An update on "Reverse vaccinology": the pathway from genomes and epitope predictions to tailored, recombinant vaccines

Marcin Michalik[1,*], Bardya Djahanshiri[2,*], Jack C. Leo[3], Dirk Linke[1]

(1) University of Oslo, Department of Biosciences, 0371 Oslo, Norway

(2) Goethe-University, Department for Applied Bioinformatics, D-60438 Frankfurt, Germany

(3) Nottingham Trent University, Department of Biosciences, Nottingham NG11 8NS, UK

[*] These authors contributed equally to the work

E-mail: dirk.linke@ibv.uio.no

Abstract

In this chapter, we review the computational approaches that have led to a new generation of vaccines in recent years. There are many alternative routes to develop vaccines based on the concept of reverse vaccinology. They all follow the same basic principles – mining available genome and proteome information for antigen candidates, and recombinantly expressing them for vaccine production. Some of the same principles have been used successfully for cancer therapy approaches. In this review, we focus on infectious diseases, describing the general workflow from bioinformatic predictions of antigens and epitopes down to examples where such predictions have been used successfully for vaccine development.

Keywords: reverse vaccinology, vaccine design, epitope prediction, surface proteins, peptide epitopes, core genome

Running head: Reverse Vaccinology

1. INTRODUCTION

The successful removal of a pathogen from the human body by the adaptive immune system requires the recognition of the pathogen's molecules as "foreign". Molecular patterns can be recognized by both branches of the mammalian immune system, innate immunity and adaptive immunity. The innate immune system recognizes widely conserved molecular features common to many pathogens (so called pathogen-associated molecular patterns), allowing this branch of the system to mount a rapid response to early signs of infection. The adaptive immune response is more specific, and the molecular patterns allowing the adaptive immune system to detect pathogens are called antigenic, or immunogenic. Furthermore, the site of the antigen to which the antigen-binding receptors actually bind is called the antigenic determinant or epitope. In most cases antigens possess several different epitopes; however, they can vary in immunogenicity, which leads to the phenomenon of so-called immunodominant epitopes.

The two branches of the immune system work on different principles. The innate immune system is inherited from the parents and is genetically fixed for life, while the adaptive branch is key to the recognition of new pathogenic structures. The adaptive immune system again is composed of two arms: the humoral and the cellular immune responses. These responses are mediated by two classes of lymphocytes, called B and T cells, respectively. B cells are able to express unique immunoglobulin receptors localized on the cell surface. These immunoglobulin receptors possess a variable antigen-binding site permitting vertebrates to specifically recognize and bind potentially billions of different epitopes. B cells are activated upon contact with an antigen, with or without the help of T-helper cells (see below). Protein antigens typically activate B cells directly ***(1)***. As soon as an antigen binds to the immunoglobulin receptor of a naïve B cell, the B-cell is stimulated to proliferate and differentiate into an antibody-producing plasma cell (effector cell) ***(2)*** with the sole task of amplifying a single type of specifically binding antibody that is able to bind its cognate antigen while circulating freely within the blood and lymph.

As soon as an immunoglobulin binds to a pathogen, the activity of the pathogen is reduced and it is marked (opsonized) for elimination by cells of the innate immune system, neutrophils and macrophages, capable of phagocytosis and subsequent killing and

degradation of the pathogen. Some B cells, however, differentiate into a different cell type, so-called memory B cells. In case of the same antigen entering the host again, these cells are promptly activated to accelerate a stronger, secondary immune response. Memory B cells have the ability to persist in the host for several years, thereby allowing a long-lasting protection ***(3)***. It is this memory of the immune system that is exploited when vaccines are used.

The cellular immune response is mediated by a second type of equally important immune cells called T cells. T cells, like B cells, are stimulated to proliferate and differentiate into the mature state by specifically binding to antigens. However, antigen recognition by T cell receptors (TCRs) is only possible if the epitopes are presented as protein fragments on the surface of cells. The presentation of protein fragments requires distinct processing pathways, which include the partial degradation of proteins within host cells. Finally, after several enzymatic processing steps, some of the resulting fragments are displayed in the context of co-simulators on the cell surface by proteins of the major histocompatibility complex (MHC) ***(4)***. Upon activation, naïve T cells can develop into two major classes of effector cells. Each of them maintains the ability to bind the same MHC-peptide complexes that had led to their activation. Cytotoxic T cells (CTLs or $CD8^+$ cells[1]) destroy nearby infected or malignant / transformed cells. T-helper cells (T_h cells, or $CD4^+$) are a decisive factor in the activation of various immune reactions of T-dependent B cells, CTLs, macrophages and dendritic cells. In analogy to B cells, subpopulations of both $CD4^+$ as well as $CD8^+$ cells are capable of differentiating into memory T cells similarly enabling long-term protection. MHC proteins are key to these processes. There are two classes of MHC proteins, named MHC Class I and Class II. While MHC Class I are found on the surface of all nucleated cells, MHC Class II are exclusively found on the surface of professional antigen presenting cells (APCs),

---

[1] CD8 and CD4 are transmembrane glycoproteins. They function as co-receptors of T cell receptors on the surface of T cells. “CD” is an abbreviation for cluster of differentiation: a superscripted plus or minus sign indicates whether this type of cell actually does or does not express the specific receptor. CTLs do not possess a CD4 receptor, and are therefore unable to bind to the MHC II-peptide complex. In contrast, T-helper cells are unable to bind MHC I as they do not express CD8 receptors on their cell surfaces.

which are part of the innate immune system, mainly dendritic cells, macrophages, and B cells. Both MHC classes have variable binding pockets which specifically bind previously processed peptides in an extended linear conformation with high affinity. In both cases, the loaded MHC receptors are subsequently translocated from the endoplasmic reticulum (where the loading takes place) to the cell surface of the APC, where these peptides are presented to bind TCRs ***(3)***. Nevertheless, there are important differences between these two classes, as explained in the following paragraphs.

Reverse vaccinology

Since the British physician Edward Jenner introduced his smallpox vaccine to the Western world in the late 18th century, classical vaccinology became one the most successful counter-measures in the constant battle against infectious diseases. In many cases, governmental programs for exhaustive vaccination were able to push the number of new infections per year of previously prevalent diseases to almost zero ***(5)***. Prominent examples include the vaccination against smallpox that effectively eradicated the disease, and against polio, where incidence rates have dropped by more than 99 per cent since the late 1980s. Despite the ongoing success of classical vaccination strategies, a number of infectious diseases have remained recalcitrant to vaccine development, largely due to the inherent constraints of classical vaccine technology.

Usually the vaccine administered is a biological suspension of either inactivated or killed cells, polysaccharide capsules or toxoids ***(6)***. However, in many cases it is challenging to prepare a potent vaccine against a specific pathogen. Non-culturable microorganisms, antigens which are not expressed *in vitro*, pathogens with antigenic determinants that can trigger detrimental autoimmune reactions, as well as extremely heterogeneous strains, are only a few of the severe difficulties classical vaccinologists are confronted with today.

Recently, a new impetus was given to current vaccine research thanks to the growing number of available complete pathogen genomes. Based on the assumption that all (protein) antigens a pathogen can express at any time are encoded in its genome (and therefore available to the scientist without cultivation), the idea is to combine bioinformatics

and biotechnology to identify protein candidates for vaccine development. As this approach begins with the genome sequence, in contrast to starting from an entire living microorganism, it is called “reverse vaccinology” ***(7, 8)***. The first projects based on this approach used genome information only to naïvely select surface-localized proteins as a pool of possible candidates for subsequent classical animal experiments. In their pioneering work for the development of a vaccine against *Neisseria meningitidis* B (MenB), R. Rappouli and colleagues collected the sequences of 570 surface-localized proteins, of which about 350 could successfully be cloned and expressed in *Escherichia coli*. The purified proteins were then used to immunize mice, and the resulting sera were subjected to various immunoassays to test for the candidate protein's efficacy as a vaccine. The researchers found 28 proteins which showed consistently positive results in all immunoassays and were able to induce antibodies with bactericidal activity ***(9)***. Furthermore, five of these candidates were also highly conserved in the genome of distantly related strains. A subset of these candidates became the basis for the development of a vaccine called "4CMenB", which contains three recombinant protein antigens combined with outer membrane vesicles derived from the meningococcal strain NZ98/254 and has obtained market authorization for the European Union in January 2013 (Bexsero, Novartis International AG ***(10)***). Reverse vaccinology has since developed enormously ***(11, 12)***, in particular by using increasingly sophisticated bioinformatic methods to mine the large quantities of information provided by pathogen genomes and proteomes. In addition, the complexity of the immune system and the vast amount of data generated from systematic characterization of the human genome and of immune cells along with clinical and epidemiological parameters have required the development of bioinformatics data structures, tools and algorithms to handle and analyze them efficiently ***(13)***. These tools have proved invaluable for reverse vaccinology.

