UNIVERSITY OF LIVERPOOL

Institute of Systems, Molecular and Integrative Biology

Improvement of selection criteria and prioritisation for neoantigen prediction

Thesis submitted in accordance with the requirements of

The University of Liverpool for the degree of Doctor in Philosophy

Phorutai Pearngam

September 2021

Abstract

A tumour-specific neoantigen-based cancer vaccine is a potentially powerful treatment option, which utilises unique mutated peptides from tumour cells to boost the immune response and selectively attack cancer cells. Thus, the characterisation of the specifically targeted peptides that can be selectively recognised by the immune system is essential for this approach. However, a major problem in neoantigen prediction is obtaining false positives, leading to poor outcomes in clinical research and practice. This thesis aims to address some of the computational issues in neoantigen prediction, including developing more reliable statistics for assessing peptide binding to MHC, and using machine learning to predict which peptides will generate an immune response. Specifically, the thesis introduces an approach for parameter estimation using the modified expectation maximisation (EM) framework with the method of moments for a two-component beta mixture model, representing the distribution of true and false scores from peptide binding prediction. The estimated parameters obtaining from the model can be further used for estimating false discovery rate (FDR) or a local peptide-level statistic such as the posterior error probability (PEP) to develop a robust method for MHC binding peptide selection. Next, the thesis introduces a new immunogenicity prediction model to classify immunogenic and non-immunogenic peptides using machine learning. A data set was assembled containing peptides classes as immunogenic and non-immunogenic peptides, and peptide features of physicochemical properties and homology features were used for constructing the Random Forest classifier for immunogenicity prediction. The two innovations were assembled into an end-to-end pipeline that provides the final probability described true MHC binding ability and the potential for immunogenicity. The final probability of MHC binding and T cell recognition provides a statistical framework to guide users in defining the appropriate thresholds, and prioritising peptides with the highest chance for being real neoantigens.

Declaration

I hereby declare that the content of this thesis corresponds to my work, which has been also submitted for the degree of Doctor in Philosophy at Chulalongkorn University, Thailand under the Double Degree Program between Chulalongkorn University and the University of Liverpool. Other sources of information were used in the text have been clearly acknowledged.

The experimental works described in Chapter 2 were obtained from a research team of Dr. Trairak Pisitkun at Chulalongkorn University System Biology Centre (CUSB), Bangkok, Thailand.

My contribution to the publication related to this research was as follow:

- Pearngam, Phorutai, Sira Sriswasdi, Trairak Pisitkun, and Andrew R. Jones. "MHCVision: estimation of global and local false discovery rate for MHC class I peptide binding prediction." *Bioinformatics* (2021); 10-1093

Acknowledgements

I would like to thank my supervisors both in Chulalongkorn University (CU) and University of Liverpool (UoL) for their extraordinary support and time dedicated to this research. I am very grateful to my supervisory team for their support and guidance throughout this journey.

I would like to express my thankful to Dr. Trairak Pisitkun for the initial idea that led to further works in this thesis, and I would also like to thank Dr. Sira Sriswasdi and Asst. Prof. Thanyada Rungrotmongkol for their support and invaluable suggestions in the topics related to mathematics and protein structure.

I would like to give my biggest thanks to Prof. Andy Jones, my primary supervisor at UoL, for his guidance throughout my PhD studies. Andy has provided me with tremendous support in my research thesis, encouraging me to develop my own ideas and understanding. He has been meeting regularly for updates and always showing an interest in my well-being. I am truly grateful for everything I have learnt under his supervision. I would like to extend my thanks to Prof. Dan Rigden for adding his knowledge of protein structure and always supports.

Likewise, I am thankful the colleagues in CUSB at CU and CBF staffs at UoL. I want also to express my appreciation for the opportunity and funding supported by the Science Achievement Scholarship of Thailand (SAST) and the Double Degree Program between the program of bioinformatics and computational biology at CU and UoL.

Finally, I am most grateful to my family and my beloved friend (PS) who always been there to give me the endless love and support me everything. They have always stood by me in difficult times and never given up on me no matter how hopeless. Their love, understanding, and support always been the inspiration and guide to pursue my goals.

Thesis Overview

Chapter 1: Introduction to neoantigen prediction and aims of thesis

The background of the adaptive immune system, including T cell and MHC interaction, cancer immunotherapy, cancer vaccines, the approach of bioinformatics for neoantigen predictions, prediction algorithms for MHC-peptide binding affinity, and T cell epitope prediction.

Chapter 2: The study of neoantigen prediction using existing bioinformatics software and public MHC-peptide binding affinity prediction tools

In this chapter, high-throughput sequencing data i.e. whole exome and RNA sequencing data were analysed using a collection of command-line tools in the Genome Analysis Toolkit (GATK) to determine specific mutations in a group of Thai cancer patients. The mutated peptides were extracted from exome DNA, and used as inputs for MHC-peptide binding predictors. The binding results and gene expression level obtained from RNA analysis were used to characterise candidate neoantigens. Moreover, the approach of molecular dynamic simulation was applied to assess the binding interaction between candidate peptides and MHC molecules. This chapter highlights the areas in which neoantigen identification workflow rely on computational methods, and where there is room for improvement, serving as the basis for the following chapters.

Chapter 3: The development of a model to estimate statistical properties from MHC-peptide binding affinity prediction

The background of statistical distribution models and the basic of the EM algorithm were initially described. In this chapter, MHC-peptide binding affinity was explored. A model, called MHCVision, that can estimate a probability for being false positive (FDR, PEP scores) for each predicted score from MHC-peptide binding prediction algorithms was developed. MHCVision was built based on mathematical models including the beta distribution model and

the EM algorithm. The script of this model was implemented using Python, as a command line application.

Chapter 4: The development of an immunogenicity prediction model for distinguishing immunogenic and non-immunogenic peptides using Random Forest

A prediction model for classifying immunogenic and non-immunogenic peptides was developed using the approach of the Random Forest algorithm. The classification model was trained from a set of features including physicochemical properties of amino acids in immunogenic and non-immunogenic peptides and a similarity between T cell epitopes and the host proteome. The model can predict the probability for each peptide to be immunogenic.

Chapter 5: A pipeline for ranking predicted neoantigens using the estimation of local FDR and immunogenicity prediction

In this chapter, a pipeline for neoantigen prioritisation was implemented by integrating the software from Chapter 3 and 4. The pipeline returns a final probability from the multiplication of true MHC binding probability (1-PEP) and immunogenic probability. The pipeline can provide probability scores related to MHC binding and immunogenicity, which could help users for ranking or selecting candidate peptides without a high risk for false positives. The results were validated using independent cancer neoantigen data sets that were not used for training in Chapter 4.

Chapter 6: General discussion, conclusion, and future work

This chapter includes the summary of the findings from the thesis, and a general discussion that extents from the models developed in this thesis. Additionally, the section describes future projects which may be generated from the current work.

Table of Contents

Chapter 3

List of Figures

List of Tables

Abbreviations

ACT	Adoptive Cell Transfer
AMBER	Assisted Model Building with Energy Refinement
APC	Antigen Presenting Cell
AUC	Area Under a Curve
BAM	Binary Alignment/Map
BLAST	Basic Local Alignment Search Tool
BQSR	Base Quality Score Recalibration
BWA	Burrows-Wheeler Aligner
CAR T cell	Chimeric Antigen Receptor T cell
CDF	Cumulative Density Function
CDR	Complementarity Determining Region
CML	Chronic Myeloid Leukemia
CTLA-4	Cytotoxic T lymphocyte associated Antigen 4
DC	Dendritic Cell
DNA	Deoxyribonucleic Acid
EBV	Epstein Barr Virus
ELISA	Enzyme-linked Immunosorbent Assay
ELISpot	Enzyme-linked Immune Absorbent spot
EM	Expectation Maximisation
Eq	Equation
ER	Endoplasmic Reticulum
FDA	Food and Drug Administration
FDR	False Discovery Rate
FPR	False Positive Rate
GATK	Genome Analysis Toolkit
GM-CSF	Granulocyte Macrophage Colony Stimulating Factor
HLA	Human Leucocyte Antigen
HLAp	HLA binding peptide
HPV	Human Papillomavirus
IC_{50}	The half maximal inhibitory concentration
IEDB	Immune Epitope Database
IMGT	The international ImMunoGeneTics information system

Indels	Insertions and Deletions
INF-γ	Interferon gamma
KL	Kullback–Leibler
KS	Kolmogorov-Smirnov
MD	Molecular Dynamics
MHC	Major Histocompatibility Complex
ML	Maximum Likelihood
MM-GBSA	Molecular Mechanics/Generalized Born Surface Area
mRNA	Messenger Ribonucleic Acid
MS	Mass Spectrometry
NGS	Next Generation Sequencing
PBMC	Peripheral Mononuclear Blood Cell
PCR	Polymerase Chain Reaction
PD-1	Program Cell Death 1
PDB	Protein Data Bank
PDF	Probability Density Function
PD-L1	Program Cell Death 1 Ligand
PEP	Posterior Error Probability
poly I:C	Polyinosine-Polycytidylic acid
RF	Random Forest
RMSD	Root Mean Square Displacement
RNA	Ribonucleic Acid
ROC	Receiver Operating Characteristic
SNV	Single Nucleotide Variant
TAP	Transporter associated with Antigen Processing
TCGA	The Cancer Genome Atlas
TCR	T Cell Receptor
TPM	Transcripts per Kilobase Million
TSV	Tab-separated Value
VCF	Variant Call Format
VEP	Ensembl Variant Effect Predictor
WES	Whole Exome Sequencing
WGS	Whole Genome Sequencing

Chapter 1

Introduction to neoantigen prediction and aims of thesis

1.1 T cells and Major histocompatibility complex (MHC) proteins

The basic context of the immune system is required for understanding the following details in this chapter. Therefore, in this section the background of the adaptive immune system involving T cells and MHC molecules is briefly described and referred to in the following sections.

1.1.1 The diversity of T cell receptors

T lymphocytes or T cells are one of the important white blood cells that play a crucial role in the adaptive immune response. They act as the primary effectors for cell-mediated immunity to confer response specificity using surface protein receptors to recognise foreign antigens [1]. There are two main classes of T cells, which are cytotoxic T cells (CD8+ cells) and helper T cells (CD4+ cells). Effector cytotoxic T cells directly kill cells that are infected with a virus or some other intracellular pathogens. In contrast, effector helper T cells help to stimulate the response of other cells which mainly are macrophages, B cells, and cytotoxic T cells [2]. A critical step in T cell development is making a functional T cell receptor (TCR). A mature T cell can have incredible diversity of TCRs that can react to a variety of random patterns, allowing the immune system to recognise many different types of pathogens. Each T cell bears about 30,000 antigen-receptor molecules on its surface, each receptor consisting of two different polypeptides chains including α and β chains. TCRs are able to bind such a wide variety of peptide-MHC complexes due to genetic recombination of gene segments creating α and β chains [3]. Both α and β chains have two regions including an amino-terminal variable (V) region and a constant (C) region, linked by disulfide bound. The V regions are encoded by separated gene segments, which are variable (V), diversity (D) and joining (J) gene segments. A random recombination of V, D, J gene segments generates high diversity of a V-region exon. The TCR α locus contains V and J gene segments while the TCR β consists of V, J and D segments. For the α chain, a V_α gene segment rearranges to a J_α segment to create a V-region

exon. Transcription and splicing of VJ_α exon to $C_\alpha$generates the mRNA that is translated to yield the TCR α chain protein. Rearrangement of three gene segments ($V_\beta, D_\beta,$ and J_β) of the β chain generates a functional VDJ_β of V-region exon that is transcribed and spliced to join to C_β, the resulting mRNA is translated to yield the TCR β chain protein [2]. Moreover, the combination of TCR α and TCR β creates more diversity of TCR proteins, those recombinant events result in an estimated 10^{15} possible different TCRs [4].

1.1.2 Major histocompatibility complex (MHC)

MHC molecules are cell surface proteins, their main function is to bind peptide fragments and present them for recognition by T cells. There are two major types of MHC proteins, which are MHC class I and class II according to types of T cells that are specific to each class. Only immune cells such as monocytes, B lymphocytes, antigen presenting cells (APCs) which are macrophages and dendritic cells (DCs), and epithelial cells can express both type of MHC molecules, while generally somatic cells can express only MHC class I molecules [5]. The human MHC is called the human leucocyte antigens (HLA) that maps to the short arm of chromosome 6 consisting of three regions including class I, class II, and class III. However, only class I and class II regions encode HLA molecules and function in the regulation of immune response [6]. For human MHC, the class I region consists of the classical HLA-A, HLA-B, and HLA-C genes, and their encoded proteins present peptides that can be recognised by cytotoxic CD8+ T cells. The class II region contains DR, DP and DQ gene families that can encode human MHC class II molecules including HLA-DRA, HLA-DRB1, HLA -DRB3, HLA-DRB4, HLA-DRB5, HLA-DQA1, HLA-DQB1, HLA-DPA1 and HLA-DPB1. Peptides displayed by MHC class II can be recognised by CD4+ T cells [7]. The diversity of human MHC alleles is high due to extensive polymorphism at most loci. The latest update of the international ImMunoGeneTics information system (IMGT) database in August 2019 contains

24,093 HLA alleles including 16,943 alleles of HLA class I, and 6,650 alleles of HLA class II [8]. Many of the alleles are exceptionally rare, carried only by a few individuals, but 1,122 alleles of HLA-A, -B, -C, -DRB, -DQA, -DQB, -DPA, and -DPB loci are common and well-documented, 415 alleles of these alleles were identified as "common" (having known frequencies) and 707 as "well-documented" base on HLA genotyping observations and available HLA haplotype data [9].

The standardised nomenclature system is typically used to define HLA polymorphism that refers to the multiple variations of allele loci. The notation system was initially designed based on HLA typing methods to detect and define HLA polymorphism such as serologic and cellular assays to DNA sequencing [10]. The current structure of nomenclature is a combination of alphanumeric characters and an asterisk (*) symbol that divides a name into two main components including the name of the locus i.e. HLA-A, -B, -C, -DRB, -DQA, -DQB, -DPA, and -DPB and the DNA sequence variant (Figure 1.1). Each HLA allele name has a unique number corresponding to up to four sets of digits separated by a colon (:) symbol. All alleles are named with at least four digits which cover the first two sets of digits. The first set described the encode HLA allele family which corresponds to the antigen group e.g. A*02, and the next set of digits after the first colon are used to define the DNA sequence variant that change in the amino acid sequence of the encoded protein which is usually assigned in a consecutive numerical order e.g. A*02:101 [11]. Longer names containing Field 3 or Field 4 are assigned if necessary, but the variations do not alternate at the protein level.

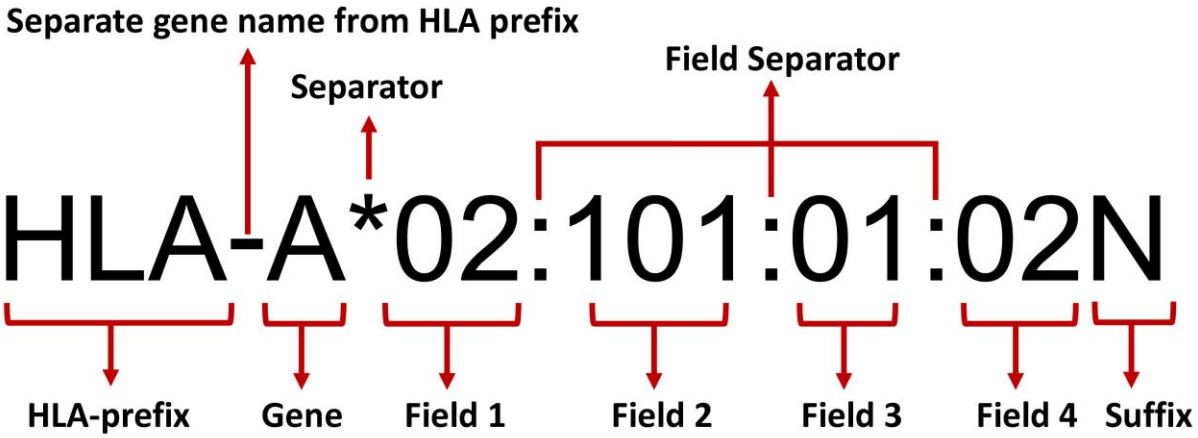

Figure 1.1 Structure of the nomenclature for HLA allele with four fields (adapted from http://hla.alleles.org).

1.1.3 MHC antigen processing and T cell recognition

MHC class I and class II have a similar function for short peptide delivery and presentation on cell surface. The difference between those two classes is the source of peptides for a step of antigen processing. Proteins that are processed in the cytosol such as intracellular proteins, tumour proteins, released proteins from viral infection or proteins from transplantation, are fragmented in the cytoplasm and presented via MHC I molecules. Cytosolic proteins are degraded by the proteasome into short peptides, then, those peptides are delivered to the endoplasmic reticulum (ER) via ER protein membrane, *transporter associated with antigen processing* (TAP) [12-14]. In the ER, peptides with specific length of 8 to 11 amino acids are potentially bound to the empty MHC I molecules. The complex of MHC I-peptide complete MHC protein folding, and the complete MHC I-peptide complexes are released from the ER and transported to the cell surface to present a peptide to CD8+ T cells [15] (Figure 1.2, top panel). MHC class II proteins can present longer peptides (11-30 amino acids) than those

presented by MHC class I molecules[16]. MHC class II molecules generally present peptides derived from the endocytic processing pathway, which are extracellular proteins or intracellular proteins degraded via the endosomal pathway [17]. The protein degradation is performed through the endosomal/lysosomal antigen-processing compartment, when the complex of MHC II molecules with the invariant chain (l_i) combine to that compartment, the mixture of proteases in the vesicle degrade the invariant chain resulting in the complex of MHC II and a fragment of l_i, called class II-associated invariant chain peptide (CLIP) [18, 19]. The enzyme called HLA-DM empties the MHC II groove by removing CLIP leading to the binding of MHC II and a peptide. The complex of the MHC II-peptide is moved to the cell surface by vesicular transportation for CD4+ T cells presentation [20, 21] (Figure 1.2, bottom panel). T cells recognise a peptide when bound to an MHC molecule. The TCR interacts with a ligand by making contacts with both MHC molecule and peptide by their TCRs, most predominantly via complementarity determining region 3 (CDR3) loops (Figure 1.3) [22].

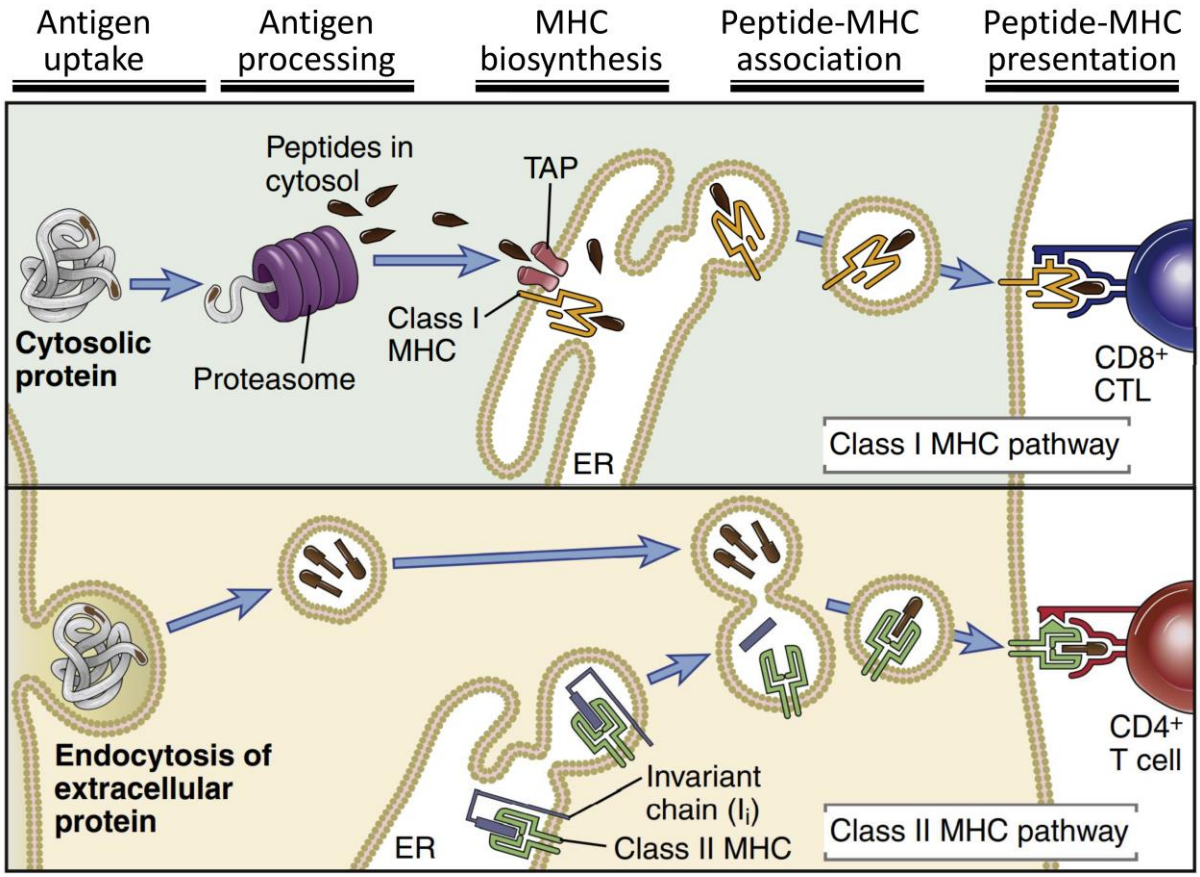

Figure 1.2 The antigen processing and MHC presentation pathways of MHC class I (top) and class II (bottom) [2].

