Synthetic and Systems Biotechnology 8 (2023) 757–770

Contents lists available at ScienceDirect

Synthetic and Systems Biotechnology

journal homepage: www.keaipublishing.com/en/journals/synthetic-and-systems-biotechnology

# Designing a multi-epitope vaccine against *Peptostreptococcus anaerobius* based on an immunoinformatics approach

Yudan Mao, Xianzun Xiao, Jie Zhang, Xiangyu Mou **, Wenjing Zhao *

*Shenzhen Key Laboratory of Systems Medicine for Inflammatory Diseases, School of Medicine, Shenzhen Campus of Sun Yat-Sen University, Shenzhen, Guangdong, 518107, China*

ARTICLE INFO

*Keywords:*
Colorectal cancer
*Peptostreptococcus anaerobius*
PCWBR2
Immunoformatics
Multi-epitope

ABSTRACT

*Peptostreptococcus anaerobius* is an anaerobic bacterium, which has been found selectively en-riched in the fecal and mucosal microbiota of colorectal cancer (CRC) patients. Emerging evidence suggest *P. anaerobius* may contribute to the development of CRC in human. In this study, we designed a multi-epitope chimeric vaccine against *P. anaerobius* PCWBR2, a recently identified adhesin that interacts directly with colon cell lines by binding α2/β1 integrin frequently overexpressed in human CRC tumors and cell lines. Immunoinformatics tools predicted six cytotoxic T lymphocyte epitopes, five helper T lymphocyte epitopes, and six linear B lymphocyte epitopes. The predicted epitopes were joined with AAY or GPGPG linkers and a previously reported TLR4 agonist was added to the vaccine construct's N terminal as an adjuvant using EAAAK linkers and the order of epitopes was optimized. Further *in silico* analysis revealed that the vaccine construct possesses satisfactory antigenicity, allergenicity, solubility, physicochemical properties, adjuvant-TLR4 molecular docking, and immune profile characteristics. Our study provided a promising design for vaccines against *P. anaerobius*.

## 1. Introduction

Colorectal cancer (CRC) is one of the most prevalent types of cancer in the world [1]. The occurrence and progression of colorectal cancer involve a complex interaction between genetic, epigenetic, and environmental factors, the latter being the major trigger for colorectal cancer [2]. A growing body of evidence has demonstrated that the gut microbiota is an important environmental factor that contributes to CRC development [3,4]. Specifically, enterotoxigenic *Bacteroides fragilis* [5], *Fusobacterium nucleatum* [6], polyketide synthase (pks+) *Escherichia coli* [7] and *Enterococcus faecalis* [8] have been demonstrated to promote the development of CRC. More recently, *Peptostreptococcus anaerobius* was found enriched in CRC patients [4,9] and playing a contributing role to tumorigenesis in mice [10]. *P. anaerobiosis* is an opportunistic pathogen that normally resides in the gastrointestinal, urogenital tract, and oral cavities [11,12] and causes infections including endocarditis, urogenital and gastrointestinal infections in certain circumstances [13,14]. *P. anaerobius* was demonstrated significantly enriched in feces and colon tissue from patients with colorectal adenomas and adenocarcinomas [14]. Moreover, within CRC patients, the abundance of *P. anaerobius* was enriched in advanced adenomas and CRC tissues compared to normal colon tissues [15]. Further study revealed *P. anaerobius* promoted cholesterol synthesis and colorectal carcinogenesis by enhancing intracellular reactive oxygen species (ROS) levels through Toll-like receptors (TLRs) 2 and 4. These results suggest that *P. anaerobius* may contribute to the development of CRC in humans.

Colonization of tumor-promoting bacteria in the colon is a prerequisite for their tumor-promoting effect, and bacterial adhesin plays an important role in the process of colonization. For example, *F. nucleatum* utilizes FadA (Fusobacterium adhesin A) to bind to the E-cadherin on epithelial and endothelial cells and Fap2 (Fusobacterial apoptosis protein 2) binds to the TIGIT on Natural Killer (NK) cells to enhance immune evasion and progression of carcinogenesis [16]. Recently, it was reported that *P. anaerobius* PCWBR2 (putative cell wall-binding repeat 2) plays a role as an adhesion by directly binding to α2/β1 integrins, a receptor that is frequently overexpressed in human CRC tumors and cell lines [10]. Given the potential contributing role of *P. anaerobius* in CRC, there would be a great need for intervention methods against these

Peer review under responsibility of KeAi Communications Co., Ltd.
* Corresponding author.
** Corresponding author.
*E-mail addresses:* mouxy5@ms.sysu.edu.cn (X. Mou), zhaowj29@ms.sysu.edu.cn (W. Zhao).

https://doi.org/10.1016/j.synbio.2023.11.004
Received 10 August 2023; Received in revised form 15 October 2023; Accepted 15 November 2023
Available online 28 November 2023

bacteria. Vaccination is a precision intervention that can be used in both the prevention and therapy of specific bacterial infections and colonization. Therefore, we aimed to develop a vaccine against *P. anaerobius* to specifically reduce its abundance in the gut. In this study, we aim to design a multi-epitope vaccine *in silico* against *P. anaerobius* adhesion PCWBR2. By using a series of immuno-informatics tools, we selected 6 CTL epitopes, 5 HTL epitopes, and 6 LBL epitopes from the amino acid sequence of PCWBR2 and constructed them into a 516-amino acid multi-epitope vaccine. Further *in silico* analysis revealed that the resulting vaccine possesses satisfactory characteristics. Our study provided a promising design for vaccines against *P. anaerobius*.

## 2. Materials and methods

The workflow summarizing the methodology used to design an effective candidate vaccine in this study is shown in Fig. 1.

### 2.1. Identification of vaccine targets

The complete amino acid sequence of PCWBR2 was retrieved from the NCBI database (https://www.ncbi.nlm.nih.gov/) in FASTA format with GenBank accession number EKX94456.1. Sequence alignment based on the PCWBR2-encoded region phylogenetic tree was built using the adjacency-join (N-J) method in Mega 11 software with 1000 Bootstrap replicates. Further, SignalP 6.0 (https://services.healthtech.dtu.dk/service.php?SignalP) server was used for the analysis of the signal peptides to differentiate the secretory and non-secretory proteins. It also indicates the positioning of cleavage sites in the proteins by making use of various artificial neural networks [17].

Subcellular localization of protein antigens was checked on PSORTb. PSORTb v3.0.2 (https://www.psort.org/psortb/) is the most precise bacterial localization prediction tool available [18] which is reported to have a measured prediction of 96 % [19]. Secretory or pathogen outer membrane proteins may be an ideal candidate target for, but not for

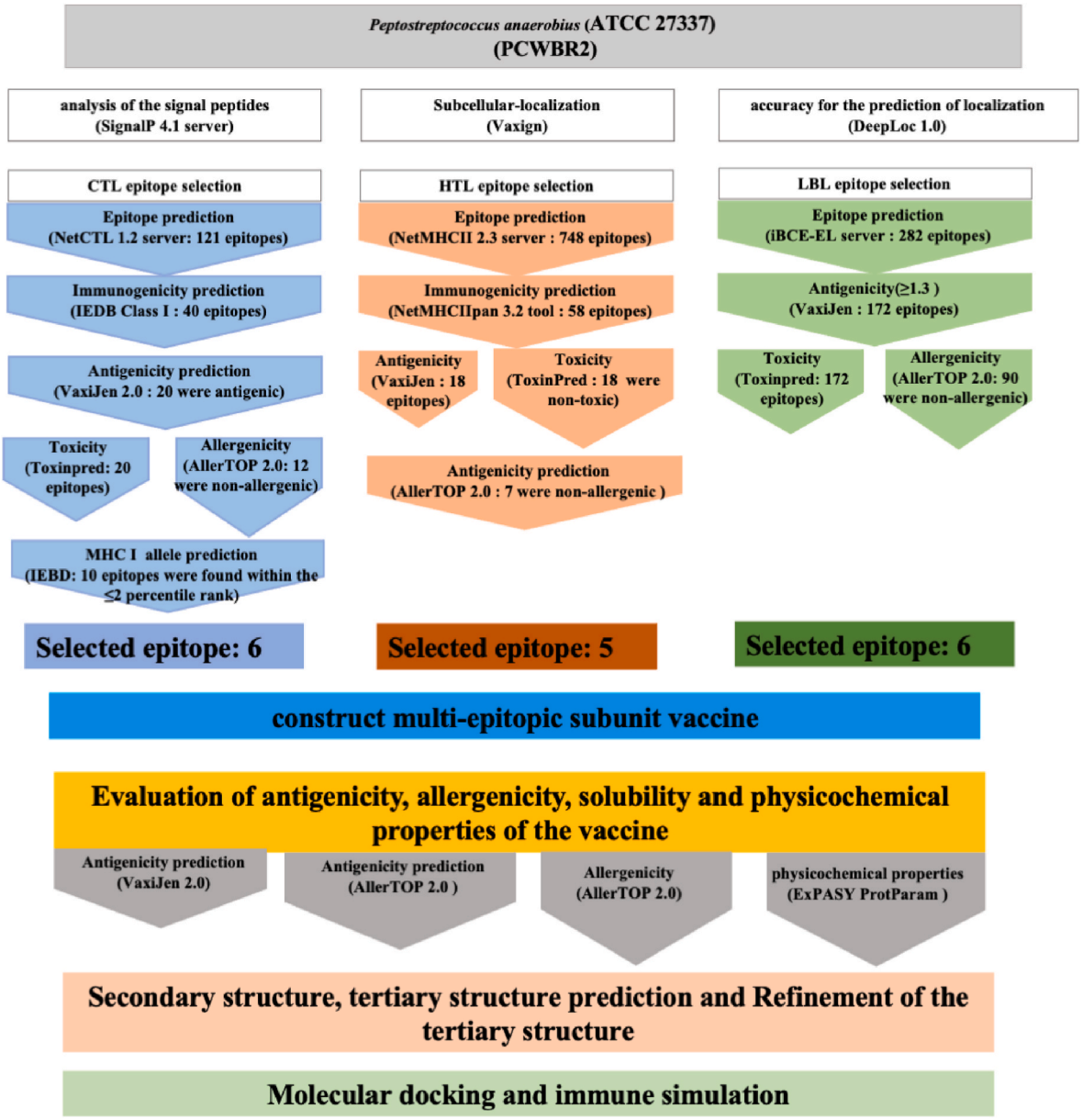


**Fig. 1.** The workflow used for designing the vaccine against *P. anaerobius*. The whole methodology is shown in four steps: the retrieval of protein sequence, prediction of T and B cell epitopes, the antigenicity, allergenicity, toxicity and physiochemical assessments of the vaccine construct, molecular docking with TLR4/MD2, and the simulation of the immune response against the vaccine.

cytoplasmic and inner membrane proteins. To achieve accuracy for the prediction of localization, the proteins were also subjected to DeepLoc 2.0 (https://services.healthtech.dtu.dk/service.php?DeepLoc-2.0) which gives the subcellular localization of eukaryotic proteins.