In this chapter, we review the computational approaches that have integrated this data, and that have led to new vaccines in recent years. We also briefly summarize the mode of action of antigens and vaccines, and how vaccines are able to provide long-term protection. We want to emphasize that there are many alternative routes to success in reverse vaccinology – thus, we focus on prominent examples of infectious diseases. We show the general workflow from bioinformatics predictions of antigens and epitopes down to examples where such predictions have been used for vaccines successfully.

Software pipelines for Reverse Vaccinology

An ideal protein vaccine candidate (PVC) has key attributes such as its accessibility by the host immune system. Identifying such proteins within a larger initial dataset, e.g. a bacterial proteome, is a recurring task in many reverse vaccinology workflows. Over the past fifteen years, several software pipelines have been designed specifically to automatize this process. Commonly, they integrate an array of tools for identifying and annotating features to the individual proteins. However, they differ in the way they exploit this information for collating an output subset of candidates. Filtering-based programs filter the proteins stepwise for those having desirable and lacking non-desirable features. Machine-learning (ML)-based programs, in contrast, have been trained a priori to correctly classify the input proteins into "candidates" and "non-candidates" based on the vector of their annotated features.

Table 1 lists the most popular pipelines together with the feature annotation tools they employ. Regardless of the method, most of the pipelines focus on the same type of features and thus, show some overlap in the tools they employ. Usually, desirable features fall into four basic categories: (i) high conservation across all strains of the pathogen (ii) (predicted) subcellular localization outside the cell/envelope (surface exposure), (iii) functional characterization (including on domain level) as a virulence factor or as a protein involved in host-pathogen-interactions, and (iv) antigenicity/immunogenicity, i.e. one or more predicted epitope(s) of cellular immune receptor classes. Likewise, non-desirable features are: (i) high sequence similarity to human or commensal bacterial proteins (or of a model host system, due to possible autoimmunity effects), and (ii) the presence of transmembrane helices which hamper the protein's purification and cloning.

In 2019, Dalsass et al. ***(14)*** benchmarked 6 pipelines for their ability to find protein vaccine candidates within the proteomes of 11 bacterial species. Intriguingly, the authors described a large variance in the number of proteins each pipeline outputs as PVCs. In addition, despite the similarities in feature prediction described earlier, the predictions were largely in disagreement (with the exception NERVE and Vaxign which are nearly identical). Moreover, none of pipelines could recover more than 76% of a set of known protective bacterial antigens extracted from the Protegen database. The best performing pipeline, Bowman/Heinson, is an optimized ML-based approach, which extends the set of features and annotation tools to include predictions for surface exposure, proteasomal cleavage, and a range of post-translational modifications as the most impactful. This suggests that careful

exploration of the feature space by increasing the number of features and annotation tools might help to increase the sensitivity. This approach is pursued by the more recently developed pipelines PanRV and especially by Vaxign-ML (module of Vaxign2) and ReVac.

The benchmark results further underline how each pipeline's performance depends on the input proteome and the feature annotation tools it employs. Filtering-based approaches rely on tool-specific, predetermined thresholds to accurately decide whether the desirable feature is present or is not. These thresholds, however, are rarely optimal for all inputs, i.e. they could be too strict for one species yet too permissive for another. Consequently, false-negative and false-positive feature predictions could lead to accumulation of less suitable PVCs in the output. Hence, avoiding parametrization with predetermined thresholds is a promising approach pursued by ML-based pipelines. Unfortunately, these approaches are still limited by the diversity, quality, and quantity of available training data as both types, protective and non-protective antigens, ideally require rigorous in vivo testing.

In conclusion, feature acumen paired with a conceptual understanding of the employed annotation tools is pivotal for choosing the appropriate pipeline in a new RV project. Only then troubleshooting problems and circumventing them by pipeline-independent analyses is possible. In the following paragraphs, we discuss the basic feature types as well as some of tools for their annotation.

		Tools and attributes used to identify key protein features					Undesirable protein features		
Pipeline	**Year**	**Conserved (Pan genome analysis)**	**Surface Localization**	**Virulence Related Function**	**Immunogenic Epitopes**	**Other**	**Host Protein Similarity**	**Many TM-helices**	**Other**
NERVE	2006 ***(15)***		pSORTb	SPAAN, BLASTp (vs. UniProt)			BLASTPp (vs. MHCPep ***(16)***)	HMM TOP	
VaxiJen	2007 ***(17)***					Physicochemical properties			
Vaxign	2010 ***(18)***	OrthoMCL	pSORTb	SPAAN	Vaxitope (MHC I, MHC II)		OrthoMCL	HMM TOP	
Jenner-predict	2013 ***(19)***	BLASTp	pSORTb	PFAM	IEDB (search)		BLASTp (vs.	HMM TOP	

							human genome)		
Vacceed	2014 ***(20)***		WoLf PSORT, SignalP, TargetP, Phobius, TMHMM		IEDB (search)			Phobius, TMHMM	
Bowman-Heinson	2017 ***(21)***		NetSurfP, NetAcet, TargetP, PSORTb, LipoP	SPAAN	GPS-CCD, GPS-ARM, Net Chop, CBTOPE (B cell), BepiPred (B cell) GPS-MBA (specific MHC II allele), PickPocket (MHC I), NetMHCPan (MHC I)	Glycosylation, Phosphorylation, PUPylation, SUMOylation, S-nitrosylation, Furin cleavage sites, Physicochemical properties		HMMTOP	
VacSol	2017 ***(22)***	DEG	PSORTb, CELLO2GO	VFDB, Mvir CELLO2GO	ABCPred (B-cell), Propred-I (MHC I), Propred (MHC II)		BLASTp (vs. human genome)	HMMTOP	
PanRV	2019 ***(23)***	Roary, BLASTp, DEG	PSORTb	VFDB, COG, UniProt	ABCPred (B-cell), Propred1 (MHC I), Propred (MHC II), VaxiJen		BlastP (vs. human genome/ gut microbiome)	HMMTOP	Mol. weight
ReVac	2019 ***(24)***	PanOCT, OrthoMCL, LS-BSR, Custom orthology prediction	LipoP, SignalP, PFAM, PSORTb, TMHMM	GO, SPAAN	IEDB (search), NetCTLpan (MHC I), IEDB-AR consensus prediction (MHC I/II),			TMHMM	IslandPath, SSR Finder

					IEDB-AR consensus prediction (linear B cell)				
Vaxign-ML (Vaxign2)	2020 ***(25)***		PSORTb, SignalP, TMHMM	SPAAN	IEDB-AR (MHC I only)	Compositional and physicochemical properties	BlastP (vs. human/ mouse/pig genome)	TMHMM	

Table 1: Software pipelines for reverse vaccinology.

Pan-genomic analysis

Apart from being a valuable approach to investigating the characteristics of a specific phylogenetic clade, pan-genomic analysis is indispensable for identifying conserved target proteins within a set of genomes of pathogenic strains within a single clade. The term first coined by Tettelin ***(27)*** is defined as the entire genomic repertoire accessible to the clade studied. It encompasses two subsets: the "core genome" and the "dispensable" or "accessory genome". While the former describes the intersection of genes (or open reading frames [ORFs)) shared by all strains of the clade, the latter comprises genes only found in subsets of strains. Such a classification is biologically meaningful as it allows us to differentiate between (core) genes considered essential for growth, and (accessory) genes encoding e.g. for supplementary pathways and functions which confer a selective advantage, such as antibiotic resistance or virulence genes that are limited to certain strains ***(28)***. Similarity between genes or proteins is usually determined by pairwise alignment. Particular thresholds are set for the percentage of sequence identity of the protein sequence over a percentage of pairwise aligned sequence length. However, depending on the phylogenetic resolution and the available quality and quantity of genomes it might be necessary to increase sensitivity. This can be done by incorporating additional methods such as orthology prediction ***(29)***, i.e. the prediction of genes among species or strains that originated by vertical descent from a single gene of their last common ancestor, as well as structural alignments. Relying solely on pairwise sequence alignments on the protein level, Tettelin

(27) chose a minimum of 50% identity over 50% of the sequence lengths, while Hiller ***(30)*** chose 70% to identify similar proteins within strains of *Streptoccocus agalacticae* and *S. pneumoniae*, respectively. At such levels of overall identity scores, it can be assumed that the identified proteins have identical functions (and are true orthologues). For the purpose of identifying target proteins it is nonetheless beneficial to choose considerably higher threshold values to exclude false positives early on in the workflow. The potential loss of immunogenic sequences due to the high threshold values is relatively low, as at least locally, epitopes need to be very highly conserved to be effective. Given the high specificity of the immune system's receptors, this is a good trade-off for the reduction of the number of proteins to analyze in subsequent steps.