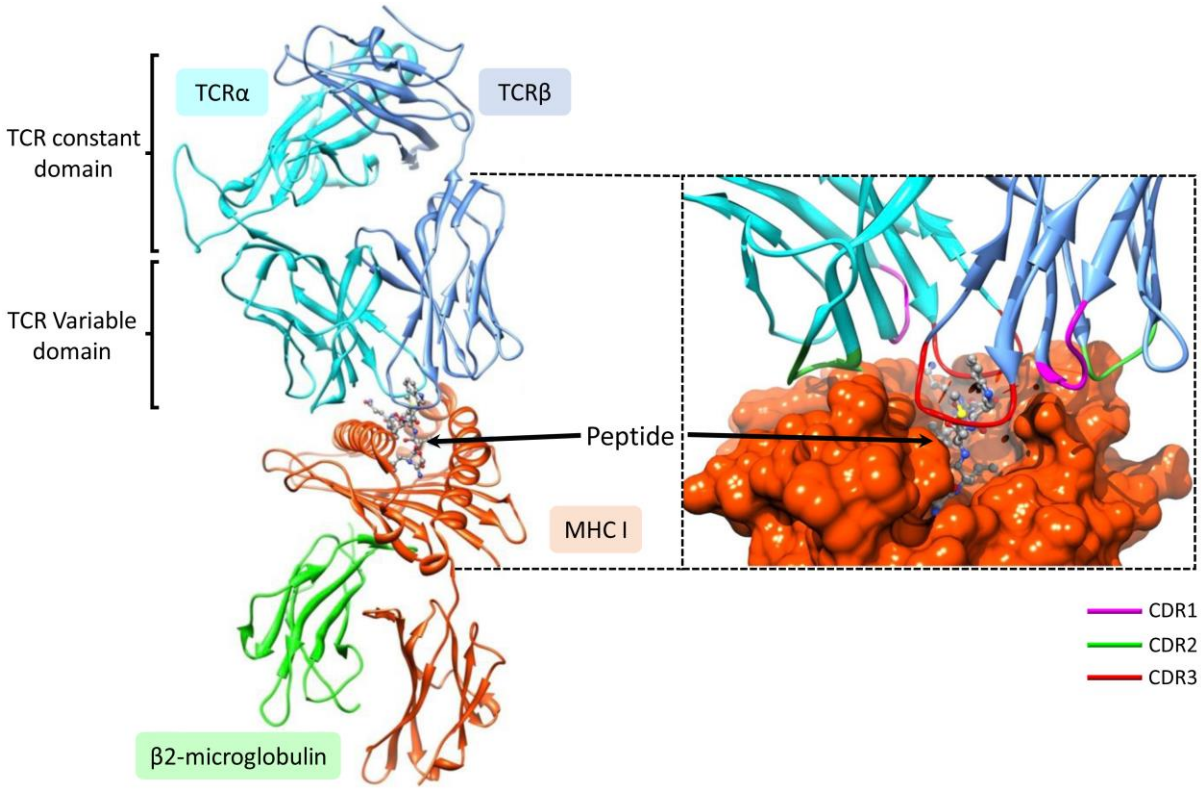

Figure 1.3 The 3D-structure of binding interaction of T cell receptor and the complex of HLA-A*02:01 presented peptide [23].

1.2 An introduction of Tumour immunology and cancer immunotherapy

Cancer immunotherapy has been developed during recent decades and has come to be a powerful approach for several types of cancer and also promising for treatment of the late stage or metastatic cancer. In contrast to the other therapeutic concepts, immunotherapy exploits the immune systems to attack cancer cells based on complementation or stimulation of the immune system specific for the individual [24]. As a result, this approach does have incomparable advantages over traditional anti-tumour therapy, which can prolong progression-free survival and overall survival. However, immunotherapy has complexity, the uncertainty of the therapeutic effect due to unpredictable factors, and the high cost of treatment [25]. This treatment may also cause severe adverse reactions due to an overactive immune system. Furthermore, several factors affect the effectiveness of cancer immunotherapy such as diversity in human immunity, due to genetics and internal microflora, tumour heterogeneity, and tumour

mutation burden [26]. Thus, the treatment responsiveness and survival rate after treatment, and prognosis of patients are uncertain. The principal of cancer immunotherapy is the concept of immune surveillance of tumours, which is the ability of the immune system to specifically identify and eliminate cancer cells that contain molecules or expressed antigens which never exist, or show aberrant expression, in normal cells. As a consequence, the lack or inhibition of immune surveillance due to an immune evasion of tumour cells can develop the cancer progression [27].Therefore, this therapeutic approach is intended to restore the ability of the immune system to combat cancer.

1.2.1 Immune surveillance of cancer and cancer immunoediting

There are three primary roles of preventing tumours through the function of the immune system. The first one is protecting the host from virus-induced cancer by suppressing viral infection processes or destroying infectious cells. Second, the balance of the immune system can regulate the immune response to avoid the inflammatory environment itself causing tumourigenesis. The third is the immune system can detect cancer cells on the basis of the presence of tumour specific antigens or molecular biomarkers indicating aberrant cells and eliminating them before they can be harmful. The last one is the role of *immune surveillance*. The concept of immune surveillance has been stated since the late 1950s, where the evidence was presented by transplant models that the host rejected tumour tissues, but normal tissue transplantation can be accepted, suggesting that the tumour-specific antigens can trigger a self-immune system [28]. The host immune system can respond to the appearance of cancer cells by the process of antigen presentation as described in Section 1.1.3. Once APCs process and present tumour-specific antigens to T cells, mature T cells survey and seek out tumour cells who express those specific antigens and eliminate them. Even in the presence of immune system functions, cancer cells still develop and do harm - the concept of *cancer immunoediting*

has been developed since 2002 to explore the relationship between cancer development and the immune system that can explain how tumour cells evade from immune surveillance [29].

The concept of cancer immunoediting consists of three processes (Figure 1.4). First, the elimination process takes in the concept of immune surveillance, growing of aberrant or transformed cells induce the inflammatory environment to recruit innate immune cells (NK cells, macrophages, and DCs) to the localisation site. The attack from the innate immune system will produce cytokines such as Interferon gamma (INF-γ) and several chemokines, the secretion of INF-γ and chemokines induce anti-tumour proliferation, apoptotic, and anti-angiogenetic mechanisms resulting in limiting cancer growth [30-32]. Then, debris from dead tumour cells are ingested by local DCs, that will process tumour associated antigens and present to naïve T cells in the lymph node. The mature T cells are effector cells, they away from the lymph node and go to the site of cancer cells and specifically recognise and eliminate cancer cells who harbour antigens that are presented by DCs. Second, the equilibrium process occurs based on natural selection, some cancer cells are eliminated, but those that have high genetic instability can generate new variants and harbour mutations that can cause escape from or resistance to the immune system. Third, in the escape process, cancer cells containing a high load of genetic mutations that survive immune surveillance (the elimination phase) can further develop cancer progression and cause detriment to the host body. From the concept of cancer immune editing, the clinically observable cancer disease indicates the failure of the natural immune system to combat cancer cells. Therefore, to exploit the immune system for effective cancer treatment, the natural immune response is needed to be re-established, which for the basis for the concept of cancer immunotherapy.

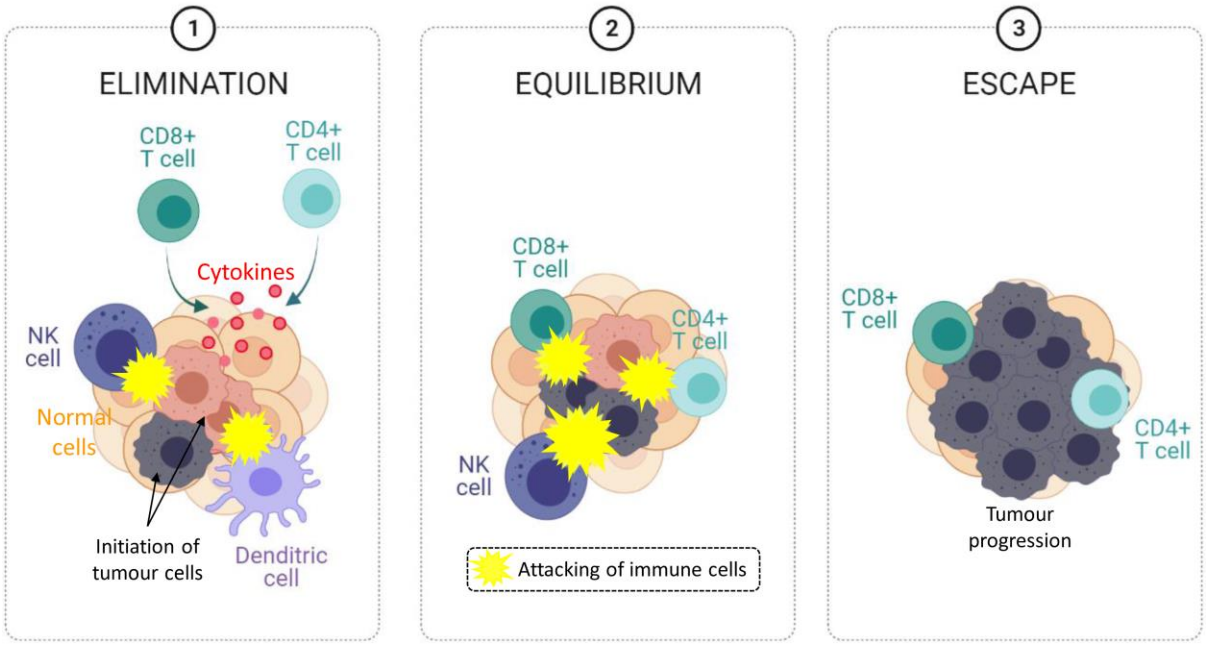

Figure 1.4 The three phases of cancer immunoediting. (1) Elimination; innate and adaptive immune cells recognise and attack transformed cells to destroy them via cytokines secretion. (2) Equilibrium; if the immune system cannot completely eliminate transformed cells, tumour cells that have surviving tumour variants can resist the attack from immune cells. (3) Escape; Tumour cells that survive from immune surveillance can evade the immune system and develop the progression.

1.2.2 The approach of immunotherapy for cancer treatment

In the present, there are several types of immunotherapies for cancer treatment, those can either help the immune system to attack cancer cells or stimulate the immune response to be active and eliminate cancer cells. Currently, there are three major types of cancer immunotherapies which are the most promising and currently developing, which are checkpoint inhibitors, adoptive cell transfer, and cancer vaccines.

a.) Immune checkpoint therapy

Program cell death 1 (PD-1) or Cytotoxic T-lymphocyte-associated antigen 4 (CTLA-4) are co-stimulatory molecules expressed on the surface of T cells. They act to amplify the initial activating signals from the interaction between TCRs and MHC presented antigens, the result

from the signal amplification can activate T cell responses. To evade immune surveillance, tumour cells express the proteins e.g. program cell death 1 ligand (PD-L1) that can bind to those co-stimulatory molecules, the binding interaction can transmit the signal to stop the killing function of T cells [33]. To revive the activation of T cells, the approach of checkpoint inhibitors, which are an antibody-based treatment, is designed to block the binding of co-stimulatory molecules and their ligands expressed by tumour cells so that the killing function of T cells is re-active to eliminate cancer cells (Figure 1.5). CTLA-4 inhibitors is the first immune checkpoint inhibitor that has been approved by the US Food and Drug Administration (FDA) in 2011 for treatment of melanoma [34]. The first PD-1/PD-L1 checkpoint inhibitors for oesophageal cancer was approved in 2014, and it has been now used for the first-line treatment of advanced non-small cell lung cancer [35].

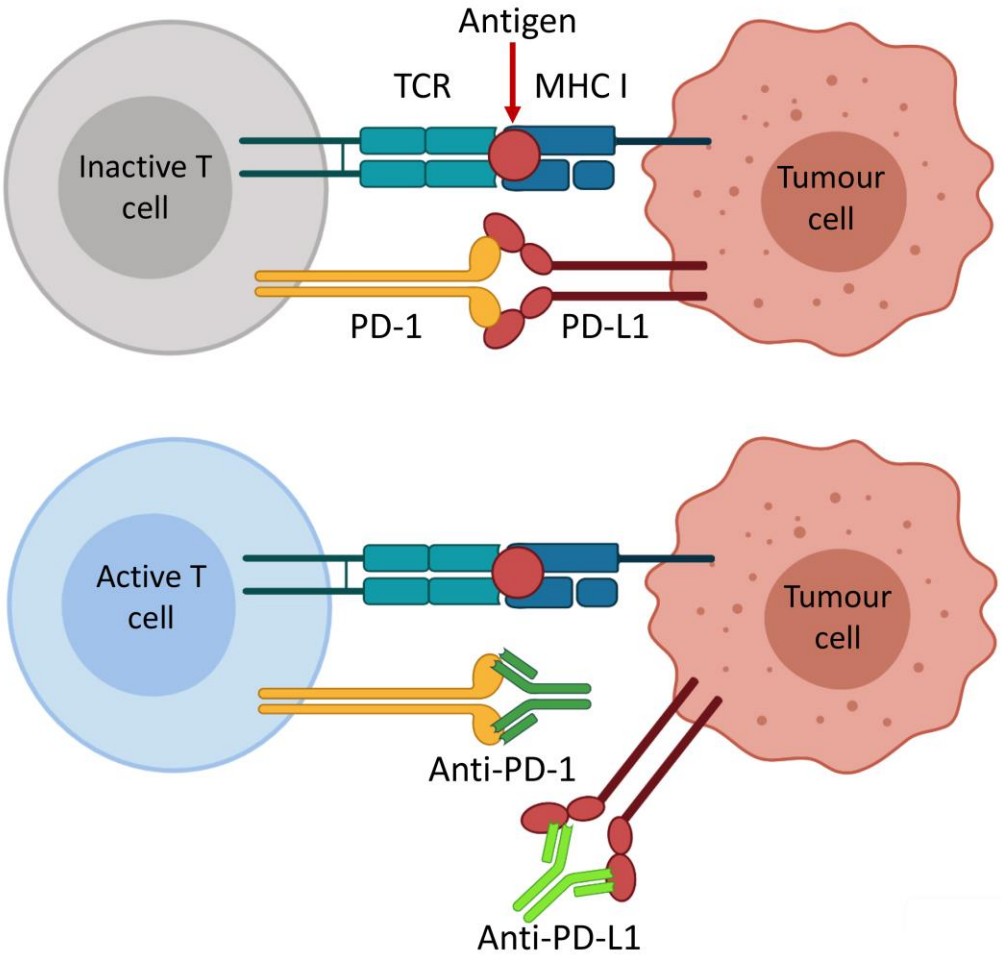

Figure 1.5 Anti-PD-1 and anti-PD-L1 therapies for re-activation of inactive T cells. The T cell receptor can recognise antigen presented by tumour cells, but the interaction of PD-1 and PD-L1 inhibits T cell activation (top panel). The monoclonal antibodies that can specifically bind to PD-1 or PD-L1 can block the binding interaction between PD-1 and PD-L1 so that an unbound PD-1 molecule can reactivate T cell responses (bottom panel).

b.) Adoptive cell transfer (ACT)

The approach of adoptive cell therapy basically utilises T cells which can directly target the specific protein expressed on a patient's cancer cells and kill them [36]. In practice, T cells are taken from a cancer patient's own blood or tumour tissue, and those T cells that can specifically recognise expressed peptides on cancer cells' surface are selected, or the protein receptors on T cell surface are engineered to make T cells more effective to target cancer cells. Then, the modified T cells are expanded in the laboratory to increase numbers and given back to the patient to attack cancer cells (Figure 1.6). The chimeric antigen receptor T cell (CAR T cell) therapy is the type of ACT which has been approved by FDA since 2017, and clinically used in lymphoma. CAR T cells target an antigen called CD19 which is especially expressed in patients with lymphoma [37].

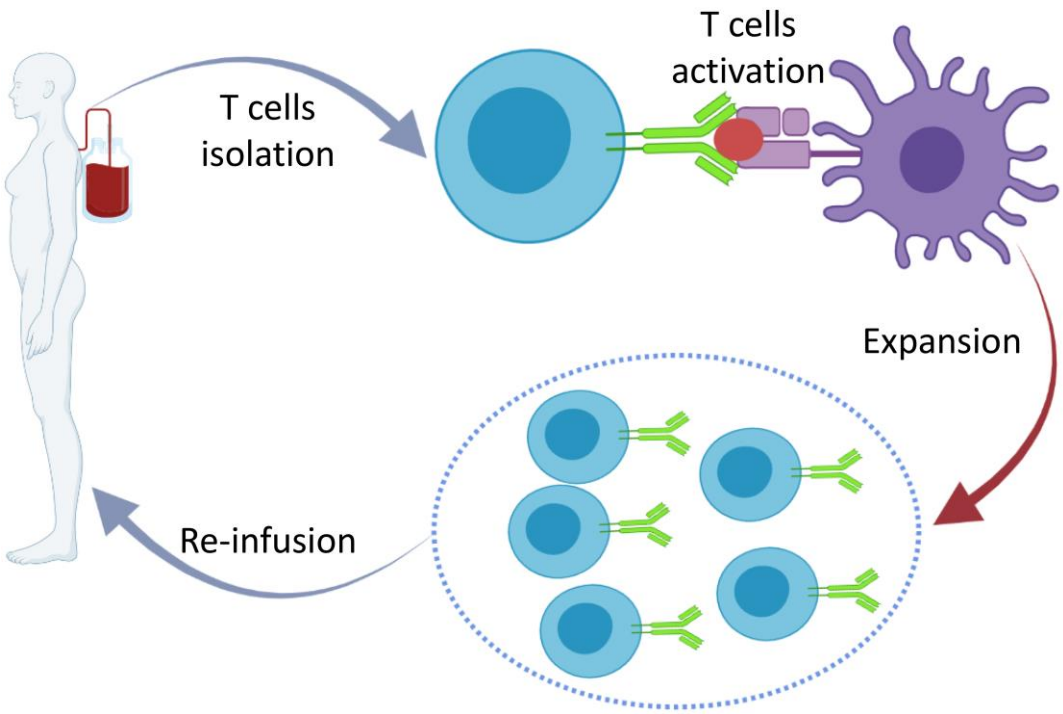

Figure 1.6 Adoptive T cells therapy. T lymphocytes are isolated from blood or tumour tissue of a cancer patient. T cells have been activated by tumour associated antigens, T cell populations that have the desired T cell receptor specificity are selected and expanded. The selected T cells are then re-infused to a cancer patient.

c.) Cancer vaccines

Vaccines for cancer treatment are not the same as vaccination for disease prevention. Cancer treatment vaccine play a role with boosting the natural immune system to exterminate cancer cells. Cancer cells are genetically unstable, resulting in them harbouring numerous somatic mutations that are a source of molecules that normal cells do not have, called *tumour specific neoantigens* [38]. With the adaptive immune response, the effector T cells can be activated by recognition and interaction with antigens presented by MHC proteins. Since neoantigens are mutated peptides that are not self-antigens, those neoantigens are possibly be presented on cancer cell surface by MHC molecules and recognised by T cells. T cells will see them as foreign peptides resulting to activation of T cell responses and subsequently kill cancer cells who express those neoantigens [39, 40]. From this context, synthetic neoantigens can be given to the patient with cancer, the antigens will stimulate the immune system to target and destroy cancer cells who express neoantigens [41]. Cancer vaccines targeting neoantigens can be formulated via various types of vaccine such as nucleic acids (DNA or RNA vaccines), dendritic cells loaded peptides (DC-based vaccine), and synthetic peptide vaccine.

1.3 An introduction of neoantigens in cancer immunotherapy

As described in the previous section, neoantigens can boost the ability of endogenous T cells of cancer patients resulting to restoring of the immune system for attacking cancer cells. Several pre-clinical and clinical studies have revealed the potential of neoantigen based cancer vaccines for the inhibition of tumour growth and tumour metastasis [42]. In this section, the generation of neoantigens from cancer cells and neoantigen-based cancer vaccines are further described.

1.3.1 Arising of tumour specific neoantigens

Epitopes that can activate host immune system leading to cancer rejection are possibly derived from two types of antigens, which are self-peptides that induce T cell tolerance, and the other type is the peptides that have never been expressed in the germline genome. The second type can be generated from either non-synonymous DNA mutations that arose during cancer development that solely create novel protein sequences that normal cells do not have, or viral peptides in virus-associated cancer types such as human papillomavirus (HPV) caused cervical cancer and Epstein Barr virus (EBV) associated cancer e.g. nasopharyngeal cancer. There are several studies that have exhibited the potential of immunogenicity and the ability of cancer suppression by neoantigens derived from non-synonymous somatic mutations indicating that the pool of neoantigens from non-synonymous somatic mutations can contribute to immunogenic peptides triggering T cell activation [43, 44]. However, the effect of cancer heterogeneity might impact the efficiency of cancer vaccines. High tumour heterogeneity is a fewer number of clonal sources neoantigen, which means it has a high risk of some cancer cells cannot be targeted by identified cancer neoantigens and can evade the immune system. The combination of traditional cancer treatment such as chemotherapy or radiotherapy might be utilised along with cancer immunotherapy to eliminate some cancer in a broad spectrum. Although, the effect of radiotherapy might instead increase sub-clonal cancer mutations, which eventually lead to a decrease in the proportion of clonal neoantigens and a further decline in the efficacy of immunotherapy [45]. Even the total number of mutation load is high, the efficacy of immunotherapy is poor. The specific mechanism for mutation targeting therapy might be required to study more. According to antigen processing described in Section 1.1.3, neoantigens derived from somatic mutation have a chance to be presented by MHC molecules to present to T cells resulting the activation T cell killing function.

