frontiers in Immunology

ORIGINAL RESEARCH
published: 26 October 2017
doi: 10.3389/fimmu.2017.01382

Identification of New Features from Known Bacterial Protective Vaccine Antigens Enhances Rational Vaccine Design

*Edison Ong[1], Mei U Wong[2] and Yongqun He[2,3]**

[1] Department of Computational Medicine and Bioinformatics, University of Michigan, Ann Arbor, MI, United States, [2] Unit for Laboratory Animal Medicine, Department of Microbiology and Immunology, University of Michigan, Ann Arbor, MI, United States, [3] Center of Computational Medicine and Bioinformatics, University of Michigan, Ann Arbor, MI, United States

With many protective vaccine antigens reported in the literature and verified experimentally, how to use the knowledge mined from these antigens to support rational vaccine design and study underlying design mechanism remains unclear. In order to address the problem, a systematic bioinformatics analysis was performed on 291 Gram-positive and Gram-negative bacterial protective antigens with experimental evidence manually curated in the Protegen database. The bioinformatics analyses evaluated included subcellular localization, adhesin probability, peptide signaling, transmembrane α-helix and β-barrel, conserved domain, Clusters of Orthologous Groups, and Gene Ontology functional annotations. Here we showed the critical role of adhesins, along with subcellular localization, peptide signaling, in predicting secreted extracellular or surface-exposed protective antigens, with mechanistic explanations supported by functional analysis. We also found a significant negative correlation of transmembrane α-helix to antigen protectiveness in Gram-positive and Gram-negative pathogens, while a positive correlation of transmembrane β-barrel was observed in Gram-negative pathogens. The commonly less-focused cytoplasmic and cytoplasmic membrane proteins could be potentially predicted with the help of other selection criteria such as adhesin probability and functional analysis. The significant findings in this study can support rational vaccine design and enhance our understanding of vaccine design mechanisms.

Keywords: vaccine design, protective antigen, reverse vaccinology, adhesin probability, subcellular localization, conserved domains, transmembrane proteins, functional analysis

OPEN ACCESS

Edited by:
Urszula Krzych, Walter Reed Army Institute of Research, United States

Reviewed by:
Stasya Zarling, Walter Reed Army Institute of Research, United States
Giampiero Pietrocola, University of Pavia, Italy

****Correspondence:***
Yongqun He
yongqunh@med.umich.edu

Specialty section:
This article was submitted to Vaccines and Molecular Therapeutics, a section of the journal Frontiers in Immunology

Received: *22 August 2017*
Accepted: *06 October 2017*
Published: *26 October 2017*

Citation:
Ong E, Wong MU and He Y (2017) Identification of New Features from Known Bacterial Protective Vaccine Antigens Enhances Rational Vaccine Design. Front. Immunol. 8:1382. doi: 10.3389/fimmu.2017.01382

INTRODUCTION

Vaccination is considered as the most effective medical intervention ever introduced in modern medicine (1) and has prevented 103 million cases of infectious diseases in the United States since 1924 (2). However, it is still difficult to develop safe and effective vaccines against many infectious diseases including tuberculosis, HIV, and malaria (3). The emerging reverse vaccinology (RV) addresses the challenge through rational vaccine design by predicting vaccine antigen based on bioinformatics analysis of pathogen genomes (4, 5). The first application of RV in Group B *meningococcus* (MenB) vaccine development predicted 350 surface-exposed proteins from MenB, and the following experiments verified 25 of them capable of inducing bactericidal antibodies (6). This finding led to the approval of the first MenB vaccine, Bexsero, for use in the Europe (7), and United States (8). The

success of Bexsero is a milestone for rational vaccine design and RV has also been applied in vaccine prediction against other challenging pathogens such as *Mycobacterium tuberculosis* (9).

Many selection criteria have been applied to vaccine antigen prediction, but a deep understanding of the rationale behind their usage is still missing. The initial RV study of MenB vaccine prediction used the subcellular localization (SCL) as a major selection criterion, given that the humoral immunity is vital to host protection against MenB and the protective antigens (PAgs) inducing antibody response are primarily located in extracellular or outer membrane (6). However, the preference of vaccine antigens in specific SCL varies across different pathogens, and SCL might not be equivalently critical for those pathogens against which cell-mediated immunity plays a major role. Another frequently used criterion is the number of transmembrane α-helices (TMHs) due to the difficulty in the isolation of proteins with more than one TMH (10). Nonetheless, it is unclear whether the number of TMH, and possibly transmembrane β-barrel (TMB), of a protein correlates with vaccine protection. Adhesin is crucial to pathogen invasion into host cells (11) but the usage of adhesin probability (AP) has not been widely appreciated. Other criteria including signal peptides, conserved domains, and biological function analysis (10) have been used in different RV tools [e.g., NERVE (12), Vaxign (13), and Jenner-predict server (14)], and machine-learning techniques are also applied to vaccine design studies (15, 16). However, the significance and association of above criteria with the protectiveness of bacterial PAgs is still lacking. The identification of such association is essential to improve vaccine antigen prediction and design studies.

The goal of this study is to systematically analyze known bacterial PAgs reported in the literature and identify underlying design mechanisms for better rational vaccine prediction. Our study uses PAgs collected from Protegen with antigen information and experimental protection evidence manually annotated from peer-reviewed articles (17). The significance and association of these Protegen PAgs are analyzed using bioinformatics tools for SLC (18), AP (19), signal peptide (20), TMH (21) and TMB (22), conserved domains (23), Clusters of Orthologous Groups (COG) (24), and gene ontology (GO) (25). This report provides a systematic analysis of protein properties and biological functions associated with known bacterial PAgs in the interest of supporting future rational vaccine prediction and design.

MATERIALS AND METHODS

Protective Antigens and Background Pan-Proteome Non-Protective Protein Sequences

Protective antigens in G^+ and G^- bacteria with supporting experimental evidence were downloaded from Protegen database (Table S1 in Supplementary Material). The most common experimental evidence is the protection results against virulent bacterial challenge in laboratory animal models. Reported assay results that correlate to protection or immune responses are also considered. Using the Gram-positive (G^+) and Gram-negative pathogen information provided along with the PAgs from Protegen, all protein-coding sequences of these pathogens were downloaded from the UniProt database (26). The taxonomy IDs reported in Protegen were queried against UniProt for possible pan-proteome sequences. The detail of taxonomy ID mapping between the reported G^+ and G^- pathogens from Protegen and their corresponding pan-proteome in UniProt is available in Table S2 in Supplementary Material. By merging all the pan-proteome protein sequences from UniProt, we obtained the background proteome for two groups used in this study: G^+ and G^- pathogen background proteomes. There is no curated dataset of non-protective G^+ and G^- proteins available in the literature. The non-protective protein datasets were generated by applying similar strategies reported in previous vaccine design studies (15, 27, 28). Specifically, the G^+ and G^- pan-proteomes downloaded from UniProt were first aligned to Protegen PAg sequences using BLAST (29). Then sequences that shared similar homology with the Protegen PAgs (*E*-value ≤10 and have a shared percent identity of 10%) were removed from the datasets. All the remaining sequences within the datasets were considered as non-protective proteins throughout the entire study. The non-protective proteins generated in this study only provide an estimated survey of the true non-protective datasets and some non-protective proteins included in this study could have never been tested for the protective capacity.

Protein Property Computations

In this paper, five types of protein properties were computed: (i) SCL, (ii) AP, (iii) signal peptide, (iv) TMH, and (v) TMB (Table S8 in Supplementary Material).

