Asian populations, respectively (Table 1).

MHC Class II Results

We excluded peptides that have been observed with a mutation probability greater than 0.001 or are predicted to have non-zero probability of being glycosylated. We used the predicted binding affinity from NetMHCIIpan-4.0 for optimization and initial evaluation. After all of our filtering steps we had 7,966 candidate peptides. With OptiVax-Unlinked, we designed a vaccine with 19 peptides that achieved 91.67% EvalVax-Unlinked population coverage (averaged over 15 geographic regions). As shown in Figure 4, the 19 vaccine peptides bind to a diverse range of alleles across the HLA-DRB, -DP, and -DQ loci. Even though less effective than OptiVax-Robust in providing a high predicted number of average peptide-HLA hits in the population, the OptiVax-Unlinked peptide set still achieves high coverage on EvalVax-Robust metrics (99.67% for $p(n \geq 1)$, 95.94% for for $p(n \geq 5)$, and 83.30% for $p(n \geq 8)$). The expected per-individual number of peptide-HLA hits for the design is 14.366, 12.711, and 9.657 for White, Black, and Asian populations, respectively (Table 1).

EvalVax Evaluation of Public Vaccine Designs for SARS-CoV-2

We used EvalVax to evaluate peptide vaccines and mega-pools proposed by other publications (Lee and Koohy, 2020; Fast et al., 2020; Poran et al., 2020; Bhattacharya et al., 2020; Baruah and Bose, 2020; Abdelmageed et al., 2020; Ahmed et al., 2020; Srivastava et al., 2020; Herst et al., 2020; Vashi et al., 2020; Akhand et al., 2020; Mitra et al., 2020; Khan et al., 2020; Banerjee et al., 2020; Ramaiah and Arumugaswami, 2020; Gupta et al., 2020; Saha and Prasad, 2020; Tahir ul Qamar et al., 2020; Singh et al., 2020; Yarmarkovich et al., 2020; Grifoni et al., 2020a; Nerli and Sgourakis, 2020; Yazdani et al., 2020; Ismail et al., 2020) on metrics including EvalVax-Unlinked and EvalVax-Robust population coverage at different per-individual number of peptide-HLA

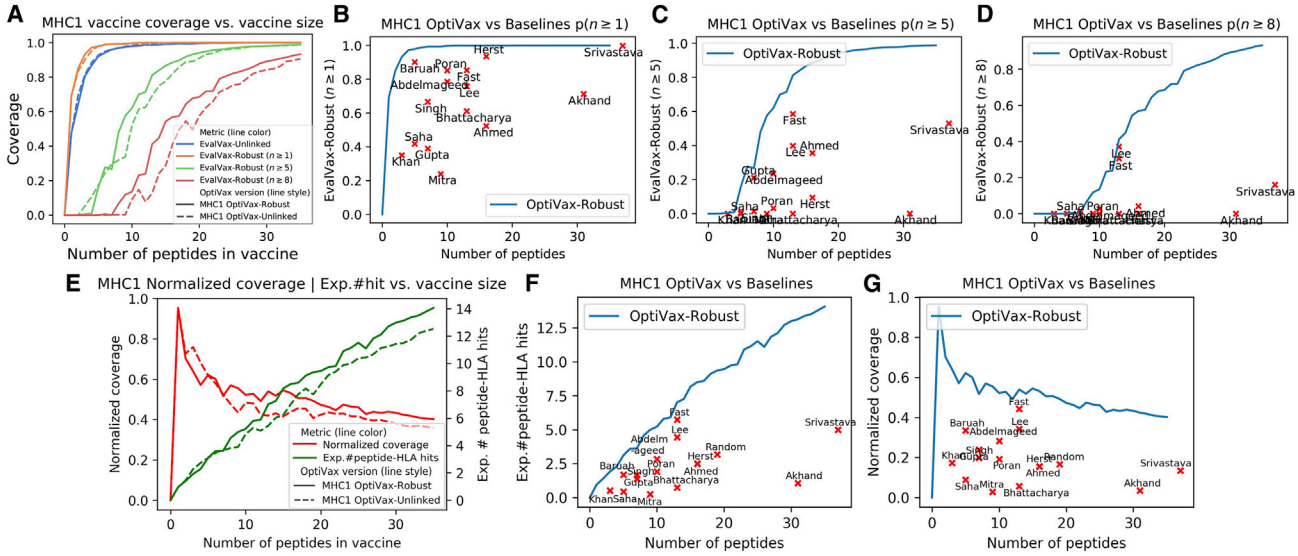


Figure 5. EvalVax Population Coverage Evaluation, Expectation of Per-Individual Number of Peptide-HLA Hits and Normalized Coverage for MHC Class I SARS-CoV-2 Vaccines

(A) EvalVax population coverage for OptiVax-Unlinked and OptiVax-Robust proposed vaccine at different vaccine sizes.

(B) EvalVax-Robust population coverage with $n \geq 1$ peptide-HLA hits per individual, OptiVax-Robust performance is shown by the blue curve and baseline performance is shown by red crosses (labeled by name of first author).

(C) EvalVax-Robust population coverage with $n \geq 5$ peptide-HLA hits.

(D) EvalVax-Robust population coverage with $n \geq 8$ peptide-HLA hits.

(E) Expected number of peptide-HLA hits vs. peptide vaccine size for OptiVax-Robust and OptiVax-Unlinked, and normalized coverage (hits divided by vaccine size) at different vaccine size.

(F) Comparison of OptiVax-Robust and baselines on expected number of peptide-HLA hits. OptiVax-Robust performance is shown by the blue curve and baseline performance is shown by red crosses.

(G) Comparison between OptiVax-Robust and baselines on normalized coverage.

hits thresholds, expected per-individual number of peptide-HLA hits in White, Black, and Asian populations, percentage of peptides that are predicted to be glycosylated, peptides observed to mutate with a probability greater than 0.001, or peptides that sit on known cleavage sites. We define ''normalized coverage'' as the mean expected per-individual number of peptide-HLA hits for a vaccine divided by the number of peptides in the vaccine.

We evaluated whole-protein vaccines by first converting them into the nonredundant peptides they display in a given haplotype population. Using a windowing strategy to enumerate all peptides in a whole-protein vaccine produces large numbers of overlapping redundant peptides that will cause EvalVax to provide optimistic and unrealistic vaccine metrics. We represented proteins as peptide vaccines by using OptiVax to create a vaccine design from the entire protein vaccine payload, without any limitations on the number of peptides in the vaccine. OptiVax eliminates highly redundant peptides during design and chooses the largest set of peptides that maximizes population coverage (STAR Methods). For example, EvalVax predicts SARS-CoV-2 S protein vaccines will have $n \geq 5$ MHC class II peptide hits in 95.99% of the population on average when simple windowing is employed resulting in 16,315 redundant peptides, and 82.72% of the population when nonredundant S is used, resulting in its representation as 102 peptides that are not glycosylated and have a mutation probability of $\leq$ 0.001 (Table 1).

Figures 5 and 6 show the comparison between OptiVax-Robust-designed MHC class I and class II vaccines at all vaccine sizes (top solution in the beam up to the given vaccine size) from 1–35 peptides (blue curves) and baseline vaccines (red crosses) proposed by other publications. We observed superior performance of OptiVax-Robust-designed vaccines on all evaluation metrics at all vaccine sizes for both MHC class I and class II. Most baselines achieved reasonable coverage at $n \geq 1$ peptide hits. However, many failed to show a high probability of higher hit counts, indicating a lack of predicted redundancy if a single peptide is not displayed. We also evaluated randomly selected peptide sets of size 19 from predicted binders of MHC class I and II, where a binder is defined as a peptide predicted to bind with $\leq$ 50 nM to more than 5 of the alleles in the MHC class. We found that a set of random binders can achieve greater coverage than some of the proposed vaccines we used as baselines.