1.3.2 Neoantigen-based cancer vaccines

Initially, the approach of neoantigen-based cancer vaccines was not much preferred as a target for cancer immunotherapy because of the genetic diversity across different patients, thus, it is difficult to develop this intervention as a "one size fits all" approach. However, in recent years, a number pre-clinical and clinical studies have been reported showing potential of neoantigen-based cancer vaccine in tumour destruction [46]. For developing the vaccination from antigens or deriving neoantigens, there are various types of vaccine formations including cell-based cancer vaccines, peptide-based cancer vaccines, and nucleic acids-based cancer vaccines (Figure 1.7). There are generally two types of cell-based cancer vaccines, which are autologous cancer cells and neoantigens loaded or transfected DCs. With the application of a DCs-based vaccine, DCs of an individual patient are isolated and loaded with synthetic peptides that are identified as neoantigens or transfected with DNA or mRNA translated neoantigens. Those neoantigens are processed and presented on DCs' surface via MHC molecules as discussed above, then neoantigen-loaded mature DCs are given to the patient [47]. However, cell-based vaccines have costly manufacturing/production and are time consuming. A peptide-based cancer vaccine is an intervention using synthetic peptides composed of about 25-30 amino acids with the region of neoantigen. The synthetic peptides can be mixed with adjuvants that improve the ability of APCs to uptake them, and provide better immune response - most clinical studies have utilised granulocyte macrophage colony-stimulating factor (GM-CSF) and polyinosine-polycytidylic acid (poly I:C) [48]. The mixture of peptides and adjuvants is given to the patient who has those specific targeting neoantigens, and it is expected that APCs will uptake those peptides and process them for presentation to T cells. Besides the peptide form, cancer vaccines can be also formulated from nucleotide sequences (i.e. DNA and mRNA) that can be encoded to predetermined neoantigens, the synthetic molecules could be engineered with immunomodulatory molecules [49]. The nucleic acids vaccines are re-infused to the patient, it must be also taken up by APCs, but not directly go to antigen processing like peptide

vaccine. The DNA vaccines are translocated to the nucleus to induct the transcription process, the resulting mRNAs (or RNA vaccines) migrate to cytoplasm, and they are translated to peptides prior to getting the process of antigen processing and MHC presentation [50].

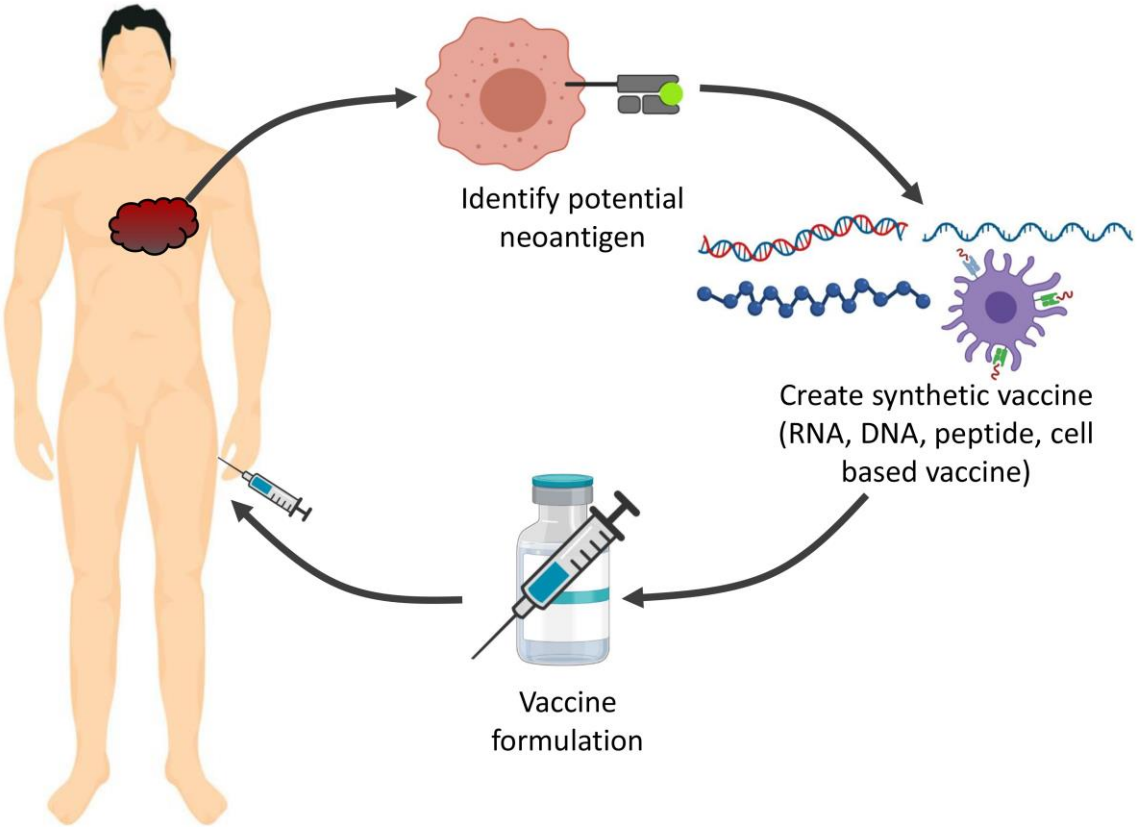

Figure 1.7 The development of cancer vaccine derived from neoantigens. There are several types of neoantigen based cancer vaccine formulation including cell-based vaccines, peptide based vaccines, and nucleic acids based vaccines.

1.3.3 Preclinical and clinical studies of cancer vaccines targeting neoantigens

The principal of personalised cancer vaccines is unlike a traditional vaccine against an immune disease, since DNA alterations across different patients have enormous diversity. The current approach is for neoantigens of an individual patient with malignant tumour to be identified; the pre-determined neoantigens are synthesised and formulated to various forms (peptides, DCs loaded, RNA, or DNA vaccines) with the appropriated adjuvants prior to administrating to the patients. In recent years, there are several studies from both pre-clinical and clinical studies that have shown success with cancer vaccines in activation of immune response and

suppressing tumour growth. Castle *et al*. performed a peptide-based cancer vaccine targeting neoantigens in a mouse melanoma model. They identified somatic mutations from whole exome sequencing (WES) from B16F0 murine melanoma and selected 50 mutated genes to generate synthetic mutated peptides. Those mutated peptides were tested for their immunogenicity using IFN-gamma ELISpot assay, there were 16 immunogenic peptides that were given to tumours transplanted mice. The results showed that two neoantigens (peptides derived from mutations of Kif18b (K739N) and Cpsf (D314N)) can reduce cancer progression and improve the survival rate [51]. The group of Yadav determined neoepitopes from MC-38 cell lines by combining the approach of mass spectrometry and prediction from WES. The selected mutated peptides including those derived from Adpgk, Reps1, and Dpagt1 proteins were injected to mice with MC-38 tumour, they found that mice with the vaccination have a decrease of tumour growth [52].

The first clinical study of a DCs-based cancer vaccine had been reported by Carreno *et al.* Somatic mutations were identified from WES, and candidate neoantigens were selected from the prediction of MHC-binding prediction algorithms. Neoantigen loaded DCs were administrated to three melanoma patients. It was found that neoantigens of DCs-based vaccine can expand the diversity of neoantigen specific T cell since the clone of T cells that are specific to neoantigens can be detected after vaccination. Moreover, this intervention can also enhance the response of existing T cells [53]. In 2017, there were two studies that demonstrated clinical trials using neoantigen-based cancer vaccine in patients with melanoma. The first one from the Dana-Faber Cancer Institute reported the efficiency of neoantigen based personalised cancer vaccine for advance stage of melanoma patients. Candidate neoantigens were identified from mutated peptides using MHC-peptide binding prediction, fully described in Section 1.4. Six melanoma patients were given mutated peptide-based vaccine, the result showed the identified neoantigens from a prediction pipeline preventing a recurrence of cancer of 4 from 6 patients

after surgery for at least two years [43]. With a similar identification strategy, the research group of Sahin prepared RNA-based vaccine instead of peptides and re-infused to 13 melanoma patients. The analysis of immunosurveillance in peripheral mononuclear blood cells (PBMC) from patients after vaccination demonstrated that the RNA coded neoantigen vaccine boosted the activation of existing T cells clones specific to neoantigens [44]. Furthermore, two recent studies in 2019 have reported the effect of a neoantigen-based cancer vaccine in glioblastoma, which is more challenging than the study of melanoma because glioblastoma typically have low mutation burden i.e. a low number of neoantigens derived from mutated proteins resulting in a less promising response in cancer immunotherapy [54]. Hilf *et al*. performed a peptide-based vaccine from synthetic peptides of pre-determined neoantigens and infused to 15 glioblastoma patients. Patients who received the neoantigen vaccination showed activation of CD4+ T cells against predicted neoantigens, and their median survival were improved for 29 months from diagnosis [55]. The other study from the group of Dana-Faber Cancer Institute, who previously studied in melanoma, personalised neoantigens from ten patients with glioblastoma were identified using the same strategy as the previous work [43]. Neoantigen-based cancer vaccines were formulated from synthetic neoepitopes mixed with appropriated adjuvants and inoculated patients after surgical resection combining conventional radiotherapy. The identified neoantigen-based vaccine elevated the number of tumour infiltrating T cells and enhanced circulation of effector CD4+ and CD8+ T cells [56]. Overall, successful evidences from clinical studies demonstrate the potential of personalised cancer vaccine targeting neoantigens as an efficient cancer treatment that can deal with great diversity of cancer disease.

1.4 The methodologies for neoantigen determination

The goal of cancer vaccine targeting neoantigens therapeutic is establishing of T cell function for attacking tumour cells because T cells are the major effector cell populations that specifically responds to tumour antigens. Therefore, the determination of neoantigens that are

specific for cancer cell is a crucial step for therapeutic success. With the advance of omics technologies, peptides that could be potent neoantigens can be determined by using the approach of immunopeptidomics via mass spectrometry or MHC-peptides binding prediction algorithms.

1.4.1 The approach of mass spectrometry (MS) based immunopeptidomics

As mentioned before, neoantigens are presented on the cell surface via MHC molecules the immunopeptidomics is used to study profiling of peptides bound HLA molecules, called HLA binding peptide (HLAp). The HLAp from tumour cells are isolated by the immunoprecipitation method using specific antibodies against HLA molecules. Normally, W6-32 is used to precipitate HLA class I, and IVA12 is used for HLA class II. HLAp complexes are then purified to elute peptides from HLA molecules, eluted peptides are prepared for liquid chromatography coupled with tandem mass spectrometry (LC-MS/MS) to identify MS/MS spectra. The MS spectra are then searched against customised protein database to identify the amino acid sequence of the eluted peptides [57]. A key step in MS workflows is the matching of spectra against a predetermined sequence database. If a sequence is not present in the database, then a match cannot be made using this methodology. Databases containing very large sets of sequences, for example including all possible sequences or mutations are costly in terms of search time and statistical power, and thus an ideal strategy is to create a database from the sequences likely present in a given individual i.e. the database for searching can be generated to include peptides carrying non-synonymous somatic mutations derived from WES of matched tumour/normal tissue from the patient, which is important to selected mutated peptides that specific for an individual patient. Most studies using MS-based method are combined with the analysis of WES or RNA sequencing data because a customised database is essential for identifying peptides containing mutated amino acids. Bassani-Sternberg *et al*. demonstrated that the combination of MS and WES data can identify 11 mutated peptides, of which two of

them were immunogenic and could trigger activation of T cells specific that neoantigens [58]. Furthermore, the combination of MS analysis and genomic sequencing analysis have been shown successful in neoantigen identification for both pre-clinical and clinical studies as described in the Section 1.3.3.

The approach of immunopeptidomics is a direct method to identify neoantigens that can be really expressed and presented via MHC proteins on cell surface. Thus, this method can reduce the risk from obtaining false neoepitopes compared to computational prediction methods [57], this issue might be crucial for planning interventions because only a handful of peptides are selected to perform immunogenic experiments and for cancer vaccine development. Besides neoantigen-derived somatic mutations, MS based immunopeptidomics also discover neoantigens that can be derived from proteasome splicing, unusual post-translational modification, and non-coding RNA [59]. However, the number of peptides identified from MS analysis depends on size of tissue sample especially in case of low mutation burden, small size of tissue samples requires the high sensitivity and accuracy for peptide identification [60]. In general, the identification of neoantigens using MS needs a large size of tissue sample for sample preparation e.g. bigger than 1 cm^3. Although, in clinical practical, the big size of tumour tissue from surgery resection is not feasible for most cases, therefore, studies of neoantigens determination using MS experiment are mostly limited in the scope of cell culture experiments [61].

1.4.2 The approach of bioinformatics in genomic sequencing and computational analysis

Using an *in silico* approach, the putative neoantigens are determined from the predicted binding affinity between mutated peptides and HLA molecules carried by an individual patient using the MHC-peptide binding affinity prediction algorithms. This approach is underpinned by next generation sequencing (NGS) data and bioinformatic software packages to generate the list of

mutated peptides and HLA alleles. Somatic mutations can be identified from WES data of matched normal and tumour cells from an individual patient, only mutations that alter protein sequences are retained. A set of short peptides containing mutated amino acid(s) are extracted from each patient's data set. Patient specific HLA alleles can be determined by either computational alignment methods or genotyping from blood samples. Since the mutated peptides are called from DNA sequencing data, RNA sequencing data is utilised to filter neoantigens derived from expressed proteins [62]. The MHC-peptide prediction-based method require a smaller size of tumour tissue sample compared to MS based immunopeptidomics method, which are thus more feasible in real clinical practice. With the advance of NGS technology, nucleotide sequencing from cell-free tumour DNA (ctDNA), released by degraded or dying cancer cells into the blood of cancer patients, could be performed, which can be feasible for both solid and non-solid tumour [63]. Analyses of mutations in ctDNA have shown high accuracy and more rapid identification of mutations. Although, standardisation of ctDNA collection, storage, sequencing techniques, and analysis methods would be critical to facilitate the wide adoption of ctDNA technology in routine clinical practice [64].

Most algorithms predict the MHC-peptide binding affinity by learning from only chemical properties between peptides and MHC molecules from *in vitro* experiment. With the continuous developments in peptidomics studies, the performance of MHC-peptide binding affinity prediction algorithms has been improved by large training data sets derived from MS experiments, which are mostly deposited in public databases e.g. the Immune Epitope Database (IEDB) [65]. However, the diversity of HLA alleles contributes to the enormous variety of their binding preferences, which cannot be completely covered by available training data at the present. Thus, those alleles with few experimental peptides for training algorithms might give low precision of predicted performance [66].

1.5 Neoantigen prediction with the approaches of bioinformatics

Tumour specific neoantigens are non-self, mutated peptides presented by MHC molecules on the tumour cell surface, which have the potential to trigger the activation of tumour specific T cells. Neoantigen identification based computational prediction involves multiple processes including somatic mutation identification, HLA determination, and MHC-peptide binding prediction, then candidate neoantigens are prioritised and selected according to their potential for being immunogenic peptides (Figure 1.8). The source of mutated peptides come from the genomic alterations during cancer cell division. To identify neoantigens from sequencing data, the WES data from normal and tumour tissue as well as RNA sequencing data are typically needed. Computational based neoantigen predication is feasible for almost solid cancer types, and this method requires only a small amount of tumour tissue for whole exome or RNA sequencing compared to MS analysis. Most computational methods focus on modelling which peptides bind to the MHC molecules such as NetMHC or MHCflurry [67, 68]. However, the sequencing technology and computational prediction can go awry due to some biases during the sequencing experiment and data training of predictive models. DNA or RNA sequencing may introduce amplification inequity and technical errors in the reads used as starting material for a source of neoantigen. Modelling MHC-peptide binding must consider the fact that humans have ~5,000 alleles encoding MHC class I molecules, with an individual patient expressing as many as six alleles, all with different peptide binding affinities. The prediction tools typically require hundreds of experimentally determined peptide-binding measurements for a particular allele to build a model with sufficient accuracy. There are several factors involved in the procession, presentation, and immunogenicity of peptides for being neoantigens, thus, the accuracy of neoantigen prediction does not only rely on a good accuracy of prediction results but also depend on the completeness of tumour tissue sample, experimental technique during sequencing and clinical data of patients that give evidence towards assessing immunogenicity.

In this section, the existing bioinformatics tools that have commonly used in neoantigen prediction are described, which include bioinformatic software for NGS analysis, programs for variant calling, the methods of *in silico*-based HLA genotyping, RNA quantification analysis, MHC-peptide binding algorithm, and the current existing pipelines for neoantigen identification.

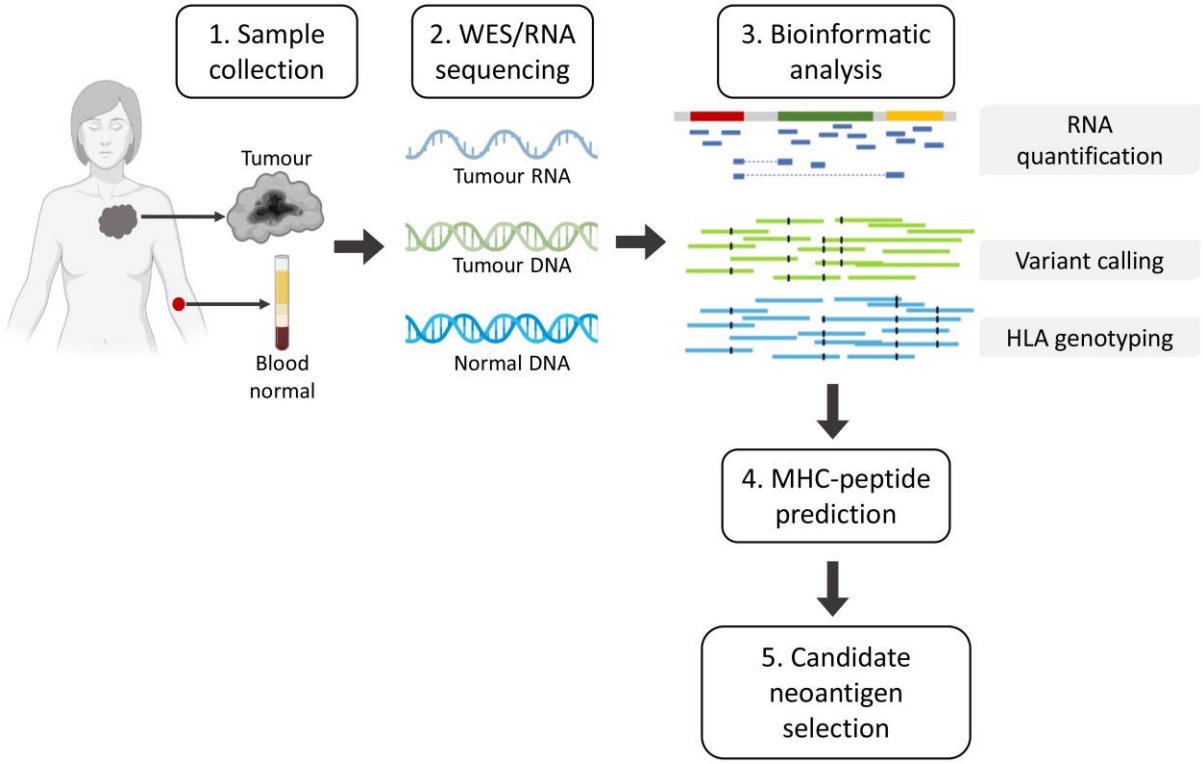

Figure 1.8 Neoantigen identification workflow from WES and RNA sequencing data with computational analysis.

1.5.1 Non-synonymous somatic mutations identification

There are many types of non-synonymous mutation causing the alteration of protein sequences such as nucleotide point mutations, frameshift mutations, insertion, deletion, and structural variants. Single nucleotide variants (SNVs), small insertions and deletions (Indels) are common genetic alterations that can be merely detected from the short-read sequencing platforms with software packages of genomic analysis [69]. However, the data from whole genome sequencing (WGS) allows for more sensitive and accurate small variants detection,

moreover, the structural variants and copy numbers can be reliably detected from WGS, which can increase the repertoire for mutated peptides [70-72]. In clinical practice, exome sequencing is preferred because only coding variants can be neoantigens, and WES assay is more feasible in term of cost, infrastructure capacity, and lower error rate compared to long read platforms such as WGS [69]. Nevertheless, the integrated use of the WES and WGS is the most suitable way to generate the best results [73]. There are three major steps for identification of non-synonymous somatic mutations, Firstly, both tumour and normal sequencing data are aligned to a reference genome, then the alignment results are processed to determine the genetic variants. Finally, mutations from a variant calling step are annotated whether they are in coding or non-coding regions (Figure 1.9).