To be even more conservative, some studies and pipelines ***(31)*** use databases to filter for so-called essential genes, i.e. genes indispensable for the survival and successful reproduction of the organism. The rationale behind this is that these genes are part of the core genome, are typically constitutively expressed, and so slowly evolving that they are highly conserved across all the strains of a pathogen. However, there are several problems with this approach in the context of reverse vaccinology. On the one hand, genome-wide identification of essential genes is labor-intensive as it requires elaborate mutagenesis or knockdown experiments. Consequently, available experimental data is scarce. Even the most used database, DEG ***(32)***, comprises only 66 genome-wide experiments on bacteria, covering an even smaller number of different species. Moreover, most studies conduct the experiments with organisms suspended in standard nutritional medium. Gene essentiality, however, is highly context-dependent and significantly influenced by the particular genome or strain studied as well as the experimental settings like medium composition, and environmental and growth conditions. Simple mapping of a target pathogen’s gene set against a database of essential genes, therefore, could result in a considerable number of incorrectly classified genes and should be interpreted cautiously.

Surface localization

To perform their functions at their native subcellular localization (SCL), newly synthesized proteins must be sorted and transported to their respective subcellular compartments. The SCL of proteins not only provides important clues to their function in the cell but is also important for judging their potential as vaccine targets. Surface-localized proteins are typically the first molecular patterns of pathogens that are in contact with the host immune system, and are generally considered the best candidates for recombinant vaccines. Determining the SCL of proteins by experimental means, such as subcellular fractionation combined with mass spectrometry, is accurate but time-consuming and expensive ***(33)***. Bioinformatics methods are an increasingly comprehensive and reliable way to determine the SCL of proteins in large datasets, as they contain defined (and thus detectable) signals in their sequence.

There are two basic types of prediction tools for subcellular localization. One predicts very specific sequence features such as signal peptides for the Sec, Tat, or lipoprotein pathways using TargetP, SignalP and related tools ***(34)*** or transmembrane segments ***(35)***. The other type predicts the exact localization of a protein by combining various localization-specific features ***(36, 37)*** or general features like amino acid composition ***(38)***, evolutionary information ***(39)***, structure conservation information ***(36)***, or gene ontology ***(40)***. The combination of different prediction tools in a pipeline increases the quality of the overall prediction significantly and can reduce false positive and false negative results ***(41)***. Last but not least, limiting the huge amount of protein sequence data to only the interesting, surface-localized vaccine candidates significantly reduces the workload for later immunogenicity prediction steps in the reverse vaccinology pipeline. Alternatively, experimental data such as proteomics approaches can be used to narrow down the number of candidates for further analysis ***(42)***.

Immunoinformatics: the prediction of epitopes

Ideal vaccine candidates are not only localized on the surface of the pathogen but will also contain multiple epitopes that elicit strong immune responses within the host organism. However, experimental identification of epitopes within a set of proteins is a very resource-

and time-intensive task making a computer-aided, complementary approach especially attractive.

While 'reverse vaccinology' describes the overall approach in opposition to classical - entirely wet-lab-based – vaccine development, a new branch of bioinformatics emerged around the same time, termed immunoinformatics or computational immunology – defined as the application of informatics techniques to molecules relevant to the immune system ***(43, 44)***. The ability to predict immunogenicity on the level of epitopes is a key tool for computer-aided vaccine design. Numerous tools exist for such predictions, for both MHC I and MHC II, as well as B cell-mediated immunity. This chapter can only provide a crude overview of the different obstacles all prediction tools face and gives a brief overview of the general strategies they pursue. As for all bioinformatics tools, it is advisable to use multiple tools in parallel and to compare the results to minimize false positive and false negative predictions. In fact, recent publications have shown that combining prediction tools to produce a consensus-like output can achieve superior predictive performances ***(45, 46)***.

MHC I and MHC II binding predictions

Generally speaking, MHC I binds and presents epitopes which are derived from proteolytically degraded intracellular proteins (e.g. from intracellular pathogens) and are 8-10 residues long. By contrast, MHC II epitopes are derived from extracellular sources (e.g. from extracellular pathogens), and are much longer on average (up to 25 residues ***(47)***). Originally, it was thought that these peptide epitopes would be recognized at least in part by their secondary structure, but structural data suggest that they are presented mostly in an extended form. Early prediction tools working under the wrong assumption accordingly gave inconsistent results ***(6)***. Additionally, MHC I and MHC II bind peptides very differently: as the molecular structure of MHC II requires longer peptides, due to its "open" binding pocket, the residues extending the binding pocket on both sides contribute to the overall peptide binding affinity ***(47, 48)***. To address this finding, modern MHC II epitope prediction tools often identify a binding core, i.e. a shorter subsequence within the longer peptide sequences of the query, which is predicted to bind to the pocket.

To use prediction tools efficiently for vaccine design, one has to consider that the human MHC molecules are encoded in a highly polymorphic locus called the human leukocyte antigen (HLA) locus on chromosome 6. There are profuse amounts of HLA alleles with different binding affinities to the same epitope sequence: more than ten thousand different human alleles have been identified and, to complicate things even further, within different populations, different alleles (i.e. variants) of the MHC genes are present in different ratios.

Various online methods are available for the prediction of epitopes, ranging from sequence-based to structure-based (using e.g. homology modelling or docking) methods. Table 2 shows a selection of sequence-based bioinformatics tools used for MHC I or MHC II predictions, which have the advantage of speed over structure-based methods and are therefore more favorable for large-scale analysis of peptides.

Authors	Method	Publication	Output
MHC I			
Bui et al.	QM	***(49)***	IC_{50} (nM)
Sidney et al.	QM	***(50)***	
Nielsen et al.	ANN	***(51)***	IC_{50} (nM)
Peters et al.	QM	***(52)***	IC_{50} (nM)
Kim et al.	QM	***(53)***	IC_{50} (nM)
Moutaftsi et al.	QM	***(54)***	Percentile rank
Nielsen et al.	ANN, Pan-specific	***(55)***	IC_{50} (nM)
Karosiene et al.	ANN, Pan-specific	***(46)***	IC_{50} (nM)
Zhang et al.	QM	***(45)***	IC_{50} (nM)
Rasmussen et al.	ANN, Pan-specific	***(56)***	$T_{1/2}$(h) and IC_{50} (nM)
O'Donnell et al.	ANN, Pan-specific	***(57)***	IC_{50} (nM)
Bassani-Sternberg et al.	Probabilistic Mixture Model	***(58)***	Binding Score
Jurtz et al.	ANN, Pan-specific	***(59)***	IC_{50} (nM)
Singh et al.	QM	***(60)***	Binding Score
MHC II			
Bui et al.	QM	***(49)***	IC_{50} (nM)
Jensen et al.	ANN, Pan-specific	***(61)***	IC_{50} (nM)
Reynisson et al.	ANN, Pan-specific	***(62)***	IC_{50} (nM)
Sidney et al.	QM	***(50)***	IC_{50} (nM)
Singh et al.	QM	***(63)***	Binding Score
Nielsen et al.	QM	***(64)***	IC_{50} (nM)
Hoof et al.	ANN, Pan-specific	***(65)***	IC_{50} (nM)
Sturniolo et al.	QM	***(38)***	IC_{50} (nM)
Wang et al.	ANN, QM	***(26)***	Probability

Racle et al.	Probabilistic Mixture Model	***(66)***	Binding Score
Chen et al.	DNN	***(67)***	Probability
Liu et al.	DNN	***(68)***	IC_{50} (nM)

Table 2 Methods: QM: quantitative matrix-based methods (QM combine a matrix-based approach with a strategy to quantify the prediction scores), A/DNN: artificial/deep neural networks, $T_{1/2}$(h): half-life of the antigen-MHC-I complex in hours at 37°C

State-of-the-art sequence-based approaches attempt to predict the binding quality of a query sequence by abstracting from the sequence information of peptides with experimentally determined binding affinities. By doing so, they are able to generate models for each individual MHC variant. Matrix-based methods try to derive position-specific binding coefficients for each residue from a database of known binders of the same length. For the prediction, each position of a query sequence is evaluated individually, yielding a score of congruousness to its respective position in the abstract model of a binding sequence. To predict the binding quality of the complete query sequence, the final score is given as the sum of the scores of the individual positions. This approach can be modified by adding weights to certain positions (so-called anchor positions) to increase their impact on the final score.

A second group of prediction tools relies on machine learning approaches or stochastic models like support vector machines, artificial neural networks or Hidden Markov models to predict the binding quality of a query sequence. Generally speaking, all of these approaches attempt to refine a model by adjusting internal parameters to the sequence information provided by a collection of known binders. Therefore, a set of known binders is used to train the model, i.e. to adjust internal parameters in such a way as to enable accurate prediction of binding quality based on empirical data (supervised learning).