### 2.2. Prediction and assessment of MHC-I (CTL) epitopes

Prediction of potential cytotoxic T lymphocytes (CTL) epitopes is a crucial and widely used step in the design of an *in-silico* vaccine [20]. NetCTL 1.2 server [21], IEDB MHC-I binding prediction tool [22], and NetMHCpan-4.1 servers [23] have been used to screen MHC-I epitopes of targeted proteins. NetCTL 1.2 server (http://www.cbs.dtu.dk/services/NetCTL/) was employed for predicting cytotoxic T lymphocytes (CTL) epitopes with 54–89 % sensitivity and 94–99 % specificity [9]. These epitopes are recognized by common HLA class I supertypes in human populations, namely: A1, A2, A3, A24, A26, B7, B8, B27, B39, B44, B58, and B62. Default settings were used (threshold, 0.75) for the estimation of CTL epitopes [21,24,25].

The immunogenicity of the CTL epitopes was predicted using the Class I immunogenicity tool of the IEDB Analysis Resource (http://tools.iedb.org/immunogenicity/) [26]. Only epitopes with a positive immunogenicity value were preserved for the following step of the assessment.

VaxiJen server (http://www.ddg-pharmfac.net/vaxijen/VaxiJen/VaxiJen.html) for antigenicity evaluation. It generates the antigenicity of the proteins without using any alignment and concentrates on the physiochemical properties of the chosen candidate [27]. Epitopes showing antigenicity prediction values $\geq 0.4$ were considered antigenic. ToxinPred 2.0 (https://webs.iiitd.edu.in/raghava/toxinpred2/batch.html) [28] and AllerTOP 2.0 (http://www.ddg-pharmfac.net/AllerTOP/) [29] servers were used, respectively, to estimate the toxicity and allergenicity of the immunogenic epitopes. The non-toxic and non-allergenic epitopes were subjected to NetMHCpan-4.1 servers (https://services.healthtech.dtu.dk/service.php?NetMHCpan-4.1) [23] to predict binding between peptides and MHC-I. A percentile rank score $\leq 2$ was considered for this study.

### 2.3. Prediction of MHC-II (HTL) epitopes

The prediction of HTL epitopes is a common component of immunoinformatics investigations for vaccine development [30–32]. The NetMHCII 2.2 server (https://services.healthtech.dtu.dk/service.php?NetMHCII-2.3) was used to predict the helper T cells' 15-mer epitopes for the selected protein sequence [33]. For predicting the linkage of peptides to the human alleles HLA-DR, HLA-DQ, and HLA-DP against a set of seven human HLAs given as HLA-DRB1 * 03:01, HLA-DRB1 * 07:01, HLA-DRB1 * 15:01, HLA-DRB3 * 01:01, HLA-DRB3 * 02:02, HLA-DRB4 * 01:01, HLA-DRB5*01:01. Furthermore, based on receptor affinity—which is often deduced from the $IC_{50}$ values—the server predicts the MHC II epitopes. The maximum binding affinity for MHC-II is shown by $IC_{50}$ values less than 50 nM, followed by $IC_{50}$ values less than 500 nM for intermediate affinity and $IC_{50}$ values less than 5000 nM for the lowest binding affinity. The percentile rank and the $IC_{50}$ score are inversely correlated. High-affinity peptides are defined as those with an $IC_{50}$ value of less than 50 nM [34]. The VaxiJen server was then used to determine the antigenicity of the epitopes. The non-toxic and non-allergenic epitopes were chosen using the ToxinPred and AllerTOP 2.0 servers in that order. Only the antigenic epitopes were chosen for the vaccine's construction based on these criteria.

### 2.4. Identification of linear B-cell epitopes

For the prediction of LBL epitopes, the iBCE-EL server (http://thegleelab.org/iBCE-EL/) was employed [35]. Only the ones positively predicted to be LBL epitopes by the server were chosen for further analysis. The VaxiJen server evaluated the antigenicity of the probable LBL epitopes. The ToxinPred and AllerTOP 2.0 servers, respectively, were used to assess the toxicity and allergenicity of epitopes with antigenicity values of 0.5.

### 2.5. Molecular docking between T lymphocyte epitopes and MHC alleles

The binding affinity of CTL and HTL epitopes to their corresponding MHC alleles was evaluated by molecular docking simulation. The MHC allele was downloaded from the RCSB PDB (https://www.rcsb.org/) [36,37] and processed using UCSF Chimera software to remove unnecessary ligands. For MHC alleles not found in the RCSB PDB, the SWISS-MODEL server (https://swissmodel.expasy.org/) [38] was used for homology modeling. The energy of the structure is then minimized using the Swiss-PdbViewer [39]. The three-dimensional form of epitopes was obtained using the PEP-FOLD 3.5 server (https://bioserv.rpbs.univ-paris-diderot.fr/services/PEP-FOLD3/) [40]. Then use the Swiss-PdbViewer to minimize their energy. Autodock Vina [41] in PyRx [42] was used for docking between epitopes and MHC molecules. Binding affinity and interaction were analyzed using Autodock Vina, PyRx, and UCSF Chimera.

### 2.6. Population coverage prediction

The geographic and racial diversity of the world had varying MHC alleles genotype and frequencies, and the polymorphism of MHC alleles influences the binding ability of epitopes to MHC-I or MHC-II molecules. In order to analyze the combined coverage of T lymphocyte epitopes to the world population, the IEDB population coverage tool (http://tools.iedb.org/population/) was used to confirm the population coverage [43].

### 2.7. Construction of multi-epitopic vaccine candidate sequence

Combining the high-scoring epitopes of B cells, CTLs, and high-affinity binding HTLs, the sequence of vaccine candidates was generated. The most recent Toll-like Receptor 4 agonist, monophosphoryl lipid A (MPLA) (QFN32858.1), has the potential to be widely utilized as an adjuvant in humans [62–64]. At the N terminal of the sequence, an EAAAK linker was utilized to attach the adjuvant, while AAY linkers were employed to connect the CTL epitopes and GPGPG linkers were used to bind the remaining B-cell and HTL epitopes. To identify and purify the protein, a 6x His tag was attached to the C terminus. The suggested vaccine's linear architecture was modeled in detail using the trRosetta web server (https://yanglab.nankai.edu.cn/trRosetta/) [44]. SAVES (https://saves.mbi.ucla.edu/) and ProSA web servers were used to validate the generated 3D model, and they were also used to predict the protein 3D model's ERRAT scores, Ramachandran plot, and Z-score [45,46].

### 2.8. Prediction of antigenicity, allergenicity, solubility, and physicochemical analysis of vaccine

The physicochemical characteristics of the vaccine were characterized by the Prot-Param tool on the Expasy server (http://web.expasy.org/protparam/) [47]. Protein-Sol (https://protein-sol.manchester.ac.uk/) was used to assess the solubility of the designed vaccine sequence, and any scaled solubility values greater than 0.45 is predicted to have a higher solubility than the average soluble *E. coli* protein from the experimental solubility dataset [48].

VaxiJen v2.0 servers were used to forecast the chimeric construct's antigenicity. The allergenicity of the multi-epitopic vaccine was predicted by AllerTOP v2.0. TMHMM software (DeepTMHMM, https://dtu.biolib.com/DeepTMHMM) [49] to predict the way of protein transmembrane, usually more than 1 transmembrane protein is difficult to clone, expression, purification, not suitable as a recombinant vaccine antigen.

### 2.9. Extrapolation of the secondary structure of the construct

PSIPRED (http://bioinf.cs.ucl.ac.uk/psipred/) was used for generating the secondary structure of the vaccine protein [50]. Next, another webserver was Phyre2 server [51] (www.sbg.bio.ic.ac.uk/~phyre2/html/page.cgi?id=index) employed to predict the secondary structure.

### 2.10. Refinement of the tertiary structure of protein and their validation

The GalaxyRefine web servers (http://galaxy.seoklab.org/cgi-bin/submit.cgi?type=REFINE) were used to improve the three-dimensional protein model received from the trRosetta server [52]. The ProSA-web server (https://prosa.services.came.sbg.ac.at/prosa.php) and SAVES server v6.0 (https://saves.mbi.ucla.edu/) have both verified the model's overall accuracy [45]. MolProbity (http://molprobity.biochem.duke.edu/) offers Ramachandran analysis [53]. RAMPAGE (http://mordred.bioc.cam.ac.uk/~rapper/rampage.php) is another freely accessible server [54].

### 2.11. Prediction of discontinuous epitopes in B cells

In a validated protein model, ElliPro (http://tools.iedb.org/ElliPro/) [55] was used to predict discontinuous epitopes of B cells. Over 90 % of B-cell epitopes are thought to be discontinuous based on a few reports. The free online server primarily combines three techniques to stabilize the protein's structure before calculating the index of residue prominence. In turn, this causes nearby residues to group together based on their PI levels. With an estimated AUC score of 0.732, ElliPro is one of the best servers in its class.

### 2.12. Molecular docking of the vaccine construct with the human TLR complex

We performed ligand-receptor docking analysis using the ClusPro2.0 website (https://cluspro.org/help.php) to evaluate the relationship between vaccinations and human Toll-like receptor-4 (4G8A) [56]. The server only shows the first 10 models by default. Finally, PYMOL2.5.3 software was used for visual analysis to calculate the protein-protein binding interactions of the vaccination designs with the human TLR4/MD2 complex.

### 2.13. Characterization of immune simulation profiling of the peptide-based vaccine

The C-ImmSim server (https://150.146.2.1/C-IMMSIM/index.php) has been used to represent the immune response profile of the designed vaccine construct [57], which defines cellular and humoral response within the mammalian immune system in the presence of antigenic or immunogenic components of bacteria, viruses, etc., at the sub-cellular level (mesoscopic scale). Other crucial variables, such as the simulation volume, random seed, and step, were set at 30, 12345, and 1000, respectively [58].

The minimum advised time between the first and second dosage for the majority of vaccines now in use is four weeks. The "time-step" scale is used by the C-ImmSim server to determine how long simulations should last. Each time step on this scale corresponds to 8 h in real life. The three injection points were placed at time steps 1, 84, and 168, respectively, for a total of 1050 time steps in the simulation [59]. The remaining settings were predetermined by default.

### 2.14. In-silico cloning

To express the designed multi-epitope structure in selected expression vectors, it is necessary to reverse-translate the structure and optimize the codon. To do this, we used the online Web server Java Codon Adaptation tool (JCAT) (http://www.jcat.de). The final construct is expressed in *E. coli* (strain K12) because the natural host *P. anaerobius* is different from that strain. Outputs obtained from the tool include the codon adaptation index (CAI) and the percentage of GC content indicating the level of protein expression. Codon usage bias is represented by CAI, where a score of 1 is considered ideal and a score greater than 0.8 is considered good and acceptable. The optimal range for good GC content is between 30 % and 70 %, as scores outside this range indicate adverse effects on transcriptional and translational performance [60]. Using SnapGene software, the designed vaccine sequence was converted into the appropriate host vector pET-28a (+).