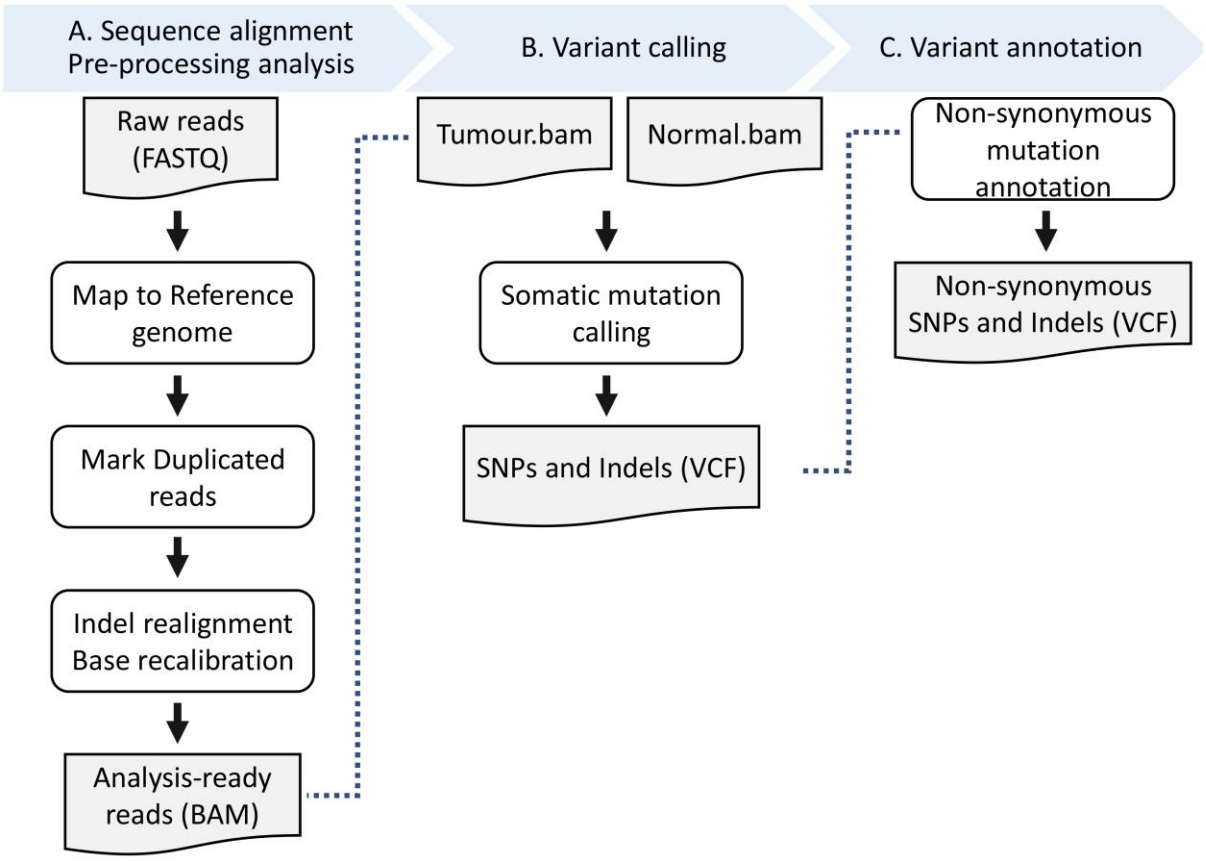

Figure 1.9 The workflow of non-synonymous somatic mutations calling from matched tumour-normal WES data.

a. Sequence alignment and pre-processing analysis

The alignment of sequencing data to a reference genome is a crucial phase in genomic sequencing analysis [62]. The raw sequencing data in the FASTQ format are initially aligned to the human reference genome using an aligner tool such as Burrows-Wheeler Aligner (BWA) [74], this tool is commonly used by GATK pipeline [75]. Besides aligner tools, a reference human genome is also important for this step, the use of a recent version of reference human assembly genome provides a logical improvement over the old versions in quality of genomic alignments and downstream analysis [76]. A result file from alignment is stored in the format of binary alignment/map (BAM) [77], then duplicated reads that originate from the same sequence, which might yield from polymerase chain reaction (PCR) step during DNA sample preparation, are removed by BAM file manipulation tools such as Samtools package, Picard, or Sambamba [77-79]. The process of quality control for BAM files, which are base quality score recalibration (BQSR) and indel realignment, is strongly recommended to be performed prior to variant calling to evaluate sequencing coverage and prevent false positive variants coming from alignment artifacts that might be called in the variant calling step [75]. Following the steps of sequence alignment and pre-processing BAM files, the analysis-ready BAM file is further used as an input file for the variant calling step.

b. Somatic mutations calling

The main propose of tumour sequencing is identifying tumour mutations excluding germline mutations that are existing in both normal and cancer cells, thus the best practice analysis for somatic mutation calling must exploit sequencing data from paired of tumour and normal samples. There are several existing variant callers that have been specifically developed for determination of somatic mutations. Among them, MuTect2 [80], Streka2 [81], and VarScan2 [82] are widely used for somatic mutation analysis with aligned data from tumour and normal

simultaneously [62]. Those tools can identify both SNVs and Indels from analysis of BAM files of a paired tumour and normal, only MuTect2 can optionally applied with unpaired tumour-only samples. There is no single tool that gives superior performance among various somatic mutation callers, hence an ensemble usage that combines the results from multiple tools might yield the best result with a balance of sensitivity and specificity [83, 84]. The analysis produces an output stored in variant call format (VCF), a text file with details of single nucleotide polymorphisms (SNPs) and Indels with various properties of that variant represented in the columns [85].

c. Non-synonymous mutations annotation

The details of variants called in a VCF file format can be further interpreted to consider the consequence of those variants such as impact on protein expression or association between variants and diseases. For neoantigen identification, the approach of transcript annotation is utilised to identify mutations that subsequently change protein sequences. The Ensembl Variant Effect Predictor (VEP) is a software tool for annotation and analysis various type of genomic variation in coding and non-coding regions. This software is critical for variant annotation and a subset prioritisation for further analysis [86].

1.5.2 Quantifying gene expression

RNA sequencing is commonly used to measure levels of transcript expression, however, the accuracy of inferring gene expression level from short sequencing reads is one of the challenging issues for quantitation of gene expression levels [87]. In the recent years, several RNA analysis tools have been developed to quantify a transcript level from short read RNA sequencing. Conceptually, short sequencing reads are assigned to their originated transcripts, and that information is used to estimate gene abundances [88]. With the traditional methods, short sequencing reads are aligned to a reference genome to identify the transcript they arise

from, then the relative gene expression levels can be inferred from reads mapping to annotated gene loci [89, 90]. Cufflinks is a popular tool-based alignment, it is mostly used for novel transcription discovery and quantification transcript levels for differential expressed gene analysis [91]. However, the methods rely on an alignment step that can be time consuming and computationally intensive. Currently, there are a number of novel tools that do not rely on the step of a reference genome alignment, so-called an alignment-free method [92, 93]. Tools based alignment-free quantify transcript levels using a k-mer counting algorithm. They work by extracting sequence reads into k-mers followed by matching of k-mers to pre-indexed transcript databases using a hash table. The common tools with alignment-free sequence analysis are Sailfish [92], Salmon [94], and Kallisto [93]. They perform ultra-fast analysis, consume less computational resources, and yield high accuracy, the benchmarking studies reported their performance are comparable [95-97]. The workflow for RNA-seq quantification is shown in Figure 1.10, the final output files from the current RNA-seq tools are generally reported as summarised read counts for each transcript or relative expression level in TPM (Transcripts Per Million).

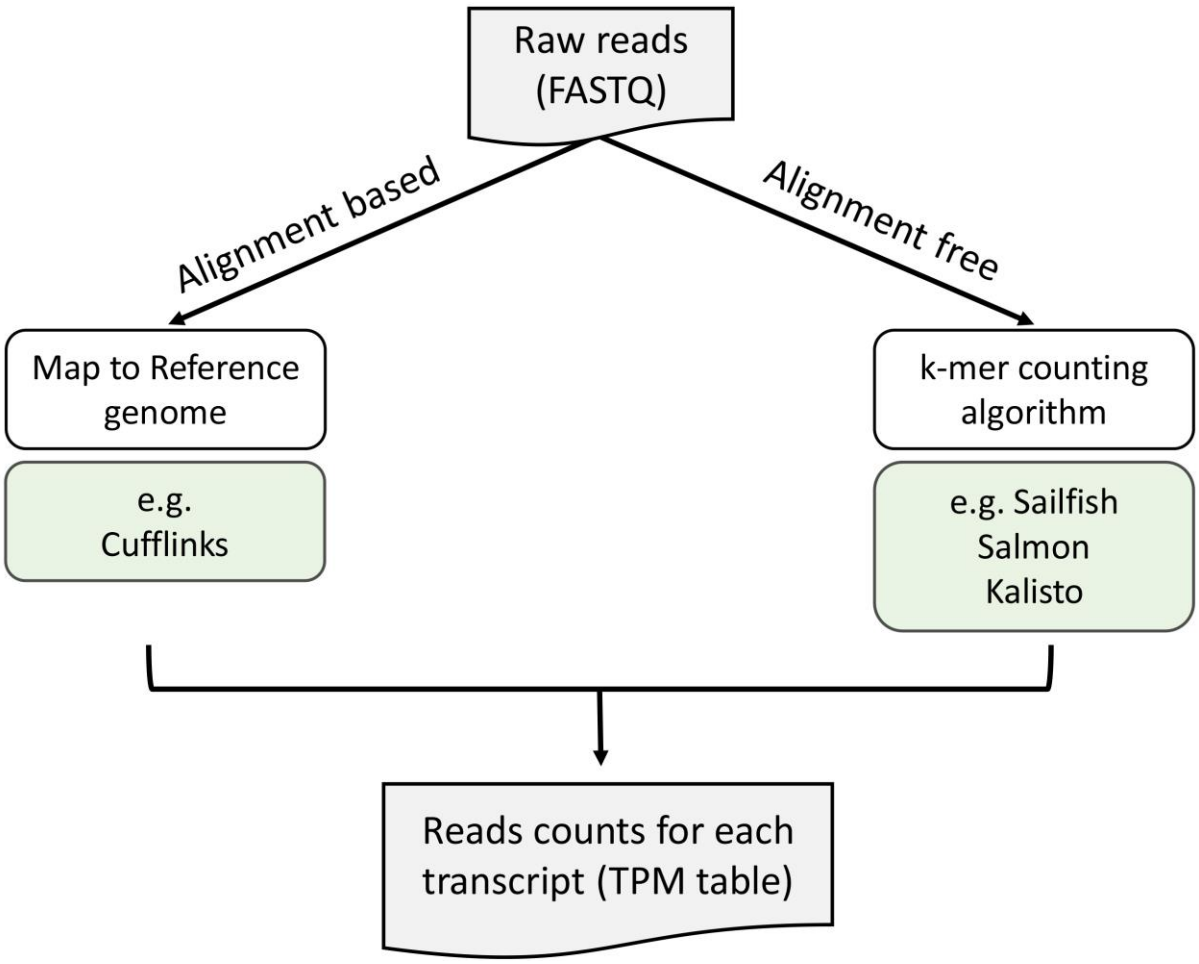

Figure 1.10 Quantification of RNA expression with alignment based and alignment free methods.

1.5.3 *In Silico* HLA class I typing using next-generation sequencing data

To apply MHC class I-peptide binding prediction tools, patient-specific HLA alleles of MHC class I including HLA-A, HLA-B, and HLA-C must be determined, the class of human MHC was fully described in Section 1.1.2. The gold standard for HLA genotyping is laboratory-based sequence specific PCR amplification [98]. Alternatively, the computational approach for HLA typing based on genomic or transcriptomic data from a peripheral blood sample or normal tissue are commonly performed to determine HLA alleles [99]. At present, there are a number of HLA class I calling algorithms that display prediction accuracy as good as results from HLA typing using DNA-based techniques [100, 101]. HLA genotyping algorithms mostly follow one of two major approaches that are an alignment-based method or an assembly-based method. The approach relying on alignment works by mapping sequencing reads to reference HLA sequences, and types of HLA are determined based on probabilistic models. The assembly-based methods assemble reads into contigs and align them to known HLA allele sequences, true HLA alleles are annotated by the best similarity score between the contig and each known HLA sequence [100]. The benchmarking study reported that OpiType [102] and PHLAT [103] display the highest accuracy with WES or RNA sequencing data, Opitype can reach up to 99% accuracy if limited by MHC class I only [100].

1.5.4 HLA class I-peptide binding affinity prediction

The binding interaction between MHC molecules and peptides plays a key role in subsequent T cell activation and triggering the adaptive immune response. In the context of neoantigen identification, the binding affinity prediction is used for the initial step for selecting candidate neoantigens for downstream experiments. Over the past few years, several binding affinity prediction algorithms based on machine learning approaches have been launched, and most are publicly available (Table 1.1). The predictors relying on machine learning distinguishing a

peptide as a binder or non-binder by generating a predicted binding affinity score using the training model based on extracted representative features. However, the presentation of peptide on MHC molecule contains several complicated steps from protein expression, protein degradation, entering to ER, compatibility of peptide and MHC, and stability of the complex [7]. Thus, only binding affinity data might not be sufficient for representing whether a given peptide to be presented as an MHC ligand. Advances in the approach of MS can provide peptidome data, generated from eluted MHC ligands of MHC-peptide complexes from *in vitro* using an immunoprecipitation technique followed by MS characterisation [104]. Recently, the existing MHC-peptide binding prediction algorithms utilise either only peptidome data or combination of MS peptides and binding affinity data, which would contain the comprehensive signal of antigen processing and presentation rather than binding affinity alone [67, 68, 105]. NetMHCpan uses an artificial neural network with trained on a data set combining data on binding affinity and MS eluted ligand data. MHCflurry is MHC class I predictor which also uses a neural networks technique for data training, the models in MHCflurry have been built by either only binding affinity data or combined with MS data. The major difference between two tools is the prediction of specific MHC alleles. While NetMHCpan uses the approach of pan-allele model which is a single model takes as input both the peptide and a representation of the MHC alleles [106], MHCflurry is an allele-specific predictor whereby training and selection of models are done separately per allele [68]. Among those publicly available predictors, the prediction tools from the NetMHC family developed by Morten's group at the Technical University of Denmark are commonly used in clinical studies. The systematic benchmark analysis reported NetMHCpan and MHCflurry display a good performance in distinguishing binding from non-binding peptides [107].

The output from both prediction tools reports similar information including the predicted half maximal inhibitory concentration (IC_{50}) value in nM unit and the predicted percentile rank (%

rank) score. Nevertheless, there is no an actual threshold for precise determining a binder, IC_{50} < 500 nM is the common threshold for binding affinity which classify that a peptide is a binder [108]. The predicted % rank scores are estimated as the rank position of a given score within a list of scores from a set of 125,000 of 8-12 mers random natural peptides (25,000 of each length), assumed to represent the distribution of false results (non-binders) or general background non-specific binding of regular human peptides. NetMHCpan4.0 documentation recommends using the % rank score rather than the predicted binding affinity since the different MHC molecules have a different preference of binding affinity. NetMHCpan4.0's developers also performed the sensitivity and specificity curve as a function of the % rank score, and a rank < 2% was selected as a threshold which has both high sensitivity and specificity [106]. MHCflurry reports the value of the predicted binding affinity, IC_{50} in nM unit, ("mhcflurry_prediction"), the low and high predicted binding affinity which come from the top 5% and the bottom 95% from different models of each allele, and the % rank scores ("prediction_percentile") that are estimated from the quantile of the affinity prediction among a large number of random peptides tested on that allele. This tool also suggests users to apply IC_{50} < 500 nM as a threshold to classify a binder and a non-binder. However, the source MHCflurry publication did not include sensitivity and specificity analysis to select the threshold for the % rank and do not recommend which the % rank threshold to be used for selecting binding peptides [68]. The topic of statistics related to MHC-peptide binding is the specific focus of Chapter 2.

Table 1.1 MHC class I-peptide binding affinity prediction tools

MHC-binding affinity prediction tool (version)	Predictive methods	Key features	Publicly available
NetMHC4.0 [109]	Artificial neural network-based algorithm, NNAlign (allele-specific)	This tool is for prediction of MHC-peptide binding affinity, it allows multiple lengths of peptides.	Yes
NetMHCpan4.1 [67]	Artificial neural network-based algorithm, NNAlign (multi-allelic)	This model is trained by expanded data from both binding affinity data and eluted peptide data identify by mass spectrometry.	Yes
NetMHCcons1.1 [110]	NetMHC, NetMHCpan: artificial neural network-based PickPocket: matrix-based	A predictor is for analysis of combinations of three MHC-peptide binding predictor tools.	Yes
MHCflurry2.0 [111]	Neural network	This tool combines new model for MHC class I binding prediction and antigen processing, which are trained by MHC I bound ligands identified by MS.	Yes
MHCnuggets2.3 [112]	Long short-term memory (LSTM) neural network	This tool can predict binding for common or rare alleles of MHC class I or class II.	Yes
MHCSeqNet [113]	Natural language processing based neural network	This tool models amino acid sequence of MHC allele and peptide as sentences with amino acids as individual words, which allow a prediction of unseen MHC alleles and peptides with any length.	Yes
EDGE [105]	Deep learning	Training data from deconvoluted specific HLA-peptide identified by MS analysis	No

1.5.5 Existing multi-step neoantigen prediction pipelines

As described above, neoantigen identification involves several steps from pre-processing data and input data preparation to binding affinity prediction. Over the past few years, several multi-step workflows relying on MHC-peptide binding predictors have been developed, those tools have built custom code to extract and generate a list of short-mutated peptides from a variant call file and perform command line-based analysis for running of MHC-peptide binding prediction tools. Furthermore, those workflows have integrated a variety of analysis methods besides binding prediction, which can help user to make a shortlist of candidate neoantigens (Table 1.2). However, those tools do not implement the steps for variant calling, HLA genotyping, and quantifying gene expression level, they usually require data input as a mutation file in VCF format, list of HLA alleles, and transcript expression level in a tabular format. Some workflows only perform binding affinity prediction and annotate gene name to input peptides, then return the result table that provide information for a pair of mutated peptide and HLA allele, such as predicted binding affinity, name of gene that peptide originate from, gene expression level, or variant types, but some of those workflows have integrated mathematical operation to compute a ranking score or machine learning models to classify levels for an individual predicted peptides that can help to select a shortlist of candidate neoantigens for downstream experiments. pVacseq is a well-documented automated pipeline for neoantigen prediction, this tool uses WES or RNA sequencing data to systematically generate the repertoire of mutated peptides, perform binding prediction, and apply the filter criteria to make a shortlist of candidate neoantigens. Aside from binding affinity and gene expression level, this program filters candidate neoepitopes from the depth coverage of sequencing reads and variant allele frequency [114]. pVacseq has recently been added in pVactools, a computational tool suite for neoantigen characterisation and vaccine designs [115]. MuPeXI requires data input similar to pVacseq and also incorporates a binding predictor from NetMHC suite. The major

difference is the process of creating a shortlist of candidate neoantigens, MuPeXI has a built-in multiplicative function to calculate a ranking score for an individual peptide deriving from the input variants. The prioritisation score is computed based on HLA-binding affinity, similarity between mutated peptides and their self-counterpart, mutant allele frequency, and gene expression levels. Moreover, MuPeXI provide the full set of identified neopeptides in a tabular format containing several informative annotations and prioritising scores that users can easily sort and filter to select candidate neoantigens [116]. However, strong binding peptides with high expression level does not ensure that they can be recognised by T cell receptor, thus, the recent study have put an effort to develop the workflow augmented with the model-based machine learning to predict immunogenicity of candidate neoantigens, such as Neopepsee [117]. The Neopepsee workflow requires the data input from RNA sequencing and extracted mutated peptides from input variants. This tool performs the binding affinity prediction using NetCTLpan [118], then peptides with predicted binding affinity scores are further determined immunogenicity using the build-in machine learning classifier, that classifies the candidate neoeptiopes into three classes including high, medium, and low according to the predicted immunogenicity.

Table 1.2 Automated pipeline for neoantigen identification tools

Neoantigen identification workflow	Used MHC-binding prediction tools	Key features
ProGeo-neo [119]	NetMHCpan4.0	This workflow is for analysis data of genomics, transcriptomics, and proteomics. The prediction of MHC-peptide binding affinity from genomic data were screened by proteomic data, gene expression, and T cell recognition epitope.
pVACtools [115]	NetMHCpan NetMHC NetMHCcons PickPocket MHCflurry MHCnuggets	This tool identifies neoantigens from a variety of somatic alterations including structural variants and prioritises with ranking scores that account from binding affinity, gene expression, sequence read coverage, agretopicity. Interactive visualisation is available.
neoANT-HILL [120]	NetMHC4.0 NetMHCpan4.0 NetMHCcons NetMHCstabpan PickPocket MHCflurry	Mutated peptides can be generated from variants from RNA sequencing data, a graphical user interface (GUI) is available.
pTuneos [121]	NetMHCpan4.0	Determination of neoantigen from predicted MHC presentation and immunogenicity, prioritisation with ranking scores that account binding affinity, sequence similarity between a pair of normal and mutant peptides, peptide, hydrophobicity score, and T cell recognition.
Neoepiscope [122]	NetMHCpan4.0 MHCnuggets MHCflurry	This tool emphasises the process of variant calling, in the context of interaction of somatic mutation and neighboring germline variants, and address variant phasing for SNVs and Indels.
ScanNeo [123]	NetMHC NetMHCpan	This workflow is for analysis RNA sequencing data to predict neoantigens derived from small to large sized Indels
antigen.garnish [124]	NetMHC NetMHCpan MHCnuggests MHCflurry	This pipeline is combination of neoantigen prediction and neoantigen quality analysis tools to predict peptide immunogenicity. Predicted affinities are averaged to generate the ensemble score.
NeoPredPipe [125]	NetMHCpan	This automated pipeline connecting commonly bioinformatic software, processing data, prediction, and summary statistics as output for downstream analysis.
retained_intron-neoantigen_pipeline [126]	NetMHCpan3.1	A computational approach to detecting intron retention events from tumour RNA sequencing data to generate peptides containing ≥1 amino acids from intron for neoantigen prediction.

Table 1.2 Automated pipeline for neoantigen identification tools (cont.)