Table 1 summarizes EvalVax results for all baselines with a vaccine peptide count of fewer than 150 peptides. We also evaluated an average of 200 random designs for MHC class I or class II containing 19 random peptides predicted to bind with $\leq$ 50 nM to more than 5 of the alleles in the MHC class. We found that the baseline methods all provide less coverage than OptiVax-derived sets, and some contain peptides predicted to be glycosylated or have a high observed mutation probability (Table 1). We also observed that some baselines

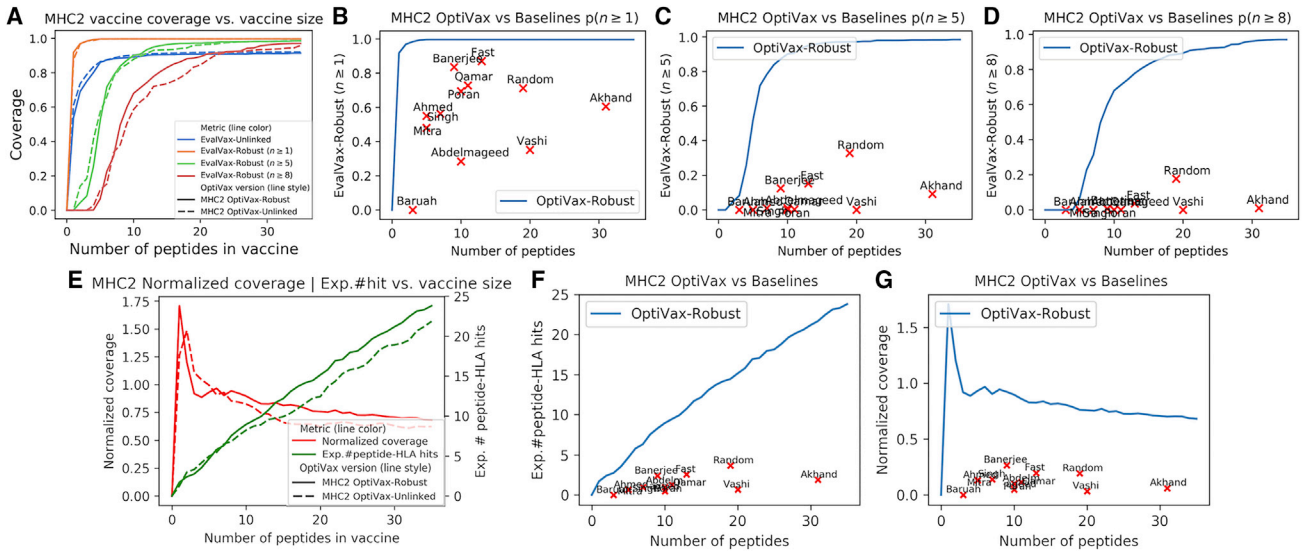


Figure 6. EvalVax Population Coverage Evaluation, Expectation of Per-Individual Number of Peptide-HLA Hits and Normalized Coverage for MHC Class II SARS-CoV-2 Vaccines

(A) EvalVax population coverage for OptiVax-Unlinked and OptiVax-Robust proposed vaccine at different vaccine sizes.

(B) EvalVax-Robust population coverage with $n \geq 1$ peptide-HLA hits per individual, OptiVax-Robust performance is shown by the blue curve and baseline performance is shown by red crosses (labeled by name of first author).

(C) EvalVax-Robust population coverage with $n \geq 5$ peptide-HLA hits.

(D) EvalVax-Robust population coverage with $n \geq 8$ peptide-HLA hits.

(E) Expected number of peptide-HLA hits versus peptide vaccine size for OptiVax-Robust and OptiVax-Unlinked, and normalized coverage (hits divided by vaccine size) at different vaccine size.

(F) Comparison of OptiVax-Robust and baselines on expected number of peptide-HLA hits. OptiVax-Robust performance is shown by the blue curve and baseline performance is shown by red crosses.

(G) Comparison between OptiVax-Robust and baselines on normalized coverage.

contained peptides that sat on cleavage sites or overlapped with self-peptides.

OptiVax Augmentation of SARS-CoV-2 S Protein Vaccines

When predicted population coverage for a whole-protein vaccine is judged to be insufficient, OptiVax can perform optimized, augmented vaccine design to suggest additional peptides to add to an existing formulation. In this mode, we used OptiVax to compute the nonredundant displayed peptide set for a protein vaccine and then used this as the initial set of peptides for OptiVax design. OptiVax then added supplemental nonredundant peptides to this initial set to improve population coverage. For example, we used OptiVax augmentation to add 26 peptides to the SARS-CoV-2 S protein vaccine to increase the predicted MHC class II population coverage for $n \geq 5$ peptide hits from 82.72% to 98.73%. For MHC class I, OptiVax augmentation added 16 peptides to the SARS-CoV-2 S protein vaccine to increase the predicted population coverage for $n \geq 5$ peptide hits from 97.4% to 99.9% (Table 1). OptiVax-derived vaccine designs, nonredundant peptide sets, and vaccine augmentations are presented in Table S3.

EvalVax Results Are Robust to Different Binding Prediction Models

We evaluated all Table 1 vaccine designs using eleven independent peptide-HLA binding prediction methods to ensure that the performance observed in Table 1 was consistent across prediction methods. For MHC class I prediction, we validated using seven methods: NetMHCpan-4.0, NetMHCpan-4.1, MHCflurry 1.6.0, PUFFIN, the mean of NetMHCpan-4.0 and MHCflurry 1.6.0 with a 50 nM cutoff on predicted affinity, and NetMHCpan-4.0 and NetMHCpan-4.1 with a 99.5% cutoff on EL ranking. For MHC class II prediction, we used four different methods for validation: NetMHCIIpan-3.2 and NetMHCIIpan-4.0, each with either a 50 nM cutoff on predicted affinity or a 98% cutoff on EL ranking. The result of all eleven EvalVax evaluation metrics for all Table 1 designs are shown in Table S2. We found that these results from all eleven evaluation methods show Table 1 contains conservative estimates of vaccine performance.

DISCUSSION

The computational design of peptide vaccines for eliciting cellular immunity is built upon the imperfect science of predicting peptide presentation by HLA molecules as a precondition for their immunogenicity. Peptide vaccine designs also need to ensure that individuals with rare HLA alleles display vaccine peptides to ensure a high rate of vaccine efficacy over the entire population.

To mitigate computational model uncertainty, we have taken a very conservative view of peptide presentation, emphasizing precision over recall. To provide coverage for individuals with rare HLA types, we use haplotype frequencies that include these

types in our evaluations. We provide an evaluation tool, EvalVax, to permit the flexible analysis of vaccine proposals on key metrics, including population coverage and the expected number of peptides displayed. Not surprisingly, OptiVax vaccine designs that are optimized with respect to EvalVax objective functions do well on the same metrics. We also find that OptiVax designs do well when evaluated on eleven computational models of peptide-HLA binding, providing encouragement that their component peptides will be displayed.

The immunogenicity of HLA displayed peptides likely varies between individuals (Croft et al., 2019), highlighting the desirability of a vaccine recipient displaying multiple vaccine peptides to increase the probability of engendering a durable immune response. EvalVax-Robust's prediction of the expected number of peptide hits for each individual provides one metric of this property, and the probability of a desired number of individual hits is optimized by OptiVax-Robust.

EvalVax can be used for vaccine designs that are focused on the expression of entire viral proteins or their subunits to evaluate the level of viral peptide-HLA presentation that is predicted to result. We note for SARS-CoV-2 in Table 1 that the S protein is limited in its predicted ability to provide robust population coverage for MHC class II display of more than four viral epitopes. This suggests that vaccines that only employ the S protein may require additional peptide components for reliable $CD4^+$ T cell activation across the entire population, and we have introduced specific augmentation methods for this purpose.

At present the World Health Organization lists 79 COVID-19 vaccine candidates in clinical or preclinical evaluation (WHO, 2020) (accessed May 16, 2020), and the precise designs of most of these vaccines are not public. We encourage the early publication of vaccine designs to enable collaboration and rapid progress toward safe and effective vaccines for COVID-19.