Some tools in both groups also include strategies to quantitatively predict the binding of a query sequence. By incorporating either position-specific affinity contributions (matrix-based approaches) or statistical regression analysis (machine learning approaches), the user can readily compare experimentally determined IC_{50} or K_d values with predicted ones. However, there are no pre-defined absolute threshold values clearly separating query sequences into either binders or non-binders. Rather, it is advisable to define cut-off values for each MHC allele individually ***(69)*** *using percentile ranks*.

It is important to note that all the tools, regardless of approach, heavily rely on experimental data on the measured binding affinities of peptide sequences for a specific MHC variant. Therefore, the quality of the prediction is determined by how well the binding space of a particular MHC variant is explored by the available data. Unfortunately, for many alleles data are scarce; this has led to the development of pan-specific methods for MHC binding prediction. These use known MHC binders to known MHC alleles to infer binding for unknown pairs. Typically, such approaches are based on structural data where alleles with similar physico-chemical attributes in the binding-pocket are classed together using machine-learning approaches ***(70–72)***.

In recent years, data from MHC ligand elution assays have emerged as a second source of training data for the development of MHC epitope predictors. Using high-throughput mass spectrometry, it is possible to detect large quantities of MHC ligands from a pool of extracted, surface presented epitopes, i.e. MHC ligands. In contrast to affinity values obtained from binding assay data, elution data does not provide a quantitative value to rank the epitopes relative to each other. Nevertheless, the large amount of data still helps to characterize binding motifs of different alleles and thereby increase the sensitivity of epitope prediction. In fact, recent developments such as NetMHC(II)pan 4.0 ***(59, 62)***, MARIA ***(67)***, MHCFlurry 2.0 ***(57)*** and MixMHC(2)Pred 2.0 ***(58, 66)***, all rely on a combination of binding assay and elution data for training, which has contributed to their significantly improved performance - especially in the more challenging prediction of MHC II epitopes - over former state-of-the-art tools.

B cell epitope binding predictions

B cell (or antibody) epitopes are 16 residues long on average but are not presented in the context of MHC molecules. Therefore, they are especially hard to predict as crystallographic studies have shown that B cell receptors (BCRs) are capable of binding discontinuous protein epitopes as well as specific peptide sequences. Epitopes are called discontinuous if they are composed of distant sequence segments which are brought into close proximity due to the protein's tertiary structure. Contemporary tools for identifying B cell epitopes can be divided into those relying solely on primary structure information and those additionally

incorporating structural data. The first group of tools calculate a prediction by considering a set of descriptors such as the propensity for a sequence segment to form a continuous, linear secondary structure, physico-chemical attributes, surface-accessibility and amino acid composition ***(73)***. In general, these tools yield reasonable accuracy for continuous (linear) epitopes, but fall short when identifying discontinuous epitopes ***(74)***. To surmount this shortcoming, prediction calculations by the second group of tools include secondary structure information, calculated surface accessibilities and/ or protrusion indices, in addition to information about the protein's three-dimensional structure and the structure of known antigen-BCR complexes. Popular sequence-based tools are *BepiPred* ***(75)*** and *BepiPred 2.0* ***(76)***, ABCpred ***(77),*** BEST ***(78)***, *LBTope* ***(79)***, and *EpiDope* ***(80)***. Commonly used structure-based tools are CBTOPE ***(81)***, ElliPro ***(82)***, Paratome ***(83)***, PEPOP ***(84)***, BEEPro ***(85)*** and DiscoTope 2.0 ***(74)***. It is even claimed that benchmarking has shown that the latter two tools are able to achieve high accuracy levels similar to MHC prediction tools ***(75)***.

Many of the tools for MHC I, II and BCR epitope prediction offer web interfaces which allow thorough testing of their predictive powers before applying them in a larger scale. A very useful analytical resource is the Immune epitope database (IEDB), funded by the National Institute of Health ***(52)***. In addition to providing a database of binding epitopes and their affinities (where available, also including elution data), the IEDB furnishes a regularly updated compilation of self-developed and newly-implemented popular prediction tools accessible via a single intuitive web interface.

Methods for using full length antigens (proteins) as vaccines

All vaccines work in a similar way: by presenting foreign antigens to the immune system in order to activate a specific immune response. The aim of vaccination is usually to induce long-term protection through memory B cells ***(86)***. The composition of vaccines can be diverse. Traditional formulations include live attenuated vaccines, which are composed of live viruses or bacteria that have been weakened in the lab to lower virulence by long-term passaging or genetic engineering (deletions in genes required for virulence) but are still able to activate the immune system. They elicit a strong response that can result in lifelong immunity with a minimal number of doses. Despite their advantages, live attenuated

vaccines can have many drawbacks. Potential problems include difficulties with storage and transportation, where inappropriate handling may cause loss of vaccine efficacy. In addition, there are cases where this type of vaccine cannot be used, e.g. when patients take anti-infective drugs, or are immunocompromised for any reason. There is a risk that attenuated vaccines can revert to a fully virulent pathogen (e.g. oral poliovirus vaccine ***(87)***). Last but not least, the attenuation process itself is lengthy and depends on random events out of the control of the researchers (examples: BCG tuberculosis vaccine, Yellow fever rotavirus vaccine) ***(88, 89)***.

An alternative method is to inactivate the pathogens before use as a vaccine. This method is safer compared to the live attenuated vaccines, but is less potent in inducing immune responses. In short, such vaccines contain pathogens killed by heat or chemical treatment (i.e. formaldehyde). Risks related to such vaccines include errors in the inactivation. Because the inactivated pathogen does not reproduce in the host organism, there is a need for one or more "boosters", i.e. administration of additional doses of the vaccine after defined intervals. (examples: Cholera vaccine, Hepatitis A vaccine, Rabies) ***(86)***.

With better biochemical and immunological methods available, it has been possible to engineer vaccine formulations by only using active antigens (rather than complete pathogens). This is referred to as a subunit vaccine. It uses only specific parts of a pathogen to immunize against disease. The search for such components is typically focused on surface-exposed or secreted antigens, which provide the best accessibility for antibodies and other immune mechanisms ***(86, 90)***. Using purified proteins as a vaccine component is a widely used technique today. With bioinformatics, it is possible to select ideal antigen candidates for subunit vaccines, which have many advantages over the "whole-pathogen" approaches ***(91)***. Subunit vaccine production is a safe process as it does not require the culturing of dangerous pathogens. The final product is also safer to use ***(92, 93)***: there is no infectious material, and thus no risk of the vaccine strain reverting to a harmful pathogen. In addition, it is possible to control all ingredients of the vaccine. Traditional vaccines induce very strong immunological responses with a very small dose; often this high response is not really necessary and does not always translate into later protection. In subunit vaccines, antigens are tested individually, and the kinds of responses they provide are known. Thus, it is in principle possible to customize vaccines for specific patient groups (for example immunocompromised patients or patients already suffering from an infectious disease) ***(88)***.

Examples of protein subunit vaccines

A vaccine against pertussis containing purified proteins was first created in 1981 in Japan by Sato and Sato, who purified the antigenic proteins by classical biochemical methods from cultures of the pathogen – with the obvious problems in biological safety and with upscaling of the procedure ***(94)***. Another example is the Hepatitis B vaccine which contains one of the proteins from the viral envelope – the Hepatitis B surface antigen (HBsAg). This was one of the first protein-based vaccines, and while at first the protein was obtained from natural human plasma, it was later successfully expressed recombinantly in yeast cells. Today, this is the production method of choice for human vaccines (Table 3) ***(95)***. Another example of a subunit vaccine on the market is the one against *Bacillus anthracis*. Although the components are still collected from pathogen cultures, which raises concerns about the safety of the procedure, the strength of the initial immune response and long-term efficacy are high ***(96)***.

Some studies have included production of plasmid-derived antigens using attenuated, avirulent *Bacillus* strains. Expressing these proteins in a *Bacillus* strain ensures properly processed and folded protein. The product is then purified from fermentation cultures and adsorbed onto an aluminum adjuvant. Preclinical studies showed that the vaccine as such is safe, well-tolerated and can induce an immune reaction with long-term immunity. Researchers are also looking for new targets using of bioinformatics, now that the complete genome of the clinical strain is available ***(97, 98)***.

Two new vaccines against Human papillomavirus (HPV) have been brought to the market recently – Cervarix and Gardasil (Silgard). Both contain proteins from the capsids of different virus strains – HPV16, 18 and HPV6, 11, 16, 18, respectively - and differ in the formulation of enhancers and adjuvants. In 2014, the US Food and Drug Administration approved another new HPV vaccine from Merck, Gardasil 9, which protects against 9 subtypes of the virus (HPV6, 11, 16, 18, 31, 33, 45, 52, and 58). These vaccines are all produced recombinantly using yeast cells (or insect cells for Cervarix) ***(99–101)***.