Neoantigen identification workflow	Used MHC-binding prediction tools	Key features
MuPeXI [116]	NetMHCpan	This pipeline automatically extracts mutated peptide sequences and returns the informative output table with a priority score for each predicted peptide.
Vaxrank [127]	NetMHC NetMHCpan NetMHCcons MHCflurry	A pipeline to determine which peptides should be used in a vaccine. This tool applied ranking scores to select putative neoantigens, and the output will be used to make long peptides.
TSNAD [128]	NetMHCpan2.8	A pipeline for extracting somatic mutations from genome analysis and predicting potential neoantigens, which could be either extracellular mutation of membrane proteins or mutated peptides presented by MHC molecules.
Neoepitope prediction [129]	NetMHCCons1.1	A pipeline for identification of putative neoantigens based on somatic missense mutations and gene fusion using whole genome sequencing data.
CloudNeo [130]	NetMHCpan	A cloud-based computational workflow for identifying patient specific tumour neoantigens for NGS sequencing data. This workflow can run on cloud platform of NCI Cancer Genomics Cloud, which provide graphical user interface.

1.6 Prediction of immunogenic T cell epitopes

The ultimate goal of neoantigen prediction is getting peptides that can be recognised by T cell receptors and activate the adaptive immune system to eliminate cancer cells. The process of antigen processing and MHC presentation allows T cells to detect antigens derived from invaded pathogens or mutated peptides expressed by cancer cells. MHC presented peptides that can trigger an immune response are described as epitopes. Even though all epitopes must be presented by MHC molecules, but not all MHC ligands can stimulate T cells activation. Most neoantigen prediction workflows currently rely on MHC-peptide binding prediction to identify neoepitopes, that step is necessary but might not sufficient to determine real neoantigens. The best validation of immunogenicity is the wet experiments i.e. cytokine secretion assays, such as ELISpot or ELISA, intracellular cytokine-staining assays, such as flow cytometry assays,

however those experiments are time and resource consuming. In recent years, there has been an expansion of databases that collect experimental data from laboratories, the computational methods for immunogenic prediction therefore become an alternative for epitope identification [131]. In this section, *in silico* methods for immunogenic MHC class I ligands prediction are emphasised.

1.6.1 Properties of immunogenic MHC class I presented peptides

A specific MHC presented peptide will be recognised by an estimated average of one in 100,000 naïve T cells [132]. The peptide-immunisation experiments have shown that about half of the presented peptides are epitopes, which means all epitopes are MHC binding peptides, but not all MHC presented peptides are immunogenic peptides [133]. The identification of epitopes is crucial to the study and understanding of cellular immune responses and great importance in vaccine development. The strength of interaction between MHC I presented peptides and TCR depends on both MHC class I molecule and the presented peptide. However, the extreme high diversity of T cells is a key factor to characterise the specificity between TCR and a peptide. According to sophisticated steps for MHC-peptide presentation and T cell recognition, as described in Section 1.1.3, the recent computational approaches consider predicting epitopes from peptide sequences. Since TCR-epitope interaction is governed by the physicochemical principles like other protein-protein interactions, thus, more immunodominant epitopes are expected to have some preferred properties that can make a stronger interaction with TCRs than non-epitopes. Within that context, physicochemical properties and amino acid characteristic of epitope and non-epitope peptides has been investigated. A set of immunogenic and non-immunogenic of MHC I presented peptides were collected and compared the amino acid frequencies and physicochemical properties of each amino acid in peptides from both sets under the hypothesis that the certain amino acids are more likely to interact with TCRs. The study showed that large and aromatic residues such as

Phenylalanine and Tryptophan were overrepresented in a set of immunogenic peptides, and a trend for overrepresentation of acidic residues was observed in immunogenic presented peptides. In addition, significant associations with immunogenicity were observed for Isoleucine, Lysine, and Methionine [134]. Moreover, the structural studies and immunogenicity studies of specific T cell clones with altered peptide ligands demonstrated that some position in a presented peptide, especially positions 4-6, are in close contact with TCR and important for specific T-cell responses [135, 136]. The amino acid profile of each position in an MHC class I presented peptide were compared in both sets of immunogenic and non-immunogenic peptides, the significant difference was found in the position 4,5, and 6, but not found a substantial difference of amino acid profile at other positions [134]. The information from those studies indicates that immunogenic MHC I presented peptides have some certain signatures for T cell recognition.

1.6.2 *In silico* prediction methods

The increasing of data repositories and advance in immunoinformatic facilitate data management and development of predictive methods for T cell epitope prediction. SYFPEITHI is one of the oldest immune epitope databases and contains more than 7000 peptides that bind to MHC class I and II molecules [137]. Although, the data in SYFPEITHI became static in 2012 due to increasing utility of IEDB that has been established since 2004. The advance in high throughput experiments results in a rapidly increase in the number of curated epitopes in 2015, and the recent update reported IEDB contains >1.6 million experiments representing the adaptive immune response to epitopes [138]. The epitope data in IEDB are not only from human and mouse but also from chimpanzee, macaque, cow, and pigs [139]. A recent report showed that IEDB stores more than 1,000,000 peptides with positive result derived from T cell assays, B cell assays, and MHC ligand assays, and more than 500,000 peptides with negative result [140]. Several bioinformatics prediction tools determine T cells epitopes by a

strength of binding interaction between peptides and MHC molecules based on the biological process that the TCR can bind to only an MHC presented peptide. These approaches rely on the fact that generation of the peptide by natural processing and subsequent HLA binding are key necessary steps for T cell immunogenicity, but HLA binding peptides might not be sufficient to be immunogenic peptides [141]. A methodology involves directly using epitope and non-epitope data to train the predictive network by learning from physicochemical properties of peptide sequences to predict if a peptide can be immunogenic. However, the determinants of epitope immunogenicity in association with their recognition by T cells remain poorly understood. Given the fact that different individuals have different TCR repertories, in theory, epitope immunogenicity should differ between individuals.

The structure requirements for the interaction of MHC presented peptide complexes and TCRs as well as the different properties in the motif between immunodominant epitopes and non-epitopes are increasing understood [134, 141]. There are currently a wide variety of sequence-based prediction methods for T cell epitope prediction, which attempt to use computational methods to discriminate epitopes and non-epitopes by the physicochemical principles and distribution of amino acids in a sequence (Table1.3). Immunogenicity has been launched in 2013, and the user interface software is available at the IEDB. This tool predicts MHC class I presented peptides into two categories for epitopes and non-epitopes. Immunogenicity was build based on the enrichment of amino acids and the importance scores of different positions of the MHC I presented peptides between immunogenic and non-immunogenic peptides [134]. NetTepi was developed based on MHC-peptide stability prediction tool, NetMHCstab [142] with the aim of creating an integrated method for T cell epitope prediction, combining MHC-peptide binding affinity, stability and T cell propensity predictions [143]. TCR classifier is developed to predict the recognition of a peptide from the sequence patterns of CDR3 region in the TCR. This model was build based on the TCR sequences of HLA-B*08, the results

demonstrated the feasibility of the approach of prediction of T cell epitope recognition based on sequence data, but does not cover other HLA alleles in practice [144]. Moreover, there is a current tool that uses a computational framework mimicking the thermodynamic interaction between peptide-MHC complexes and public TCR clonotypes, termed TCR-peptide contact potential profiling (CPP), generates probabilistic estimates of immunogenicity [145]. INeo-Epp is the current T cell epitope prediction tool, which has web-based user interface. This tool combined several factors involved in physicochemical properties of amino acids such as accessibility, molecular weight, molecular structure, hydrophobicity, polarity, entropy, and charge as well as MHC-peptide binding affinity for training epitope and non-epitope peptide data set to develop the T cell epitope classification model [146].

Table 1.3 T cell-MHC class I epitope prediction tools

Tool	Predictive method	Features for model learning
Immunogenicity [134]	Immunogenicity score	Molecular weight, charge, the importance of positions in a peptide
NetTepi [143]	Integrated predicted scores from NetMHCcons NetMHCstab Immunogenicity	Peptide-MHC binding affinity and stability, T cell propensity
TCR-classifier [144]	Random forest classifiers	Properties of the CDR3: sequence length, absolute count of each amino acid, basicity, hydrophobicity, helicity, isoelectric point, and mutation rate
Repitope [145]	randomised trees (ERT) algorithm	Peptide length, amino acid existence, peptide description, and TCR-peptide contact potential profiling (CPP)
INeo-Epp [146]	random forest classifier	Amino acid physicochemical property, MHC-peptide binding affinity, peptide entropy, predicted immunogenicity score from Immunogenicity

1.7 Aim of the thesis

The main objective of this research was to develop models that help to improve the criteria to select and prioritise candidate neoantigens based on the prediction of MHC-peptide binding affinity and immunogenicity prediction. This research work was specifically aimed to develop new software that can give an accurate probability for a peptide to be a genuine neoantigen.

1.7.1 Global and local false discovery rate (FDR) estimation model for MHC-peptide binding affinity prediction

As mentioned in Section 1.4, the ability of MHC binding is widely used to determine neoantigen since MHC-peptide presentation is a necessary step for T cell recognition. However, the existing MHC-peptide binding prediction tools provide a predicted binding affinity or an estimated score relying on distribution of pre-set of negative data. The uncertainty of neoantigen prediction based on insufficient statistical values is discussed in Chapter 2. Therefore, the first goal of this research was to develop a model that can estimate global and local false discovery rate for an individual predicted MHC-peptide binding affinity score. This work is summarised in Chapter 3.

1.7.2 MHC class I immunogenicity classification model

As described in Section 1.6, a key to identify if a peptide is neoantigen it must be an immunogenic epitope. Even MHC binding is a necessary step for T cell recognition, but it is not sufficient to determine whether an MHC presented peptide is an immunogenic peptide. The criteria for candidate neoantigen selection performed in Chapter 2 rely not only MHC-peptide binding affinity, but also consider other biological factors such as gene expression, mutational site, and orientation of side chain of mutated amino acid. However, they do not have a statistical basis to determine immunogenicity. Thus, the second goal of this research was to build a model

to classify peptides to be epitopes or non-epitopes using a framework of machine learning, this work is summarised in Chapter 4.

1.7.3 The pipeline for ranking HLA class I neoantigens based on true MHC binding affinity and immunogenicity prediction

To serve the main objective of this research, the final goal was developing a new software that can produce accurate statistics for neoantigen selection and prioritisation, built by integrating models from Chapter 3 and 4, summarised in Chapter 5.

Chapter 2

The study of neoantigen prediction using existing bioinformatics software and public MHC-peptide binding affinity prediction tools

Author contributions

All whole exome and RNA sequencing data in this chapter were kindly provided by a research team at Chulalongkorn University System Biology Centre (CUSB), Bangkok, Thailand. The preparation of DNA and RNA from patient samples were done by research staffs at CUSB. All the computational analyses were performed by the thesis author PP, with input and direction from the supervisory team.

2.1 Introduction

There are several neoantigen identification pipelines that have been launched in recent years as shown in the Table 1.2. Those tools perform neoantigen identification based on genomic sequencing data and MHC-peptide prediction. Those existing automated workflows generally requires a VCF file, list of MHC alleles and a table of gene expression levels as inputs. A list of mutated peptides of multiple lengths are usually generated via customised code of those existing tools prior to taking those peptides and MHC types to the MHC-peptide binding prediction tool. Most current workflows launched from 2017 onward have an amended prioritisation model to return a ranking score for each peptide that is useful for selecting candidate neoantigens for downstream experiments. In this chapter, the identification of neoantigens from WES and/or RNA sequencing data from colon cancer patients was performed using existing bioinformatic tools and a neoepitope prediction workflow called MuPeXI [116] to demonstrate the proof of concept and test the tools. The common mutations existing among the cohort in this study were also analysed to see if some mutations could be common in colon cancer patients. The common mutations could be the potential for developing a "vaccine warehouse". Besides cancer vaccines, information on mutated genes from neoantigen identification carried by patients can be exploited to develop a specific antibody to the common mutation found among patients [147]. Beyond the sequence analysis, a structural based analysis was also demonstrated to explore the quality of the predictions; peptides with high predicted binding affinity from the MuPeXI prediction were selected for the energetic analysis based on the structure model of MHC-peptides using the molecular dynamics (MD) simulation technique.

In summary, the aims of this chapter are to demonstrate the practicability of approaches for neoantigen prediction using both sequence analysis with a publicly available pipeline and a structural based analysis. The chapter also explores the rate of false positive answers from

MHC-peptide binding predictors i.e. NetMHCpan4.1 and MHCflurry, as a potential source of error in the analysis.

2.2 Materials and methods

2.2.1 Patient samples and sample preparation for DNA/RNA sequencing

The biospecimen samples were collected from nine colorectal cancer patients. The sample collection and preparation were done by research staff at Chulalongkorn University System Biology Centre (CUSB), Bangkok, Thailand. Tumour and blood samples were obtained from study participants in King Chulalongkorn Memorial Hospital, Bangkok, Thailand. Tumour tissue and blood from each patient were collected immediately after surgery resection. The method of Ficoll-Paque (GE Healthcare, United States) density gradient centrifugation was used to isolate PBMCs from a blood sample, PBMCs were cryopreserved with 10 µl RNAlater solution (Qiagen, Germany). Tumour tissue was chopped to a size of 1 mm^3, then 0.5 mL RNAlater solution was added prior to storage at -80 °C. The fresh frozen tumour tissues were minced into small pieces, and then 20 µl Proteinase K (Qiagen, Germany) was added. The mixture of tissue and proteinase K were incubated at 56 °C with agitation for 30 min. After an incubation period, the samples were centrifuged at 12,000 g for 3 min, the supernatant fraction was collected and transferred to two RNase free microcentrifuge tubes for DNA and RNA isolation. Tumour and normal DNA molecules were isolated from tissue lysate and PBMCs by DNA isolation kit (Qiagen DNeasy kit, Germany). The amount of DNA was quantified using DNA Quantitation Kit with the fluorescent technique (Merk, United States), to ensure there was sufficient DNA for exome sequencing: higher than 200 ng in 50 µl. RNA from tissue lysate was extracted using RNeasy kit (Qiagen, Germany), the quality control of RNA for the sequencing experiment was performed by the sequencing company. The library of DNA and RNA were prepared, and nucleotide sequencing was performed by the Vishuo company

(Vishuo Biomedical, Singapore). Briefly, genomic DNA was extracted and ligated with barcode Illumina sequencing adapters, then DNA sequences were amplified, and short reads with 100-200 nucleotides were enriched. Whole exome capture was performed using an Agilent SureSelect Human All Exon V6. The libraries were then qPCR quantified, pooled, and sequenced with 150 base-paired-end reads using HiSeq 2000 sequencers (Illumina, United States).

2.2.2 Preparation of input data for MuPeXI

MuPeXI requires peptide sequences and MHC types (i.e. HLA allele) as mandatory inputs, but gene expression levels are optional. Thus, WES and RNA sequencing data are needed to process to determine non-synonymous somatic mutations, HLA alleles, and gene expression level (Figure 2.1).

a.) Sequencing data pre-processing

The FASTQ file of WES data was initially processed to remove adapter sequences by Cutadapt, and Fastqc was used to evaluate the data quality [148, 149]. Then, sequencing data without adapter reads were aligned with the reference genome (GRch38) using BWA with mem option (BWA-mem) [74], and a read group was defined for each sample for the variant calling step. Picard tools with MarkDuplicates option was used to remove redundant reads [78]. Indel realignment and base recalibration were performed with GATK workflow [150].

b.) Determination of non-synonymous somatic mutation

The analysis-ready reads of tumour and normal samples from a step above was used to detect somatic mutations by Mutect2 [80]. Somatic mutations including point mutations, small insertions/deletions and frameshift mutations were identified, then list of somatic mutations with detail in columns were return in a VCF file format. Non-synonymous mutations in the VCF file were further annotated using VEP [151].

c.) HLA genotyping

HLA class I alleles were determined from normal WES using the HLA typing tool, called Athlates [152]. The algorithm aligns normal WES to the reference of HLA class I sequences, including A, B, and C loci, from the IMGT database [8]. The alignment coverage and Hamming distance are calculated to a similarity score. Two HLA alleles with the first two sets of digits of each locus that have the highest percent coverage and lowest Hamming distance were selected.

d.) Gene expression quantification

RNA sequencing data was quantified as the level of gene expression in Transcripts per Kilobase Million (TPM) by Kallisto, which returns expression level of each gene in a table with a tab-separated values (TSV) format [93].

2.2.3 Neoantigen identification using MuPeXI pipeline and candidate neoantigen prioritisation

Somatic mutations in a VCF file, list of HLA alleles, and TPM level of each transcript in a TSV file were used as inputs for the neoantigen prediction pipeline, MuPeXI [116] (Figure 2.1). This workflow was designed to predict neoantigens based on binding affinity between peptides and specific HLA alleles using NetMHCpan3.0 predictor [153], and a ranking score for an individual peptide is computed by a built-in multiplicative function, which calculates a prioritisation score based on predicted binding affinity, similarity between mutated peptides and their self-counterpart, mutant allele frequency, and gene expression level. MuPeXI returns the full set of identified neoantigens in a tabular format containing several informative annotations and ranking scores. The predicted results with available gene expression level, candidate neoantigens were chosen by the criteria of $IC_{50} < 500$ nM and TPM ≥ 1.

However, in this study, RNA sequencing data are not available in some patients, in that case, only predicted IC_{50} values is the only data available to characterise binding and non-binding peptides. In the clinical study, there are about 20-30 candidate peptides further selected for downstream experiments, therefore, using predicted binding affinity scores without gene expression level is not sufficient to rule out non-neoantigens. Therefore, to create a shortlist of candidate peptides for the cases that do not have gene expression information, the prioritisation system is set on the basis of binding affinity and the potential for being an immunogenic peptide relying on a difference from self-peptides, following approaches define in the work from Ott P.A. *et al.*, in 2017 [43]. The potential epitopes with predicted $IC_{50} < 500$ nM were chosen for inclusion based on a pre-defined set of criteria in the following rank order: (1) The binding epitopes with frameshift, insertion, or deletion. These mutation types alter more than one amino acid in the peptides, which is likely to make greater differences between mutated peptides and their self-counterparts; (2) The increased binding affinity epitopes with somatic single nucleotide variations at the HLA anchor residues (position of 2 and 9), which can imply that peptides have never been presented by MHC molecule to T cells; (3) The very high MHC binding peptide (less than 150 nM) with somatic single nucleotide variations at non-anchored residues.

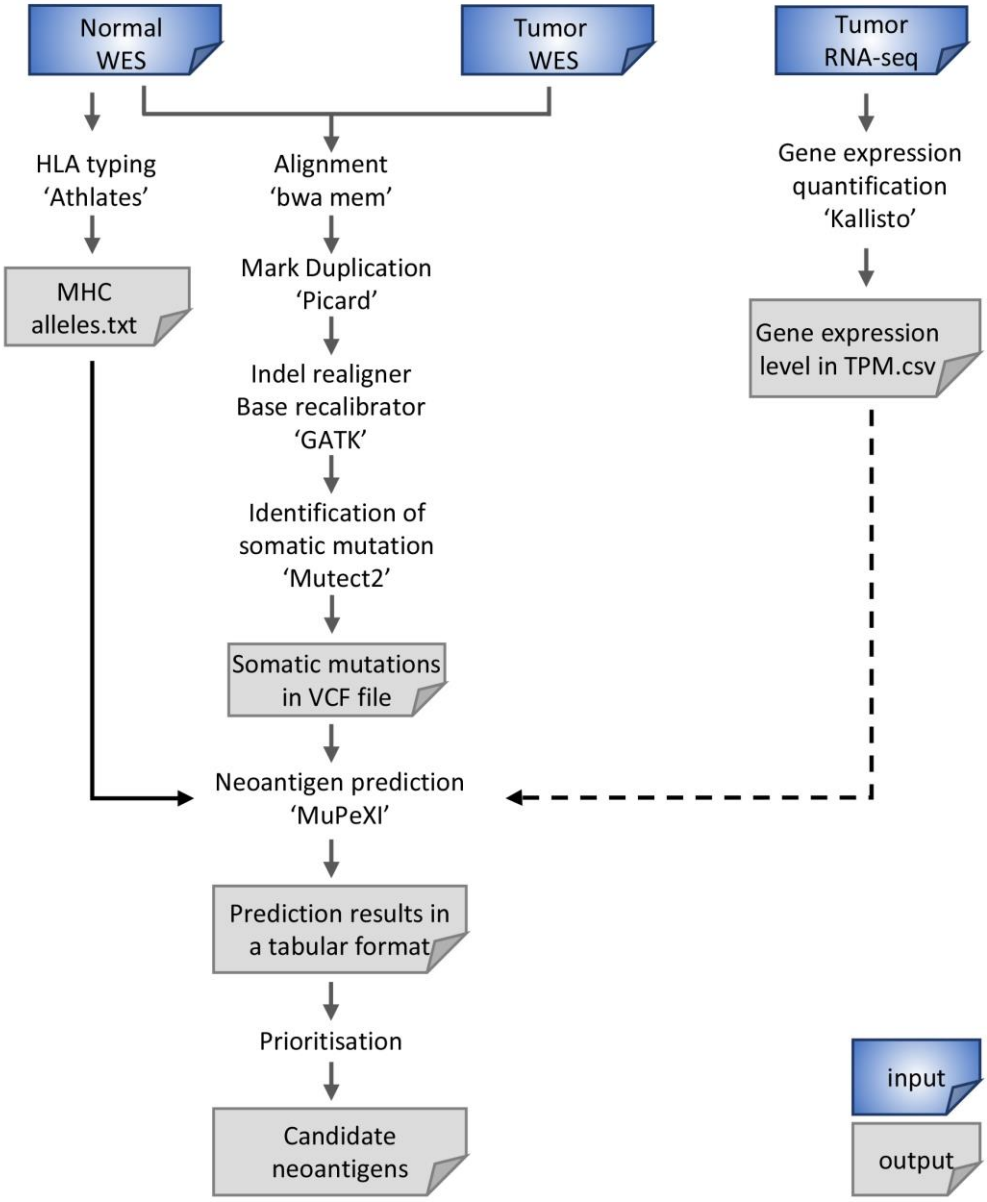

Figure 2.1 Analysis workflow of neoantigen identification based on genomic sequencing data and MHC-peptide binding affinity prediction.