## 3. Results

### 3.1. General characterization of Peptostreptococcus anaerobes PCWBR2 sequence

The amino acid sequence of *Peptostreptococcus anaerobes* PCWBR2 was acquired from the NCBI database for the construction of a multi-epitopic vaccine. Based on the PCWBR2 amino acid sequence alignment and the evolutionary relationships between *P. anaerobius* and other species, we created a phylogenetic tree (Supplementary Fig. 1). BLASTP analysis showed that *P. anaerobius* PCWBR2 was moderate identified with *P. stomatis* (42.53 %), *P. russellii* (41.76 %), *P. canis* (40.2 %), *P. porci* (40.38 %), and low identity with *Clostridioides mengenotii* (33.88 %), *Clostridioides difficile* (34.6), *Romboutsia maritimum* (33.2 %) and *Asaccharospora irregularis* (34.3 %). Following the splitting of the signal peptide, SignalP 6.0 predicted the functional sequences of the proteins, and the findings show that the proteins possess secretory signal peptides (Supplementary Table 1). The precise location of the protein within the cell wall was predicted using PSORTb (Supplementary Table 2). DeepLoc predicted the localization of the protein according to the protein localized in the extracellular compartment (Supplementary Table 3). Secretory or outer membrane proteins of pathogens may be an ideal candidate target, but not cytoplasmic and inner membrane proteins. We employed two suggested parameters for Vaxign prediction: adhesin probability >0.51 and surface-exposed proteins (sub-cellular location in the cell wall, outer membrane, or extracellular space) [61]. For this protein PCWBR2, the vaxign-ML Score was 99.1(>90 %), the location was in Cellwall (probability = 92 %), the adhesin probability was 0.671 (>0.51), and there was 1 transmembrane helix, with no protein sequence similarity to humans, mice, and pigs (Supplementary Table 4). Adhesion-like molecules are often ideal vaccine targets.

### 3.2. Prediction of cytotoxic T lymphocytes (CTL) epitopes of PCWBR2

With a prediction score better than 0.75 and a high likelihood of binding to the MHC class-I alleles, the NetCTL v1.2 server predicted 121 possible CTL epitopes for the MHC class-I supertype. The IEDB class I immunogenicity tool indicated that 40 of them would be immunogenicity positive. Only 20 of the 40 immunogenic epitopes that underwent VaxiJen analysis were able to surpass the antigenicity criterion of 0.4. Twelve of these 20 epitopes were determined by the ToxinPred and AllerTOP 2.0 servers, respectively, to be non-allergenic and non-toxic. All were determined to be non-toxic by ToxinPred and twelve of these 20 epitopes were determined by the AllerTOP 2.0 servers to be non-allergenic. Out of these 12 epitopes, the IEDB MHC-I allele binding prediction tool returned 10 epitopes within the ≤2 percentile rank. Out of which, most antigenic 6 epitopes (SKDGYPIVL, RYDTNLAIL, KVNDKVWTV, YPNVVVERL, KTELNVTDK, and GQAGVRNIV) had been chose for vaccine construction (Fig. 1, Table 1).

### 3.3. Prediction of helper T lymphocytes (HTL) epitopes of PCWBR2

Based on the $IC_{50}$ scores, the NetMHCII 2.3 web server suggested 748 MHC-II epitopes with the highest binding to the human alleles. The NetMHCIIpan 3.2 web server returned 58 high binding epitopes within

**Table 1**
Identified CTL epitopes using different tools with their immunogenicity, antigenicity, toxicity, and allergenicity.

| Protein | Peptide Sequence | Prediction Score | Immunogenicity Score | Antigenicity | Toxicity | Allergenicity | Percentile Rank |
|---|---|---|---|---|---|---|---|
| EKX94456.1 | SKDGYPIVL | 2.1233 | 0.1637 | 0.8299 | Non-Toxin | non-allergen | 0.71 |
| | RYDTNLAIL | 1.4825 | 0.1403 | 0.4937 | Non-Toxin | non-allergen | 0.21 |
| | KVNDKVWTV | 1.2326 | 0.0587 | 0.6878 | Non-Toxin | non-allergen | 0.01 |
| | YPNVVVERL | 1.1532 | 0.23324 | 0.8348 | Non-Toxin | non-allergen | 0.02 |
| | KTELNVTDK | 0.8322 | 0.09962 | 1.0485 | Non-Toxin | non-allergen | 0.54 |
| | GQAGVRNIV | 0.7626 | 0.20802 | 0.7169 | Non-Toxin | non-allergen | 0.99 |

the ≤50 nm affinity and strong binder threshold ≤2.00. 18 epitopes of them were antigenic according to the VaxiJen threshold (0.5). According to the ToxinPred server, 18 antigenic epitopes were non-toxic. The AllerTOP 2.0 server had determined that just seven of them were non-allergenic. For the creation of the vaccine, the five most antigenic epitopes (IDLHDVVTDKTELNV, DLHDVVTDKTELNV, GGHEYIVTDAAKASY, LHDVVTDKTELNVTD, and GYPIVLADGKLNADQ) out of the seven had been selected for the construction of the vaccine (Table 2).

### 3.4. Prediction of linear B lymphocytes (LBL) epitopes of PCWBR2

To destroy certain disease invaders, B cells release specialized antibodies [62]. They guarantee long-term immune protection by differentiating into long-lived plasma cells and memory B lymphocytes [35,63]. Linear B lymphocyte epitopes (LBL) play a crucial part in the development and manufacturing of vaccines. These epitopes are referred to as antigenic determinants that the immune system can recognize. Additionally, the B cells bind to these antigen fragments and trigger an immunological response [64]. Reliable B cell epitope prediction using multiple computer approaches is crucial for vaccine design [65]. iBCE-EL server predicted 282 probable LBL epitopes with a prediction score greater than 0.8. Of those, 172 were antigenic according to the VaxiJen standard (0.5). The ToxinPred server predicted that all 172 of these epitopes were non-toxic, and the AllerTOP 2.0 server suggested that 90 of them were non-allergenic. Out of the 90, the 6 epitopes with the highest antigenicity (QYKVTNNEGIGGDDYKKTDD, YKVTNNEGIGGDDYKKTDDI, GKQYKVTNNEGIGGDDYKKT, KQYKVTNNEGIGGDDYKKTD, ALGKQYKVTNNE-GIGGDDYK, LGKQYKVTNNEGIGGDDYKK) were chosen to construct the vaccine (Table 3).

### 3.5. Molecular docking between T lymphocyte epitopes and MHC alleles

Six CTL epitopes and five HTL epitopes recognized 18 MHC alleles in total. Except for only three CTL epitopes (SKDGYPIVL, KTELNVTDK, and GQAGVRNIV) and all the HTL epitopes, the rest had more than one binding allele (Tables 4–5). Some had even as high as 6 (YPNVVVERL) or 8 (KVNDKVWTV) alleles (Table 4). Out of these 11 epitopes and their 8 allelic partners, we decided to perform molecular docking between each epitope and one of their corresponding alleles for representative docking analysis. The docking of T lymphocyte epitopes and their MHC allele is shown in Table 6. These 6 alleles and their PDB IDs were HLA-A*24:02 (PDB ID- 7JYV) [66], HLA-A*02:01 (PDB ID-4U6Y) [67], HLA-A*11:01 (PDB ID-7S8S) [68], HLA-A*02:06 (PDB ID-3OXR) [69], HLA-DRB1*03:01(PDB ID-6BIX), and HLA-DRB3*01:01(PDB ID-2Q6W) [70]. The rest 2 alleles (HLA-B*40:01, HLA-B*53:01) were modeled using SWISS-MODEL. Table 6 shows the results of the molecular docking obtained from the AutoDock Vina. The docking score for the Autodock Vina is Binding affinity, with smaller values indicating stronger binding. There is usually a rule of thumb: Binding affinity $> -4$ kcal/mol indicates very weak or no binding; Binding affinity $<= -7$ kcal/mol indicates moderate binding; Binding affinity $<= -7$ kcal/mol indicates strong binding. The results showed that 6 epitopes had strong binding to their corresponding alleles and the epitope GGHEYIVTDAAKASY showed the strongest affinity (−8.9 kcal/mol) for its corresponding MHC allele.

### 3.6. Population coverage prediction

Combined population coverage of the CTL and HTL epitopes used in this vaccine construction was analyzed with IEDB population coverage analysis. The distribution of their 18 corresponding MHC alleles was assessed in 17 geographic regions and 101 countries identified in the IEDB database. The global coverage of our vaccine stands at 90.26 % (Supplementary Fig. 2).

### 3.7. Construction of multi-epitope vaccine against PCWBR2

The six CTL epitopes, five HTL epitopes, and six LBL epitopes identified in Tables 1–3 were constructed for multi-epitope chimeric vaccines. CTL epitopes were joined with AAY linkers, while HTL and LBL epitopes were linked with GPGPG linkers. These linkers help prevent forming of potential junctional epitopes and promote immune processing [71]. Additionally, MPLA is attached to the amino terminus of the peptide sequence and via the EAAAK linker. MPLA is a TLR4 agonist and has been proven as an efficient adjuvant [72]. A 6x His tag was inserted for protein identification and purification at the C-terminus of the vaccine sequence.

To optimize the order of CTL, HTL, and LBL epitopes, we made six constructs as shown in Supplementary Table 5, and the efficiency of the three-dimensional structures of all constructs was examined according to Z-score, ERRAT, and Ramachandran plot analysis. As shown in Supplementary Table 6, the order of CTL, HTL, and LBL yielded the highest score and we decided to follow with this construct (illustrated in Fig. 2).

It is reported that partial filtration occurs in the kidney of particles between 7 and 70 kDa. To avoid filtration by kidney, we increase the molecular weight by attaching albumin or IgG Fc to the vaccine construct 6 as an adjuvant, respectively, and then evaluated it for its physicochemical properties, antigenicity, and toxicity. Preliminary results show that the antigenicity of the albumin/IgG Fc-attached vaccines yields an antigenicity score of 0.6284 and 0.9691, respectively (Supplementary Tables 7–9), which are much lower than that of MPLA as an adjuvant. The results showed that MPLA had some advantages over

**Table 2**
Identified HTL epitopes using IEDB MHC-II binding tool and NetMHCIIpan server with their toxicity, antigenicity, and GRAVY value.

| Gene name | Epitope | Affinity | Antigenicity | Toxicity | Allergenicity |
|---|---|---|---|---|---|
| EKX94456.1 | DLHDVVTDKTELNVT | 1.6 | 0.8743 | Non-Toxin | non-allergen |
| | LHDVVTDKTELNVTD | 1.6 | 0.8461 | Non-Toxin | non-allergen |
| | GGHEYIVTDAAKASY | 1.5 | 0.8469 | Non-Toxin | non-allergen |
| | GYPIVLADGKLNADQ | 0.15 | 0.5611 | Non-Toxin | non-allergen |
| | IDLHDVVTDKTELNV | 1.9 | 1.0831 | Non-Toxin | non-allergen |

**Table 3**
Identified LBL epitopes using iBCE-EL server with their toxicity, antigenicity, and GRAVY value.

| Gene name | Epitope | Antigenicity | Toxicity | Allergenicity |
|---|---|---|---|---|
| EKX94456.1 | QYKVTNNEGIGGDDYKKTDD | 2.0481 | Non-Toxin | non-allergen |
| | YKVTNNEGIGGDDYKKTDDI | 1.8804 | Non-Toxin | non-allergen |
| | GKQYKVTNNEGIGGDDYKKT | 1.84 | Non-Toxin | non-allergen |
| | KQYKVTNNEGIGGDDYKKTD | 1.8322 | Non-Toxin | non-allergen |
| | ALGKQYKVTNNEGIGGDDYK | 1.5427 | Non-Toxin | non-allergen |
| | LGKQYKVTNNEGIGGDDYKK | 1.5124 | Non-Toxin | non-allergen |