For SCL computation, all sequences were computed for tentative SCL locations by running through PSORTb v3.0 program (18). Briefly, PSORTb uses Bayesian network to integrate different SCL location prediction modules such as support vector machine, SCL-BLAST, and motif-based modules. The program predicts and assigns score for each possible SLC locations of the input sequence, and the location with the highest score is returned. In this study, the default setting was used besides specifying the G^+ or G^- of input sequences.

The AP of all sequences was computed using SPAAN program with default setting (19). SPAAN calculates probability of being adhesin for an input sequence using neural network with five features including amino-acid frequencies, multiplet frequencies, dipeptide frequencies, charge composition, and hydrophobic composition. Sachdeva et al. reported 89% sensitivity and 100% specificity when the cutoff value AP ≥ 0.51 was used (19), and therefore the same threshold was applied in this study.

Prediction of signal protein secretion of all sequences was calculated by SignalP 4.1 standalone version (20), which is built solely on neural network to discriminate signal peptides from transmembrane regions. The discrimination score (D-score) computed by SignalP provides a value for protein secretion. As suggested by SignalP[1], the threshold value of D-score of 0.45 for G^+ and 0.51 for G^- provides the best sensitivity in signal peptide detection. In this study, the suggested cutoff values were used and

[1]http://www.cbs.dtu.dk/services/SignalP/performance.php.

the default configuration was applied besides specifying the G^+ or G^- of input sequences.

The TMH was computed using TMHMM 2.0 (21) with default settings and the number of TMH of the input G^+ and G^- pathogen sequences was reported. In brief, the tool uses hidden Markov model to predict transmembrane state of the input sequences and the Krogh et al. reported 97–98% prediction sensitivity (21).

The TMB was computed using PROFtmb tool, which is also a hidden Markov model-based prediction program (22). Only TMB of G^- pathogen sequences were computed because classical G^+ bacteria do not contain β-barrel membrane proteins (30). Based on the performance evaluation of the PROFtmb on discriminating transmembrane vs. non-transmembrane β-barrel using the whole protein dataset by Bigelow et al. (22), a cutoff value of ≥0.6 accuracy was chosen in order to achieve a balance with coverage.

Protein Sequence Property Computations

The PAg sequences, non-protective protein and background proteome sequences were functionally annotated with (i) Pfam conserved domains, (ii) COG functional classifications, and (iii) GO biological process (BP), molecular function (MF) and cellular component (CC) terms (Table S8 in Supplementary Material).

The PfamScan tool was used to annotate the conserved domains in all PAg, non-protective proteins and background proteomes. The sequences were aligned using the downloaded Pfam-A domain hidden Markov models (23).

The sequences of all PAgs were scanned for COG clusters using HMMER[2] with the hidden Markov models downloaded from the EggNog 4.5 database (31). Each input sequence was initially assigned with one ENOG identifier, which was then mapped to the corresponding COG cluster. For background proteomes and non-protective proteins, the COG cluster identifiers were retrieved directly from the UniProt database.

The PAg sequences were submitted to Argot2 web server for GO annotation prediction (32). The GO information of non-protective proteins and background proteomes were directly downloaded from UniProt database.

Statistical Analysis

Unless specified, the statistical significance of the association between reported PAgs and computed protein properties including SCL, AP, signal peptide, TMH, and TMB were calculated using one-way Fisher's exact test since we were only interested in over-representation of properties in PAgs only. For the *ad hoc* analysis of specific property (e.g., SCL prediction), the significance of individual sub-property (e.g., individual SCL locations such as extracellular, cell wall, cytoplasmic membrane, and cytoplasm in G^+ bacteria) were further examined by performing one vs. other Fisher's exact test and the resulting p-value was adjusted by applying Bonferroni correction.

The over-representation of conserved domains, COG clusters, and GO BP, MF, CC terms among Protegen PAgs were tested using Fisher's exact test and adjusted using Benjamini–Hochberg–Yekutieli procedure. In addition, the significant (adjusted $p \leq 0.05$) GO terms (BP, MF, CC) were visualized in hierarchical format using GOfox (33). GOfox[3] laid out GO terms using the internal hierarchical GO structure simplification algorithm since GO enrichment analysis tends to generate a large list of enriched GO terms (33).

RESULTS

Three sets of data were collected and generated for the bioinformatics analysis. Our study specifically analyzed frequently used PAg prediction features, including SCL, AP, signal peptide, TMH and TMB, conserved domain, and biological function analysis.

Collection of Protective Vaccine Antigens, Background and Non-Protective Proteins

After removal of identical sequences, the curated Protegen dataset contained 81 and 210 non-redundant vaccine PAgs from 14 Gram-positive (G^+) and 34 Gram-negative (G^-) bacteria, respectively (Table S1 in Supplementary Material). The corresponding pan-proteomes of these G^+ and G^- pathogens were downloaded from the UniProt database (26) as the background proteomes, which included 39,397 G^+ and 73,371 G^- peptide sequences (Table S2 in Supplementary Material). A set of non-protective proteins were selected from background proteome as described in Materials and Methods and other RV studies (15, 16, 27, 28), and contained 4,954 G^+ and 5,478 G^- pathogen peptide sequences.

Subcellular Localization Analysis

Our analysis found that 44.4% and 19.8% of PAgs in G^+ bacteria located in extracellular space and cell wall, respectively (**Figure 1A**). In comparison, only 1.7% and 1.2% of the G^+ non-protective proteins were extracellular and cell wall proteins, respectively (**Figure 1B**). Our statistical analysis showed significant over-representation of PAgs in these two SCLs ($p < 0.01$). In G^- bacteria, 15.7%, 30.0%, and 8.1% of PAgs were extracellular, outer membrane, and periplasmic proteins, respectively (**Figure 1C**). Compared with the corresponding SCL proportions in G^- non-protective proteins (0.4, 0.4, and 0.9%) (**Figure 1D**), these three locations were significantly over-represented in PAgs ($p < 0.01$). In non-protective proteins, most proteins (78.3% in G^+ and 67.7% in G^-) were localized in the cytoplasmic or cytoplasmic membrane (**Figures 1B,D**) but these two SCL locations also accounted for 26.8% G^+ and 31.1% G^- of the reported PAgs (**Figures 1A,C**). The SCL predictions of background proteome were shown in Figure S1 in Supplementary Material.

To confirm the SCL analysis results, we also analyzed signal peptides using SignalP (20), which predicted the presence of signal sequences of the majority of synthesized proteins designated to secretory pathways. The distribution histograms of the calculated score for PAgs, non-protective proteins, and background proteomes were plotted (Figure S2 in Supplementary Material). The signal peptide scores of extracellular (both G^+ and G^-) or surface-exposed

[2] http://hmmer.org/.

[3] http://gofox.hegroup.org/.