2.2.4 Molecular dynamics (MD) simulation

The candidate peptides identified by the MuPeXI workflow were used in the MD experiment. A set of 9mers peptides presented by HLA-A*02:01 from a patient who express HLA-A*02:01 alleles which is Sample 6 (see Table 2.3) that have high predicted binding affinity and high gene expression level were selected for MD analysis, totalling seven candidate peptides. The 3D structure of HLA-A*02:01 with a 9mers peptide was downloaded from Protein Data Bank

(PDB), the structure ID is 3QEQ [154]. The two peptides with low binding affinity (predicted IC_{50} > 40,000 nM) and having negative charge residues at anchor positions were selected as a negative control group (Table 2.1). The complex of HLA-A*02:01 and each peptide were constructed by changing chemical elements of amino acids in the template peptide to the mutated peptide in Table 2.1 using the Discovery studio 2.5 [155]. Finally, there are ten systems of HLA-A*02:01/peptide complex for performing MD simulation, which are an HLA-A*02:01/template, seven complexes of HLA-A*02:01/candidate peptide, and two complexes of HLA-A*02:01/negative control peptide. All complexes were defined with the protonation state of each residues at pH 7.4 via the PROPKA3 [156].

In the step of MD simulation run, the module called Leap in Assisted Model Building with Energy Refinement (AMBER) version 14 was used to add the missing atoms and hydrogen atoms [157]. The complex structures were solvated in a 25A° radius, with TIP3P model for water molecules. The isothermal-isobaric (NPT) ensemble with constrained number of atoms (N), pressure (P) and temperature (T) was applied in a periodic boundary. The SANDER module in AMBER 14 was used to minimise all water molecules and protein complexes. MD simulation was performed using the pmemd.cuda module in AMBER 14 for all complexes, and snapshots were stored every 0.2 ps during a trajectory of 100 ns. For the analysis part, the MD trajectories in the production phase were collected for analysis of complex stability, binding free energy, and binding free energy decomposition. The approach of Molecular Mechanics/Generalized Born Surface Area (MM-GBSA) was used to estimate the binding free energy of a ligand to protein [158].

Table 2.1 Peptides for molecular dynamic simulation

Sample	Peptide	Predicted IC_{50} (nM)
Tem	AAGIGILTV	2254.35
c1	YMNDINCRM	12.4
c2	LLLGLLLFL	16.5
c3	SLPQLTHEV	25.3
c4	ILHHLGQEV	94.4
c5	LLGGTALLL	121.2
c6	SMTVRTTPV	179.2
c7	VMHNYRNLV	234.2
nc1	KEERDDDTD	49326.2
nc3	PRVRDNYRD	49203.1

Tem = a peptide from crystal structure (3QEQ); c = candidate peptides from MuPeXI prediction; nc = non-candidate peptides

2.2.5 Random peptide data sets

The random peptides generated from the human proteome were assumed to model as false data points (non-binding peptides) for the study of MHC-peptide binding predictors. Since 9mers have the highest preference for MHC class I binding groove, a set of random 9mers peptides was created. The human proteome from UniProt (www.uniprot.org) database [159] was processed to produce 9mers peptides via a sliding window approach, different by one amino acid at a time. The total unique 9mers peptides from human proteome is 12 million. The five data sets of random peptides were generated by random selecting 10,000 peptides per data set from whole human peptides data set.

2.2.6 MHC-peptide prediction using NetMHCpan4.1 and MHCflurry

Each random data set was predicted against the 79 alleles of MHC class I, which are supported in both MHCflurry and NetMHCpan4.1. An input file was prepared from each random data set with each HLA allele, and those input files were predicted the MHC-peptide binding affinity by NetMHCpan4.1 and MHCflurry from a stand-alone software package by the command line for each specific HLA allele. For NetMHCpan4.1, the prediction was performed with the

default model. With using MHCflurry, the model of MHCflurry 1.2.0, which has been built by binding affinity data combined with MS data for model selection, was utilised as the predictor. To select predicted binding peptides, the fixed thresholds were used to cut off the predicted results. The NetMHCpan4.1 documentation recommends using the % rank rather than the IC_{50} value, but most studies select the binding peptides based on the predicted IC_{50} [67, 68]. Therefore, both values were used to distinguish binder from non-binder peptides. With using the IC_{50}, the threshold value is <500 nM, whereas the threshold of the % rank is <2%. Finally, there were two sets of the binder result, i.e. selecting binders with (i) less than 500 nM and (ii) less than 2% rank, per allele per prediction tool. The number of peptides that passed the criteria were calculated as the percentage of binders.

2.3 Results

2.3.1 The neoantigen identification-based sequence analysis using the publicly available tools

Variant calling analysis from matched tumour and normal WES data showed different number of somatic mutations among different patients. The number of non-synonymous somatic mutations across nine colorectal cancer patients ranged between 10 to 400 mutations indicating that even amongst the same type of cancer, the diversity of genetic mutation is individualised. Most mutations come from missense mutations which alter only one amino acid whereas the small insertions/deletions or frameshift mutations rarely occurred (Table 2.2). The alleles of HLA class I including A, B, and C loci of an individual patient were identified from normal WES data using the HLA genotype algorithm as described in the section 2.2.2. Each patient has at least three different alleles for A, B, and C loci and a maximum of six different alleles; two alleles per locus (Table 2.3). Among nine samples, the highest frequency HLA class I

alleles for A, B, and C loci are A*33:03, B*40:01, B*46:01, and C*01:02 respectively (Figure 2.2).

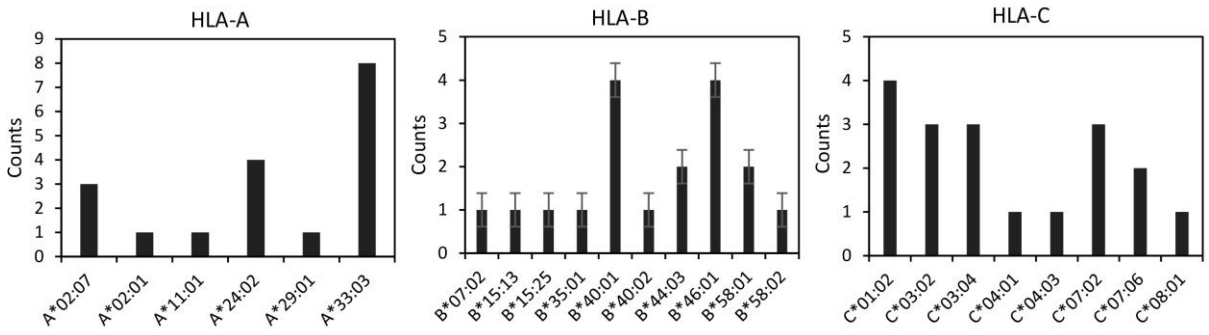

Figure 2.2 The frequency of HLA class I alleles from nine colorectal cancer patients

A variant file and list of HLA alleles of each patient were taken as inputs to MuPeXI. For only samples 6 and 8 were RNA expression level available. The program returned the results table including predicted IC_{50} from mutated peptides and their normal counterparts, variant information, RNA expression level in TPM, and ranking score calculated from the built-in model in MuPeXI software. However, seven of nine samples do not have RNA sequencing data. As discussed above, a predicted binding affinity score alone is not sufficient for making a shortlist of candidate neoantigens. Therefore, the prioritisation criteria described in Section 2.2.3 was used to filter and select candidate neoantigens on the basis of capability of MHC binding and the immunogenicity potential characterised by similarity between mutated peptides and their self-counterparts. It was found that the number of candidate neoantigens that pass filtered criteria vary across different patients (Table 2.4). It can also be observed that transcriptomic data can help to reduce numbers of predicted neoantigens by excluding non-expressed peptides and create a shortlist of candidate peptides, for example, Sample 3 and Sample 6 have similar number of mutations, but a ratio between number of candidates to the number of total mutations of Sample 3 is twice that from Sample 6. Moreover, the result showed that the number of candidate neoantigens seem to be approximately proportional with

the number of non-synonymous somatic mutations. The linear relationship between number of mutations against number of candidate neoantigens displayed a good correlation ($R^2 = 0.978$), although this correlation analysis did not include data of Sample 6 and 8 because the criteria of candidate neoantigen selection for those data have a step of filtering by gene expression level but other samples do not have RNA sequencing data (Figure 2.3). Thus, the number of candidate neoantigens that selected from different criteria could not be compared. Next, shared mutated genes across nine patients were explored, among those samples, the mutated genes that were found in more than one sample were selected. There were 12 mutated genes that were shared by two or more samples. Two of them were TP53 and APC which are well known cancer driver genes in colorectal cancer [160], and mutations of TP53 were found in 4 of 9 cases with different mutations sites. Only APC and ZNF808 shared the same mutated residues in two samples which are frameshift mutation of APC in Sample 4 and 7 and point mutation of ZNF808 in Sample 6 and 7 (Table 2.5). Finding common mutated peptides across different patients might have good potential for developing a "warehouse vaccine", suitable for many patients, although we do not see much evidence for this potential in our data.

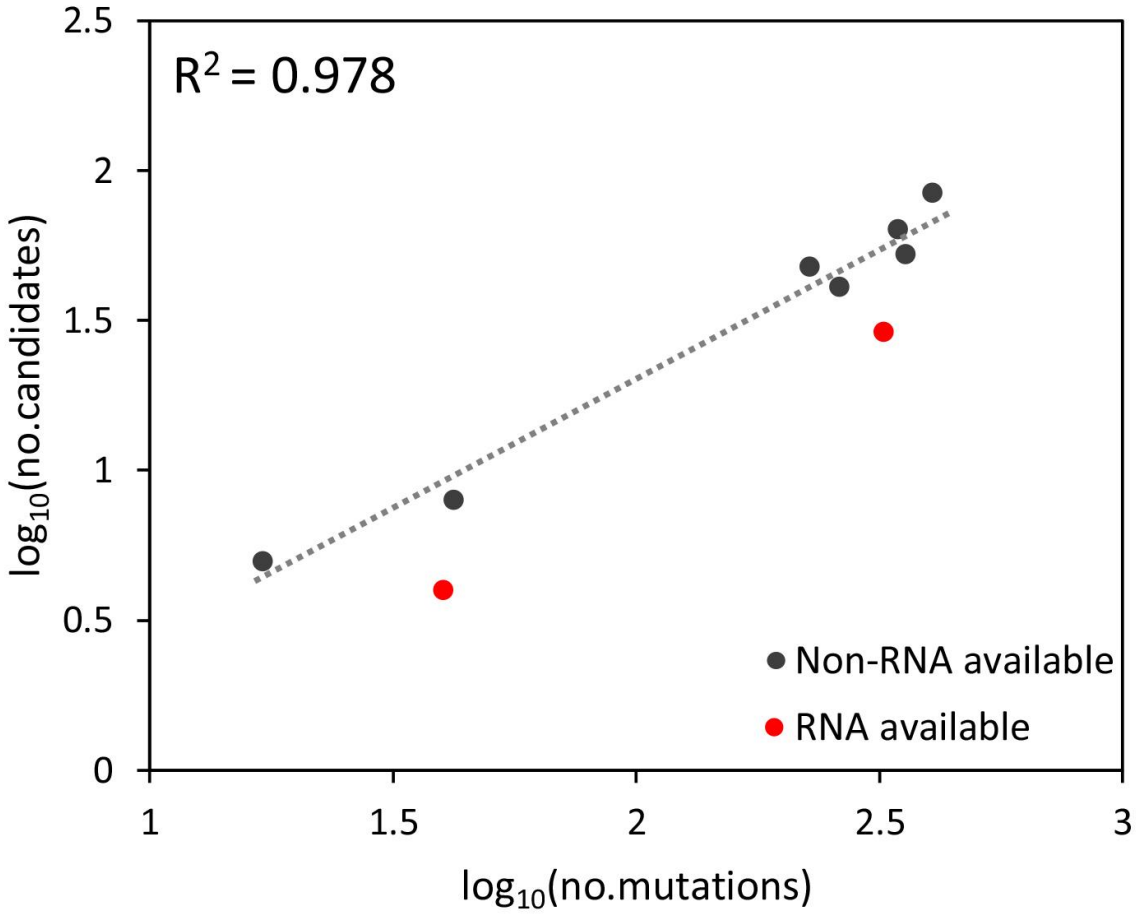

Figure 2.3 The scatter plot between number of non-synonymous mutations and predicted candidate neoantigens. The linear regression model was fit to those data points without RNA Sequencing data because the criteria of candidate neoantigens selection from data with and without RNA sequencing data are different.

Table 2.2 Number of non-synonymous somatic mutation of nine colorectal cancer patients

Sample	Missense Mutation	Insertion/Deletion	Frameshift mutation	Total mutations
1	330	11	16	357
2	12	1	4	17
3	327	7	11	345
4	210	1	16	227
5	236	10	15	261
6	304	12	7	323
7	327	43	35	405
8	38	2	0	40
9	40	2	0	42

Table 2.3 HLA class I alleles of nine colorectal cancer patients

Sample	**HLA-A**	**HLA-B**	**HLA-C**
1	HLA-A02:07 HLA-A02:07	HLA-B46:01 HLA-B40:02	HLA-C01:02 HLA-C03:04
2	HLA-A33:03 HLA-A33:03	HLA-B46:01 HLA-B07:02	HLA-C01:02 HLA-C07:02
3	HLA-A24:02 HLA-A33:03	HLA-B44:03 HLA-B15:25	HLA-C07:06 HLA-C04:03
4	HLA-A11:01 HLA-A29:01	HLA-B46:01 HLA-B35:01	HLA-C01:02 HLA-C04:01
5	HLA-A33:03 HLA-A24:02	HLA-B40:01 HLA-B40:01	HLA-C03:04 HLA-C07:02
6	HLA-A33:03 HLA-A02:01	HLA-B58:01 HLA-B15:13	HLA-C03:02 HLA-C08:01
7	HLA-A33:03 HLA-A24:02	HLA-B40:01 HLA-B58:01	HLA-C03:04 HLA-C03:02
8	HLA-A33:03 HLA-A24:02	HLA-B40:01 HLA-B58:02	HLA-C03:02 HLA-C07:02
9	HLA-A33:03 HLA-A02:07	HLA-B46:01 HLA-B44:03	HLA-C01:02 HLA-C07:06

Table 2.4 The number of candidate neoantigens of nine colorectal cancer patients

Sample	**Total mutations**	**Candidate neoantigens**	**A ratio of candidates to total mutation**	**RNA data available**
1	357	53	0.15	No
2	17	5	0.29	No
3	345	64	0.19	No
4	227	48	0.21	No
5	261	41	0.16	No
6	323	29	0.09	Yes
7	405	85	0.21	No
8	40	4	0.10	Yes
9	42	8	0.19	No

Table 2.5 Shared mutated genes among nine cancer patients

		Sample								
Gene	Cancer Driver Gene	1	2	3	4	5	6	7	8	9
TP53	yes	R174H		C275F			L130R	A161T		
TTN	no	R33134C						C213764R		
AFF2	no	I1023M			L1034I					
IFT122	no	A662E						E655del		
PCDHA8	no	R498Q				K124R				
KDM4E	no	R100H		Q42R						
PTGFR	no			V106A	R133W					
OBSCN	no			Y3606H	A3300V					
SLITRK5	no			V678L					G59D	
ADCY10	no							K900N	A1131T	
APC	yes				E1554fm			E1554fm		
ZNF808	no						R474T	R474T		

del = deletion, in = insertion, fm = frameshift mutation

2.3.2 The analysis of MHC-peptide binding based on structure analysis

The approach of MD simulation can assess the binding energy and other physicochemical properties between protein structures and ligands. In this study, this technique is applied to investigate the binding strength of MHC molecules and their ligands. A set of predicted candidate neoantigens presented by HLA-A*02:01 from a sequence analysis were selected to perform MD simulation. The binding energy was computed from MD analysis aiming to validate the predicted IC_{50} from MHC-peptide predictor in MuPeXI. There were ten complexes of HLA-A*02:01/peptide as shown in the Table 2.1. The MD simulation of the prepared ten system models was performed as following the protocol described in the section 2.2.4. Root-mean-square displacement (RMSD) calculation was performed to monitor conformational stability during the MD simulation. RMSD of candidate peptides are similar to the structure of the template from PDB (~1 A°) and reach equilibrium around the last 10 ns whereas the RMSD of negative control (non-candidate peptides) reached ~4 A° indicating that the stability of binding structure between MHC and non-candidate peptides is not as good as for those candidates (Figure 2.4). To calculate binding free energy of the MHC-peptide complex, snapshots from the production phase i.e. MD trajectories from the last 10 ns, were captured to

estimate the binding free energy using the MM-GBSA technique. In MM-GBSA, the binding free energy is evaluated as a sum of a conformation energy terms in the MM part (ΔE_{MM}), a solvation free energy term (ΔG_{sol}) that is computed using electrostatic field, and the entropy terms at a constant temperature ($-T\Delta S$) (Eq.1.1).

$$\Delta G_{bind-GBSA} = \Delta E_{MM} + \Delta G_{sol} + (-T\Delta S) \quad \text{(1.1)}$$

In the MM part, a conformation energy is a summation of the electrostatic interaction energy (ΔE_{ele}) and the van der Waals (ΔE_{vdW}) interaction between a ligand and its surroundings. In the GBSA part, the solvation energy is contributed by a summation of polar ($\Delta G_{sol-ele}$) and non-polar (ΔG_{sol-np}) energy terms. The energy components are shown in Table 2.6. The total binding free energy between HLA-A*02:01 and candidates 1, 6, and 7 are similar to the energy of a peptide obtained from a crystallisation of a complex of HLA-peptide (tem), indicated by red boxes, suggesting that the complexes of candidates 1, 6, and 7 HLA-A*02:01 molecule have a potential for a favourable protein-ligand interactions. However, candidates 2, 3, 4, and 5 displayed binding free energies similar to the negative controls and higher than the template suggesting that they might have a poor interaction with HLA-A*02:01 molecule.

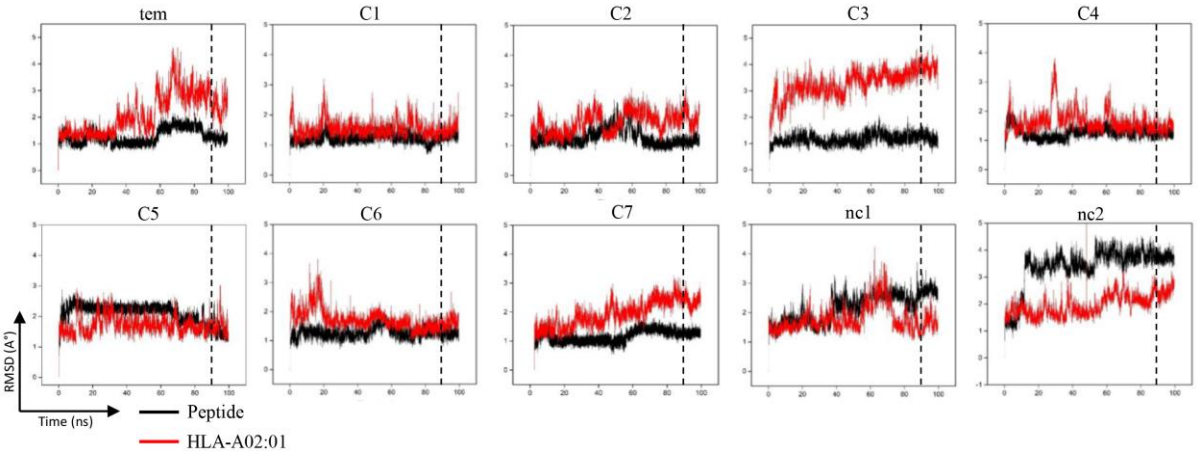

Figure 2.4 Root mean square deviation (RMSD) of HLA-A*02:01 and peptide complexes of 100 ns simulation. The dashed line marks 90 ns. Tem = template from crystal structure (3QEQ); c = candidate peptides from MuPeXI prediction; nc = non-candidate peptides.

Table 2.6 The binding free energy and energy components (kcal/mol) for the ten complexes of HLA-A*02:01/Peptide.