**Table 4**
Selected CTL and their corresponding MHC alleles.

| CTL epitopes | MHC I binding alleles |
|---|---|
| SKDGYPIVL | HLA-B*40:01 |
| RYDTNLAIL | HLA-A*24:02, HLA-A*23:01 |
| KVNDKVWTV | HLA-A*02:06, HLA-A*02:01, HLA-A*02:03, HLA-A*32:01, HLA-A*68:02, HLA-A*30:01, HLA-B*08:01, HLA-A*31:01 |
| YPNVVVERL | HLA-B*53:01, HLA-B*35:01, HLA-B*51:01, HLA-B*07:02, HLA-B*08:01, HLA-A*68:02 |
| KTELNVTDK | HLA-A*11:01 |
| GQAGVRNIV | HLA-A*02:06 |

**Table 5**
Selected HTL epitopes and their corresponding MHC alleles.

| HTL epitopes | MHC II binding alleles |
|---|---|
| DLHDVVTDKTELNVT | HLA-DRB1*03:01 |
| LHDVVTDKTELNVTD | HLA-DRB1*03:01 |
| GGHEYIVTDAAKASY | HLA-DRB3*01:01 |
| GYPIVLADGKLNADQ | HLA-DRB1*03:01 |
| IDLHDVVTDKTELNV | HLA-DRB1*03:01 |

**Table 6**
CTL epitopes and their corresponding MHC alleles chosen for docking analysis.

| CTL epitopes | MHC I alleles | Binding affinity (kcal/mol) |
|---|---|---|
| SKDGYPIVL | HLA-B*40:01 | −6.9 |
| RYDTNLAIL | HLA-A*24:02(**7JYV**) | −8.0 |
| KVNDKVWTV | HLA-A*02:01(4U6Y) | −8.4 |
| YPNVVVERL | HLA-B*53:01 | −7.3 |
| KTELNVTDK | HLA-A*11:01(7S8S) | −6.7 |
| GQAGVRNIV | HLA-A*02:06(3OXR) | −8.0 |
| DLHDVVTDKTELNVT | HLA-DRB1*03:01(6BIX) | −6.7 |
| LHDVVTDKTELNVTD | HLA-DRB1*03:01 | −6.4 |
| GGHEYIVTDAAKASY | HLA-DRB3*01:01(2Q6W) | −8.9 |
| GYPIVLADGKLNADQ | HLA-DRB1*03:01 | −5.3 |
| IDLHDVVTDKTELNV | HLA-DRB1*03:01 | −7.6 |

human albumin and IgG Fc. The antigenicity of the MPLA-conjugated vaccine might be strong enough to overcome the disadvantage of being filtered by the kidney, a hypothesis supported by previous observations including a 29-kDa tuberculosis vaccine MP3RT that elicits a satisfactory immune response in wet lab [73]. It would be interesting to test all three adjuvants above in wet lab.

To investigate whether the permutation of epitopes within each subtypes affects the evaluation of vaccine constructs, we rearranged the epitopes within each of the CTL, HTL, and LBL subtypes based on vaccine construct 6, respectively (Supplementary Tables 10–12), and then evaluated their antigenicity through the VaxiJen 2.0 online server and stereochemical stability through the Ramachandran plot and ERRAT analyses, respectively. The antigenicity analysis showed that only permutation of the CTL epitopes affected the antigenicity of the vaccine construct, and the effect is slight. The permutation of epitopes within the HTL or LBL did not affect the antigenicity (Supplementary Tables 13–15). The stereochemical stability analysis showed that although the Ramachandran plot and ERRAT values were affected by epitope permutations, the original vaccine 6 construct is still the best (Supplementary Tables 16–18).

### 3.8. Prediction of antigenicity, allergenicity, solubility, and physicochemical properties of the vaccine

ExPASY ProtParam estimated that the final chimeric construct would have a molecular weight (MW) of 56 kDa and consist of 516 amino acids. The theoretical isoelectric point value (pI) of the protein was calculated to be 5.55. Proteins are thought to have a slightly acidic nature based on this score. The subject protein's half-life was shown to be 30 h in mammalian reticulocytes in vitro, more than 20 h in yeast, and more than 10 h in *E. coli* in vivo. Since values greater than 40 imply instability, the protein's estimated instability index of 12.28 places it in the stable model category. GRAVY, the grand average of hydropathicity for the protein was predicted to be −0.652. The GRAVY value of the hydrophilic peptide was negative, and the GRAVY value of the hydrophobic peptide was positive. Only the chimera vaccine's hydrophilic epitopes could stimulate the host cells' immunological response [74]. The protein is hydrophilic and may easily interact with the water molecules, as indicated by the negative value. Therefore, while designing vaccines, only hydrophilic epitopes are considered. The protein was determined to be soluble in its expression based on Protein-Sol's estimation [75], with a solubility score of 0.543. The protein's thermostability was confirmed by an estimate of its aliphatic index of 64.05.

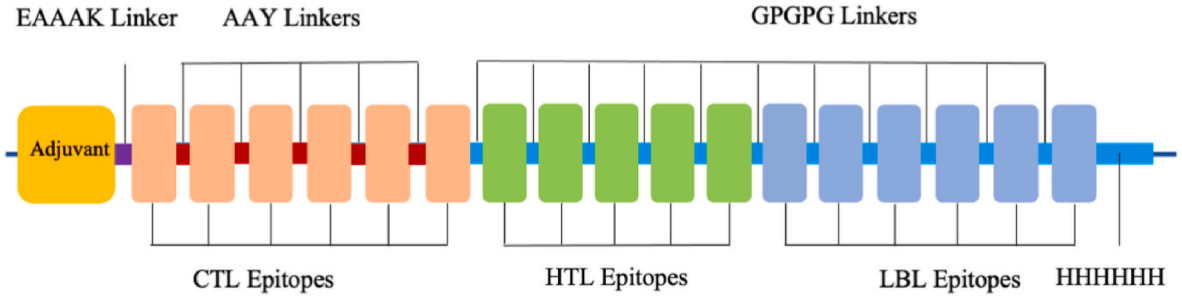


**Fig. 2.** Schematic presentation of the final multi-epitope vaccine protein. The EAAAK linker (purple) connects the 516 amino acids long protein sequence, which has an adjuvant (orange) at the amino-terminal end. GPGPG linkers (blue) are used to link B-cell epitopes and HTL epitopes, while AAY linkers (red) are used to link CTL epitopes. The Carboxy terminus is purified and identified by the addition of a 6x-His tag.

The VaxiJen 2.0 online server predicted the antigenicity of the vaccine design attached with an adjuvant, and by using a threshold of 0.4, it provided a satisfactory antigenicity score of 1.0193 with the bacterial model. The vaccine candidate's antigenicity was also examined without considering the adjuvant component, for which VaxiJen provided scores of 1.2989. A non-antigenic epitope with a predicted antigenicity value of less than 0.4 cannot elicit an immunological response from the host cell. These findings suggest that the vaccination protein sequence, whether it is combined with an adjuvant, is naturally antigenic. The vaccination sequence was projected to be non-allergenic in nature by AllerTOP v.2 online servers in both the presence and absence of the adjuvant.

Results of the physicochemical analyses, antigenicity, and allergenicity have been listed in Table 7.

### 3.9. Prediction of secondary structure extrapolation of the vaccine

To forecast the secondary structure of the vaccine protein, the PSIPRED algorithm was employed. Fig. 3 depicts the secondary structure, which consists of 4 % of residues, 20 % beta-strand, and 7 % alpha helix.

### 3.10. Refinement of the tertiary structure model and validation of the vaccine

The unprocessed chimeric vaccine's tertiary structure model (Fig. 4A) from the trRosetta server was initially improved using GalaxyRefine, which offered five models. Model 1 was chosen from among these models based on several factors, including GDT-HA (0.9532), RMSD (0.406), and MolProbity (2.087) (Fig. 4B). According to the calculations, the clash score was 13.2, and the poor rotamers score was 0.9. Thus, this model was chosen in the end to serve as the chimeric model for subsequent research.

We further investigated the stereochemical stability through the Ramachandran plot and ERRAT analyses. The Ramachandran plot analysis for the refined vaccine predicted a score of 95.3 %, wherein >90 % of the constituent amino acid residues were detected in the favored regions, demonstrating that the optimized vaccine structure has good stereochemical quality. Additionally, utilizing SAVES, it was discovered that 3.9 % of the residues were present in extra permissible locations and just 0.8 % in disallowed regions (Fig. 4C). An ERRAT score of 90.878 further validates the protein's structural validity (Fig. 4E). A value higher than 50 denotes a high-quality model. The ProSA-web server calculates the overall quality score for the input structure and displays it alongside all the protein structures that are currently known.

**Table 7**
Antigenicity, allergenicity, solubility, and other physicochemical property evaluations of the primary protein sequence of the multi-epitopic potent vaccine candidate.

| Sl. No. | Features | Assessment | Remark |
|---|---|---|---|
| 1 | Number of amino acids | 516 | Suitable |
| 2 | Molecular weight | 56165.26 | Average |
| 3 | Chemical formula | $C_{2537}H_{3777}N_{675}O_{767}S_5$ | – |
| 4 | Theoretical pI | 5.55 | Slightly acidic |
| 5 | Total number of negatively charged residues (Asp + Glu) | 63 | – |
| 6 | Total number of positively charged residues (Arg + Lys) | 50 | – |
| 7 | Total number of atoms | 7761 | – |
| 8 | Instability Index (II) | 12.28 | Stable |
| 9 | Aliphatic index (AI) | 64.05 | Thermostable |
| 10 | Grand Average of hydropathicity (GRAVY) | −0.652 | Hydrophilic |
| 11 | solubility | 0.543 | soluble |
| 12 | Antigenicity | 1.0193 | Antigenic |
| 13 | Allergenicity | non-allergen | Non-allergen |

The black dot in Fig. 4D represents the vaccine protein Z score, which was calculated by the ProSA server and is −3.77. The study's validation methods all pointed to the proposed vaccine candidate's high quality.

### 3.11. Prediction of discontinuous epitopes in B cells

A total of 292 residues were identified by the prediction of four discontinuous B-cell epitopes, with scores ranging from 0.529 to 0.737. According to Fig. 5 and Supplementary Table 19, these epitopes ranged in size from ten to two hundred and fifty-seven residues.

### 3.12. Prediction of molecular docking of the vaccine with the human TLR4

A mammalian Toll-Like Receptor family member of pathogen pattern recognition molecules, TLR4 is a 100 kDa type I transmembrane glycoprotein [76]. The 25 kDa secreted protein MD-2, often referred to as ESOP-1, is necessary for TLR4-mediated reactions to bacterial lipopolysaccharide (LPS) [77]. Immune responses to microbial infection are strongly elicited by the human TLR4/MD2 complex [78]. From the ClusPro2.0 results, the lowest binding energy required for the vaccine to TLR4 receptors was −1265.5 kcal/mol (visualized in Fig. 6). Here, molecular docking reveals the interaction of a candidate vaccine with TLR4 on the surface of human cells. To compare the binding energy of the vaccine with the human TLR4, we used albumin as a control. The vaccine-albumin binding energy is −597.4 kcal/mol which is much higher than that of vaccine-TLR4, demonstrating that the vaccine-TLR4 binding is relatively stable.