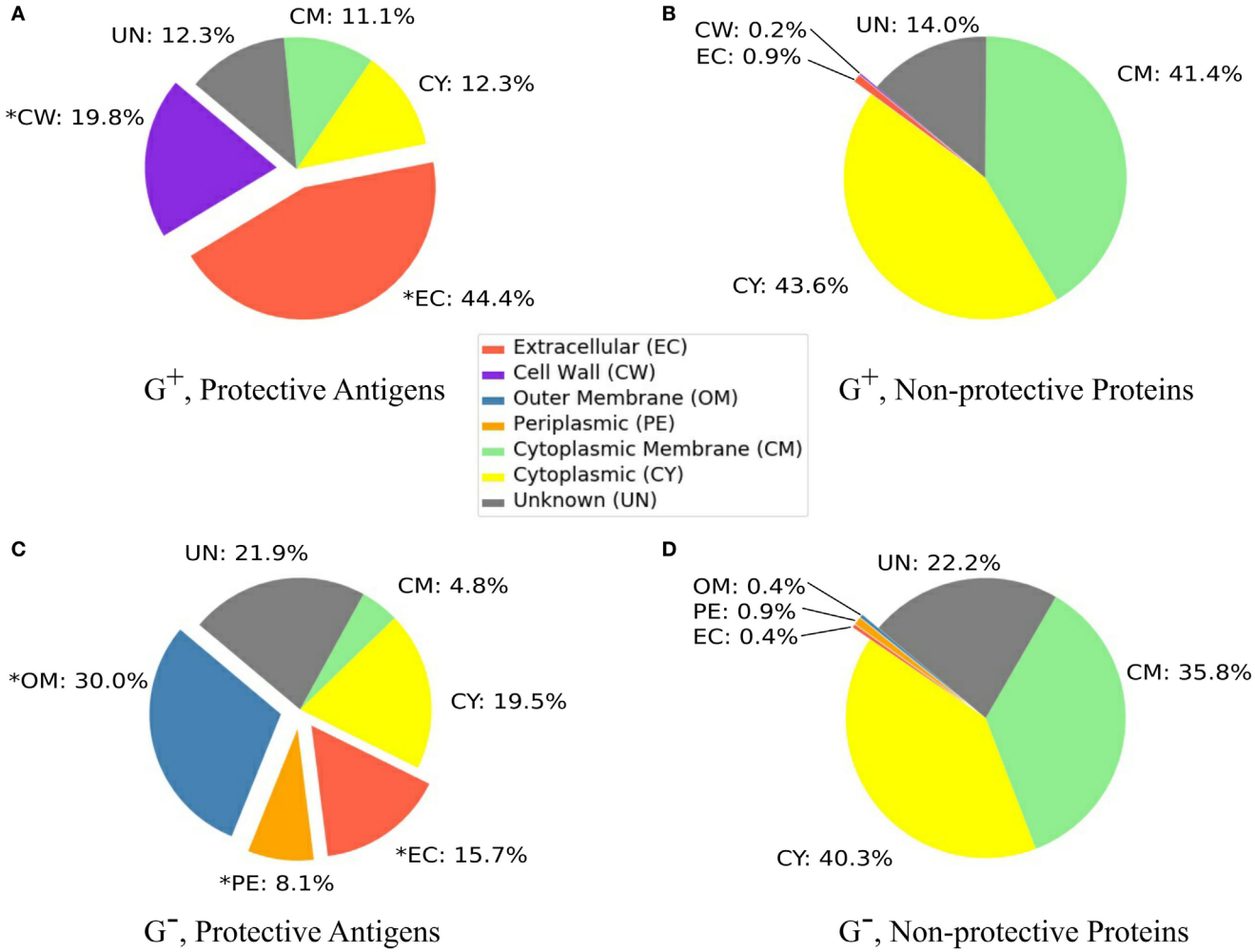

FIGURE 1 | Subcellular localization profiles. G^+ and G^- bacterial protective antigens showed significantly higher ($p < 0.01$) proportions of EC (G^+ and G^-), CW (G^+), and PE and OM (G^-) **(A,C)** compared with the non-protective proteins **(B,D)** (abbreviations shown in the middle color key legend). Only PAgs and non-protective proteins are displayed and the composition of background proteomes and non-protective proteins are very alike (Figure S1 in Supplementary Material). The significant over-representation of PAgs' subcellular localization prediction compared with non-protective proteins is indicated with "*" ($p < 0.01$).

proteins (cell wall for G^+ and outer membrane for G^-) showed that a large fraction of PAgs was predicted to be secreted signal peptides (Figure S2 and Table S3 in Supplementary Material).

Adhesin Probability Analysis

Adhesins are proteins critical for bacterial pathogens to invade host cells and cause infections (11). Over half of the PAgs could be identified with AP (56.8% of G^+ and 52.8% of G^-) using the suggested cutoff of no <0.51 (19). The AP of proteins with different SCLs also had different patterns (**Figure 2**). Specifically, comparing PAgs (**Figures 2B,E**) and non-protective proteins (**Figures 2C,F**), PAgs with SCL locations other than cytoplasmic membrane and cytoplasm generally showed increasing trend in AP. There were 87.5% G^+ PAgs in the cell wall and 82.5% G^- PAgs in outer membrane that were also adhesins, compared with 37.5% G^+ and 20% G^- non-protective proteins in the cell wall and outer membrane, respectively (**Figure 2**; Table S4 in Supplementary Material). This high preference of surface-exposed proteins (cell wall for G^+ and outer membrane for G^-) with high AP was significant ($p < 0.01$, **Figure 2**) and illustrated the importance of SCL and AP as two major criteria in vaccine design. Additionally, 90.0% and 54.3% of the PAgs in G^+ and G^- bacteria with unknown SCL were in fact predicted to be adhesins. Therefore, utilizing AP with SCL could potentially overcome the limitation of excluding "Unknown" SCL and avoid inaccuracy generated by individual SCL prediction tool. For PAgs located at the cytoplasmic membrane and cytoplasm, the computed AP also showed different patterns between G^+ and G^- (**Figures 2B,E**). G^+ PAgs in cytoplasmic membrane were more likely adhesins (77.8%), while in G^- only 20.0% were adhesins (Table S4 in Supplementary Material). For cytoplasm, PAgs were both unlikely adhesins (0% for G^+ and 4.9% for G^-, Table S4 in Supplementary Material). AP