	Template	Candidate 1	Candidate 2	Candidate 3	Candidate 4	Candidate 5	Candidate 6	Candidate 7
MM								
ΔE_{ele}	-286.30±1.43	-329.98±2.05	-70.94±6.19	131.71±4.89	-62.78±12.22	-121.96±7.21	-257.14±5.83	-318.78±1.40
ΔE_{vdW}	-76.34±0.31	-92.90±0.32	-18.33±2.09	-2.41±0.84	-34.64±2.69	-37.27±2.35	-66.56±1.75	-91.58±0.31
ΔE_{MM}	-362.64±1.43	-422.88±1.99	-89.27±8.27	129.31±5.72	-97.42±14.88	-159.23±9.526	-323.70±7.51	-410.36±1.35
GBSA								
ΔG_{sol-np}	303.01±1.19	358.90±1.75	76.22±6.79	-128.40±5.14	75.07±13.02	134.62±7.98	282.60±6.46	350.46±1.22
$\Delta G_{sol-ele}$	-12.10±0.03	-14.5±0.03	-2.88±0.323	-0.33±0.12	-5.24±0.41	-5.79±0.36	-10.65±0.28	-13.43±0.02
ΔG_{sol}	290.91±1.18	344.32±1.731	73.34±6.456	-128.73±5.02	69.83±12.62	128.83±7.62	271.95±6.19	337.03±1.22
$\Delta G_{total-GBSA}$	-71.74±0.40	-78.56±0.40	-15.93±1.82	0.58±0.71	-27.58±2.27	-30.41±1.93	-51.75±1.38	-73.33±0.39
$-T\Delta S$	-45.47	-54.26	-8.61	-62.82	4294.40	-15.36	-53.59	-52.65
$\Delta G_{bind-GBSA}$	-117.20	-132.82	-24.54	-62.24	4266.81	-45.76	-105.34	-125.97

	Non-candidate 1	Non-candidate 2
MM		
ΔE_{ele}	-321.16±13.97	-371.17±6.99
ΔE_{vdW}	-49.84±1.91	-36.99±0.77
ΔE_{MM}	-370.99±15.82	-408.16±6.30
GBSA		
ΔG_{sol-np}	352.56±15.08	380.55±6.44
$\Delta G_{sol-ele}$	-8.22±0.31	-6.00±0.12
ΔG_{sol}	344.34±14.77	374.56±6.56
$\Delta G_{total-GBSA}$	-26.65±1.09	-33.61±0.49
$-T\Delta S$	-46.66	-34.68
$\Delta G_{bind-GBSA}$	-73.31	-68.29

Furthermore, the relative binding affinity of HLA-A*02:01 and each peptide was examined by per-residue energy decomposition using an implicit solvent model. The total binding free energy excluding the entropic contribution was plotted per-residue to illustrate the peptide-HLA binding pattern. The binding free energy of an individual amino acid within the core 9mers was compared at identical position among all ten peptides. The residues at position 2 and the C-terminus in a peptide are anchor residues that contribute binding interactions to HLA-A*02:01 binding groove. The result showed amino acids at the p2, p8, and p9 positions have lower binding energy than other position, and the candidates 1, 6, and 7 have lower binding energy at those positions compared to other candidates and negative controls (Figure 2.5A). A structural superimposition over all complexes taken from the last MD snapshot was performed. Examination of the side chain directions of the residues at p2 and p9 positions shows that the side chains of template and candidate 1, 6, and 7 were orientated towards the binding cleft of HLA-A*02:01. In contrast, the direction of non-candidates was out of the

groove of HLA molecule (Figure 2.5B). The binding energy results agrees with the predicted binding affinity from a sequence analysis for only three from seven peptides suggesting that those three candidate neoantigens are more convincing as true neoantigens than the other four peptides. However, MD simulation is a technique that analyses the physical movement of atoms and molecules in a simulated circumstance for a fixed period of time as trajectories can go in different directions from the same starting point. Hence, the binding free energy might not fully represent the genuine interactions of macromolecules in the real biological scenario.

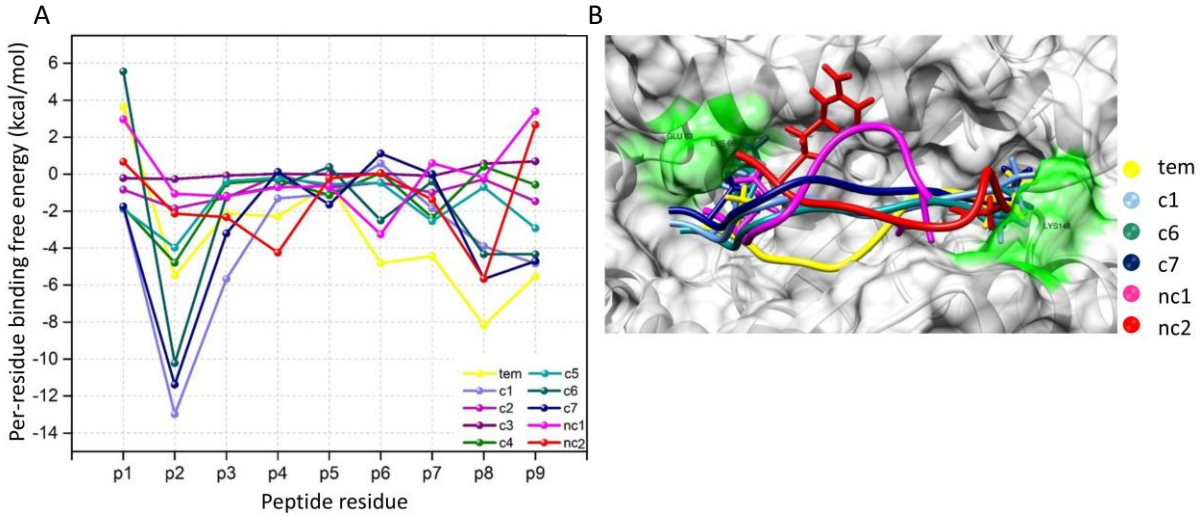

Figure 2.5 Per-residue free energy decomposition values of HLA-A*02:01/peptide complexes. (A) The binding free energy per residue of 9mers peptides. (B) The superimposition of candidate peptides and negative peptides in HLA-A*02:01 binding pocket. Tem = template from crystal structure (3QEQ); c = candidate peptides from MuPeXI prediction; nc = non-candidate peptides.

Besides the analysis of binding strength, a structural analysis can also provide a visualisation of the binding structure of a protein and ligand. The last MD snapshots of candidates1, 6, and 7 are shown in Figure 2.6. The structure models showed that the side chains of mutated residues in candidates 1 and 7 are oriented towards the solvent interface (indicated by the red boxes) whereas the side chain of the mutated residue of candidate 6 is buried in the binding cleft suggesting that mutated peptides of candidate 1 and 7 probably have good potential to be

recognised by T cell receptors: the orientation of side chains of mutated residues towards the surface might makes a peptide to be more prominently recognisable as non-self, which is promising for T cell recognition. While it is not straightforward to scale up MD simulation to high-throughput data, these results suggest that MD can play a role in suggesting improved peptide candidates for vaccine development. Therefore, the following analyses in this study focus on the methods of prediction of MHC-peptide binding affinity for solving neoantigen identification tasks.

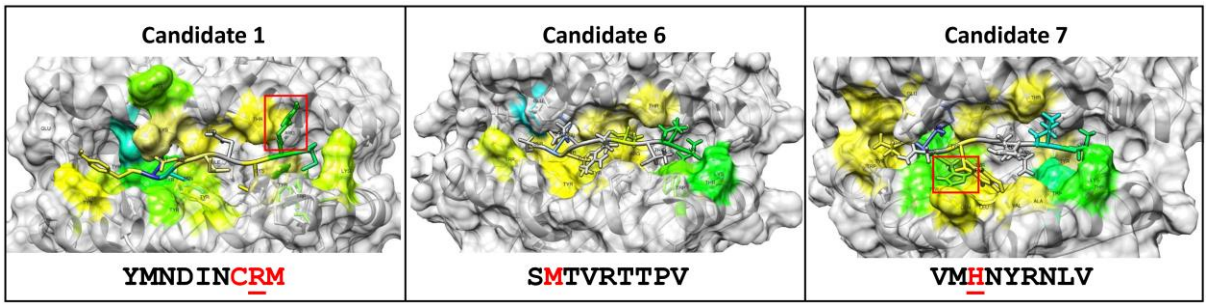

Figure 2.6 The orientation of side chain of the mutated amino acids for three selected candidates.

2.3.3 Analysis of the predicted scores with existing MHC class I-peptide binding prediction tools

From the analysis above, the prediction of binding affinity between MHC and peptide sequences is the most feasible in clinical practice and suitable for dealing with high throughput data such as genomic or transcriptomic sequencing data. The binding peptides are commonly determined by predicted IC_{50} < 500 nM or predicted rank score < 2%. From the analysis in the Section 2.3.1, the predicted binding affinity and gene expression level are the main contributing variables for candidate neoantigen selection. Nevertheless, binding affinity predictions still may carry a high risk for getting false positives especially in HLA alleles lacking training data for prediction tools. As such, we analysed the false positive rate for different HLA alleles to further understand this phenomenon. The overlaying of predicted IC_{50} of candidates from

Sample 6 and random peptides (predicted against same set of HLA alleles of Sample 6), was created to explore the distribution of predicted MHC binding affinity of putative candidate neoantigens and random peptides. The result displayed some overlapping between the scores from candidate neoantigens and from random peptides indicating imperfect separation of true and random binders (Figure 2.7). In this section, the prediction behaviour of NetMHCpan and MHCflurry for random peptides against various HLA alleles was studied to explore the prediction of random background and estimate false positive rate from random peptides for different types of HLA alleles.

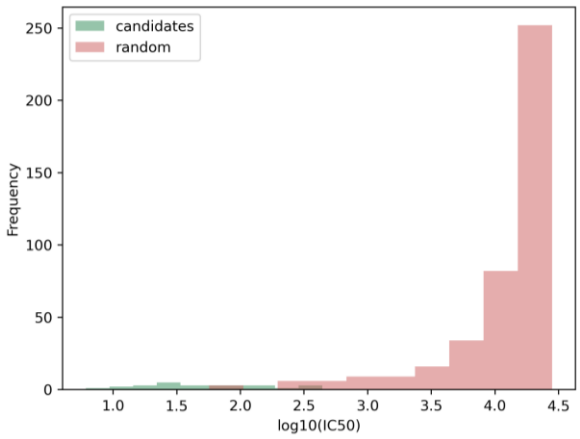

Figure 2.7 The overlaid distribution of predicted binding affinity scores (log_{10} IC_{50}) from candidate neoantigens of Sample 6 and random peptides. Those peptides were predicted against HLA alleles carried by Sample 6 using NetMHCpan4.1.

2.3.3.1 The analysis of random background

A set of 9mers random peptides was used as input for NetMHCpan4.1 and MHCflurry for predicting against different 79 HLA alleles, which are supported by MHCflurry and NetMHCpan. Those predictors are commonly used in several neoantigen identification workflows, and their performance are comparable [66]. The thresholds of 500 nM and 2% rank score were used to characterise “binding” peptides i.e. expected true positive binders. For each allele, the binding peptides were counted and calculated to the percent binder of a random set. The results showed that different MHC molecules have different number of random binders

cut off by a fix threshold in both prediction tools (Figure 2.8). Among those alleles, the percent random binder of NetMHCpan4.1 using predicted IC_{50} (<500 nM) as a threshold ranged from 0.01% to 7.32%, while using the predicted % rank score < 2% as a threshold, the percent random binder ranged from 1.88% to 6.29%. For MHCflurry prediction, the percent random binder selected by IC_{50} < 500 nM ranged from 0.01% to 7.12%. With using the rank score < 2% with MHCflurry predicted results, the percent random binder ranged from 0.97% to 15.65%. This result can imply different binding preferences for different MHC molecules. However, given that NetMHCpan4.1 documentation suggests using the predicted % rank score to select binders rather than the IC_{50}, the counts of random binders among different alleles still vary considerably at this threshold.

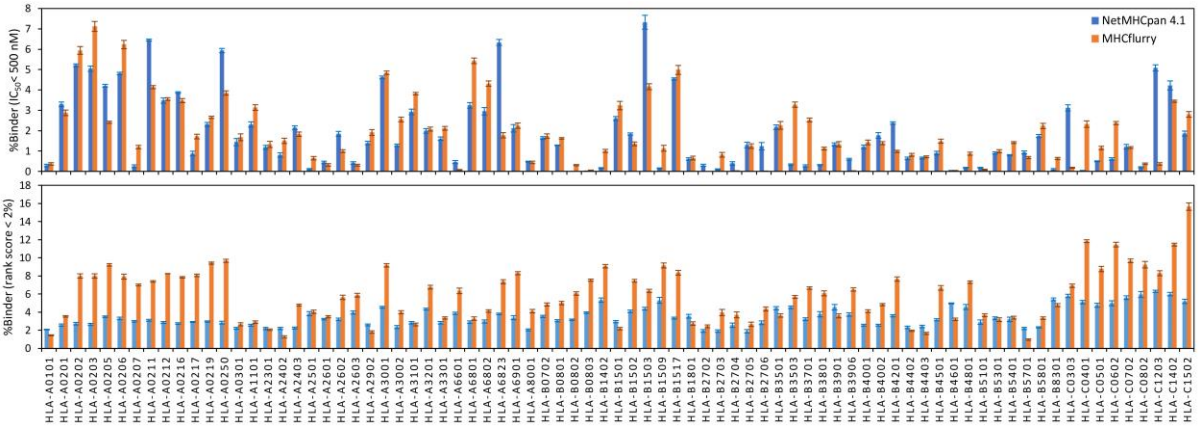

Figure 2.8 The percent random binder of specific alleles. The upper panel is the binders those were selected by the IC_{50} less than 500 nM. The lower panel is the binders those were selected by the % rank < 2%

2.3.3.2 Determination of predicted binding affinity at the top 1%

A method to find the value for both predicted IC_{50} and predicted % rank score, corresponding to an approximate top 1% of ranked scores was next applied. The predicted IC_{50} of each HLA allele were ranked from low to high. The value at 100^{th} ranked position out of 10,000 for each allele was selected as the threshold, that value was denoted as 1% false positive rate (FPR) threshold per allele. It should be noted that peptides that pass 1% threshold might not genuinely

false binding peptides, but we use the term FPR to imply the proportion of random peptides passing a given threshold, which could be non-binders or "random" real binders. The predicted IC_{50} at the top 1% FPR from prediction using NetMHCpan4.1 and MHCflurry displayed high variation across 79 HLA alleles, and most of them are not close to 500 nM, marked as the red line (Figure 2.9). For NetMHCpan4.1, the predicted IC_{50} giving the 1% FPR threshold ranges from 7 nM (A*02:11) to 9,795 nM (B*08:02). With MHCflurry, the predicted IC_{50} giving the 1% FPR threshold ranges from 12 nM (A*02:03) to 11,248 nM (B*27:02). Moreover, within a set of alleles in the same HLA supertype, which have the same preference of amino acids at the anchor positions, also exhibited great variability in the predicted IC_{50} giving the 1% FPR threshold, such as the superfamily of HLA-A*2. Those results above indicate that different HLA alleles have different predicted binding affinity scores even though they are likely to bind to same (or highly similar) peptides. As the input was the same set of random peptides, the results certainly showed that different alleles can bind the peptides at 1% FPR either lower or higher than 500 nM. If a single fixed threshold is used for any allele, results from some alleles will contain more false positives, while some sets will lose true positives. Therefore, each specific HLA allele should have their own threshold that would allow the same FPR. This topic is the focus of Chapter 3.

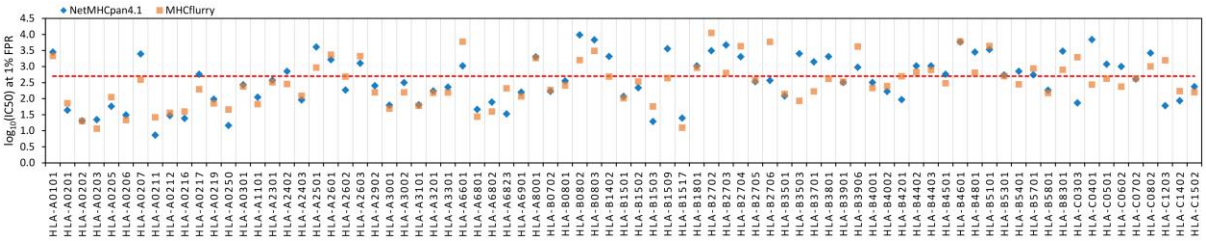

Figure 2.9 The predicted IC_{50} corresponding to 1% FPR across 79 HLA alleles. The predicted binding affinity scores were obtained from the prediction between random peptides against to 79 HLA alleles using NetMHCpan4.1 and MHCflurry. The red line marked at the value of $\log_{10}$(500 nM).

2.4 Discussion

The current technology of genomics sequencing and bioinformatics allows the identification of tumour-specific mutation in protein sequences that play a role as neoantigens for cancer vaccines. Thus, identification of neoantigens is crucial for cancer therapeutics-based cancer vaccine. In this chapter, the analysis of neoantigen identification was performed via the application of genomic analysis with packages of bioinformatic software and structure analysis with a molecular dynamic simulation technique. In this study, the tissue and blood samples for DNA and RNA sequencing were obtained from nine colorectal patients from King Chulalongkorn Memorial Hospital, Bangkok, Thailand. This part has been achieved through collaboration with a research team at the Faculty of Medicine, Chulalongkorn University, they kindly shared sequencing data sets from their cohort to use as input data for performing neoantigen prediction pipelines. Using prediction methods based on sequencing data, there are several factors that can affect the accuracy of the identification (for workflow see Figure 2.1). The quality of tumour tissue is an initial factor that results to a quality of sequencing depth. The high depth of sequencing data can contribute more accuracy in a set of tumour specific-mutations that are a source for neoantigens. Besides the biological factors, accurate variant calling analysis with bioinformatic software is critical since falsely identified mutations increase the risk of getting false neoantigens. This study utilised GATK analysis that is a standard pipeline for identifying SNPs, small Indels in DNA data [150]. Since the first step that can rule out non-neoantigens is the prediction of binding affinity between the mutated peptide and a patient specific HLA, hence, the precision of determining MHC from a patient is also important. MuPexI uses NetMHCpan3.0 for MHC-peptide binding prediction, NetMHCpan has been accepted as a gold standard for MHC-peptide binding prediction in the present [161]. In this study, Kallisto was used to quantified gene expression level, the benchmarking with

standard RNA data showed its performance is fast and accuracy is as good as existing tools [93].

Neoantigens are highly person-specific, and mutations can occur in any genes besides common cancer driver genes, thus neoantigen identification of individual must be tailored made. As described above, good quality of sequencing data is firstly important for a neoantigen prediction pipeline. However, in some cancer cases, the tissue sample from surgery might not feasible or not enough for making a good quality sequencing data. Furthermore, different cancer types have various level of mutation burden, low mutation burden obstructs the neoantigen identification, consistent with Figure 2.3 that showed a direct proportion between numbers of non-synonymous mutations and identified candidate neoantigens. Several studies have put the effort to investigate common mutations to create cancer vaccine as off-the-shelf therapies. There have been approaches to mine data from publicly data in The Cancer Genome Atlas (TCGA) to explore the common somatic mutations present in each tumour type [162]. The analysis from that research found that TP53 mutation is highly found across breast, head and neck and colon cancer [163]. That result agrees with our finding that neoantigens from TP53 mutations are found in 4 of 9 colon cancer patients suggesting that shared mutation-specific tumour could be possibly used for generic vaccine development.

Besides the method relying on sequencing data and binding prediction algorithms, this chapter demonstrated a proof of concept of structure-based analysis for scoring MHC-peptide binding energy. The results from MD simulation analysis provided insight into the energetic binding between MHC molecule and peptide. That information can reasonably explain the interaction of MHC-peptides by physicochemical properties of amino acids in a peptide and in a binding groove of MHC molecule. The advantage of structure analysis is that results can be visualised, which allow us to observe the side chain direction of mutated residue(s) that can further infer the potential for immunogenicity of candidate peptides. It has previously reported that peptides

with their mutated residues orientated towards the solvent are likely to be immunogenic peptides because they are well captured by T cell receptor binding region [52]. However, the approach of MD simulation might not genuinely represent binding interaction in the real biological environment and not be suitable for high throughput screening in practice because it consumes high computational resources. Moreover, there are only a few MHC types that have a crystal structure deposited in PDB. For other alleles, using a predicted structure might increase the risk for getting inaccurate results. With those limitations, it might not appropriate for using MD results to validate predicted results from a sequence analysis. Therefore, MHC-peptide binding prediction algorithms relying on sequence analysis was emphasised in this study, and in later chapters of the thesis.

In the phase of binding affinity prediction, not only mutated gene expression data but the detail of MHC types of patients is also required. In the present, the performance of binding prediction algorithms relies on experimental data of peptide-MHC binding affinity deposited in the IEDB [65], then the accuracy of prediction result might be biased due to lacking data of some MHC types. The diversity of MHC molecules is extremely polymorphic due to extensive polymorphism at most loci, and expression of MHC molecules may have evolved through diversity of pathogen specific immune system. Hence, some haplotypes might be common in specific for some ethnic groups, which might not be common in deposited data in available databases [9]. Each MHC type has a binding preference to specific peptides, those uncommon types might be inadequate experimental binding affinity data in the database, thus, the algorithms might provide false predicted scores due to lack of training data. The analysis in the Figure 2.9 supported the hypothesis above, at the 1% FPR the predicted IC_{50} has high variation across different HLA alleles, even in those are in the same supertype i.e. the predicted IC_{50} of all alleles in the family HLA-A*2 supposed to be similar, but the result displayed high variation among them. In this study, a set of short peptides derived from human proteomes was used to

perform analysis, each analysis was done from five different sets of 9mers peptides (10,000 peptides per set). The results showed very small variation from the five different sets, indicating that sampling random peptides can represent a population (Figure 2.8). However, to prevent the bias from true MHC binders that might be in a set of random peptides, the peptides that have a preference of anchored position for a given HLA allele must be removed before predicting that data and HLA. Furthermore, the predicted scores at 1% FPR are diverse from the fixed threshold of 500 nM indicating that using the fixed threshold for any HLA alleles might not be appropriate and cannot control a false positive rate in the predicted binding peptides. This topic is the focus for the following chapter.