### 3.13. Characterization of the immune profile of the vaccine

To analyze the immune response generated by the final chimeric vaccine construct, the immune simulator C-ImmSim generated simulations that matched the real response formed by the immune system. As expected, high levels of the immune response are produced after the secondary and tertiary immune responses, containing HTL, CTL, and other associated immune cells (DC cells, macrophages (MA), etc.) (Fig. 7). As shown in Fig. 7A, high levels of immunoglobulin IgG + IgM and IgG1 + IgG2 were produced after the tertiary immune stimulation. Memory formation in the immune system after repeated immune stimulation (Fig. 7B). Additionally, it was shown that certain B-cell isotypes persisted for extended periods, suggesting the possibility of isotype switching that may be scaled up at 650–700 cells/mm3 (Fig. 7B and C). Concerning each population's memory development, an enhanced response was seen in the CTL and HTL populations (Fig. 7D–F). Furthermore, the T cells showed a great deal of diversity among the T cell population both during the resting and active phases (Fig. 7F). The successful administration of the vaccine candidate effectively elevated added regulatory elements of the immune system (e.g., interleukins and cytokines) (Fig. 7I). Additionally, whereas dendritic cell activity was determined to be steady (Fig. 7H), an increase in macrophage activity was seen (Fig. 7G). IFN-g and IL-2 concentrations were likewise very high (Fig. 7I). Additionally, a lower Simpson index (D) suggested that different immunological responses might be possible. These findings suggest that the *P. anaerobius* vaccine being developed could be regarded as a powerful peptide-based next-generation vaccine for inducing a strong immune response to fight *P. anaerobius* infection.

### 3.14. In silico cloning

By using *in silico* cloning, the cloning and expression effectiveness of the multi-epitope vaccination construct in the expression vector were examined. The Java codon adaptation tool was used to optimize the codon usage of vaccine constructs in *E. coli* (strain K12) [60]. The length of the codon sequence that was optimized was 1548 nucleotides. The multi-epitope vaccine's optimized nucleotide sequence had a codon

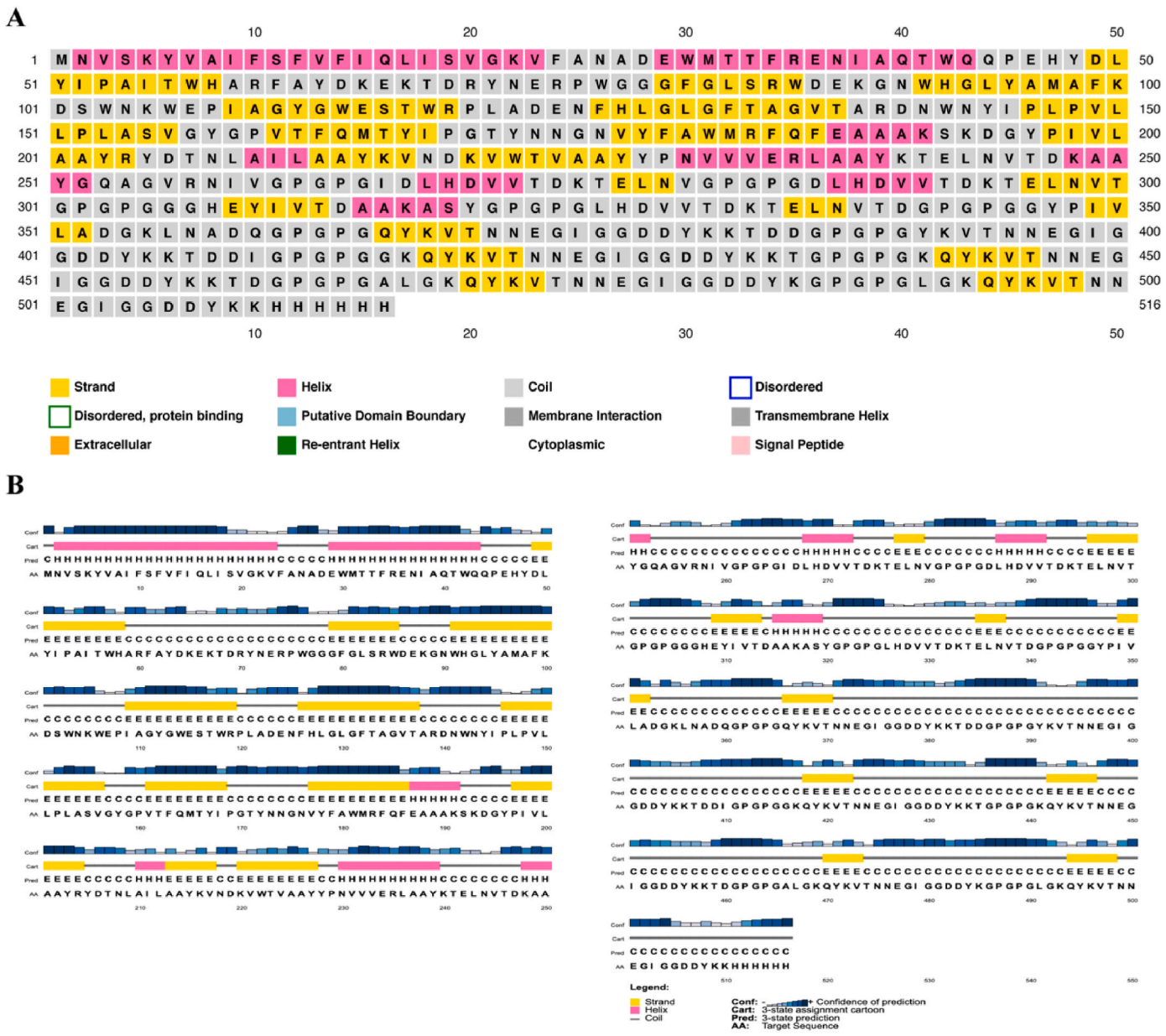


**Fig. 3.** The secondary structure analysis of the result of the multi-epitopic vaccine by the PESIPRED webserver.

adaptation index (CAI) of 0.98 and a GC content of 52.73 %, respectively, the possibility of highly efficient expression of the designed vaccine in the host (Supplementary Fig. 3). Finally, restriction cloning was performed to insert an adaptive codon sequence (multi-epitope vaccine sequence) in the *E. coli* pET28a (+) vector for optimal gene expression using the SnapGene tool (Fig. 8).

## 4. Discussion

Vaccines have been one of the greatest public health achievements of the last century, saving an estimated 2 to 3 million lives each year from deadly pathogens [79]. In recent years, a series of tumor-promoting bacteria were identified, which expanded the horizon of potential vaccine applications. These vaccines, termed oncomicrobe vaccine can prevent tumor initiation or alleviate tumor development (see review [80]). At present, there is no effective vaccine for *P. anaerobius*. This study aimed to design a multi-epitope vaccine against *P. anaerobius* using immunoinformatics and reverse vaccinology strategies. Immunoinformatics offers the advantages of low-cost and rapid identification and screening of epitopes and design of vaccines [59] and reverse vaccinology (RV) is a strategy that overcomes the limitations of forward vaccinology methods. RV has been widely applied to many deadly pathogens, leading to the development of the first successful *Neisseria meningitidis* vaccine [81], and in a study of an immunoinformatics-designed vaccine against *Mycobacterium tuberculosis*, the predicted epitopes were consistent with the actual epitopes from animal experiments [73]. The essence of immunoinformatics and RV is the prediction of T-cell and B-cell epitopes. Constructed vaccines containing only these predicted epitopes will avoid unnecessary antigen loading and reduces the chance of anaphylaxis [82].

Why native PCWBR2 on the surface of *P. anaerobius* failed to elicit a proper immune response to eliminate *P. anaerobius* in the gut is an interesting question. Recent studies have showed that regulatory T (Treg) cells and T helper 17($T_h17$) cells were increased in $Apc^{min/+}$ mice when infected with *P. anaerobic*, and that Treg and Th17 cells were highly correlated with intestinal inflammation and CRC [83]. *P. anaerobius* affects many kinds of immune cells, especially the immune-suppressive myeloid-derived suppressor cells (MDSCs),

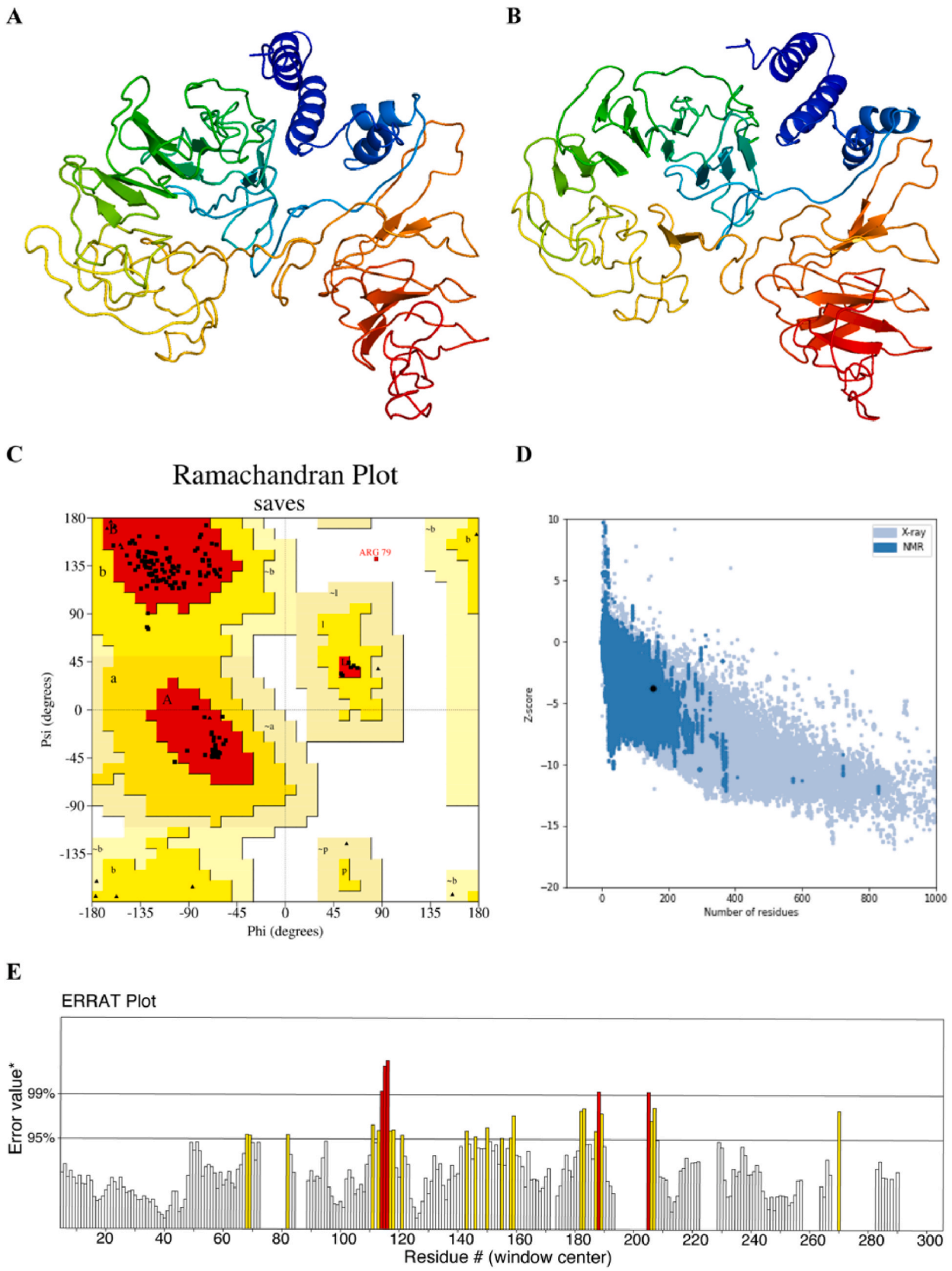


**Fig. 4. Protein modeling refinement and validation.** (A) The final 3D model of the multi-epitope vaccine was obtained after homology modeling on I-TASSER. (B) The refined model was obtained via GalaxyRefine. Validation of the refined model with (C) Ramachandran plot analysis showing 95.3 %, 3.9 %, and 0.8 % of protein residues in favored, allowed, and disallowed (outlier) regions respectively. (D) ProSA-web, giving a Z-score of −3.77. (E) Protein structure validation by ERRAT score of 90.878, where the X-axis denotes residues and Y axis denotes error values.