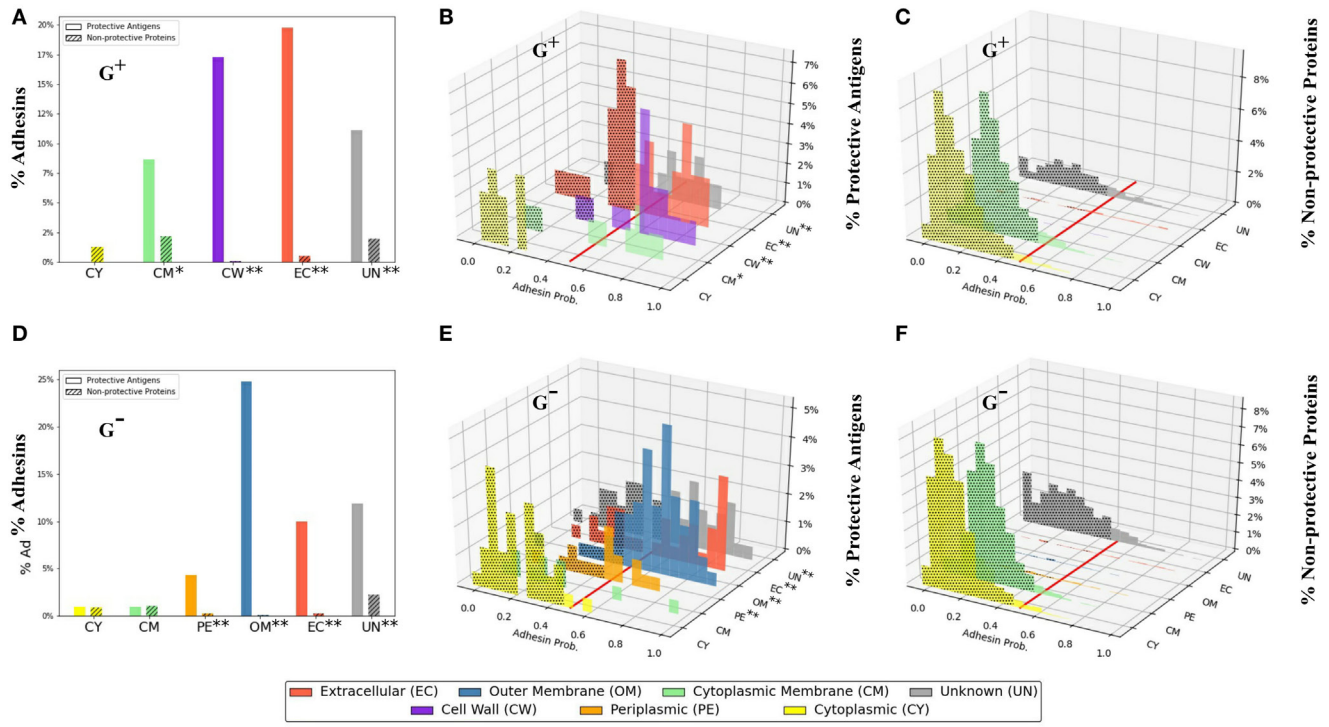

FIGURE 2 | Profiles of adhesin probabilities of protective antigens (PAgs) and non-protective proteins with different subcellular localizations. The top three panels **(A–C)** show G⁺ pathogens, and the bottom three panels show G⁻ pathogens. Specifically, the first column **(A,D)** represents the overall percentages of adhesin probabilities. The second column (**B,E**) and third column **(C,F)** show adhesin probability distributions of PAg and non-protective proteins, respectively. The red line in **(B,C,E,F)** indicates adhesin probability cutoff of no <0.51. Overall, PAgs have significantly higher ($p < 0.01$) percentages in EC (G⁺ and G⁻), CW (G⁺), and PE and OM (G⁻) (abbreviations shown in bottom color key legend). Interestingly, CM in G⁺ is also significant ($p < 0.05$) when coupled with adhesin probability. See the text for detailed discussion. The significant over-representation of PAgs' adhesin probabilities at different subcellular localizations compared with non-protective proteins is indicated with "*" ($p < 0.05$) or "**" ($p < 0.01$).

prediction of background proteome is also shown in Figure S3 in Supplementary Material.

Transmembrane α-Helix and β-Barrel

We analyzed and compared the TMH profiles between PAgs and non-protective proteins. Specifically, none of the PAgs located at the cell wall (G⁺), outer membrane, or periplasm (G⁻) had more than one TMH (**Figure 3A**). There were two G⁻ PAgs with more than 10 TMH (lipoprotein signal peptidases in *Brucella melitensis* and L-lactate permease in *Neisseria meningitides*). The β-barrel analysis was only performed for G⁻ pathogens because classical G⁺ bacteria do not contain β-barrel membrane proteins (30). Using the probability cutoff of 0.60, our study found that 12.9% of Gram-negative PAgs predicted to have TMB compared with <0.001% in non-protective proteins (**Figure 3B**).

Conserved Domain Analysis

Conserved domains represent functional units in proteins and some domains are more frequently associated with PAgs (14, 34). Our analysis identified eight conserved domains that were only frequently found among reported PAgs (**Table 1**). These domains included "autotransporter β-domain," "outer membrane protein β-barrel domain," "fimbrial protein," "TonB-dependent receptor plug domains," "OmpH-like outer membrane protein," and "extended signal peptide of type V secretion system." The full list of all predicted conserved domains and their frequencies in PAgs and non-protective proteins can be found in Table S5 in Supplementary Material.

Functional Analysis

The functional annotations were analyzed using the COG and GO. COG includes 26 functional clusters (24). Our COG analysis of PAgs identified 16 COG functional categories that were significantly enriched (adjusted $p < 0.05$) in PAgs (**Figure 4**; Table S6 in Supplementary Material). Four COG clusters "cell wall/membrane envelope biogenesis," "cell motility," "signal transduction mechanisms," and "extracellular structures" were notably enriched in PAgs.

We also analyzed enriched GO terms from the three GO branches: biological process (BP), molecular function (MF), and cellular component (CC) (25). Eighteen GO BP terms were found significantly enriched (adjusted $p < 0.05$) in bacterial PAgs, including "pathogenesis" as the most significantly enriched term among PAgs in bacterial pathogens (**Figure 5**; Table S7 in Supplementary Material). BPs related to pathogen invasion (e.g., "cell adhesion" and "proteolysis") and terms related to transporter (e.g., "transmembrane transport") were significantly over-represented among PAgs. Twenty GO MF terms were

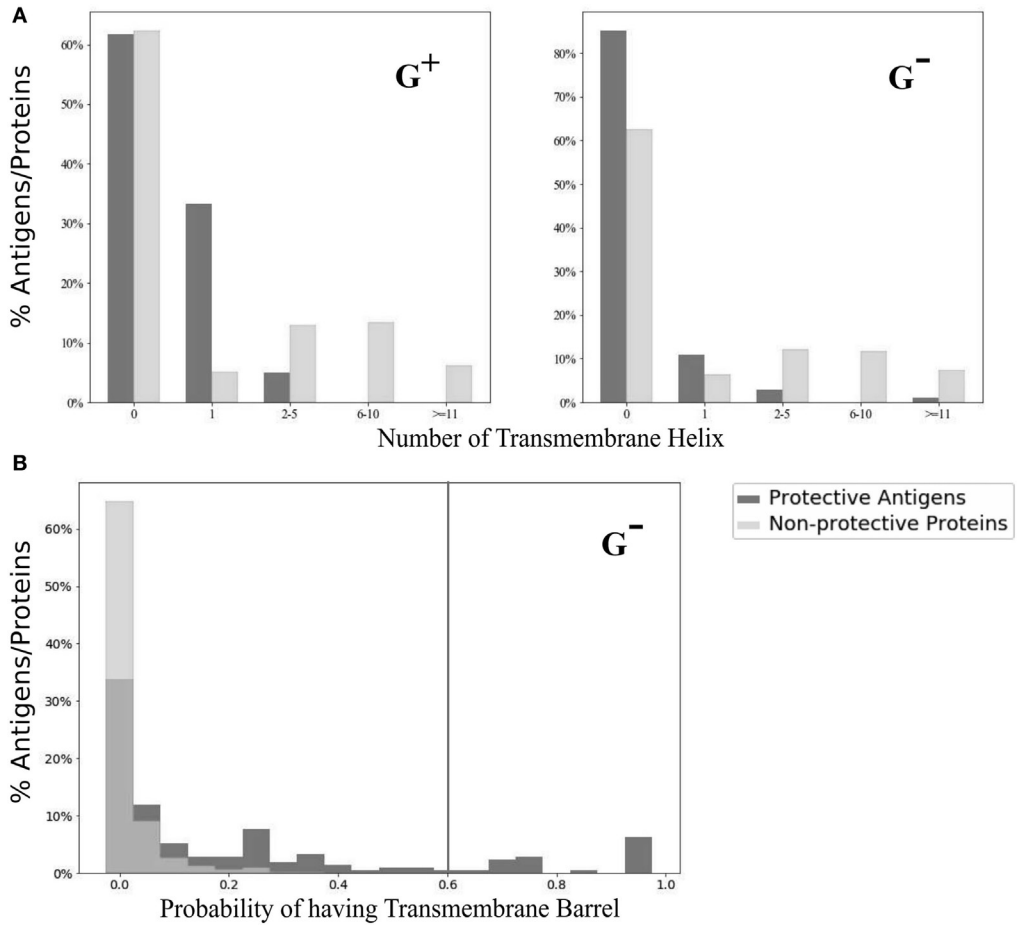

FIGURE 3 | Transmembrane α-helix (G^+ and G^-) and β-barrel (G^- only) profiles. As compared with non-protective proteins, there were much higher percentages of protective antigens with zero or one transmembrane α-helix **(A)**. For transmembrane β-barrel **(B)**, only two (0.0004%) out of all non-protective proteins had probability higher than the designated cutoff (indicated as black vertical line) described in method.

TABLE 1 | Frequent Pfam-A conserved domains among reported PAgs.