All of our software and data are freely available as open source to allow others to use and extend our methods.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- RESOURCE AVAILABILITY
  - Lead Contact
  - Materials Availability
  - Data and Code Availability
- METHOD DETAILS
  - Peptide filtering
  - EvalVax
  - OptiVax
  - Computational Peptide-HLA Prediction Models
  - Criteria for Predicted Binding
- QUANTIFICATION AND STATISTICAL ANALYSIS

SUPPLEMENTAL INFORMATION

Supplemental Information can be found online at https://doi.org/10.1016/j.cels.2020.06.009.

ACKNOWLEDGEMENTS

Michael Birnbaum, Brooke Huisman, and Jonathan Krog provided helpful discussions. Viral sequences are from GISAID (see Table S4: GISAID acknowledgements). This work was supported in part by Schmidt Futures and NIH grant R01CA218094 to D.K.G.

HLA haplotype frequencies were generated from previously published NGS data generated in the Carrington lab and we thank our many collaborators. This project was funded in part with federal funds from the Frederick National Laboratory for Cancer Research, under contract no. HHSN261200800001E. The content of this publication does not necessarily reflect the views or policies of the Department of Health and Human Services, nor does mention of trade names, commercial products, or organizations imply endorsement by the U.S. Government. This Research was supported in part by the Intramural Research Program of the NIH, Frederick National Lab, Center for Cancer Research. The views expressed in this article do not necessarily reflect the official policy or position of the National Institutes of Health, the Department of the Navy, the Department of Defense, or any other agency of the US Government.

AUTHOR CONTRIBUTIONS

G.L., B.C., and D.G. contributed to problem definition and solution; G.L. designed and implemented the optimization procedure, with advice from B.C. and D.G.; T.B. analyzed viral genome mutation rates, and S.J. analyzed self peptides. M.V. and M.C. analyzed and provided haplotype data; G.L., B.C., and D.G. wrote the paper.

DECLARATION OF INTERESTS

David Gifford is a Founder and shareholder of ThinkTx.

Received: May 15, 2020
Revised: May 31, 2020
Accepted: June 18, 2020
Published: July 27, 2020

REFERENCES

Abdelmageed, M.I., Abdelmoneim, A.H., Mustafa, M.I., Elfadol, N.M., Murshed, N.S., Shantier, S.W., and Makhawi, A.M. (2020). Design of multi epitope-based peptide vaccine against E protein of human 2019-nCoV: an immunoinformatics approach. bioRxiv. https://doi.org/10.1101/2020.02.04.934232.

Ahmed, S.F., Quadeer, A.A., and McKay, M.R. (2020). Preliminary identification of potential vaccine targets for the COVID-19 coronavirus (SARS-CoV-2) based on SARS-CoV immunological studies. Viruses *12*, 254.

Akhand, M.R.N., Azim, K.F., Hoque, S.F., Moli, M.A., Joy, B.D., Akter, H., Afif, I.K., Ahmed, N., and Hasan, M. (2020). Genome based evolutionary study of SARS-CoV-2 towards the prediction of epitope based chimeric vaccine. bioRxiv. https://doi.org/10.1101/2020.04.15.036285.

Banerjee, A., Santra, D., and Maiti, S. (2020). Energetics based epitope screening in SARS CoV-2 (COVID 19) spike glycoprotein by immuno-informatic analysis aiming to a suitable vaccine development. bioRxiv. https://doi.org/10.1101/2020.04.02.021725.

Baruah, V., and Bose, S. (2020). Immunoinformatics-aided identification of T cell and B cell epitopes in the surface glycoprotein of 2019-nCoV. J. Med. Virol. *92*, 495–500.

Bhattacharya, M., Sharma, A.R., Patra, P., Ghosh, P., Sharma, G., Patra, B.C., Lee, S.S., and Chakraborty, C. (2020). Development of epitope-based peptide vaccine against novel coronavirus 2019 (SARS-COV-2): immunoinformatics approach. J. Med. Virol. *92*, 618–631.

Bui, H.H., Sidney, J., Dinh, K., Southwood, S., Newman, M.J., and Sette, A. (2006). Predicting population coverage of T-cell epitope-based diagnostics and vaccines. BMC Bioinformatics *7*, 153.

Buus, S., Lauemøller, S.L., Worning, P., Kesmir, C., Frimurer, T., Corbet, S., Fomsgaard, A., Hilden, J., Holm, A., and Brunak, S. (2003). Sensitive

quantitative predictions of peptide-MHC binding by a 'Query by Committee' artificial neural network approach. Tissue Antigens *62*, 378–384.

Chicz, R.M., Urban, R.G., Lane, W.S., Gorga, J.C., Stern, L.J., Vignali, D.A., and Strominger, J.L. (1992). Predominant naturally processed peptides bound to HLA-DR1 are derived from MHC-related molecules and are heterogeneous in size. Nature *358*, 764–768.

UniProt Consortium (2019). UniProt: a worldwide hub of protein knowledge. Nucleic Acids Res. *47*, D506–D515.

Coutard, B., Valle, C., de Lamballerie, X., Canard, B., Seidah, N.G., and Decroly, E. (2020). The spike glycoprotein of the new coronavirus 2019-nCoV contains a furin-like cleavage site absent in CoV of the same clade. Antiviral Res. *176*, 104742.

Croft, N.P., Smith, S.A., Pickering, J., Sidney, J., Peters, B., Faridi, P., Witney, M.J., Sebastian, P., Flesch, I.E.A., Heading, S.L., et al. (2019). Most viral peptides displayed by class I MHC on infected cells are immunogenic. Proc. Natl. Acad. Sci. USA *116*, 3112–3117.

Elbe, S., and Buckland-Merrett, G. (2017). Data, disease and diplomacy: GISAID's innovative contribution to global health. Global Chall. *1*, 33–46.

Excoffier, L., and Slatkin, M. (1995). Maximum-likelihood estimation of molecular haplotype frequencies in a diploid population. Mol. Biol. Evol. *12*, 921–927.

Fast, E., Altman, R.B., and Chen, B. (2020). Potential T-cell and B-cell epitopes of 2019-nCoV. bioRxiv. https://doi.org/10.1101/2020.02.19.955484.

Finkel, Y., Mizrahi, O., Nachshon, A., Weingarten-Gabbay, S., Yahalom-Ronen, Y., Tamir, H., Achdout, H., Melamed, S., Weiss, S., Isrealy, T., et al. (2020). The coding capacity of SARS-CoV-2. bioRxiv. https://doi.org/10.1101/2020.05.07.082909.

Grifoni, A., Sidney, J., Zhang, Y., Scheuermann, R.H., Peters, B., and Sette, A. (2020a). A sequence homology and bioinformatic approach can predict candidate targets for immune responses to SARS-CoV-2. Cell Host Microbe *27*, 671–680.e2.

Grifoni, A., Weiskopf, D., Ramirez, S., Mateus, J., Dan, J., Moderbacher, C., Rawlings, S., Sutherland, A., Premkumar, L., Jadi, R., et al. (2020b). Targets of T cell responses to SARS-CoV-2 coronavirus in humans with COVID-19 disease and unexposed individuals. Cell *181*, 1489–1501.

Gupta, E., Mishra, R.K., and Niraj, R.R.K. (2020). Identification of potential vaccine candidates against SARS-CoV-2, a step forward to fight novel coronavirus 2019-nCoV: A reverse vaccinology approach. bioRxiv. https://doi.org/10.1101/2020.04.13.039198.

Gupta, R., Jung, E., and Brunak, S. (2004). Prediction of N-glycosylation sites in human proteins. http://www.cbs.dtu.dk/services/NetNGlyc/.

Hadfield, J., Megill, C., Bell, S.M., Huddleston, J., Potter, B., Callender, C., Sagulenko, P., Bedford, T., and Neher, R.A. (2018). Nextstrain: real-time tracking of pathogen evolution. Bioinformatics *34*, 4121–4123.

hapferret (2020). hapferret. https://github.com/nilsboar/hapferret.