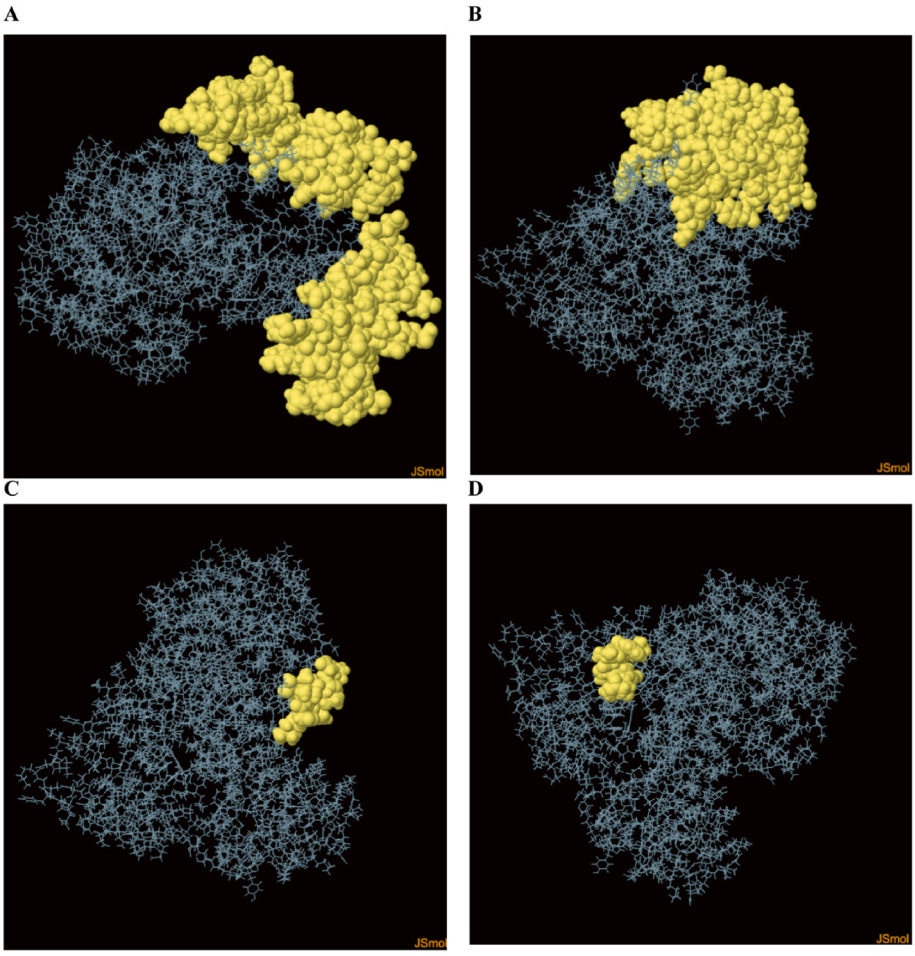


**Fig. 5. Discontinuous B-cell epitopes predicted by the ElliPro.** (**A–D**) Three-dimensional representation of conformational or discontinuous epitopes of the highest antigenic chimeric protein of *Peptostreptococcus anaerobius*. The epitopes are represented by yellow surfaces, and the bulk of the protein is represented in grey sticks.

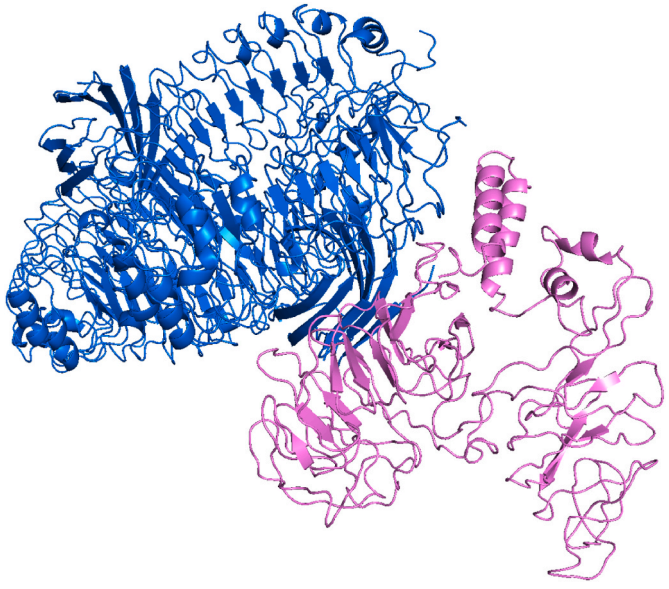

**Fig. 6.** Docked complex of vaccine construct with TLR4 complex. TLR-4. The blue color shows each receptor while the pink color is the final vaccine structure.

tumor-associated macrophages (TAMs), and granulocytic tumor-associated neutrophils (GTANs), which inhibits immune function and promote tumor progression [10]. The immunosuppressive cytokines IL-6 and IL-10 produced by MDSCs directly inhibits the activity of CD4+T cells [84], a cell type that participates in humoral immunity against bacterial pathogens. The above research may have explained how *P. anaerobius* form an immunosuppressive microenvironment in the tumor and evade immunity. Therefore, a vaccine against *P. anaerobius* is needed to elicit a proper immune response to eliminate the bacteria.

In this study, we designed a PCWBR2 subunits-based vaccine against *P. anaerobius*. PCWBR2 interacts with α2/β1 integrins, an overexpressed receptor in human CRC tumors (ref [10]). Although there is a small possibility that the designed vaccine can also interact with the α2/β1 integrins in vitro, when the vaccine enters the circulation system, it is expected to be engulfed by antigen-presenting cells, and therefore the vaccine is unlikely to reach the intestinal epithelium and bind to the colonic epithelial cells. Nevertheless, it would be essential to test in wet lab whether the designed vaccine interact with the α2/β1 integrins in vitro in future studies.

To summarize, we executed an immunoinformatics-assisted design of a multi-epitope vaccine against *P. anaerobius* adhesion PCWBR2. The resulting vaccine exhibits satisfactory characteristics and deserves to be validated experimentally in the future.

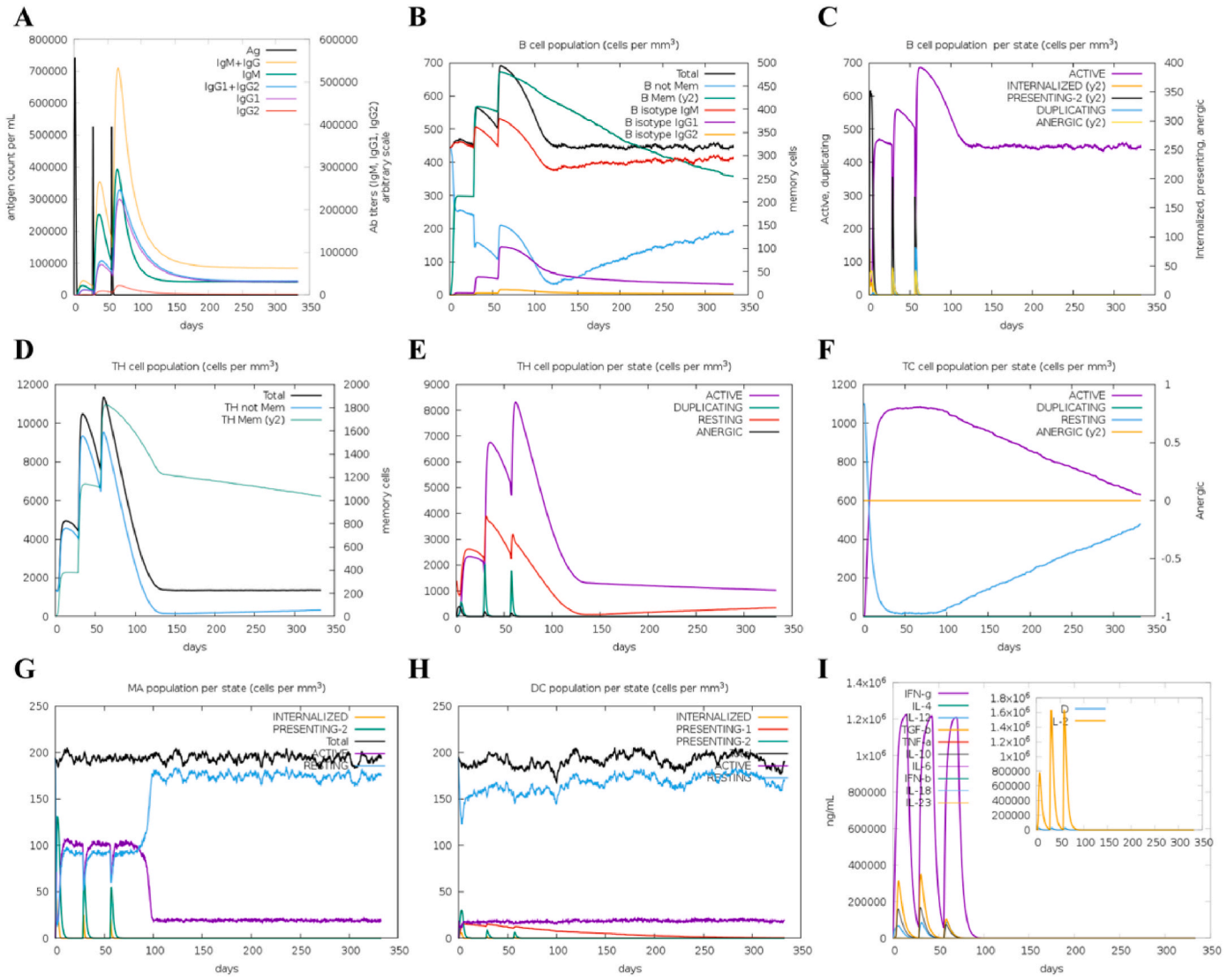


**Fig. 7.** The immune response of humans after injection of our *P. anaerobius* vaccine construct. The study noted different results from the *In silico* simulation and machine learning approaches. (A) The simulation shows the contract can elevate immunoglobulins. The noted elevation of immunoglobulins at different concentrations of antigen. (B) The study indicates the population of B lymphocytes (IgM, IgG1, and IgG2) after three injections of our vaccine construct. (C) The figure depicts the analysis outcome of the population per entity-state (i.e., showing counts for active, presenting on class-II, internalized the Ag, duplicating and antigenic by the different color variants. (D) It shows the total count of the TH cell population along with memory cells and is subdivided into isotypes IgM, IgG1, and IgG2 after injection of our vaccine. (E) It shows the population per entity-state of Helper T cell count in the resting and active states after injection of our vaccine. (F) The figure illustrates the Cytotoxic T lymphocyte population in different states; resting and active, in the time (days) after injection of our vaccine construct. (G) It shows the population of macrophages after vaccination. (H) The behavior of the population of Dendritic cells in the active and resting states after injection of our vaccine. (I) The concentration of cytokines and interleukins with Simpson index. Act = active, Intern = the Ag, Pres II = presenting on MHC II, Dup = in the mitotic cycle, Anergic = anergic, Resting = not active.