Pfam domain description	Protective antigen count
Autotransporter β-domain	11
Outer membrane protein β-barrel domain	10
Fimbrial protein	10
ATPase family associated with various cellular activities (AAA)	9
TonB-dependent receptor plug domain	8
Outer membrane protein (OmpH-like)	5
ABC transporter	5
Extended signal peptide of Type V secretion system	5

The most over-represented (PAg count ≥5) Pfam-A conserved domains among reported PAgs were listed. The top two frequently found Pfam-A conserved domains among reported PAgs were β-barrel domains, which support the positive selection of transmembrane β-barrel in PAg prediction. In addition to β-barrel domains, proteins with over-represented conserved domains were more likely related to the pathogenesis of bacteria including pathogen colonization and invasion, and therefore could be used as a good indicator of PAg prediction.

significantly enriched (adjusted $p < 0.05$), including those related to invasion (e.g., "peptidase activity") and transportation (e.g., "transferase activity" and "receptor activity"). Fifteen GO CC terms were significantly enriched (adjusted $p < 0.05$). In agreement with the SCL prediction results, extracellular or surface-exposed CC terms were significantly over-represented among reported PAgs. In addition, CC terms that were related to bacterial colonization and invasion within host such as "bacterial-type flagellum filament," "pilus," "host cell part," "host cell plasma membrane," and "host cell junction" were also enriched, suggesting PAgs' role in the interactions between bacteria and the host cells.

DISCUSSION

Although extensive research has been conducted, modern vaccine research and development still faces challenges of quick and accurate development of vaccines in response to major infectious diseases [e.g., tuberculosis (3)], outbreaks [e.g., Ebola and Zika virus (35, 36)], and new drug-resistant pathogens (37). Our efforts to develop vaccines using traditional methods have not been successful to address these challenges. The future success of effective vaccine development relies on powerful rational vaccine design

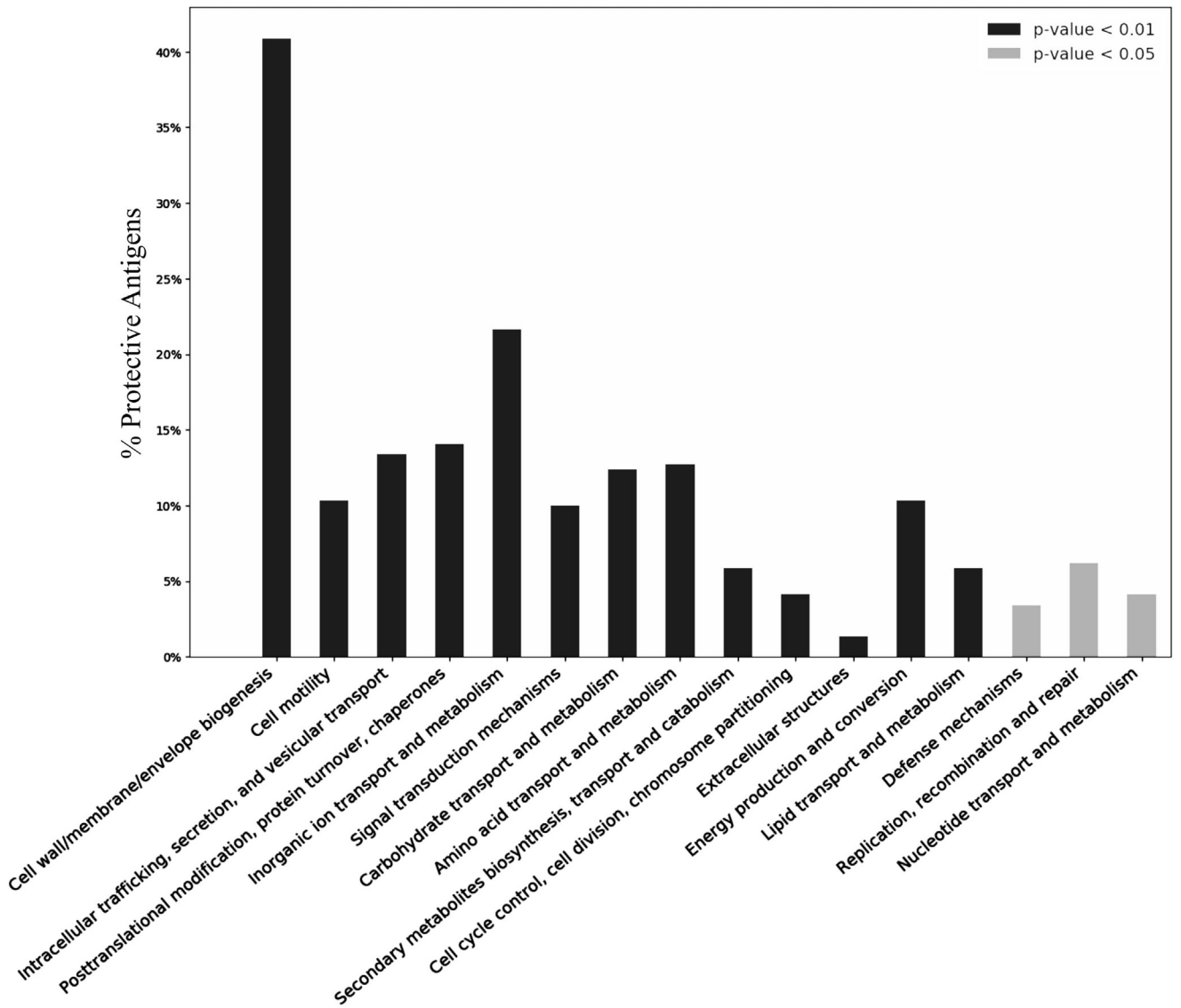

FIGURE 4 | Over-representation of COG clustering profiles among reported protective antigens (PAgs). Over 40% of the reported PAgs belong to the cluster "Cell wall/membrane/envelop biogenesis," which agrees with common knowledge of using surface-exposed proteins as a key criterion in vaccine antigen prediction. Other COG clusters related to pathogen motility, secretion, signal transduction, and transportation are also significantly enriched in PAgs as compared with non-protective proteins. See the text for detailed discussion. The significant over-representation of PAgs' COG clusters compared with non-protective proteins is colored with gray ($p < 0.05$) and black ($p < 0.01$).

including reverse and structural vaccinology (1) and based on our deeper understanding of vaccination mechanism. By systematically studying and comparing bacterial PAgs and non-protective proteins, our comprehensive bioinformatics study analyzed key criteria for vaccine design including various protein properties and biological functions. The summarized characteristics in this study are specifically used for bacterial model PAg prediction and might not hold true for viral or parasitic pathogens. The results of this study confirmed and provided details on the usage of these prediction criteria, including SCL, AP, signal peptides, TMH and TMB, conserved domains, and biological function annotations, for RV prediction against bacterial pathogen. Most importantly, our results suggested new insights toward rational vaccine prediction and design.