Helmberg, W., Dunivin, R., and Feolo, M. (2004). The sequencing-based typing tool of dbMHC: typing highly polymorphic gene sequences. Nucleic Acids Res. *32*, W173–W175.

Herst, C.V., Burkholz, S., Sidney, J., Sette, A., Harris, P.E., Massey, S., Brasel, T., Cunha-Neto, E., Rosa, D.S., Chao, W.C.H., et al. (2020). An effective CTL peptide vaccine for Ebola zaire based on survivors' CD8+ targeting of a particular nucleocapsid protein epitope with potential implications for COVID-19 vaccine design. Vaccine *38*, 4464–4475.

Ismail, S., Ahmad, S., and Azam, S.S. (2020). Immuno-informatics characterization sars-cov-2 spike glycoprotein for prioritization of epitope based multivalent peptide vaccine. bioRxiv. https://doi.org/10.1101/2020.04.05.026005.

Jensen, K.K., Andreatta, M., Marcatili, P., Buus, S., Greenbaum, J.A., Yan, Z., Sette, A., Peters, B., and Nielsen, M. (2018). Improved methods for predicting peptide binding affinity to MHC class II molecules. Immunology *154*, 394–406.

Jurtz, V., Paul, S., Andreatta, M., Marcatili, P., Peters, B., and Nielsen, M. (2017). NetMHCpan-4.0: improved peptide–MHC class I interaction predictions integrating eluted ligand and peptide binding affinity data. J. Immunol. *199*, 3360–3368.

Khan, A., Alam, A., Imam, N., Siddiqui, M.F., and Ishrat, R. (2020). Design of an epitope-based peptide vaccine against the severe acute respiratory syndrome Coronavirus-2 (SARS-CoV-2): A vaccine informatics approach. bioRxiv. https://doi.org/10.1101/2020.05.03.074930.

Lee, C.H.J., and Koohy, H. (2020). In silico identification of vaccine targets for 2019-nCoV. F1000Res. *9*, 145.

Mitra, D., Shekhar, N., Pandey, J., Jain, A., and Swaroop, S. (2020). Multi-epitope based peptide vaccine design against SARS-CoV-2 using its spike protein. bioRxiv. https://doi.org/10.1101/2020.04.23.055467.

Moise, L., Gutierrez, A., Kibria, F., Martin, R., Tassone, R., Liu, R., Terry, F., Martin, B., and De Groot, A.S. (2015). iVAX: an integrated toolkit for the selection and optimization of antigens and the design of epitope-driven vaccines. Hum. Vaccin. Immunother. *11*, 2312–2321.

Nerli, S., and Sgourakis, N.G. (2020). Structure-based modeling of sars-cov-2 peptide/hla-a02 antigens. bioRxiv. https://doi.org/10.1101/2020.03.23.004176.

Nielsen, M., Lundegaard, C., Worning, P., Lauemøller, S.L., Lamberth, K., Buus, S., Brunak, S., and Lund, O. (2003). Reliable prediction of T-cell epitopes using neural networks with novel sequence representations. Protein Sci. *12*, 1007–1017.

O'Donnell, T., Rubinsteyn, A., and Laserson, U. (2020). A model of antigen processing improves prediction of MHC I-presented peptides. bioRxiv. https://doi.org/10.1101/2020.03.28.013714.

O'Donnell, T.J., Rubinsteyn, A., Bonsack, M., Riemer, A.B., Laserson, U., and Hammerbacher, J. (2018). MHCflurry: open-source class I MHC binding affinity prediction. Cell Syst. *7*, 129–132.e4.

World Health Organization. (2020). DRAFT landscape of COVID-19 candidate vaccines. https://www.who.int/blueprint/priority-diseases/key-action/novel-coronavirus-landscape-ncov.pdf.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., et al. (2011). Scikit-learn: machine learning in Python. J. Mach. Learn. Res. *12*, 2825–2830.

Peters, B., Nielsen, M., and Sette, A. (2020). T cell epitope predictions. Annu. Rev. Immunol. *38*, 123–145.

Poran, A., Harjanto, D., Malloy, M., Rooney, M.S., Srinivasan, L., and Gaynor, R.B. (2020). Sequence-based prediction of vaccine targets for inducing T cell responses to SARS-CoV-2 utilizing the bioinformatics predictor RECON. bioRxiv. https://doi.org/10.1101/2020.04.06.027805.

Prachar, M., Justesen, S., Steen-Jensen, D.B., Thorgrimsen, S.P., Jurgons, E., Winther, O., and Bagger, F.O. (2020). COVID-19 vaccine candidates: prediction and validation of 174 SARS-CoV-2 epitopes. bioRxiv. https://doi.org/10.1101/2020.03.20.000794.

Ramaiah, A., and Arumugaswami, V. (2020). Insights into cross-species evolution of novel human coronavirus 2019-nCoV and defining immune determinants for vaccine development. bioRxiv. https://doi.org/10.1101/2020.01.29.925867.

Reynisson, B., Alvarez, B., Paul, S., Peters, B., and Nielsen, M. (2020a). NetMHCpan-4.1 and NetMHCIIpan-4.0: improved predictions of MHC antigen presentation by concurrent motif deconvolution and integration of MS MHC eluted ligand data. Nucleic Acids Res. *48*, W449–W454.

Reynisson, B., Barra, C., Kaabinejadian, S., Hildebrand, W.H., Peters, B., and Nielsen, M. (2020b). Improved prediction of MHC II antigen presentation through integration and motif deconvolution of mass spectrometry MHC eluted ligand data. J. Proteome Res. *19*, 2304–2315.

Rist, M.J., Theodossis, A., Croft, N.P., Neller, M.A., Welland, A., Chen, Z., Sullivan, L.C., Burrows, J.M., Miles, J.J., Brennan, R.M., et al. (2013). HLA peptide length preferences control CD8+ T cell responses. J. Immunol. *191*, 561–571.

Saha, R., and Prasad, B.V. (2020). In silico approach for designing of a multi-epitope based vaccine against novel coronavirus (SARS-COV-2). bioRxiv. https://doi.org/10.1101/2020.03.31.017459.

Sette, A., Vitiello, A., Reherman, B., Fowler, P., Nayersina, R., Kast, W.M., Melief, C.J., Oseroff, C., Yuan, L., Ruppert, J., et al. (1994). The relationship between class I binding affinity and immunogenicity of potential cytotoxic T cell epitopes. J. Immunol. *153*, 5586–5592.

Singh, A., Thakur, M., Sharma, L.K., and Chandra, K. (2020). Designing a multi-epitope peptide-based vaccine against SARS-CoV-2. bioRxiv. https://doi.org/10.1101/2020.04.15.040618.

Srivastava, S., Verma, S., Kamthania, M., Kaur, R., Badyal, R.K., Saxena, A.K., Shin, H.J., Kolbe, M., and Pandey, K. (2020). Structural basis to design multi-epitope vaccines against Novel coronavirus 19 (COVID19) infection, the ongoing pandemic emergency: an in silico approach. bioRxiv. https://doi.org/10.1101/2020.04.01.019299.

Tahir ul Qamar, M., Rehman, A., Ashfaq, U.A., Awan, M.Q., Fatima, I., Shahid, F., and Chen, L.L. (2020). Designing of a next generation multiepitope based vaccine (MEV) against SARS-COV-2: immunoinformatics and in silico approaches. bioRxiv. https://doi.org/10.1101/2020.02.28.970343.

Trolle, T., McMurtrey, C.P., Sidney, J., Bardet, W., Osborn, S.C., Kaever, T., Sette, A., Hildebrand, W.H., Nielsen, M., and Peters, B. (2016). The length distribution of class I–restricted T cell epitopes is determined by both peptide supply and MHC allele–specific binding preference. J. Immunol. *196*, 1480–1487.