In ongoing Phase III clinical trials (NCT01563263), promising results have been obtained for a vaccine against *Pseudomonas aeruginosa* (IC43) (Table 4). This is an outer membrane protein-based vaccine containing an OprF/OprI fusion with a His tag. The product is expressed in *E.coli* from a plasmid. The vaccine gives good immune responses with and

without an alum adjuvant ***(102–104)***.

- No need to culture dangerous pathogens
- Problems with toxic or oncogenic parts of the pathogen, or with antigens potentially causing allergies or auto-immune diseases, can be avoided
- Proteins can be altered by adding different chemical groups to improve immunogenicity, stability or solubility
- Quality of the final vaccine is higher and is more reproducible
- Distribution and storage is improved (high stability e.g. in freeze-dried form)
- No risk of reversion to a more virulent strain (in contrast to live attenuated vaccines)
- Using computational and bioinformatics methods potentially lowers the costs of initial research
- Production methods are comparatively easy to scale up

*Table 3 Advantages of protein-based vaccines **(92, 105)**.*

Methods for using predicted epitopes/peptides as vaccines

Producing complete proteins in a stable form for vaccines or other purposes is not always straightforward. Many potential vaccine targets are membrane proteins, are otherwise insoluble, or are prone to degradation or aggregation. Short peptide epitopes taken from vaccine target proteins are a promising alternative, as they can still be efficiently recognized and displayed by either MHC I or MHC II. In some cases, reducing a subunit vaccine to a single epitope has the additional advantage of removing deleterious further epitopes; examples where this can be important are epitopes that can cause cross-reactivity leading to autoimmune responses.

In principle, an unlimited number of defined peptide epitopes can be combined to create multi-epitope vaccines. To obtain such epitopes, both reverse vaccinology approaches based on bioinformatics predictions (see above), or more traditional techniques based on antisera can be used to fish for epitopes ***(88)***. One approach to using predicted peptide epitopes is to fuse them to a previously chosen protein scaffold as a carrier. This scaffold can itself play additional important roles in enhancing the immunological response, e.g. due to the

presence of T-helper cell epitopes in its own sequence. A distinct advantage of this method is that multiple epitopes from different target proteins or even from diverse pathogen strains can be combined to obtain wider spectrum of protection ***(92)***. Production and handling can also be improved in the process as the scaffold can be chosen according to desired properties (water solubility, non-toxicity, stability at room temperature, etc.).

An example for using predicted epitopes conjugated to a carrier scaffold is an ongoing study using *Aeromonas hydrophila* epitopes from outer membrane proteins (OmpF, OmpC) with the heat-labile enterotoxin B (LTB) of *Escherichia coli* as a scaffold ***(106)***. LTB has been reported to be an efficient adjuvant capable of eliciting a strong immune response ***(73)***. In four out of five cases (five different fusions), the authors found that the recombinant fusion proteins induce antibody production. The antisera generated by this process were able to recognize the native proteins from which epitopes were taken. All epitopes in the study were predicted as B cell epitopes using bioinformatics approaches and tools as described above.

There are also potential problems with using peptide-based epitopes as vaccines: removing an epitope from its native context risks losing immunogenic efficacy and as a result, general response to the vaccination ***(105)***. Examples for such context-dependent recognition by the immune system are the loss of secondary structure, or the fact that especially B cell antigens are known to be mostly (90%) discontinuous, non-linear antigens - they derive from different protein regions localized closely in space due to the three-dimensional structure. Such conformation-dependent recognition cannot always be achieved using only a linear peptide/epitope ***(107)***. Using suitable scaffold proteins for peptide epitopes can solve some of these problems, e.g. by adding sequences which will enhance binding and the stability of the peptide-MHC complex ***(108)***. Another option for optimization is to modify epitopes using β-amino acids instead of natural ones, which can increase the binding affinity to MHC dramatically. Such recombinant epitopes maintain the properties of natural epitopes because the side chains of the amino acids are identical between the α– and β–type. However, the modification improves resistance to proteases as the epitopes do not have the same peptide backbone, so that the epitope is protected from digestion before it is loaded on the MHC. Even changing one amino acid to its β– variant has dramatic effects on the overall stability of the peptide ***(109–111)***.

Another, less well understood disadvantages of using subunit vaccines is that they can be less efficient in inducing long lasting immunity ***(112, 113)***. Peptide vaccines often lack T-helper epitopes, especially when just a mix of peptides is used as a vaccine ***(114)***. To improve the response, vaccine formulations are modified with different immunostimulants (adjuvants) and also by conjugation of the peptides to carrier proteins which will enhance immunogenicity and immune system activation ***(88)***. The most common general adjuvant is an aluminum salt that can be found in many existing vaccines and is still used in new formulations in clinical trials and in pre-clinical phases ***(115)***. Many novel adjuvants are being tested currently, with the aim of finding adjuvants that are safe to use, can enhance the immune response of even of weakly binding peptides or proteins, and can play a direct role as a delivery system at the same time. Typically, these are different types of emulsions (water-in-oil and oil-in-water), e.g. MF59 which is composed of squalene (licensed for influenza vaccines in Europe), polymeric particles like PLA (polylactic acid) or PLGA (poly[lactide-co-glycolide] acid), liposomes (which can protect peptides from enzymatic digestion, keep the folded structure of antigen and elicit a high cellular immune response), virus-like particles (VLPs, self-assembling proteins which mimic the conformation of native viruses), inorganic nanoparticles, and carbon nanotubes ***(105, 116, 117)***. Other adjuvants include flagellin-based adjuvants, lipopolysaccharide, and other bacterial structures that co-stimulate the immune system ***(91)***, as well as complete avirulent (and thus safe) living cells expressing the foreign antigen on the surface. Examples include the use of a type III secretion system) ***(118)***) ***(119, 120)***, and the use of autodisplay systems based on type-V secretion systems ***(121, 122)***. As described above, in cases where the immunological memory is not lifelong, there is a need for additional “boosting” to increase and maintain the protectivity of a vaccine ***(123)***.

The most recent developments in reverse vaccinology include personalized vaccines, which are aimed at specific patient groups or even individuals. This is particularly relevant for anti-cancer vaccines, where the targets (cancer cells) are highly variable from patient to patient. As an example, GAPVAC, with a promising results from Phase I clinical trials ***(124)***, is a vaccine that uses patient-specific genes expressed in brain tumors and is based on peptides as well as cancer-specific mutations ***(125)***. A similar study is currently being performed using HEPAVAC, a patient-specific vaccine against liver cancer ***(126)***.

Currently, no peptide-based vaccine is licensed for human use, but there are currently over 400 clinical trials of peptide vaccines in progress ***(92, 127)***. A number of promising examples of peptide-based and other subunit vaccines are shown in Table.

There is an obvious need for more basic research and clinical trials and especially for long-term studies to demonstrate that reverse vaccinology approaches can yield vaccines that are potentially safer and at least as efficient as traditional vaccines. With increasing numbers of antibiotic-resistant bacteria, and with old and new viral diseases such as Ebola, Middle-East Respiratory Syndrome (MERS), most recently SARS-CoV-2, and others emerging or re-emerging, tailored vaccines are promising solutions to the continuous problem of infectious diseases. The great potential of patient-specific vaccines, especially for use in cancer therapy, where traditional approaches cannot be used at all, has barely been tapped.

Vaccine	Target	Notes	Stage	Active compound	Ref
Improvac	boar taint	Stimulation of the (pig) immune system to produce antibodies that ultimately block and reverse the accumulation of compounds responsible for boar taint.	On market (animal use)	synthetic incomplete analogue of gonadotropin-releasing factor (GnRF) (without hormone activity) linked with carrier protein	***(128)***
Recombitec WNV	West Nile virus	Combination of existing canarypox vaccine (ALVAC) with genes expressing two proteins from West Nile virus	On market (animal use)	*prM*/*E* genes	***(129)***
Vacc-4x	HIV	Synthetic peptides targeting HIV protein p24	Phase III	Peptides with adjuvants	***(130)***
Vacc-C5	HIV	Synthetic peptides targeting HIV glycoprotein gp120 (C5)	Phase II/III	Peptides with adjuvants	***(130)***
RECOMBIVAX HB	Hepatitis B virus	Recombinantly produced HBsAg protein in yeast cells	On market	Protein with aluminum adjuvant	***(101)***
IC43	*Pseudomonas aeruginosa*	Recombinant outer membrane protein-based vaccine	Phase II/III	OprF/OprI hybrid vaccine with N-terminal His tag	***(102, 104, 127)***
NDV-3	*Candida* sp.	Recombinant vaccine	Phase I/II	agglutinin-like sequence 3 protein (Als3p) from *Candida albicans* with aluminum hydroxide adjuvant	***(131, 132)***
SA4Ag	*Staphylococcus aureus*	Recombinant vaccine containing 2 different capsular polysaccharides and 2 surface proteins	Phase I/II	Polysaccharides CP5 and CP8; recombinant surface protein clumping factor A (rmClfA) and recombinant manganese transporter protein C (rP305A)	***(127, 133)***