In accordance with secreted extracellular or surface-exposed antigens commonly known to be PAgs, our study observed the differences among the SCL profiles of G^+ and G^- bacterial PAgs. In terms of extracellular proteins, G^+ bacterial PAgs had a much higher percentage (44%) being PAgs than G^- bacterial PAgs (15.7%). We also found a strong correlation between the presence of secretory signal peptides and PAgs. Approximately half of the PAgs (over 45% in both G^+ and G^-) were predicted to be signal peptides (Table S3 in Supplementary Material; **Figure 4**). Coupling the selection of SCL and signal peptides, particularly in

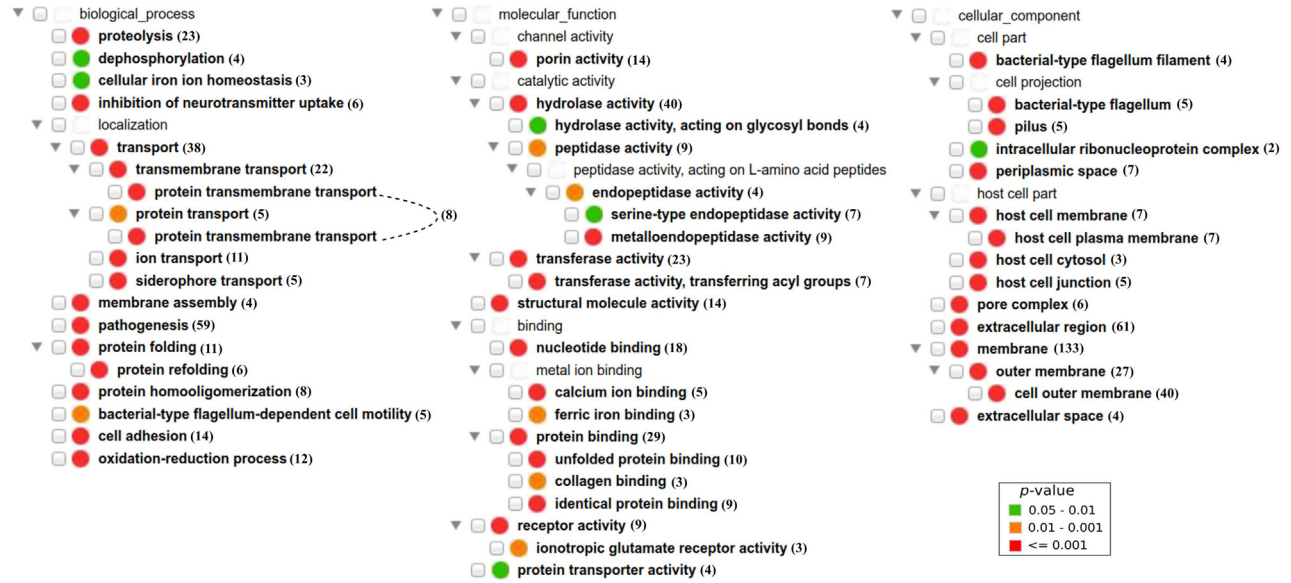

FIGURE 5 | Over-representation of GO BP, MF, and CC term profiles among reported protective antigens (PAgs) and visualized using the GOfox tool. The number next to each GO term indicates the number of PAg with the corresponding GO functional annotation. Similar to COG clustering, GO terms that are related to pathogen motility, secretion, signal transduction, and transportation are also significantly enriched in PAgs as compared with non-protective proteins. The GO CC terms also supported the high preference of extracellular, surface-exposed (cell wall in G^+ and outer membrane in G^-), and periplasmic (G^-) PAgs. The significant over-representation of PAgs' GO terms compared with non-protective proteins is color-coded following the legend in the lower right corner.

G^+ bacterial pathogens, pose a viable option for a more precise PAg prediction. On the other hand, 19.8% cell wall proteins in G^+ and 30.0% outer membrane proteins in G^- bacteria were surface-exposed PAgs (**Figures 1A,C**). The G^+ bacterial PAgs showed higher preference in extracellular proteins, while both G^+ and G^- bacterial PAgs shared similar proportions as surface-exposed proteins.

Moreover, 8.1% G^- PAgs were in the periplasm, a subcellular location that vaccine researchers often ignore due to lack of direct interaction with the host immune cells. Hence, the percentage of periplasmic PAgs was significant ($p < 0.05$, **Figure 1C**) and over-represented (**Figure 5**) when taking the non-protective periplasmic proteins (0.9%) into consideration. It is possible that G^- bacterial periplasmic proteins can be released extracellularly after being packed within outer membrane vesicles and can induce strong immune responses (38, 39). These periplasmic proteins can be potentially a good source of PAg candidates when coupling with other selection criteria such as functional analysis.

The results of our study highlight the importance of AP and its effect in improving RV prediction when combined with SCL. Adhesin is critical for bacterial invasion and is capable of inducing strong immune responses (11). Adhesins can also function as enzymes and mediate a main part of bacterial pathogenesis (40). The majority of vaccine design studies do not incorporate AP in their selection pipeline (6, 15, 16, 27), and AP as a selection criterion is currently underused and poorly investigated in the vaccine development field. Our study managed to identify over 50% of the PAgs with AP as the only criterion. The prediction of coupling SCL and AP was even more significant, with the identification of over 80% cell wall (G^+) and outer membrane (G^-) PAgs (Table S4 in Supplementary Material). By addressing the importance of adhesin playing an important role in vaccine development, we hope to promote the AP as a viable option in future vaccine design studies.

The functional analysis of adhesive PAgs in our study proposes a mechanistic explanation of their roles in pathogen colonization and invasion. Cell motility is one of the most important steps in host colonization and invasion, and the bacterial movement requires structure such as flagellum and pillus for cell adhesion and colonization (41), and cell motility related COG clusters and GO terms were significantly enriched (**Figures 4** and **5**). Pilli are composed of fimbrial and other proteins (41), and the Pfam domain "fimbrial protein" was highly conserved among the reported PAgs (**Table 1**). GO BP term "proteolysis" and GO MF terms "peptidase activity" (**Figure 5**) were also found to be significant in the functional analysis. For instance, *Yersinia pestis* can produce the surface protease to mediate invasion into host endothelial cells (42). The pili, fimbri, and protease mentioned earlier can occur as one of the various architectures of adhesins (40). Given these important roles of adhesins, more investigations of adhesins as potential PAgs and how they induce protective immunity are much deserved.

Our study showed two distinct correlation patterns of the PAgs protectiveness to the TMH and TMB. The TMH is more abundant in cytoplasmic or inner membranes, and the TMB type is more likely located in bacterial outer membranes (43). Our study confirmed that TMH proteins with more than one TMH were not typically used for vaccine development (10) (**Figure 3A**; Figure S5 in Supplementary Material). The two exceptional proteins

that had more than 10 TMHs, which were *Brucella* lipoprotein signal peptidase and *Neisseria meningitides* L-lactate permease. *Brucella* lipoprotein signal peptidase is a known virulence factor, which is involved in lipopolysaccharides biosynthesis (44). The *N. meningitides* L-lactate permease is a protein required by *N. meningitides* during bacteraemic infection and induces protective immunity in systemic *meningococcal* infection (45). Different from TMH, our study indicated that the presence of TMB was associated with significantly higher portions ($p < 0.01$ from χ^2 test) of G^- PAgs (**Figure 3B**). In particular, none of the G^- outer membrane non-protective proteins was predicted to have TMB. Our results suggested the use of TMH as a negative and TMB as a positive selection criterion in future vaccine development.

Although not usually considered as PAgs, large portions (26.8% G^+ and 31.1% G^-) of cytoplasmic and cytoplasmic membrane proteins were found to be PAgs (**Figures 1A,C**). Compared with a much larger size of cytoplasmic and cytoplasmic membrane non-protective proteins, this fraction of PAgs was not significant. However, the ignorance of proteins located at these two SCLs might hinder the productivity of effective PAg prediction. Cytoplasmic and cytoplasmic membrane proteins might not induce humoral immune response due to their SCLs, but these proteins often time can be potent inducers of cell-mediated immunity. For example, the cytoplasmic catalase-peroxidase protein in *M. tuberculosis*, which contributes to intracellular survival within host macrophage by protecting against reactive oxygen species (46), is able to induce protective immunity (47). How to accurately predict cytoplasmic PAgs remains a big challenge but it can be potentially addressed using multiple features such as AP, conserved domains, COG clusters, and GO terms. Particularly in G^+ cytoplasmic membrane, PAgs showed significant over-representation ($p < 0.05$) when coupled with AP prediction. Conserved domains have been reported as a viable option in the PAg prediction (14). In our study, many conserved domains were frequently found among PAgs and each domain might link to important bacterial biological functions within the host such as "TonB-dependent receptor plug domain." As a strategy in antibiotics resistance is the bacterial efflux pumps (48), TonB-dependent receptor is a G^- bacterial protein responsible for the transportation of large ion complex and has been identified as potent vaccine PAgs (49). The over-represented COG clusters and GO terms among the reported PAgs suggested a viable alternative to overcome the challenge of identifying cytoplasmic and cytoplasmic membrane PAgs and complement to current vaccine prediction studies.