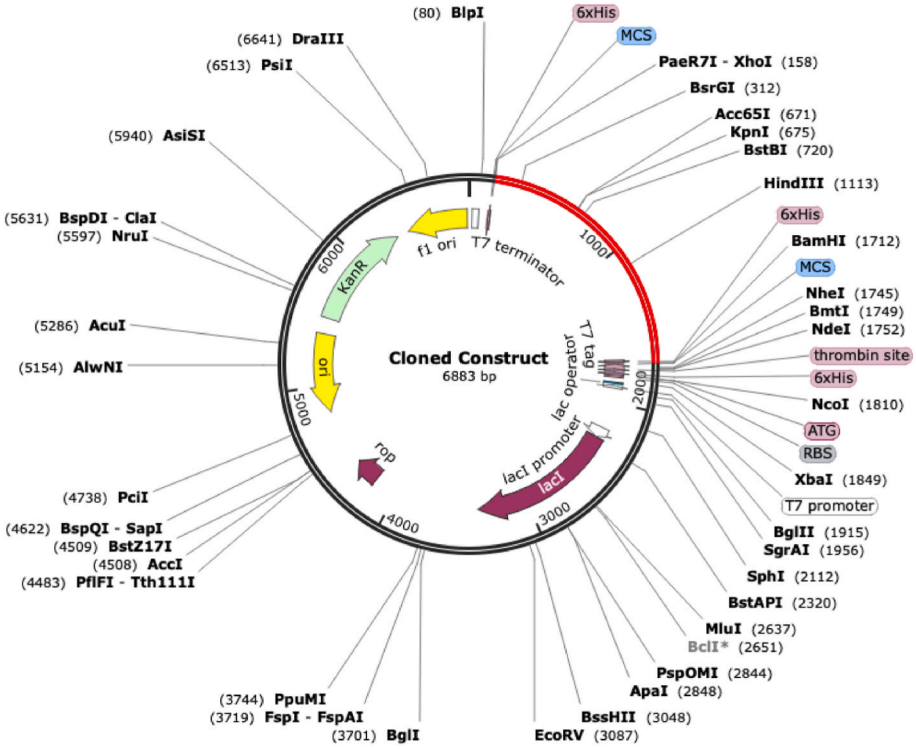


**Fig. 8.** In-silico restriction cloning of the vaccine construct in pET28a (+) vector. The red part shows the gene sequence of the designed vaccine, and the black part shows the skeleton of the *E. coli* vector.

## 5. Conclusions

Given *P. anaerobius*' role in tumorigenesis, carcinogenic bacteria have been eliminated using a variety of techniques to prevent cancer or delay the formation of tumors. In our study, antigenic, immunogenic, non-toxic T-cell, and B-cell epitopes have been selected from vaccine targets and designed as a unique multi-epitope vaccine against *P. anaerobius* using a variety of immunoinformatics tools. Designed vaccine candidates have shown exciting results in computational studies and thus the vaccines hold promise for eliminating *P. anaerobius* and opening new areas of research to develop vaccines against CRC.

## Funding

This work has been supported by the National Key Research and Development Program of China (Grant No. 2020YFA0907800), the National Natural Science Foundation of China (Grant No. 32000096), the Shenzhen Science and Technology Innovation Program (KQTD20200820145822023), the Program of Shenzhen Key Laboratory of Systems Medicine for Inflammatory Diseases (ZDSYS20200606100803007), and the Foshan Science and Technology Innovation Program (2120001010795).

## Data availability statement

All data generated or analyzed during this study are included in this article.

## CRediT authorship contribution statement

**Yudan Mao:** Formal analysis, Writing – original draft, designed and performed all the experiments, analyzed the results, and drafted the manuscript, performed the experiment, prepared tables and figures. **Xianzun Xiao:** Formal analysis, Writing – original draft, designed and performed all the experiments, analyzed the results, and drafted the manuscript. **Jie Zhang:** performed the computational work. **Xiangyu Mou:** Writing – review & editing, reviewed the results and edited the manuscript, Conceptualization, Formal analysis, Writing – review & editing, Supervision, conceptualized and design the study, analyzed the results, edited as well as finalized the manuscript and supervised the study. **Wenjing Zhao:** Writing – review & editing, reviewed the results and edited the manuscript.

## Declaration of competing interest

All authors disclosed no relevant relationships.

## Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.synbio.2023.11.004.

## References

[1] Ferlay J, Soerjomataram I, Dikshit R, Eser S, Mathers C, Rebelo M, et al. Cancer incidence and mortality worldwide: sources, methods and major patterns in GLOBOCAN 2012. Int J Cancer 2015;136:E359–86.
[2] Nistal E, Fernández-Fernández N, Vivas S, Olcoz JL. Factors determining colorectal cancer: the role of the intestinal microbiota. Front Oncol 2015;5:220.
[3] Zeller G, Tap J, Voigt AY, Sunagawa S, Kultima JR, Costea PI, et al. Potential of fecal microbiota for early-stage detection of colorectal cancer. Mol Syst Biol 2014;10:766.
[4] Yu J, Feng Q, Wong SH, Zhang D, Liang QY, Qin Y, et al. Metagenomic analysis of faecal microbiome as a tool towards targeted non-invasive biomarkers for colorectal cancer. Gut 2017;66:70–8.
[5] Wu S, Rhee KJ, Albesiano E, Rabizadeh S, Wu X, Yen HR, et al. A human colonic commensal promotes colon tumorigenesis via activation of T helper type 17 T cell responses. Nat Med 2009;15:1016–22.
[6] Rubinstein MR, Wang X, Liu W, Hao Y, Cai G, Han YW. Fusobacterium nucleatum promotes colorectal carcinogenesis by modulating E-cadherin/β-catenin signaling via its FadA adhesin. Cell Host Microbe 2013;14:195–206.
[7] Cuevas-Ramos G, Petit CR, Marcq I, Boury M, Oswald E, Nougayrède JP. Escherichia coli induces DNA damage in vivo and triggers genomic instability in mammalian cells. Proc Natl Acad Sci U S A 2010;107:11537–42.