Vashi, Y., Jagrit, V., and Kumar, S. (2020). Understanding the B and T cells epitopes of spike protein of severe respiratory syndrome coronavirus-2: a computational way to predict the immunogens. bioRxiv. https://doi.org/10.1101/2020.04.08.013516.

Walls, A.C., Park, Y.J., Tortorici, M.A., Wall, A., McGuire, A.T., and Veesler, D. (2020). Structure, function, and antigenicity of the SARS-CoV-2 spike glycoprotein. Cell *181*, 281–292.e6.

Wang, Q., Qiu, Y., Li, J.Y., Zhou, Z.J., Liao, C.H., and Ge, X.Y. (2020). A unique protease cleavage site predicted in the spike protein of the novel pneumonia coronavirus (2019-nCoV) potentially related to viral transmissibility. Virol. Sin. 1–3.

Wolfert, M.A., and Boons, G.J. (2013). Adaptive immune activation: glycosylation does matter. Nat. Chem. Biol. *9*, 776–784.

Wrapp, D., Wang, N., Corbett, K.S., Goldsmith, J.A., Hsieh, C.L., Abiona, O., Graham, B.S., and McLellan, J.S. (2020). Cryo-EM structure of the 2019-nCoV spike in the prefusion conformation. Science *367*, 1260–1263.

Yarmarkovich, M., Warrington, J.M., Farrel, A., and Maris, J.M. (2020). Identification of SARS-CoV-2 vaccine epitopes predicted to induce long-term population-scale immunity. Cell. Rep. Med. *1*, 100036.

Yazdani, Z., Rafiei, A., Yazdani, M., and Valadan, R. (2020). Design an efficient multi-epitope peptide vaccine candidate against sars-cov-2: an in silico analysis. bioRxiv. https://doi.org/10.1101/2020.04.20.051557.

Zeng, H., and Gifford, D.K. (2019). Quantification of uncertainty in peptide-MHC binding prediction improves high-affinity peptide selection for therapeutic design. Cell Syst *9*, 159–166.e3.

Zhang, L., Lin, D., Sun, X., Curth, U., Drosten, C., Sauerhering, L., Becker, S., Rox, K., and Hilgenfeld, R. (2020a). Crystal structure of SARS-CoV-2 main protease provides a basis for design of improved α-ketoamide inhibitors. Science *368*, 409–412.

Zhang, Y., Zhao, W., Mao, Y., Wang, S., Zhong, Y., Su, T., Gong, M., Lu, X., Cheng, J., and Yang, H. (2020b). Site-specific N-glycosylation characterization of recombinant SARS-CoV-2 spike proteins using high-resolution mass spectrometry. bioRxiv. https://doi.org/10.1101/2020.03.28.013276.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Deposited Data		
HLA haplotype population frequency data	This paper, Mendeley Data	Mendeley Data: https://doi.org/10.17632/cfxkfy9zp4.1
SARS-CoV-2 vaccine designs	This paper	Table S3
SARS-CoV-2 proteome	GISAID (Elbe and Buckland-Merrett, 2017)	Sequence entry Wuhan/IPBCAMS-WH-01/2019
SARS-CoV proteome	UniProt (Consortium, 2019)	UniProt: UP000000354 (Proteome ID)
HLA population frequency data	dbMHC, as obtained from the IEDB Population Coverage Tool download (Bui et al., 2006)	http://tools.iedb.org/population/
Human proteome	UniProt (Consortium, 2019)	UniProt: UP000005640 (Proteome ID)
SARS-CoV-2 experimental peptide stability data (Immunitrack)	(Prachar et al., 2020)	Data S1. COVID19-Intavis-Immunitrack-dataset: https://www.immunitrack.com/wp/wp-content/uploads/Covid19-Intavis-Immunitrack-datasetV2.xlsx
SARS-CoV-2 cleavage regions: ORF1a and ORF1b	UniProt (Consortium, 2019)	UniProt: P0DTD1; https://covid-19.uniprot.org/uniprotkb/P0DTD1#Protein%20Processing
SARS-CoV-2 cleavage regions: Spike (S)	(Wang et al., 2020)	Figure 1
Additional SARS-CoV-2 proteomes for mutation analysis	GISAID (Elbe and Buckland-Merrett, 2017)	Acknowledgements and detailed GISAID accessions in Table S4 (this paper)
Experimental data of Spike N-glycosylation: Cryo-EM	(Walls et al., 2020)	Table 2
Experimental data of Spike N-glycosylation: tandem mass spectrometry	(Zhang et al., 2020b)	https://www.biorxiv.org/content/10.1101/2020.03.28.013276v1.upplementary-material, Figures S5A and S5B
Baseline vaccine MHC I: (Srivastava et al., 2020)	(Srivastava et al., 2020)	Figure 2
Baseline vaccine MHC I: (Herst et al., 2020)	(Herst et al., 2020)	Data S1; Table 6
Baseline vaccine MHC I: (Herst et al., 2020)-top16	(Herst et al., 2020)	Table 4
Baseline vaccine MHC I: (Baruah and Bose, 2020)	(Baruah and Bose, 2020)	Table 1
Baseline vaccine MHC I: (Fast et al., 2020)	(Fast et al., 2020)	Table 2
Baseline vaccine MHC I: (Poran et al., 2020)	(Poran et al., 2020)	Table S5
Baseline vaccine MHC I: (Vashi et al., 2020)	(Vashi et al., 2020)	Table 5
Baseline vaccine MHC I: (Abdelmageed et al., 2020)	(Abdelmageed et al., 2020)	Table 2
Baseline vaccine MHC I: (Lee and Koohy, 2020)	(Lee and Koohy, 2020)	Table 4
Baseline vaccine MHC I: (Akhand et al., 2020)	(Akhand et al., 2020)	Table 2
Baseline vaccine MHC I: (Singh et al., 2020)	(Singh et al., 2020)	Table 2
Baseline vaccine MHC I: (Bhattacharya et al., 2020)	(Bhattacharya et al., 2020)	Table 2
Baseline vaccine MHC I: (Ahmed et al., 2020)	(Ahmed et al., 2020)	Table 2
Baseline vaccine MHC I: (Saha and Prasad, 2020)	(Saha and Prasad, 2020)	Table 1
Baseline vaccine MHC I: (Gupta et al., 2020)	(Gupta et al., 2020)	Table 4a
Baseline vaccine MHC I: (Khan et al., 2020)	(Khan et al., 2020)	Sub-Section 2 in Results of the Main Text
Baseline vaccine MHC I: (Mitra et al., 2020)	(Mitra et al., 2020)	Table 1(C)
Baseline vaccine MHC I: (Nerli and Sgourakis, 2020)	(Nerli and Sgourakis, 2020)	Tables S1 and S2

(*Continued on next page*)