The findings in this study can be translated into a predictive framework with different approaches to improve existing methods and achieve better identification and validation of novel PAgs. Even though traditional rule-based prediction has been successful in multiple studies (6, 9) and also applied in many tools (12–14), this type of "all-or-nothing" selection might fail to capture the relationship among different criteria (16). For example, a potential cytoplasmic or cytoplasmic membrane PAg would be immediately discarded from a study that includes surface-exposing SCL as one of the criterion. As indicated in our findings, the cytoplasmic or cytoplasmic membrane PAg could be predicted by incorporating other criteria such as AP, conserved domains, and biological functions. As a natural solution, a combinatory strategy has been proposed that assigns each criterion with a weight and synthesizes multiple criteria in a composite way such as weighted metrics (50). Candidate proteins that have low score in a set of rules could still achieve a reasonable score and are compensated by another set of selection criteria. Another advance technique is to apply machine-learning methods such as support vector machine, random forest, and neural network as described in many previous studies (15, 16, 28, 34). Even though the machine-learning-based prediction can overcome the "all-or-nothing" scenario, these methods have not captured all the significant features as reported in this study. For example, AP and conserved domains are not implemented in current ML-based prediction (15, 16, 28) except the preliminary study by Xiang and He (34), and none of these studies incorporated TMB and biological functional analysis into their prediction pipeline. The additional features given from our findings showed promising improvement on current machine-learning methods.

Based on the new discoveries reported in this study, we plan to explore the possibility of integrating these significant criteria along with other including MHC-epitope binding and structure on protein selection as vaccine candidates to improve our Vaxign software program (13). Even though our analysis focused on bacterial model, some criteria such as AP, signal peptide, transmembrane proteins, and pathogenesis-related conserved domains and biological functions can be extended to viral or parasitic PAgs prediction after further verification and analysis. The better understanding of the association between individual criterion and PAgs, as well as the inter-relation among different criteria, will provide new opportunities for more accurate and rational vaccine design, leading to better prevention and control of various infectious diseases.

AUTHOR CONTRIBUTIONS

EO and YH conceived and designed the study. EO and MW collected protective antigens and background proteome sequences. EO performed bioinformatics analyses of the protein properties and functional annotations. EO, MW, and YH prepared the manuscript. All authors participated in result interpretation, paper editing, discussion, and approved the paper publication.

FUNDING

This work was supported by the US National Institutes of Health (grant number 1R01AI081062).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at http://www.frontiersin.org/article/10.3389/fimmu.2017.01382/full#supplementary-material.