[8] Huycke MM, Abrams V, Moore DR. Enterococcus faecalis produces extracellular superoxide and hydrogen peroxide that damages colonic epithelial cell DNA. Carcinogenesis 2002;23:529–36.
[9] Tsoi H, Chu ESH, Zhang X, Sheng J, Nakatsu G, Ng SC, et al. Peptostreptococcus anaerobius induces intracellular cholesterol biosynthesis in colon cells to induce proliferation and causes dysplasia in mice. Gastroenterology 2017;152:1419–1433. e5.
[10] Long X, Wong CC, Tong L, Chu ESH, Ho Szeto C, Go MYY, et al. Peptostreptococcus anaerobius promotes colorectal carcinogenesis and modulates tumour immunity. Nat Microbiol 2019;4:2319–30.
[11] Riggio MP, Lennon A. Development of a PCR assay specific for Peptostreptococcus anaerobius. J Med Microbiol 2002;51:1097–101.
[12] Murphy EC, Frick IM. Gram-positive anaerobic cocci–commensals and opportunistic pathogens. FEMS Microbiol Rev 2013;37:520–53.
[13] van der Vorm ER, Dondorp AM, van Ketel RJ, Dankert J. Apparent culture-negative prosthetic valve endocarditis caused by Peptostreptococcus magnus. J Clin Microbiol 2000;38:4640–2.
[14] Ng J, Ng LK, Chow AW, Dillon JA. Identification of five Peptostreptococcus species isolated predominantly from the female genital tract by using the rapid ID32A system. J Clin Microbiol 1994;32:1302–7.
[15] Nakatsu G, Li X, Zhou H, Sheng J, Wong SH, Wu WK, et al. Gut mucosal microbiome across stages of colorectal carcinogenesis. Nat Commun 2015;6:8727.
[16] Lopez LR, Bleich RM, Arthur JC. Microbiota effects on carcinogenesis: initiation, promotion, and progression. Annu Rev Med 2021;72:243–61.
[17] Teufel F, Almagro Armenteros JJ, Johansen AR, Gíslason MH, Pihl SI, Tsirigos KD, et al. SignalP 6.0 predicts all five types of signal peptides using protein language models. Nat Biotechnol 2022;40(7):1023–5.
[18] Yu NY, Wagner JR, Laird MR, Melli G, Rey S, Lo R, et al. PSORTb 3.0: improved protein subcellular localization prediction with refined localization subcategories and predictive capabilities for all prokaryotes. Bioinformatics 2010;26:1608–15.
[19] He Y, Xiang Z, Mobley HL. Vaxign: the first web-based vaccine design program for reverse vaccinology and applications for vaccine development. J Biomed Biotechnol 2010;2010:297505.
[20] Baruah V, Bose S. Immunoinformatics-aided identification of T cell and B cell epitopes in the surface glycoprotein of 2019-nCoV. J Med Virol 2020;92:495–500.
[21] Larsen MV, Lundegaard C, Lamberth K, Buus S, Lund O, Nielsen M. Large-scale validation of methods for cytotoxic T-lymphocyte epitope prediction. BMC Bioinf 2007;8:424.
[22] Kim Y, Ponomarenko J, Zhu Z, Tamang D, Wang P, Greenbaum J, et al. Immune epitope database analysis resource. Nucleic Acids Res 2012;40:W525–30.
[23] Reynisson B, Alvarez B, Paul S, Peters B, Nielsen M. 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 2020;48:W449. w54.
[24] Peters B, Bulik S, Tampe R, Van Endert PM, Holzhütter HG. Identifying MHC class I epitopes by predicting the TAP transport efficiency of epitope precursors. J Immunol 2003;171:1741–9.
[25] Larsen MV, Lundegaard C, Lamberth K, Buus S, Brunak S, Lund O, et al. An integrative approach to CTL epitope prediction: a combined algorithm integrating MHC class I binding, TAP transport efficiency, and proteasomal cleavage predictions. Eur J Immunol 2005;35:2295–303.
[26] Calis JJ, Maybeno M, Greenbaum JA, Weiskopf D, De Silva AD, Sette A, et al. Properties of MHC class I presented peptides that enhance immunogenicity. PLoS Comput Biol 2013;9:e1003266.
[27] Doytchinova IA, Flower DR. VaxiJen: a server for prediction of protective antigens, tumour antigens and subunit vaccines. BMC Bioinf 2007;8:4.
[28] Gupta S, Kapoor P, Chaudhary K, Gautam A, Kumar R, Raghava GP. In silico approach for predicting toxicity of peptides and proteins. PLoS One 2013;8: e73957.
[29] Dimitrov I, Bangov I, Flower DR, Doytchinova I. AllerTOP v.2–a server for in silico prediction of allergens. J Mol Model 2014;20:2278.
[30] Zahroh H, Ma'rup A, Tambunan US, Parikesit AA. Immunoinformatics approach in designing epitope-based vaccine against meningitis-inducing bacteria (Streptococcus pneumoniae, Neisseria meningitidis, and Haemophilus influenzae type b). Drug Target Insights 2016;10:19–29.
[31] Kumar Pandey R, Ojha R, Mishra A, Kumar Prajapati V. Designing B- and T-cell multi-epitope based subunit vaccine using immunoinformatics approach to control Zika virus infection. J Cell Biochem 2018;119:7631–42.
[32] Khan M, Khan S, Ali A, Akbar H, Sayaf AM, Khan A, et al. Immunoinformatics approaches to explore Helicobacter Pylori proteome (Virulence Factors) to design B and T cell multi-epitope subunit vaccine. Sci Rep 2019;9:13321.
[33] Jensen KK, Andreatta M, Marcatili P, Buus S, Greenbaum JA, Yan Z, et al. Improved methods for predicting peptide binding affinity to MHC class II molecules. Immunology 2018;154:394–406.
[34] Nielsen M, Lund O, NN-align. An artificial neural network-based alignment algorithm for MHC class II peptide binding prediction. BMC Bioinf 2009;10:296.
[35] Do RK, Hatada E, Lee H, Tourigny MR, Hilbert D, Chen-Kiang S. Attenuation of apoptosis underlies B lymphocyte stimulator enhancement of humoral immune response. J Exp Med 2000;192:953–64.
[36] Burley SK, Berman HM, Bhikadiya C, Bi C, Chen L, Di Costanzo L, et al. RCSB Protein Data Bank: biological macromolecular structures enabling research and education in fundamental biology, biomedicine, biotechnology and energy. Nucleic Acids Res 2019;47:D464. d74.
[37] Burley SK, Bhikadiya C, Bi C, Bittrich S, Chen L, Crichlow GV, et al. RCSB Protein Data Bank: powerful new tools for exploring 3D structures of biological macromolecules for basic and applied research and education in fundamental biology, biomedicine, biotechnology, bioengineering and energy sciences. Nucleic Acids Res 2021;49:D437. d51.
[38] Schwede T, Kopp J, Guex N, Peitsch MC. SWISS-MODEL: an automated protein homology-modeling server. Nucleic Acids Res 2003;31:3381–5.
[39] Guex N, Peitsch MC, Schwede T. Automated comparative protein structure modeling with SWISS-MODEL and Swiss-PdbViewer: a historical perspective. Electrophoresis 2009;30(Suppl 1):S162–73.
[40] Lamiable A, Thévenet P, Rey J, Vavrusa M, Derreumaux P, Tufféry P. PEP-FOLD3: faster de novo structure prediction for linear peptides in solution and in complex. Nucleic Acids Res 2016;44:W449–54.
[41] Trott O, Olson AJ. AutoDock Vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. J Comput Chem 2010;31:455–61.
[42] Dallakyan S, Olson AJ. Small-molecule library screening by docking with PyRx. Methods Mol Biol 2015;1263:243–50.
[43] Bui HH, Sidney J, Dinh K, Southwood S, Newman MJ, Sette A. Predicting population coverage of T-cell epitope-based diagnostics and vaccines. BMC Bioinf 2006;7:153.
[44] Yang J, Anishchenko I, Park H, Peng Z, Ovchinnikov S, Baker D. Improved protein structure prediction using predicted interresidue orientations. Proc Natl Acad Sci U S A 2020;117:1496–503.
[45] Wiederstein M, Sippl MJ. ProSA-web: interactive web service for the recognition of errors in three-dimensional structures of proteins. Nucleic Acids Res 2007;35: W407–10.
[46] Colovos C, Yeates TO. Verification of protein structures: patterns of nonbonded atomic interactions. Protein Sci 1993;2:1511–9.
[47] Wilkins MR, Gasteiger E, Bairoch A, Sanchez JC, Williams KL, Appel RD, et al. Protein identification and analysis tools in the ExPASy server. Methods Mol Biol 1999;112:531–52.
[48] Hebditch M, Carballo-Amador MA, Charonis S, Curtis R, Warwicker J. Protein-Sol: a web tool for predicting protein solubility from sequence. Bioinformatics 2017;33: 3098–100.
[49] Hallgren J, Tsirigos KD, Pedersen MD, Almagro Armenteros JJ, Marcatili P, Nielsen H, et al. DeepTMHMM predicts alpha and beta transmembrane proteins using deep neural networks. bioRxiv 2022. https://www.biorxiv.org/content/10.1101/2022.04.08.487609v1.
[50] McGuffin LJ, Bryson K, Jones DT. The PSIPRED protein structure prediction server. Bioinformatics 2000;16:404–5.
[51] Kelley LA, Mezulis S, Yates CM, Wass MN, Sternberg MJ. The Phyre2 web portal for protein modeling, prediction and analysis. Nat Protoc 2015;10:845–58.
[52] Heo L, Park H, Seok C. GalaxyRefine: protein structure refinement driven by side-chain repacking. Nucleic Acids Res 2013;41:W384–8.
[53] Chen VB, Arendall 3rd WB, Headd JJ, Keedy DA, Immormino RM, Kapral GJ, et al. MolProbity: all-atom structure validation for macromolecular crystallography. Acta Crystallogr D Biol Crystallogr 2010;66:12–21.
[54] Lovell SC, Davis IW, Arendall 3rd WB, de Bakker PI, Word JM, Prisant MG, et al. Structure validation by Calpha geometry: phi,psi and Cbeta deviation. Proteins 2003;50:437–50.
[55] Ponomarenko J, Bui HH, Li W, Fusseder N, Bourne PE, Sette A, et al. ElliPro: a new structure-based tool for the prediction of antibody epitopes. BMC Bioinf 2008;9: 514.
[56] Kozakov D, Hall DR, Xia B, Porter KA, Padhorny D, Yueh C, et al. The ClusPro web server for protein-protein docking. Nat Protoc 2017;12:255–78.
[57] Rapin N, Lund O, Bernaschi M, Castiglione F. Computational immunology meets bioinformatics: the use of prediction tools for molecular binding in the simulation of the immune system. PLoS One 2010;5:e9862.
[58] Bhattacharya M, Chatterjee S, Nag S, Dhama K, Chakraborty C. Designing, characterization, and immune stimulation of a novel multi-epitopic peptide-based potential vaccine candidate against monkeypox virus through screening its whole genome encoded proteins: an immunoinformatics approach. Trav Med Infect Dis 2022;50:102481.
[59] Ahammad I, Lira SS. Designing a novel mRNA vaccine against SARS-CoV-2: an immunoinformatics approach. Int J Biol Macromol 2020;162:820–37.
[60] Grote A, Hiller K, Scheer M, Münch R, Nörtemann B, Hempel DC, et al. JCat: a novel tool to adapt codon usage of a target gene to its potential expression host. Nucleic Acids Res 2005;33:W526–31.
[61] Ong E, Wang H, Wong MU, Seetharaman M, Valdez N, He Y. Vaxign-ML: supervised machine learning reverse vaccinology model for improved prediction of bacterial protective antigens. Bioinformatics 2020;36:3185–91.
[62] Chokephaibulkit K, Chien YW, AbuBakar S, Pattanapanyasat K, Perng GC. Use of animal models in studying roles of antibodies and their secretion cells in dengue vaccine development. Viruses 2020;12.
[63] Parker DC. T cell-dependent B cell activation. Annu Rev Immunol 1993;11:331–60.
[64] El-Manzalawy Y, Dobbs D, Honavar VG. Silico prediction of linear B-cell epitopes on proteins. Methods Mol Biol 2017;1484:255–64.
[65] El-Manzalawy Y, Dobbs D, Honavar V. Predicting linear B-cell epitopes using string kernels. J Mol Recogn 2008;21:243–55.
[66] Hensen L, Illing PT, Bridie Clemens E, Nguyen THO, Koutsakos M, van de Sandt CE, et al. CD8(+) T cell landscape in Indigenous and non-Indigenous people restricted by influenza mortality-associated HLA-A*24:02 allomorph. Nat Commun 2021;12: 2931.
[67] Hassan C, Chabrol E, Jahn L, Kester MG, de Ru AH, Drijfhout JW, et al. Naturally processed non-canonical HLA-A*02:01 presented peptides. J Biol Chem 2015;290: 2593–603.

[68] Habel JR, Nguyen AT, Rowntree LC, Szeto C, Mifsud NA, Clemens EB, et al. HLA-A*11:01-restricted CD8+ T cell immunity against influenza A and influenza B viruses in Indigenous and non-Indigenous people. PLoS Pathog 2022;18:e1010337.
[69] Liu J, Chen KY, Ren EC. Structural insights into the binding of hepatitis B virus core peptide to HLA-A2 alleles: towards designing better vaccines. Eur J Immunol 2011; 41:2097–106.
[70] Parry CS, Gorski J, Stern LJ. Crystallographic structure of the human leukocyte antigen DRA, DRB3*0101: models of a directional alloimmune response and autoimmunity. J Mol Biol 2007;371:435–46.
[71] Saadi M, Karkhah A, Nouri HR. Development of a multi-epitope peptide vaccine inducing robust T cell responses against brucellosis using immunoinformatics based approaches. Infect Genet Evol 2017;51:227–34.
[72] Gosavi M, Patil HP. Evaluation of monophosphoryl lipid A as an adjuvanted for inactivated chikungunya virus. Vaccine 2022;40:5060–8.
[73] Cheng P, Xue Y, Wang J, Jia Z, Wang L, Gong W. Evaluation of the consistence between the results of immunoinformatics predictions and real-world animal experiments of a new tuberculosis vaccine MP3RT. Front Cell Infect Microbiol 2022;12:1047306.
[74] Ali M, Pandey RK, Khatoon N, Narula A, Mishra A, Prajapati VK. Exploring dengue genome to construct a multi-epitope based subunit vaccine by utilizing immunoinformatics approach to battle against dengue infection. Sci Rep 2017;7: 9232.
[75] Ikai A. Thermostability and aliphatic index of globular proteins. J Biochem 1980; 88:1895–8.
[76] Pålsson-McDermott EM, O'Neill LA. Signal transduction by the lipopolysaccharide receptor, Toll-like receptor-4. Immunology 2004;113:153–62.
[77] Gangloff M, Gay NJ. MD-2: the Toll 'gatekeeper' in endotoxin signalling. Trends Biochem Sci 2004;29:294–300.
[78] Miyake K. Endotoxin recognition molecules, Toll-like receptor 4-MD-2. Semin Immunol 2004;16:11–6.
[79] Delany I, Rappuoli R, De Gregorio E. Vaccines for the 21st century. EMBO Mol Med 2014;6:708–20.
[80] Holt RA. Oncomicrobial vaccines: the potential for a Fusobacterium nucleatum vaccine to improve colorectal cancer outcomes. Cell Host Microbe 2023;31:141–5.
[81] Bambini S, Rappuoli R. The use of genomics in microbial vaccine development. Drug Discov Today 2009;14:252–60.
[82] Chauhan V, Rungta T, Goyal K, Singh MP. Designing a multi-epitope based vaccine to combat Kaposi Sarcoma utilizing immunoinformatics approach. Sci Rep 2019;9: 2517.
[83] Grivennikov SI, Wang K, Mucida D, Stewart CA, Schnabl B, Jauch D, et al. Adenoma-linked barrier defects and microbial products drive IL-23/IL-17-mediated tumour growth. Nature 2012;491:254–8.
[84] Gabrilovich DI, Nagaraj S. Myeloid-derived suppressor cells as regulators of the immune system. Nat Rev Immunol 2009;9:162–74.