Continued

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Baseline vaccine MHC I: (Yarmarkovich et al., 2020)	(Yarmarkovich et al., 2020)	Table S5 (65 33-mers, to which we applied sliding windows of lengths 8–10 to obtain the peptide set considered for MHC class I)
Baseline vaccine MHC I: (Yazdani et al., 2020)	(Yazdani et al., 2020)	Table 1 (peptides created by sliding windows of length 8–10)
Baseline vaccine MHC I: (Ismail et al., 2020)	(Ismail et al., 2020)	Section 3.2 "MEPVC Designing" (peptides created by sliding windows of length 8–10)
Baseline vaccine MHC II: (Ramaiah and Arumugaswami, 2020)	(Ramaiah and Arumugaswami, 2020)	Table S2 - "Unique Mean HBA T-Cell Epitopes" for each protein Subunit
Baseline vaccine MHC II: (Fast et al., 2020)	(Fast et al., 2020)	Table 2
Baseline vaccine MHC II: (Banerjee et al., 2020)	(Banerjee et al., 2020)	Table 3
Baseline vaccine MHC II: (Akhand et al., 2020)	(Akhand et al., 2020)	Table 2
Baseline vaccine MHC II: (Poran et al., 2020)	(Poran et al., 2020)	Table S7
Baseline vaccine MHC II: (Singh et al., 2020)	(Singh et al., 2020)	Table 2
Baseline vaccine MHC II: (Ahmed et al., 2020)	(Ahmed et al., 2020)	Table 2
Baseline vaccine MHC II: (Tahir ul Qamar et al., 2020)	(Tahir ul Qamar et al., 2020)	Table 1
Baseline vaccine MHC II: (Mitra et al., 2020)	(Mitra et al., 2020)	Table 1(B)
Baseline vaccine MHC II: (Abdelmageed et al., 2020)	(Abdelmageed et al., 2020)	Table 3
Baseline vaccine MHC II: (Vashi et al., 2020)	(Vashi et al., 2020)	Table 6
Baseline vaccine MHC II: (Baruah and Bose, 2020)	(Baruah and Bose, 2020)	Table 2
Baseline vaccine MHC II: (Yarmarkovich et al., 2020)	(Yarmarkovich et al., 2020)	Table S5 (65 33-mers, to which we applied sliding windows of lengths 13–25 to obtain the peptide set considered for MHC class II)
Baseline vaccine MHC II: (Yazdani et al., 2020)	(Yazdani et al., 2020)	Table 1 (peptides created by sliding windows of length 13–25)
Megapool MHC I: (Grifoni et al., 2020a)	(Grifoni et al., 2020a)	Table S6
Megapool MHC II: (Grifoni et al., 2020a)	(Grifoni et al., 2020a)	Table S3
Software and Algorithms		
OptiVax	This paper, GitHub	https://github.com/gifford-lab/optivax
EvalVax	This paper, GitHub	https://github.com/gifford-lab/optivax
NetMHCpan-4.0	(Jurtz et al., 2017)	http://www.cbs.dtu.dk/services/NetMHCpan-4.0/
NetMHCpan-4.1	(Reynisson et al., 2020a)	http://www.cbs.dtu.dk/services/NetMHCpan-4.1/
NetMHCIIpan-4.0	(Reynisson et al., 2020b)	http://www.cbs.dtu.dk/services/NetMHCIIpan-4.0/
NetMHCIIpan-3.2	(Jensen et al., 2018)	http://www.cbs.dtu.dk/services/NetMHCIIpan-3.2/
MHCflurry 1.6.0	(O'Donnell et al., 2020)	Version 1.6.0, https://github.com/openvax/mhcflurry
PUFFIN	(Zeng and Gifford, 2019); https://github.com/gifford-lab/PUFFIN	GitHub commit a63f6c563b7e2f7b04eac28a6cf09d8078ac3a2a with pre-trained model
Hapferret	https://github.com/nilsboar/HapFerret	GitHub commit e2381dc567cec97373acb49c09f167e46ea0bb53
Nextstrain	(Hadfield et al., 2018); https://github.com/nextstrain/ncov	GitHub commit 639c63f25e0bf30c900f8d3d937de4063d96f791
NetNGlyc	(Gupta et al., 2004)	http://www.cbs.dtu.dk/services/NetNGlyc/

RESOURCE AVAILABILITY

Lead Contact

Further information and requests for resources should be directed to and will be fulfilled by the Lead Contact, David K. Gifford (gifford@mit.edu).

Materials Availability

This study did not generate new materials.

Data and Code Availability

The code and peptide-HLA predictions generated during this study are available at https://github.com/gifford-lab/optivax. Original data for HLA haplotype frequency data have been deposited to Mendeley Data: https://dx.doi.org/10.17632/cfxkfy9zp4.1.

METHOD DETAILS

Peptide filtering

Removal of Highly Mutable Peptides

We eliminate peptides that are observed to mutate above an input threshold rate to improve coverage over all SARS-CoV-2 variants and reduce the chance that the virus will mutate and escape vaccine-induced immunity in the future. When possible, we select peptides that are observed to be perfectly conserved across all observed SARS-CoV-2 viral genomes. Peptides that are observed to be perfectly conserved in thousands of examples may be functionally constrained to evolve slowly or not at all. If functional data are available, they can be used to supplement observed viral genome mutation rates by increasing mutation rates over functionally non-constrained residues.

For SARS-CoV-2, we obtained the most up to date version of the GISAID database (Elbe and Buckland-Merrett, 2017) (as of 2:02pm EST May 13, 2020, see Table S4: GISAID acknowledgements) and used Nextstrain (Hadfield et al., 2018) to remove genomes with sequencing errors, translate the genome into proteins, and perform multiple sequence alignments (MSAs). We retrieved 24,468 sequences from GISAID, and 19,288 remained after Nextstrain quality processing. After quality processing, Nextstrain randomly sampled 34 genomes from every geographic region and month to produce a representative set of 5,142 genomes for evolutionary analysis. Nextstrain definition of a ''region'' can vary from a city (e.g., ''Shanghai'') to a larger geographical district. Spatial and temporal sampling in Nextstrain is designed to provide a representative sampling of sequences around the world.

The 5,142 genomes sampled by Nextstrain were then translated into protein sequences and aligned. We eliminated viral genome sequences that had a stop codon, a gap, an unknown amino acid (because of an uncaled nucleotide in the codon), or had a gene that lacked a starting methionine, except for ORF1b which does not begin with a methionine. This left a total of 4,690 sequences that were used to compute peptide level mutation probabilities. For each peptide, the probability of mutation was computed as the number of non-reference peptide sequences observed divided by the total number of peptide sequences observed.

Removal of Cleavage Regions

SARS-CoV-2 contains a number of post-translation cleavage sites in ORF1a and ORF1b that result in a number of nonstructural protein products. Cleavage sites for ORF1a and ORF1b were obtained from UniProt (Consortium, 2019) under entry P0DTD1. In addition, a furin-like cleavage site has been identified in the spike protein (Wang et al., 2020; Coutard et al., 2020). This cleavage occurs before peptides are loaded in the endoplasmic reticulum for class I or endosomes for class II. Any peptide that spans any of these cleavage sites is removed from consideration. This removes 3,739 peptides out of the 154,996 we consider across windows 8–10 (class I) and 13–25 (class II) (~2.4%).

Removal of Glycosylated Peptides

We eliminate all peptides that are predicted to have N-linked glycosylation as it can inhibit MHC class I peptide loading and T cell recognition of peptides (Wolfert and Boons, 2013; Wrapp et al., 2020). In addition, we do not know how well existing peptide prediction methods function on glycosylated peptides. Finally, any use of peptides that are natively glycosylated in a virus would likely require that vaccine peptides be identically glycosylated to enable T cell recognition of vaccine primed memory. The use of non-glycosylated vaccine peptides in this case has resulted in vaccine failures (Wolfert and Boons, 2013).

Glycosylation is a post-translational modification that involves the covalent attachment of carbohydrates to specific motifs on the surface of the protein. We identified peptides that may be glycosylated with the NetNGlyc N-glycosylation prediction server (Gupta et al., 2004). We verified these predictions for the spike protein by ensuring they were in the same locations as those found using experimental data of spike N-glycosylation from Cryo-EM (Walls et al., 2020) and tandem mass spectrometry (Zhang et al., 2020b). A majority of the potential N-glycosylation sites (16 out of 22) were identified in both experimental studies, and further supported by homologous regions with glycosylation found in SARS-CoV (Walls et al., 2020). We found that all 22 experimentally identified real or likely N-glycosylation sites from the SARS-CoV-2 spike protein were predicted to be glycosylated with non-zero probability by NetNGlyc. Therefore, we eliminated all peptides where NetNGlyc predicted a non-zero N-glycosylation probability in any residue. This resulted in the elimination of 18,957 of the 154,996 peptides considered (~12%).

Self-epitope Removal

T cells are selected to ignore peptides derived from the normal human proteome, and thus we remove any self peptides from consideration for a vaccine. In addition, it is possible that a vaccine might stimulate the adaptive immune system to react to a self peptide that was presented at an abnormally high level, which could lead to an autoimmune disorder. All peptides from SARS-CoV-2 were scanned against the entire human proteome downloaded from UniProt (Consortium, 2019) under Proteome ID UP000005640. A total of 48 exact peptide matches (46 8-mers, two 9-mers) were discovered and eliminated from consideration.