REFERENCES

1. Rappuoli R, Pizza M, Del Giudice G, De Gregorio E. Vaccines, new opportunities for a new society. *Proc Natl Acad Sci U S A* (2014) 111:12288–93. doi:10.1073/pnas.1402981111
2. van Panhuis WG, Grefenstette J, Jung SY, Chok NS, Cross A, Eng H, et al. Contagious diseases in the United States from 1888 to the present. *N Engl J Med* (2013) 369:2152–8. doi:10.1056/NEJMms1215400
3. WHO. *MDG 6: Combat HIV/AIDS, Malaria and Other Diseases*. Geneva: WHO (2014).
4. Rappuoli R. Reverse vaccinology. *Curr Opin Microbiol* (2000) 3:445–50. doi:10.1016/S1369-5274(00)00119-3
5. Adu-Bobie J, Capecchi B, Serruto D, Rappuoli R, Pizza M. Two years into reverse vaccinology. *Vaccine* (2003) 21:605–10. doi:10.1016/S0264-410X(02)00566-2
6. Pizza M, Scarlato V, Masignani V, Giuliani MM, Aricò B, Comanducci M, et al. Identification of vaccine candidates against serogroup B meningococcus by whole-genome sequencing. *Science* (2000) 287:1816–20. doi:10.1126/science.287.5459.1816
7. Vernikos G, Medini D. Bexsero® chronicle. *Pathog Glob Health* (2014) 108:305–16. doi:10.1179/2047773214Y.0000000162
8. Folaranmi T, Rubin L, Martin SW, Patel M, MacNeil JR. Use of serogroup B meningococcal vaccines in persons aged >/=10 years at increased risk for serogroup B meningococcal disease: recommendations of the advisory committee on immunization practices, 2015. *MMWR Morb Mortal Wkly Rep* (2015) 64:608–12.
9. Baldwin SL, Reese VA, Huang PWD, Beebe EA, Podell BK, Reed SG, et al. Protection and long-lived immunity induced by the ID93/GLA-SE vaccine candidate against a clinical *Mycobacterium tuberculosis* isolate. *Clin Vaccine Immunol* (2016) 23:137–47. doi:10.1128/CVI.00458-15
10. He Y, Rappuoli R, De Groot AS, Chen RT. Emerging vaccine informatics. *J Biomed Biotechnol* (2010) 2010:218590. doi:10.1155/2010/218590
11. Ribet D, Cossart P. How bacterial pathogens colonize their hosts and invade deeper tissues. *Microbes Infect* (2015) 17:173–83. doi:10.1016/j.micinf.2015.01.004
12. Vivona S, Bernante F, Filippini F. NERVE: new enhanced reverse vaccinology environment. *BMC Biotechnol* (2006) 6:35. doi:10.1186/1472-6750-6-35
13. He Y, Xiang Z, Mobley HLT. Vaxign: the first web-based vaccine design program for reverse vaccinology and applications for vaccine development. *J Biomed Biotechnol* (2010) 2010:297505. doi:10.1155/2010/297505
14. Jaiswal V, Chanumolu SK, Gupta A, Chauhan RS, Rout C. Jenner-predict server: prediction of protein vaccine candidates (PVCs) in bacteria based on host-pathogen interactions. *BMC Bioinformatics* (2013) 14:211. doi:10.1186/1471-2105-14-211
15. Bowman BN, McAdam PR, Vivona S, Zhang JX, Luong T, Belew RK, et al. Improving reverse vaccinology with a machine learning approach. *Vaccine* (2011) 29:8156–64. doi:10.1016/j.vaccine.2011.07.142
16. Goodswen SJ, Kennedy PJ, Ellis JT. A novel strategy for classifying the output from an in silico vaccine discovery pipeline for eukaryotic pathogens using machine learning algorithms. *BMC Bioinformatics* (2013) 14:315. doi:10.1186/1471-2105-14-315
17. Yang B, Sayers S, Xiang Z, He Y. Protegen: a web-based protective antigen database and analysis system. *Nucleic Acids Res* (2011) 39:1073–8. doi:10.1093/nar/gkq944
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. doi:10.1093/bioinformatics/btq249
19. Sachdeva G, Kumar K, Jain P, Ramachandran S. SPAAN: a software program for prediction of adhesins and adhesin-like proteins using neural networks. *Bioinformatics* (2005) 21:483–91. doi:10.1093/bioinformatics/bti028
20. Petersen TN, Brunak S, von Heijne G, Nielsen H. SignalP 4.0: discriminating signal peptides from transmembrane regions. *Nat Methods* (2011) 8:785–6. doi:10.1038/nmeth.1701
21. Krogh A, Larsson B, von Heijne G, Sonnhammer EL. Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. *J Mol Biol* (2001) 305:567–80. doi:10.1006/jmbi.2000.4315
22. Bigelow HR, Petrey DS, Liu J, Przybylski D, Rost B. Predicting transmembrane beta-barrels in proteomes. *Nucleic Acids Res* (2004) 32:2566–77. doi:10.1093/nar/gkh580
23. Punta M, Coggill PC, Eberhardt RY, Mistry J, Tate J, Boursnell C, et al. The Pfam protein families databases. *Nucleic Acids Res* (2012) 40:D290–301. doi:10.1093/nar/gkp985
24. Tatusov RL, Galperin MY, Natale DA, Koonin EV. The COG database: a tool for genome-scale analysis of protein functions and evolution. *Nucleic Acids Res* (2000) 28:33–6. doi:10.1093/nar/28.1.33
25. Blake JA, Christie KR, Dolan ME, Drabkin HJ, Hill DP, Ni L, et al. Gene ontology consortium: going forward. *Nucleic Acids Res* (2015) 43:D1049–56. doi:10.1093/nar/gku1179
26. The UniProt Consortium. The Universal Protein Resource (UniProt). *Nucleic Acids Res* (2008) 35:D193–7. doi:10.1093/nar/gkl929
27. Doytchinova IA, Flower DR. VaxiJen: a server for prediction of protective antigens, tumour antigens and subunit vaccines. *BMC Bioinformatics* (2007) 8:4. doi:10.1186/1471-2105-8-4
28. El-Manzalawy Y, Dobbs D, Honavar V. Predicting protective bacterial antigens using random forest classifiers. *Proceedings of the ACM Conference on Bioinformatics, Computational Biology and Biomedicine*. Orlando, FL: ACM (2012). p. 426–33. doi:10.1145/2382936.2382991
29. Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, et al. BLAST plus: architecture and applications. *BMC Bioinformatics* (2009) 10:1. doi:10.1186/1471-2105-10-421
30. Wimley WC. The versatile beta-barrel membrane protein. *Curr Opin Struct Biol* (2003) 13:404–11. doi:10.1016/S0959-440X(03)00099-X
31. Huerta-Cepas J, Szklarczyk D, Forslund K, Cook H, Heller D, Walter MC, et al. EGGNOG 4.5: a hierarchical orthology framework with improved functional annotations for eukaryotic, prokaryotic and viral sequences. *Nucleic Acids Res* (2016) 44:D286–93. doi:10.1093/nar/gkv1248
32. Falda M, Toppo S, Pescarolo A, Lavezzo E, Camillo BD, Facchinetti A, et al. Argot2: a large scale function prediction tool relying on semantic similarity of weighted gene ontology terms. *BMC Bioinformatics* (2012) 13:S14. doi:10.1186/1471-2105-13-S4-S14
33. Ong E, He Y. GOfox: semantics-based simplified hierarchical classification and interactive visualization to support GO enrichment analysis. *CEUR Workshop Proc* (2015) 1515:1–2.
34. He Y, Xiang Z. Bioinformatics analysis of bacterial protective antigens in manually curated Protegen database. *Procedia Vaccinol* (2012) 6:3–9. doi:10.1016/j.provac.2012.04.002
35. Leligdowicz A, Fischer WA II, Uyeki TM, Fletcher TE, Adhikari NK, Portella G, et al. Ebola virus disease and critical illness. *Crit Care* (2016) 20:217. doi:10.1186/s13054-016-1325-2
36. Saiz JC, Vazquez-Calvo A, Blazquez AB, Merino-Ramos T, Escribano-Romero E, Martín-Acebes MA. Zika virus: the latest newcomer. *Front Microbiol* (2016) 7:496. doi:10.3389/fmicb.2016.00496
37. Kling HM, Nau GJ, Ross TM, Evans TG, Chakraborty K, Empey KM, et al. Challenges and future in vaccines, drug development, and immunomodulatory therapy. *Ann Am Thorac Soc* (2014) 11:S201–10. doi:10.1513/AnnalsATS.201401-036PL
38. Collins BS. Gram-negative outer membrane vesicles in vaccine development. *Discov Med* (2011) 12:7–15.
39. Godlewska R, Kuczkowski M, Wyszyńska A, Klim J, Derlatka K, Woźniak-Biel A, et al. Evaluation of a protective effect of in ovo delivered *Campylobacter jejuni* OMVs. *Appl Microbiol Biotechnol* (2016) 100:8855–64. doi:10.1007/s00253-016-7699-x
40. Patel S, Mathivanan N, Goyal A. Bacterial adhesins, the pathogenic weapons to trick host defense arsenal. *Biomed Pharmacother* (2017) 93:763–71. doi:10.1016/j.biopha.2017.06.102
41. Ramos HC, Rumbo M, Sirard JC. Bacterial flagellins: mediators of pathogenicity and host immune responses in mucosa. *Trends Microbiol* (2004) 12:509–17. doi:10.1016/j.tim.2004.09.002
42. Lähteenmäki K, Kukkonen M, Korhonen TK. The Pla surface protease/adhesin of *Yersinia pestis* mediates bacterial invasion into human endothelial cells. *FEBS Lett* (2001) 504:69–72. doi:10.1016/S0014-5793(01)02775-2
43. Schulz GE. The structure of bacterial outer membrane proteins. *Biochim Biophys Acta* (2002) 1565:308–17. doi:10.1016/S0005-2736(02)00577-1
44. Zygmunt MS, Hagius SD, Walker JV, Elzer PH. Identification of *Brucella melitensis* 16M genes required for bacterial survival in the caprine host. *Microbes Infect* (2006) 8:2849–54. doi:10.1016/j.micinf.2006.09.002

45. Sun Y, Li Y, Exley RM, Winterbotham M, Ison C, Smith H, et al. Identification of novel antigens that protect against systemic meningococcal infection. *Vaccine* (2005) 23:4136–41. doi:10.1016/j.vaccine.2005.03.015
46. Ng VH, Cox JS, Sousa AO, MacMicking JD, McKinney JD. Role of KatG catalase-peroxidase in mycobacterial pathogenisis: countering the phagocyte oxidative burst. *Mol Microbiol* (2004) 52:1291–302. doi:10.1111/j.1365-2958.2004.04078.x
47. Li Z, Howard A, Kelley C, Delogu G, Collins F, Morris S. Immunogenicity of DNA vaccines expressing tuberculosis proteins fused to tissue plasminogen activator signal sequences. *Infect Immun* (1999) 67:4780–6.
48. Blair JM, Webber MA, Baylay AJ, Ogbolu DO, Piddock LJ. Molecular mechanisms of antibiotic resistance. *Nat Rev Microbiol* (2015) 13:42–51. doi:10.1038/nrmicro3380
49. Ni Z, Chen Y, Ong E, He Y. Antibiotic resistance determinant-focused *Acinetobacter baumannii* vaccine designed using reverse vaccinology. *Int J Mol Sci* (2017) 18:458. doi:10.3390/ijms18020458
50. Lopera-Madrid J, Osorio JE, He Y, Xiang Z, Adams LG, Laughlin RC, et al. Safety and immunogenicity of mammalian cell derived and modified-vaccinia ankara vectored African swine fever subunit antigens in swine. *Vet Immunol Immunopathol* (2017) 185:20–33. doi:10.1016/j.vetimm.2017.01.004

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

The reviewer SZ and handling Editor declared their shared affiliation.