PreviThrax	*Bacillus anthracis*	Recombinant protective antigen protein	Phase II	purified recombinant protective antigen protein	***(134)***
Respiratory Syncytial Virus (RSV) Vaccine	Respiratory Syncytial Virus (RSV)	F glycoprotein produced recombinantly in insect cells with a recombinant baculovirus	Phase II	Purified recombinant RSV F oligomers	***(135)***
Cenv3	Hepatitis C	Selected 3 peptides from 2 envelope proteins. Each was synthesized in 8 multiple antigenic peptides (MAPs)	Phase II	3 envelope peptides derived from 2 envelope proteins E1 and E2	***(127, 136)***
NeuroVax	Multiple Sclerosis	Vaccine contains three peptides which correspond to potentially pathogenic TCR's on T cells (which are over-expressed in 90% of multiple sclerosis patients)	Phase II/III	3 TCR peptides in aqueous solution and IFA	***(137)***
IC41	Hepatitis C	Vaccine contains 5 peptides derived from hepatitis C virus genotype 1 core. There are 4 cytotoxic T lymphocyte (CTL) epitopes and 3 helper epitopes.	Phase I/II	5 synthetic peptides with Poly-L-arginine as adjuvant	***(127, 138)***

Table 4. Selected list of ongoing clinical trials with subunit vaccines

References

1. Janeway CAJ, Travers P, Walport M, et al. (2001) Immunobiology, Garland Science, New York.

2. Alberts B, Johnson A, Walter P, et al. (2007) Molecular Biology of the Cell, Taylor & Francis.

3. Neumann J (2008) Immunbiologie, Springer-Lehrbuch, Berlin, Heidelberg.

4. Saha B (2001) Encyclopedia of life sciences, John Wiley & Sons, Ltd, Chichester, UK.

5. WHO UNICEF World Bank (2009) State of the world's vaccines and immunization, World Health Organization, Geneva.

6. Flower DR (2009) Bioinformatics for Vaccinology, John Wiley & Sons, Ltd, Chichester, UK.

7. Rinaudo CD, Telford JL, Rappuoli R, et al. (2009) Vaccinology in the genome era, The Journal of Clinical Investigation 119: 2515–2525.

8. Seib KL, Zhao X, Rappuoli R (2012) Developing vaccines in the era of genomics: A decade of reverse vaccinology, Clinical Microbiology and Infection 18: 109–116.

9. Pizza M, Scarlato V, Masignani V, et al. (2000) Identification of vaccine candidates against serogroup B meningococcus by whole-genome sequencing., Science 287: 1816–1820.

10. Medicinal Products and Human Use. Bexsero. Technical report, European Medicines Agency, http://www.ema.europa.eu/docs/en_GB/document_library/EPAR_.

11. Rappuoli R, Bottomley MJ, D'Oro U, et al. (2016) Reverse vaccinology 2.0: Human immunology instructs vaccine antigen design, The Journal of Experimental Medicine jem.20151960.

12. Burton DR (2017) What Are the Most Powerful Immunogen Design Vaccine Strategies? Reverse Vaccinology 2.0 Shows Great Promise., Cold Spring Harbor perspectives in biology 9:.

13. Hegde NR, Gauthami S, Sampath Kumar HM, et al. (2018) The use of databases, data mining and immunoinformatics in vaccinology: where are we?, Expert opinion on drug discovery 13: 117–130.

14. Dalsass M, Brozzi A, Medini D, et al. (2019) Comparison of Open-Source Reverse Vaccinology Programs for Bacterial Vaccine Antigen Discovery., Frontiers in immunology 10: 113.

15. Vivona S, Bernante F, Filippini F (2006) NERVE: new enhanced reverse vaccinology environment., BMC biotechnology 6: 35.

16. Brusic V (1998) MHCPEP, a database of MHC-binding peptides: update 1997, Nucleic Acids Research 26: 368–371.

17. Doytchinova IA, Flower DR (2007) VaxiJen: a server for prediction of protective antigens, tumour antigens and subunit vaccines., BMC bioinformatics 8: 4.

18. He Y, Xiang Z, Mobley HLT (2010) Vaxign: the first web-based vaccine design program for reverse vaccinology and applications for vaccine development., Journal of biomedicine & biotechnology 2010: 297505.

19. Jaiswal V, Chanumolu SK, Gupta A, et al. (2013) Jenner-predict server: prediction of protein vaccine candidates (PVCs) in bacteria based on host-pathogen interactions., BMC bioinformatics 14: 211.

20. Goodswen SJ, Kennedy PJ, Ellis JT (2014) Vacceed: a high-throughput in silico vaccine candidate discovery pipeline for eukaryotic pathogens based on reverse vaccinology., Bioinformatics (Oxford, England) 30: 2381–3.

21. Heinson AI, Gunawardana Y, Moesker B, et al. (2017) Enhancing the Biological Relevance of Machine Learning Classifiers for Reverse Vaccinology., International journal of molecular sciences 18:.

22. Rizwan M, Naz A, Ahmad J, et al. (2017) VacSol: a high throughput in silico pipeline to predict potential therapeutic targets in prokaryotic pathogens using subtractive reverse vaccinology., BMC bioinformatics 18: 106.

23. Naz K, Naz A, Ashraf ST, et al. (2019) PanRV: Pangenome-reverse vaccinology approach for identifications of potential vaccine candidates in microbial pangenome., BMC bioinformatics 20: 123.

24. D'Mello A, Ahearn CP, Murphy TF, et al. (2019) ReVac: a reverse vaccinology computational pipeline for prioritization of prokaryotic protein vaccine candidates., BMC genomics 20: 981.

25. Ong E, Wang H, Wong MU, et al. (2020) Vaxign-ML: supervised machine learning reverse vaccinology model for improved prediction of bacterial protective antigens, Bioinformatics 36: 3185–3191.

26. Goodswen SJ, Kennedy PJ, Ellis JT (2021) Computational Antigen Discovery for Eukaryotic Pathogens Using Vacceed., Methods in molecular biology (Clifton, N.J.) 2183: 29–42.

27. Tettelin H, Masignani V, Cieslewicz MJ, et al. (2005) Genome analysis of multiple pathogenic isolates of Streptococcus agalactiae: Implications for the microbial "pan-genome," Proceedings of the National Academy of Sciences of the United States of America 102: 13950–13955.

28. Vernikos G, Medini D, Riley DR, et al. (2015) Ten years of pan-genome analyses, Current Opinion in Microbiology 23: 148–154.

29. Nichio BTL, Marchaukoski JN, Raittz RT (2017) New Tools in Orthology Analysis: A Brief Review of Promising Perspectives, Frontiers in Genetics 8:.

30. Hiller NL, Janto B, Hogg JS, et al. (2007) Comparative Genomic Analyses of Seventeen Streptococcus pneumoniae Strains: Insights into the Pneumococcal Supragenome , Journal of Bacteriology 189: 8186–8195.

31. Vilela Rodrigues TC, Jaiswal AK, Sarom A de, et al. (2019) Reverse vaccinology and subtractive genomics reveal new therapeutic targets against Mycoplasma pneumoniae : a causative agent of pneumonia, Royal Society Open Science 6: 190907.

32. Luo H, Lin Y, Gao F, et al. (2014) DEG 10, an update of the database of essential genes that includes both protein-coding genes and noncoding genomic elements., Nucleic acids research 42: D574-80.

33. Thein M, Sauer G, Paramasivam N, et al. (2010) Efficient subfractionation of gram-negative bacteria for proteomics studies, Journal of Proteome Research 9: 6135–6147.

34. Emanuelsson O, Brunak S, Heijne G von, et al. (2007) Locating proteins in the cell using TargetP, SignalP and related tools, Nature Protocols 2: 953–971.

35. Punta M, Forrest LR, Bigelow H, et al. (2007) Membrane protein prediction methods, Methods 41: 460–474.

36. Su EC-Y, Chiu H-S, Lo A, et al. (2007) Protein subcellular localization prediction based on compartment-specific features and structure conservation, BMC bioinformatics 8: 330.

37. Yu NY, Wagner JR, Laird MR, et al. (2010) PSORTb 3.0: Improved protein subcellular localization prediction with refined localization subcategories and predictive capabilities for all prokaryotes, Bioinformatics 26: 1608–1615.

38. Yu C-S, Chen Y-C, Lu C-H, et al. (2006) Prediction of protein subcellular localization, Proteins 64: 643–651.

39. Rashid M, Saha S, Raghava GP (2007) Support Vector Machine-based method for predicting subcellular localization of mycobacterial proteins using evolutionary information and motifs, BMC Bioinformatics 8: 337.

40. Chou KC, Shen HB (2006) Large-scale predictions of gram-negative bacterial protein subcellular locations, Journal of Proteome Research 5: 3420–3428.