Removal of Undesired Proteins

OptiVax can design vaccines using peptides from specific viral or oncogene proteins of interest by removing peptides from undesired proteins from the candidate pool. Grifoni et al. (2020b) tested T cell responses from COVID-19 convalescent patients and found that peptides from the S, M, and N proteins of SARS-CoV-2 produce the dominant $CD4^+$ and $CD8^+$ responses when compared to other SARS-CoV-2 proteins. We used OptiVax to produce additional SARS-CoV-2 vaccines comprised of peptides drawn from only the S, M, and N proteins.

EvalVax

EvalVax-Robust Considers Linkage Disequilibrium of HLA Genes

EvalVax-Robust computes the distribution of per individual peptide-HLA binding hits over a given population. It accounts for the significant linkage disequilibrium (LD) between HLA loci and uses haplotype frequencies for population coverage estimates. We expect that a vaccine will be more effective if more of its peptides are displayed by an individual's HLA molecules, and thus EvalVax-Robust computes the probability of having at least N predicted peptide-HLA binding hits for each individual in the population.

Assuming for each of the HLA-A, -B, -C loci there are M_A, M_B, M_C alleles, respectively, for a given haploid $A_iB_jC_k$, the haplotype frequency is defined as $G(i,j,k)$ and $\sum_{i=1}^{M_A}\sum_{j=1}^{M_B}\sum_{k=1}^{M_C}G(i,j,k) = 1$. We assume independence of inherited haplotypes and compute the frequency of a diploid genotype as:

$$F_{i_1j_1k_1i_2j_2k_2} = F\left(A_{i_1}B_{j_1}C_{k_1}, A_{i_2}B_{j_2}C_{k_2}\right) = G(i_1,j_1,k_1)G(i_2,j_2,k_2) \quad \text{(Equation 1)}$$

For each allele a, $e(a)$ denotes the number of peptides predicted to bind to the allele with $\leq$ 50nM affinity, which we call the number of peptide-HLA hits. Then for each possible diploid genotype we compute the total number of peptide-HLA hits of the genotype as the sum of $e(a)$ of the unique alleles in the genotype (there can be 3-6 unique alleles depending on the zygosity of each locus):

$$C_{i_1j_1k_1i_2j_2k_2} = C\left(A_{i_1}B_{j_1}C_{k_1}, A_{i_2}B_{j_2}C_{k_2}\right) = \sum_{a\in\left\{A_{i_1},B_{j_1},C_{k_1}\right\}\cup\left\{A_{i_2},B_{j_2},C_{k_2}\right\}} e(a) \quad \text{(Equation 2)}$$

We then compute the frequency of having exactly k peptide-HLA hits in the population as:

$$P(n=k) = \sum_{i_1=1}^{M_A}\sum_{j_1=1}^{M_B}\sum_{k_1=1}^{M_C}\sum_{i_2=1}^{M_A}\sum_{j_2=1}^{M_B}\sum_{k_2=1}^{M_C} F_{i_1j_1k_1i_2j_2k_2}\mathbb{1}\left\{C_{i_1j_1k_1i_2j_2k_2} = k\right\} \quad \text{(Equation 3)}$$

We define the population coverage objective function for EvalVax-Robust as the probability of having at least N peptide-HLA hits in the population, where the cutoff N is set to the minimum number of displayed vaccine peptides desired:

$$P(n\geq N) = \sum_{k=N}^{\infty} P(n=k) \quad \text{(Equation 4)}$$

When we evaluate metrics on a world population, we equally weight population coverage estimations over three population groups (White, Black, and Asian) as the final objective function. In addition to the probability of having at least N peptide-HLA hits per individual, we also evaluate the expected number of per individual peptide-HLA hits in the population, which provides insight on how well the vaccine is displayed on average.

EvalVax-Unlinked Computes Population Coverage by at least One Peptide-HLA Hit

When haplotype frequencies are not available for a population, we can evaluate a vaccine using HLA allele frequencies that assume independence and compute the probability that at least one peptide binds to any of the alleles at any of the loci. To encourage a diverse set of peptides to bind to a single HLA allele, we use the predicted binding probability of a peptide to an allele instead of using a binary indicator of binding. This permits multiple peptides to contribute to the probability score at each allele. Considering K loci $\{L_1,\ldots,L_K\}$, for each locus there are M_k alleles $a_1,\ldots,a_{M_k}$ and the allele frequency is defined as $G_k(a_i)$ and $\sum_{i=1}^{M_k}G_k(a_i) = 1$. Given a set of N peptides $\mathcal{P} = \{P_1,P_2,\ldots,P_N\}$, for each allele (of locus L_k) the predicted binding probability to peptide P_n is $e_k^n(a_i)$. Assuming no competition between peptides, the probability that allele a_i ends up having at least one peptide bound is:

$$e_k(a_i) = 1 - \prod_{n=1}^{N}\left(1 - e_k^n(a_i)\right) \quad \text{(Equation 5)}$$

We define the diploid frequency of alleles as $F_k(a_i,a_j) = G_k(a_i)G_k(a_j)$, and we conservatively assume that a homozygous diploid locus does not improve the chance of peptide presentation over a single copy of the locus. Thus, the probability that a diploid genotype has at least one peptide bound is defined as:

$$B_k(a_i, a_j) = \begin{cases} 1 - (1 - e_k(a_i))(1 - e_k(a_j)), & \text{if} \quad i \neq j \\ e_k(a_i), & \text{if} \quad i = j \end{cases} \quad \text{(Equation 6)}$$

Therefore, the probability that a person in the given population displays at least one peptide in the set $\mathcal{P}$ at a particular locus L_k is calculated by:

$$F_k(\mathcal{P}) = \sum_{i=1}^{M_k} \sum_{j=1}^{M_k} F_k(a_i, a_j) B_k(a_i, a_j) \quad \text{(Equation 7)}$$

To combine different loci assuming no linkage disequilibrium, the probability that a person in the given population has at least one locus that binds to at least one peptide from $\mathcal{P}$ is defined as:

$$P(\mathcal{P}) = 1 - \prod_{k=1}^{K} (1 - F_k(\mathcal{P})) \quad \text{(Equation 8)}$$

which is the evaluation metric for EvalVax-Unlinked.

We conservatively only consider peptides with predicted binding affinity $\leq$ 50 nM. We set values of $e_k^n(a_i)$ weaker than 50 nM predicted binding affinity to zero. This constraint on peptide binding is in addition to peptide filtering described above. When we evaluate on a world population, we equally weight population coverage estimates over 15 geographic regions (see Results for the list of regions) as the final objective function.

OptiVax

OptiVax-Robust Searches for a Peptide Set with High Expected Number of Per-individual Peptide-HLA Hits

OptiVax-Robust uses beam search to find a minimal set of peptides that reaches a target population coverage probability at a threshold of n predicted peptide-HLA hits for each individual. We start from an empty set of peptides and $n = 0$, and iteratively expand the solution by one peptide at a time and retain the top k solutions until the population coverage probability for the current n reaches the target population coverage probability threshold for that n. We then repeat the same process for $n + 1$. If it not possible to reach the target population coverage probability threshold for n then the current coverage is accepted and we repeat the process for $n + 1$. At the expense of increased computational cost, beam search improves upon greedy optimization by considering k possible solutions at each step. During each iteration, the population coverage probability threshold at the present n controls the robustness of coverage. Increasing the target population coverage probability increases the difficulty of the optimization task. The iterative process stops when the target population coverage at the desired n is achieved. In early rounds of optimization, OptiVax uses a high population coverage probability to provide better individual coverage. In subsequent rounds, the target population coverage probability is reduced on a fixed schedule.

OptiVax-Unlinked Searches for a Peptide Set that Covers a Population

OptiVax-Unlinked uses beam search to find a minimal set of peptides that reaches a desired population coverage probability that each individual on average displays at least one vaccine peptide. We iteratively expand solutions in the beam by adding one peptide at a time to reach the population coverage objective, and keep the top k solutions over all possible expansions in the beam.