41. Paramasivam N, Linke D (2011) Clubsub-P: Cluster-based subcellular localization prediction for gram-negative bacteria and archaea, Frontiers in Microbiology 2: 218.

42. Dunston CR, Herbert R, Griffiths HR (2015) Improving T cell-induced response to subunit vaccines: opportunities for a proteomic systems approach, Journal of Pharmacy and Pharmacology.

43. Flower DR, Doytchinova IA (2002) Immunoinformatics and the prediction of

immunogenicity., Applied bioinformatics 1: 167–76.

44. Groot AS De, Sbai H, Aubin C Saint, et al. (2002) Immuno-informatics: Mining genomes for vaccine components., Immunology and cell biology 80: 255–69.

45. Zhang H, Lund O, Nielsen M (2009) The PickPocket method for predicting binding specificities for receptors based on receptor pocket similarities: Application to MHC-peptide binding, Bioinformatics 25: 1293–1299.

46. Karosiene E, Lundegaard C, Lund O, et al. (2012) NetMHCcons: A consensus method for the major histocompatibility complex class i predictions, Immunogenetics 64: 177–186.

47. Wang P, Sidney J, Dow C, et al. (2008) A systematic assessment of MHC class II peptide binding predictions and evaluation of a consensus approach, PLoS Computational Biology 4: e000048.

48. Zhang L, Udaka K, Mamitsuka H, et al. (2012) Toward more accurate pan-specific MHC-peptide binding prediction: A review of current methods and tools, Briefings in Bioinformatics 13: 350–364.

49. Bui H-H, Sidney J, Peters B, et al. (2005) Automated generation and evaluation of specific MHC binding predictive tools: ARB matrix applications., Immunogenetics 57: 304–14.

50. Sidney J, Assarsson E, Moore C, et al. (2008) Quantitative peptide binding motifs for 19 human and mouse MHC class I molecules derived using positional scanning combinatorial peptide libraries, Immunome Research 4: 2.

51. Nielsen M, Lundegaard C, Worning P, et al. (2003) Reliable prediction of T-cell epitopes using neural networks with novel sequence representations., Protein science : a publication of the Protein Society 12: 1007–17.

52. Peters B, Sette A (2005) Generating quantitative models describing the sequence specificity of biological processes with the stabilized matrix method, BMC Bioinformatics 6: 132.

53. Kim Y, Sidney J, Pinilla C, et al. (2009) Derivation of an amino acid similarity matrix for peptide:MHC binding and its application as a Bayesian prior, BMC Bioinformatics 10: 394.

54. Moutaftsi M, Peters B, Pasquetto V, et al. (2006) A consensus epitope prediction approach identifies the breadth of murine T(CD8+)-cell responses to vaccinia virus., Nature biotechnology 24: 817–819.

55. Nielsen M, Lundegaard C, Blicher T, et al. (2007) NetMHCpan, a method for quantitative predictions of peptide binding to any HLA-A and -B locus protein of known sequence., PloS one 2: e796.

56. Rasmussen M, Fenoy E, Harndahl M, et al. (2016) Pan-Specific Prediction of Peptide-

MHC Class I Complex Stability, a Correlate of T Cell Immunogenicity., Journal of immunology (Baltimore, Md. : 1950) 197: 1517–24.

57. O'Donnell TJ, Rubinsteyn A, Laserson U (2020) MHCflurry 2.0: Improved Pan-Allele Prediction of MHC Class I-Presented Peptides by Incorporating Antigen Processing, Cell Systems 11: 42-48.e7.

58. Bassani-Sternberg M, Chong C, Guillaume P, et al. (2017) Deciphering HLA-I motifs across HLA peptidomes improves neo-antigen predictions and identifies allostery regulating HLA specificity, PLOS Computational Biology 13: e1005725.

59. Jurtz V, Paul S, Andreatta M, et al. (2017) NetMHCpan-4.0: Improved Peptide-MHC Class I Interaction Predictions Integrating Eluted Ligand and Peptide Binding Affinity Data., Journal of immunology (Baltimore, Md. : 1950) 199: 3360–3368.

60. Singh H, Raghava GPS (2003) ProPred1: prediction of promiscuous MHC Class-I binding sites, Bioinformatics 19: 1009–1014.

61. Jensen KK, Andreatta M, Marcatili P, et al. (2018) Improved methods for predicting peptide binding affinity to MHC class II molecules., Immunology 154: 394–406.

62. Reynisson B, Alvarez B, Paul S, et al. (2020) 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 research 48: W449–W454.

63. Singh H, Raghava GPS (2001) ProPred: prediction of HLA-DR binding sites, Bioinformatics 17: 1236–1237.

64. Nielsen M, Lundegaard C, Blicher T, et al. (2008) Quantitative predictions of peptide binding to any HLA-DR molecule of known sequence: NetMHCIIpan., PLoS computational biology 4: e1000107.

65. Hoof I, Peters B, Sidney J, et al. (2009) NetMHCpan, a method for MHC class I binding prediction beyond humans., Immunogenetics 61: 1–13.

66. Racle J, Michaux J, Rockinger GA, et al. (2019) Robust prediction of HLA class II epitopes by deep motif deconvolution of immunopeptidomes, Nature Biotechnology 37: 1283–1286.

67. Chen B, Khodadoust MS, Olsson N, et al. (2019) Predicting HLA class II antigen presentation through integrated deep learning, Nature Biotechnology 37: 1332–1343.

68. Liu Z, Jin J, Cui Y, et al. (2019) DeepSeqPanII: an interpretable recurrent neural network model with attention mechanism for peptide-HLA class II binding prediction, bioRxiv 817502.

69. Paul S, Weiskopf D, Angelo M a, et al. (2013) HLA Class I Alleles Are Associated with Peptide-Binding Repertoires of Different Size, Affinity, and Immunogenicity, The Journal of Immunology 191: 5831–5839.

70. Doytchinova IA, Guan P, Flower DR (2004) Identifiying human MHC supertypes using bioinformatic methods, Journal of immunology 172: 4314–4323.

71. Sidney J, Peters B, Frahm N, et al. (2008) HLA class I supertypes: a revised and updated classification, BMC immunology 9: 1.

72. Doytchinova IA, Flower DR (2005) In silico identification of supertypes for class II MHCs, Journal of immunology 174: 7085–7095.

73. Ponomarenko J V., Regenmortel MHV van (2009) B-cell epitope prediction, In: Gu, J. and Bourne, P.E. (eds.) Structural Bioinformatics, Wiley-Blackwell.

74. Kringelum JV, Lundegaard C, Lund O, et al. (2012) Reliable B Cell Epitope Predictions: Impacts of Method Development and Improved Benchmarking, PLoS Computational Biology 8: e1002829.

75. Larsen JEP, Lund O, Nielsen M (2006) Improved method for predicting linear B-cell epitopes, Immunome research 2: 2.

76. Jespersen MC, Peters B, Nielsen M, et al. (2017) BepiPred-2.0: improving sequence-based B-cell epitope prediction using conformational epitopes., Nucleic acids research 45: W24–W29.

77. Saha S, Raghava GPS (2006) Prediction of continuous B-cell epitopes in an antigen using recurrent neural network, Proteins: Structure, Function and Genetics 65: 40–48.

78. Gao J, Faraggi E, Zhou Y, et al. (2012) BEST: Improved prediction of B-cell epitopes from antigen sequences, PLoS ONE 7: e40104.

79. Singh H, Ansari HR, Raghava GPS (2013) Improved Method for Linear B-Cell Epitope Prediction Using Antigen's Primary Sequence, PLoS ONE 8: e62216.

80. Collatz M, Mock F, Barth E, et al. (2020) EpiDope: a deep neural network for linear B-cell epitope prediction, Bioinformatics.

81. Ansari HR, Raghava GP (2010) Identification of conformational B-cell Epitopes in an antigen from its primary sequence., Immunome research 6: 6.

82. Ponomarenko J, Bui H-H, Li W, et al. (2008) ElliPro: a new structure-based tool for the prediction of antibody epitopes, BMC bioinformatics 9: 514.

83. Kunik V, Ashkenazi S, Ofran Y (2012) Paratome: An online tool for systematic identification of antigen-binding regions in antibodies based on sequence or structure, Nucleic Acids Research 40: W521–W524.

84. Moreau V, Fleury C, Piquer D, et al. (2008) PEPOP: computational design of immunogenic peptides, BMC bioinformatics 9: 71.

85. Lin SY, Cheng C, Su EC (2013) Prediction of B-cell epitopes using evolutionary information and propensity scales, BMC Bioinformatics 14: S10.

86. Patronov A, Doytchinova I (2013) T-cell epitope vaccine design by immunoinformatics., Open biology 3: 120139.

87. Shimizu H, Thorley B, Paladin FJ, et al. (2004) Circulation of Type 1 Vaccine-Derived Poliovirus in the Philippines in 2001, Journal of Virology 78: 13512–13521.