OptiVax Improves Vaccine Sequence Diversity

OptiVax reduces vaccine sequence redundancy by not selecting peptides with closely related sequences for a vaccine formulation. This issue arises because sliding a window over a proteome produces overlapping sequences that are very similar in HLA binding characteristics. When any version of OptiVax selects a peptide during optimization, it eliminates from further consideration all unselected peptides that are within three (MHC class I) or five (MHC class II) edits on a sequence distance metric from the selected peptide. The distance metric computation aligns two peptides not allowing gaps and mismatches and the distance metric is the sum of the lengths of any end overhangs where the opposite peptide sequence is absent.

Computational Peptide-HLA Prediction Models

Computational Models

For MHC class I design, we use an ensemble that outputs the mean predicted binding affinity of NetMHCpan-4.0 (Jurtz et al., 2017) and MHCflurry 1.6.0 (O'Donnell et al., 2020, 2018). We find this ensemble increases the precision of binding affinity estimates over the individual models on available SARS-CoV-2 experimental data (Table S1). For MHC class II design, we use NetMHCIIpan-4.0 (Reynisson et al., 2020b). For evaluation, we use our ensemble estimate of binding (MHC class I), as well as use binding predictions from a wide range of prediction algorithms (MHC class I: NetMHCpan-4.0 (Jurtz et al., 2017), NetMHCpan-4.1 (Reynisson et al., 2020a), MHCflurry 1.6.0 (O'Donnell et al., 2020), PUFFIN (Zeng and Gifford, 2019); MHC class II: NetMHCIIpan-3.2 (Jensen et al., 2018), NetMHCIIpan-4.0 (Reynisson et al., 2020b), PUFFIN (Zeng and Gifford, 2019)) to ensure that all methods agree that we have a good peptide vaccine. We validate our computational models on a dataset of SARS-CoV-2 peptides evaluated for stability (Prachar et al., 2020). We find scoring peptides by predicted binding affinity maximizes AUROC as compared to alternative scoring methods, and selecting peptides using a 50 nM binding affinity threshold maximizes precision in classification of stable binders compared to alternative binding criteria (Table S1). Our ensemble of NetMHCpan-4.0 and MHCflurry further increases AUROC and precision over individual predictors.

All models take as input a (HLA, peptide) pair and output predicted peptide-HLA binding affinity (IC50) on a nanomolar scale. For both MHC class I and class II models, we consider peptides to be binders if the predicted HLA binding affinity is $\leq$ 50 nM (Sette et al., 1994). This provides a conservative threshold to increase the probability of peptide display. Where our methods require a probability of peptide-HLA binding (as in Equation 5), affinity predictions are capped at 50000nM and transformed into $[0, 1]$ using a logistic transformation, $1 - \log_{50000}(\text{aff})$, where larger values correspond to greater likelihood of eliciting an immunogenic response (Sette et al., 1994; Buus et al., 2003; Nielsen et al., 2003). The $\leq$ 50 nM binding affinity threshold corresponds to a threshold of $\geq$ 0.638 after logistic transformation. We explored other criteria to classify peptides as binders and found using predicted binding affinity with a 50 nM threshold to satisfy percentile rank criteria and maximize precision on available SARS-CoV-2 experimental data (Table S1).

Criteria for Predicted Binding

NetMHCpan-4.0 (Jurtz et al., 2017) and NetMHCIIpan-4.0 (Reynisson et al., 2020b) output predicted binding affinity (BA), percentile rank of predicted BA compared to a set of random natural peptides, and percentile rank of an eluted ligand (EL) score compared to a set of random natural peptides. Default parameters for these methods suggest EL percentile rank thresholds of 0.5% and 2% rank for classifying peptides as strong and weak binders, respectively, for MHC class I and thresholds of 2% and 10% for strong and weak binders, respectively, for MHC class II.

To identify binders for our vaccine designs, we used a 50 nM predicted binding affinity threshold. We found binders selected with this criterion are also considered binders under alternative criteria based on percentile rank. Across our set of all candidate SARS-CoV-2 MHC class I peptides, we found 91.0% of peptide-HLA hits with $\leq$ 50 nM predicted binding affinity by NetMHCpan-4.0 were also considered binders using BA percentile rank $\leq$ 0.5% (100.0% have BA percentile rank $\leq$ 2%). Using percentile rank for EL scores, 67.6% of peptide-HLA hits with $\leq$ 50 nM predicted binding affinity have EL percentile rank $\leq$ 0.5% (92.6% have EL percentile rank $\leq$ 2%). Across all candidate SARS-CoV-2 MHC class II peptides, we found 86.1% of peptide-HLA hits with $\leq$ 50 nM predicted binding affinity by NetMHCIIpan-4.0 were also considered binders using BA percentile rank $\leq$ 2% (100.0% have BA percentile rank $\leq$ 10%). Using percentile rank for EL scores, 26.1% of peptide-HLA hits with $\leq$ 50 nM predicted binding affinity have EL percentile rank $\leq$ 2% (63.1% have EL percentile rank $\leq$ 10%).

Binders selected using percentile rank metrics were generally not considered binders under a 50 nM predicted binding threshold. Across our set of all candidate SARS-CoV-2 MHC class I peptides, we found 17.5% of peptide-HLA hits with EL percentile rank $\leq$ 0.5% have $\leq$ 50 nM predicted binding affinity by NetMHCpan-4.0. Across all candidate SARS-CoV-2 MHC class II peptides, we found 11.3% of peptide-HLA hits with EL percentile rank $\leq$ 2.0% have $\leq$ 50 nM predicted binding affinity by NetMHCIIpan-4.0.

Validation on SARS-CoV-2 Experimental Stability Data

We evaluate peptide-HLA binding predictions on a set of experimentally assessed SARS-CoV-2 peptides whose peptide-HLA complex stability was assessed in vitro across 11 MHC allotypes (5 HLA-A, 1 HLA-B, 4 HLA-C, 1 HLA-DRB1) (Prachar et al., 2020). Prachar et al. (2020) suggests peptides with at least 60% of the stability of a reference peptide in a NeoScreen assay are likely high affinity binders. For MHC class I alleles, the dataset contains 912 unique peptide-HLA pairs, of which 185 peptides are considered stable ($\geq$ 60% stability). For MHC class II, the dataset contains 93 total peptides, of which 22 are stable. We use our computational models to predict peptide-HLA binding and evaluate them using various binding criteria against the experimental peptide stability measurement (Table S1). We compare classification performance using different binding criteria as described above and find in general that classifying binders using predicted binding affinity maximizes AUROC and a 50 nM binding affinity threshold maximizes precision (Table S1). We find our mean ensemble of NetMHCpan-4.0 and MHCflurry further improves classification AUROC and precision over the individual models for predicting MHC class I epitopes. On MHC class II data, we note NetMHCIIpan-4.0 achieves AUROC 0.848 and precision 0.625 using a 500 nM threshold (Table S1). While NetMHCIIpan-4.0 with a 50 nM threshold does not identify any peptides in this dataset as binders, we use this stricter threshold in our vaccine designs as it is more conservative and less likely to admit false positive binders. In general, we find performance of PUFFIN with a 50 nM binding threshold comparable to alternative methods on both MHC class I and class II data and use PUFFIN as part of our vaccine design evaluation.

QUANTIFICATION AND STATISTICAL ANALYSIS

Classification performance of peptide-MHC scoring models was calculated using scikit-learn (Pedregosa et al., 2011) in Python using the *sklearn.metrics.roc_auc_score* (AUROC), *sklearn.metrics.average_precision_score* (Average Precision), *sklearn.metrics.precision_score* (Precision), and *sklearn.metrics.classification_report* (Sensitivity and Specificity) functions. AUROC and average precision are computed using raw predictions, and the remaining metrics are computed using binarized predictions based on the respective binding criteria.