88. Moyle PM (2015) Progress in Vaccine Development, Curr. Protoc. Microbiol. 36: 1–17.

89. Centers for Disease Control and Prevention (2012) Epidemiology and Prevention of Vaccine-Preventable Diseases, Public Health Foundation, Washington DC.

90. Plotkin S (2014) History of vaccination., Proceedings of the National Academy of Sciences of the United States of America 2014: 1–5.

91. Moyle PM, Toth I (2013) Modern Subunit Vaccines: Development, Components, and Research Opportunities, ChemMedChem 8: 360–376.

92. Purcell AW, McCluskey J, Rossjohn J (2007) More than one reason to rethink the use of peptides in vaccine design., Nature reviews. Drug discovery 6: 404–414.

93. Moyle PM, Toth I (2008) Self-adjuvanting lipopeptide vaccines, Current medicinal chemistry 15: 506–516.

94. Sato Y, Sato H (1999) Development of Acellular Pertussis Vaccines, Biologicals 27: 61–69.

95. Michel M-L, Tiollais P (2010) Hepatitis B vaccines: protective efficacy and therapeutic potential., Pathologie-biologie 58: 288–295.

96. Cybulski RJ, Sanz P, O'Brien AD (2009) Anthrax vaccination strategies, Molecular Aspects of Medicine 30: 490–502.

97. Chun JH, Hong KJ, Cha SH, et al. (2012) Complete genome sequence of Bacillus anthracis H9401, an isolate from a korean patient with anthrax, Journal of Bacteriology 194: 4116–4117.

98. Keitel W a (2006) Recombinant protective antigen 102 (rPA102): profile of a second-generation anthrax vaccine., Expert review of vaccines 5: 417–430.

99. McKee SJ, Bergot A-S, Leggatt GR (2015) Recent progress in vaccination against human papillomavirus-mediated cervical cancer, Reviews in Medical Virology 25: 54–71.

100. Khallouf H, Grabowska A, Riemer A (2014) Therapeutic Vaccine Strategies against Human Papillomavirus, Vaccines 2: 422–462.

101. Merck, http://www.merck.com.

102. Rello J, Krenn C-G, Locker G, et al. (2017) A randomized placebo-controlled phase II

study of a Pseudomonas vaccine in ventilated ICU patients, Critical Care 21: 22.

103. Vincent J-L (2014) Vaccine development and passive immunization for Pseudomonas aeruginosa in critically ill patients: a clinical update., Future microbiology 9: 457–63.

104. Westritschnig K, Hochreiter R, Wallner G, et al. (2014) A randomized, placebo-controlled phase I study assessing the safety and immunogenicity of a Pseudomonas aeruginosa hybrid outer membrane protein OprF/I vaccine (IC43) in healthy volunteers., Human vaccines & immunotherapeutics 10: 170–83.

105. Skwarczynski M, Toth I (2014) Recent advances in peptide-based subunit nanovaccines, Nanomedicine 9: 2657–2669.

106. Sharma M, Dixit A (2015) Identification and immunogenic potential of B cell epitopes of outer membrane protein OmpF of Aeromonas hydrophila in translational fusion with a carrier protein, Applied Microbiology and Biotechnology.

107. Regenmortel MHV Van (1996) Mapping Epitope Structure and Activity: From One-Dimensional Prediction to Four-Dimensional Description of Antigenic Specificity, Methods 9: 465–472.

108. Sette A, Fikes J (2003) Epitope-based vaccines: An update on epitope identification, vaccine design and delivery, Current Opinion in Immunology 15: 461–470.

109. Guichard G, Zerbib A, Gal FA Le, et al. (2000) Melanoma peptide MART-1(27-35) analogues with enhanced binding capacity to the human class I histocompatibility molecule HLA-A2 by introduction of a β-amino acid residue: Implications for recognition by tumor-infiltrating lymphocytes, Journal of Medicinal Chemistry 43: 3803–3808.

110. Reinelt S, Marti M, Dédier S, et al. (2001) β-Amino Acid Scan of a Class I Major Histocompatibility Complex-restricted Alloreactive T-cell Epitope, Journal of Biological Chemistry 276: 24525–24530.

111. Webb AI, Dunstone MA, Williamson NA, et al. (2005) T Cell Determinants Incorporating β-Amino Acid Residues Are Protease Resistant and Remain Immunogenic In Vivo, The Journal of Immunology 175: 3810–3818.

112. Brito LA, Malyala P, O'Hagan DT (2013) Vaccine adjuvant formulations: A pharmaceutical perspective, Seminars in Immunology 25: 130–145.

113. Pulendran B, Ahmed R (2011) Immunological mechanisms of vaccination, Nature immunology 12: 509–517.

114. Berti F, Adamo R (2013) Recent mechanistic insights on glycoconjugate vaccines and future perspectives, ACS Chemical Biology 8: 1653–1663.

115. Plotkin S a. (2009) Vaccines: The fourth century, Clinical and Vaccine Immunology 16: 1709–1719.

116. Azmi F, Fuaad AAHA, Skwarczynski M, et al. (2014) Recent progress in adjuvant discovery for peptide-based subunit vaccines, Human Vaccines and Immunotherapeutics 10: 778–796.

117. Lua LHL, Connors NK, Sainsbury F, et al. (2014) Bioengineering virus-like particles as vaccines, Biotechnology and Bioengineering 111: 425–440.

118. Wieser A, Magistro G, Nörenberg D, et al. (2012) First multi-epitope subunit vaccine against extraintestinal pathogenic Escherichia coli delivered by a bacterial type-3 secretion system (T3SS)., International journal of medical microbiology : IJMM 302: 10–8.

119. Bumann D, Hueck C, Aebischer T, et al. (2000) Recombinant live Salmonella spp. for human vaccination against heterologous pathogens, FEMS Immunology and Medical Microbiology 27: 357–364.

120. Garmory HS, Leary SEC, Griffin KF, et al. (2003) The use of live attenuated bacteria as a delivery system for heterologous antigens., Journal of drug targeting 11: 471–479.

121. Nicolay T, Vanderleyden J, Spaepen S (2015) Autotransporter-based cell surface display in Gram-negative bacteria., Critical reviews in microbiology 41: 109–23.

122. Berg van Saparoea HB van den, Houben D, Jonge MI de, et al. (2018) Display of Recombinant Proteins on Bacterial Outer Membrane Vesicles by Using Protein Ligation., Applied and environmental microbiology 84:.

123. Demento SL, Siefert AL, Bandyopadhyay A, et al. (2011) Pathogen-associated molecular patterns on biomaterials: a paradigm for engineering new vaccines, Trends in biotechnology 29: 294–306.

124. Hilf N, Kuttruff-Coqui S, Frenzel K, et al. (2019) Actively personalized vaccination trial for newly diagnosed glioblastoma, Nature 565: 240–245.

125. GAPVAC, http://gapvac.eu/.

126. HepaVac, http://www.hepavac.eu/.

127. A service of the U.S. National Institutes of Health, https://clinicaltrials.gov/.

128. Improvac, http://improvac.com.

129. Garch H El, Minke JM, Rehder J, et al. (2008) A West Nile virus (WNV) recombinant canarypox virus vaccine elicits WNV-specific neutralizing antibodies and cell-mediated immune responses in the horse., Veterinary immunology and immunopathology 123: 230–9.

130. Bionorpharma, http://www.bionorpharma.com.

131. NovaDigm Therapeutics, http://www.novadigm.net/.

132. Schmidt CS, White CJ, Ibrahim AS, et al. (2012) NDV-3, a recombinant alum-adjuvanted vaccine for Candida and Staphylococcus aureus, is safe and immunogenic in healthy adults, Vaccine 30: 7594–7600.

133. Anderson AS, Miller A a., Donald RGK, et al. (2012) Development of a multicomponent Staphylococcus aureus vaccine designed to counter multiple bacterial virulence factors, Human Vaccines and Immunotherapeutics 8: 1585–1594.

134. Emergent Biosolutions, http://emergentbiosolutions.com/.

135. Raghunandan R, Lu H, Zhou B, et al. (2014) An insect cell derived respiratory syncytial virus (RSV) F nanoparticle vaccine induces antigenic site II antibodies and protects against RSV challenge in cotton rats by active and passive immunization., Vaccine 32: 6485–92.

136. El-Awady MK, Gendy M El, Waked I, et al. (2013) Immunogenicity and safety of HCV E1E2 peptide vaccine in chronically HCV-infected patients who did not respond to interferon based therapy, Vaccine.

137. Immune Response BioPharma, Inc., http://www.immuneresponsebiopharma.com.

138. Wedemeyer H, Schuller E, Schlaphoff V, et al. (2009) Therapeutic vaccine IC41 as late add-on to standard treatment in patients with chronic hepatitis C., Vaccine 27: 5142–51.