2.5 Conclusions

The results provided by this chapter demonstrated the use of a practical workflow for neoantigen identification as well as the behaviour of MHC-peptide prediction algorithms. The neoantigen identification can be generated by a sequencing analysis approach with WES and RNA sequencing data using bioinformatic software for variant calling, identifying MHC types, quantifying gene expression levels, and MHC-peptide binding prediction algorithms. The approach of protein structure-based analysis was also promising to quantify binding strength between MHC and a peptide. Finally, the predicted behaviour of the gold standard MHC-peptide binding prediction algorithms including NetMHCpan and MHCflurry was explored. The predicted scores at 1% FPR across different HLA alleles are greater or lower than the fixed threshold suggesting using the fixed threshold might not deliver a stable ratio of true and false positives. Therefore, the statistical values that can describe the probability of predicted scores for being true or false positive is essential to improve the criteria for binding peptide selection, which will be discussed further in the following chapter.

Chapter 3

The development of a model to estimate statistical properties from MHC-peptide binding affinity prediction

Author contributions

All data in this chapter is published in *Bioinformatics*, where the thesis author PP is the first author. **Pearngam, P**., Sriswasdi, S., Pisitkun, T. and Jones, A.R., 2021. MHCVision: estimation of global and local false discovery rate for MHC class I peptide binding prediction. *Bioinformatics*.

3.1 Introduction

In Chapter 2, we demonstrated that through the use of random peptides, one can estimate a concept similar to FPR i.e. the proportion of false observations passing a given threshold, from MHC-peptide binding results. Such a statistic could be somewhat useful for selecting peptides for onward analysis, and removing allele-specific differences in the proportion of random peptides that pass an ad hoc score threshold e.g. <500 nM affinity. In fact, it is arguable whether the proportion of random peptides passing a given threshold is an accurate FPR, since it might cover both genuinely false positives i.e. peptides that will not be bound by the given MHC molecule, as well as random true binders. Nevertheless, for most uses of peptide binding prediction results, more useful concepts relate to the local or global False Discovery Rate (FDR) than the FPR. The local FDR, the posterior error probability (PEP) associated with each predicted value, which describes the actual probability that a given peptide will not bind to a given MHC molecule (and 1-PEP gives us the probability that it will bind). Moreover, the global FDR is widely used as a standard threshold in other scientific disciplines using large data sets, for deciding how to apply a threshold that present a good balance between sensitivity (proportion of true positives from all true) and reporting false observations. To estimate local or global FDR perfectly, one would need to know which data points are genuinely true and false (real or not real binders), which in practice will never be the case (since this is what we wish to predict). A typical approach to calculate PEP (and the converse is the true probability of prediction, 1-PEP) and global FDR, is commonly performed via fitting two distributions to the assumed true and false distribution of scores and estimating the relative density at a given data point for PEP values, whilst the FDR can be estimated by the relative ratio of accumulated numbers of true and false at a given data point (Figure 3.1).

In this chapter, the data distribution of predicted scores of MHC-peptide binding affinity was studied, and the model for parameter estimation was investigated to finding the best estimated

parameters generating distributions of predicted data sets, which are further utilised for estimating FDR and PEP values. The context of the expectation maximisation (EM) algorithm for parameter estimation is described first. In the section 3.3.1, the normal distribution and the beta distribution were explored to fit the predicted data to determine the best model representing predicted data distribution. To estimate the true and false results, the parameters of data distribution are needed to estimate. Section 3.3.2 demonstrated the framework of the EM algorithm with using the method of moments for beta mixture parameter estimation. The estimated parameters were used to calculate FDR and PEP by a cumulative density function (CDF) and a probability density function (PDF) of beta distribution as shown in the section 3.3.3.

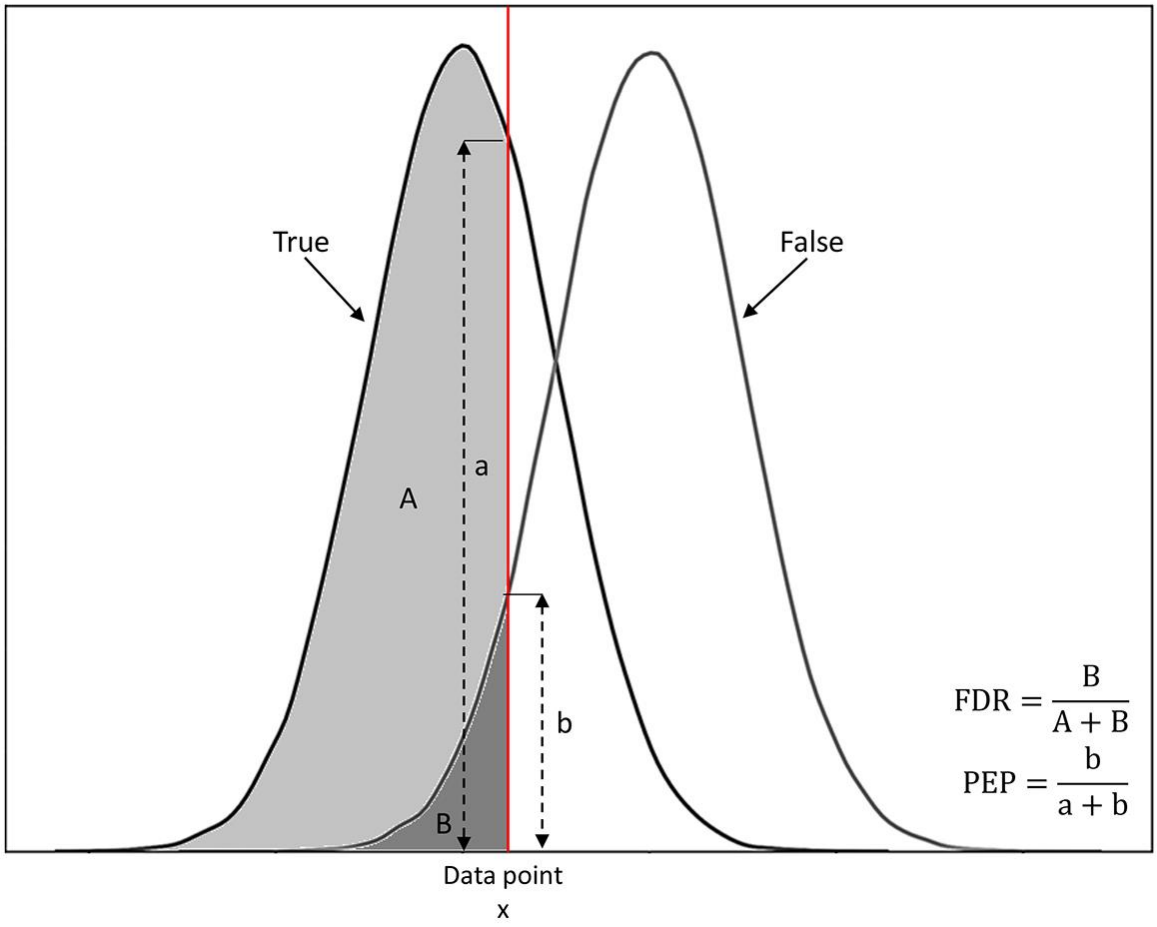

Figure 3.1 The calculation of PEP and FDR from true and false results.

3.1.1 The Expectation Maximisation (EM) algorithm for parameter estimation

The EM algorithm, introduced by Dempster *et al.* in 1977 [164], is an efficient iterative method to compute the Maximum Likelihood (ML) estimation in statistical models with the presence of unobserved latent variables. Each iteration cycles between two processes including the E-step (Expectation) and the M-step (Maximisation). The E-step attempts to estimate the missing data or latent variables given the observed data and current estimated parameters of the distribution. Then, the M-step tries to optimise the parameters of the model by maximising the likelihood function with the assumption that the missing data are known, where the missing data are placed from the estimation in the E-step. The application of EM algorithm is widely used for estimating missing data for clustering in a mixture model, or in ML estimation, moreover, the EM approach is commonly used for estimating parameters of the distribution [165]. In this section, the use of EM algorithm for parameter estimation is highlighted.

To describe what probability each random variable has in the whole data, the model family and parameters for a distribution must be known for the situation of interest. However, in reality, the generative source of data might be uncertain i.e. the model family and parameters representing that distribution are not known, thus it is essential to explore and predict the parameter values and the statistical model that can well describe data distribution. Parameter estimation is a branch of statistics that contributes tools using observed data to estimate parameters of a distribution, so-called estimator [166]. By leveraging the EM algorithm, the approach of ML estimation is a well-known method and commonly used. The EM algorithm iteratively switches back and forth between the two steps of the E-step M-step to optimise the estimate parameters by likelihood maximisation. If the estimate parameters or the likelihood are not getting to convergence, the new parameters from the M steps will return to the E-step. Finally, these two steps are repeated until the estimate model get the convergence [167]. To find a maximum likelihood solution, it generally involves the derivatives of the likelihood

function that requires taking all the unknown values, the parameters, and the latent variables, and together with solving the resulting equations [164].

Besides the conventional method with likelihood function, the estimation with method of moments is also widely used for the parameter estimation approach. The method of moments estimators is simple and in closed form. In statistics, method of moments, introduced by Karl Pearson in 1984 [168], is an approach for population parameters estimation such as mean or variance. This approach estimates the parameters of a distribution model by matching the moments of the data set with that candidate model. In the first moment condition, it expresses the expected values of random variables under the parameters calculated from population as functions of the population moments. Then, those equations are set as equal to the sample moment. The number of those equations is equivalent to the number of parameters that are desired to estimate, and they are solved for estimating the parameters of interest [169]. With the EM framework, method of moments can apply in the estimation step by replacing the likelihood function. Instead of maximising likelihood in the estimation step, the parameters are optimised based on method of moment until convergence [170]. One of the problems related to parameter estimation is we may not know which types of statistical model would be best represent the distribution of data. Hence, the distribution models representing observed data must be determined, once the best fitting distribution has been identified, the parameters of that function can be then estimated. In the following section, the parameters estimation for a mixture distribution of normal and non-normal (beta) distributions using EM algorithm with ML estimation are described.

3.1.2 The parameter estimation using EM for a mixture of normal distributions

The normal distribution (also known as Gaussian distribution) is the single most important distribution in natural and social sciences to represent real-valued random variables. It is a type

of continuous probability distribution that is described by a distribution with a symmetrical bell shape which is parametrised by two parameters that are mean (μ) and standard variation (σ) (Figure 3.2) [171]. The area under the curve is the same and most of the values occur in the middle of the curve. The mean controls the location of the central peak, while the variance controls the width of the distribution [172]. The general form of its PDF and CDF with the form of error function (*erf(x)*) are shown as the Eq. 3.1 and 3.2, respectively.

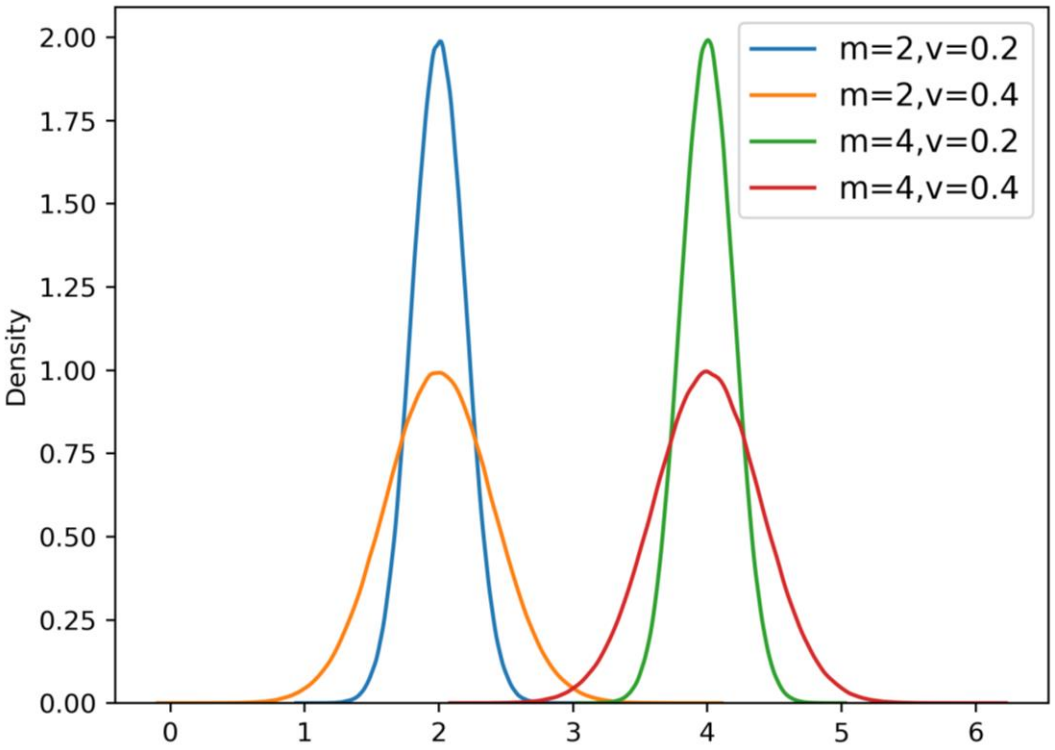

Figure 3.2 The distribution shapes generated by the normal distribution with different of mean and variance values.

$$N_{\mu,\sigma}(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp[-\frac{(x-\mu)^2}{2\sigma^2}] \quad (3.1)$$

$$F_{\mu,\sigma}(x) = \frac{1}{2}[1 + \text{erf}\left(\frac{x-\mu}{\sigma\sqrt{2}}\right)] \quad (3.2)$$

There are two distribution parameters including μ and σ, π is the mathematical constant ~3.1415, the random variables in normal distribution can be any real number, $-\infty < x < \infty$. However, some observed data might arise from more than one generation process, such a distribution is represented by *mixture distributions*, which are contributed from the average

weight of two or more PDFs. The general form of a Gaussian mixture model is the Eq. 3.3, where c is the number of components, each component is contributed by the weighting parameter (w), $0 < w_j < 1$. The model parameters including μ_j, σ_j and w_j are described in term of θ.

$$f_{\theta(x)} = \Sigma_{j=1}^{c} w_j \cdot N_{\mu_j,\sigma_j}(x) \quad (3.3)$$

A Gaussian mixture model is commonly used as a parametric probability density function for a distribution of continuous measurement. The model parameters are usually estimated using the iterative EM algorithm that aim to obtain the maximum likelihood estimate of Gaussian parameters. Given the data have N observations, $X = \{x_1,\ldots,x_N\}$, the likelihood for the Gaussian parameters of μ and σ displays in the Eq. 3.4, and the log-likelihood function is generally used because it is convenient for derivative calculation (Eq. 3.5).

$$L(\mu,\sigma|X) = \sigma^{-N}\left(\sqrt{2\pi}\right)^{-N} \prod_{n=1}^{N} \exp\left[-\frac{(x_n-\mu)^2}{2\sigma^2}\right] \quad (3.4)$$

$$l(\mu,\sigma|X) = \log[L(\mu,\sigma|X)] = -N\log(\sigma) - N\log\left(\sqrt{2\pi}\right) - \frac{1}{2\sigma^2}\sum_{n=1}^{N}(x_n-\mu) \quad (3.5)$$

For parameter estimation using the EM algorithm, we define $X = \{x_1,\ldots,x_N\}$, as observed variables from a Gaussian mixture model, which have N data points and K components, and $Z = \{z_1,\ldots,z_N\}$, is denoted as a set of latent variables corresponding to each data point of each component. The EM algorithm attempts to find the maximum likelihood estimates for model parameters with latent variables, thus the complete log-likelihood derived from the posterior of the latent variables of the data X with the expression of normal distribution parameters (Eq. 3.6).

$$\log\left(P(X,Z|\mu,\sigma,w) = \sum_{n=1}^{N}\sum_{k=1}^{K} z_{nk}\log\left[N(x_n;\mu_k,\sigma_k)w_k\right]\right. \quad (3.6)$$

In the E-step, the current values of estimated parameters at the iteration *i* are used to calculate the expected value of the latent variables (Eq. 3.7). Therefore, the expected value of the complete log-likelihood is shown as the Eq. 3.8.

$$\mathrm{E}_{p(Z|X,\mu_k^i,\sigma_k^i,w_k^i)}[z_{nk}] = \frac{N\left(x_n;\mu_k^i,\sigma_k^i\right)w_k^i}{\sum_{k=1}^{K} N(x_n;\mu_k^i,\sigma_k^i)w_k^i} \quad (3.7)$$

$$\mathrm{E}[\log\left(P(X,Z|\mu,\sigma,w)\right] = \sum_{n=1}^{N}\sum_{k=1}^{K} E[z_{nk}]\log[N(x_n;\mu_k,\sigma_k)w_k] \quad (3.8)$$

In the M-step, the expectation of the complete log-likelihood needs to be maximised to update parameters for the next iteration. The estimated parameters are solved from the derivative of the expected complete log-likelihood with the respect to μ_k, σ_k, and w_k (Eq. 3.9-3.11).

$$\mu_k^{i+1} = \frac{\sum_{n=1}^{N} E[z_{nk}]x_n}{\sum_{n=1}^{N} E[z_{nk}]} = \frac{1}{N_k}\sum_{n=1}^{N} z_{nk}x_n \quad (3.9)$$

$$\sigma_k^{i+1} = \frac{\sum_{n=1}^{N} E[z_{nk}]\left(x_n-\mu_k^{i+1}\right)^2}{\sum_{n=1}^{N} E[z_{nk}]} = \frac{1}{N_k}\sum_{n=1}^{N} z_{nk}\left(x_n-\mu_k^{i+1}\right)^2 \quad (3.10)$$

$$w_k^{i+1} = \frac{\sum_{n=1}^{N} E[z_{nk}]}{N} = \frac{N_k}{N} \quad (3.11)$$

3.1.3 The parameter estimation using EM for a mixture of beta distributions

The beta distribution is a flexible model, with a continuous probability distribution that takes values in the unit interval of 0 to 1. The beta distribution is widely used in statistical analysis and data science (including bioinformatics applications) to model the behaviour of random variables that naturally takes values between 0 and 1 such as relative frequencies, probabilities, absolute correlation coefficients [173]. The beta distribution is parametrised by two positive shape parameters that are denoted by α and β, that materialises as proponents of the random variable and control the shape of the distribution. The two parameters (α and β) must be positive numbers, and they can produce a variety of shapes depending on whether α = β, α < β, or α > β (Figure 3.3). The beta probability density on [0, 1] forms as the Eq.3.12, and the CDF for the beta distribution is also formed as the incomplete beta function ratio, that is normally denoted by I_x (Eq. 3.13).

$$f_{\alpha,\beta}(x) = \frac{x^{\alpha-1}(1-x)^{\beta-1}}{B(\alpha,\beta)}, B(\alpha,\beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)} \text{ and } \Gamma \text{ is the gamma function} \quad (3.12)$$

$$F_{\alpha,\beta}(x) = I_x(\alpha,\beta) = \frac{B_{\alpha,\beta}(x)}{B(\alpha,\beta)}, B_{\alpha,\beta}(x) \text{ is the incomplete beta function} \quad (3.13)$$

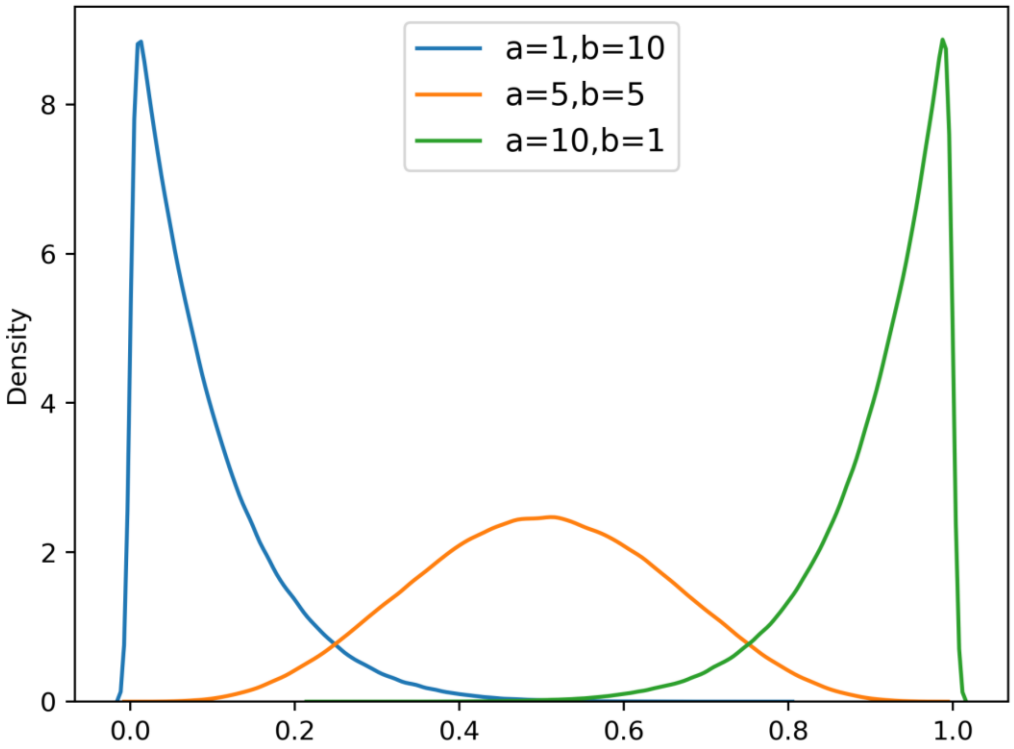

Figure 3.3 The distribution shapes generated by the beta distribution with different values of α and β parameters.

Even a single beta model can create various distribution shapes depending on different combinations of the parameters of α and β, the mixtures of beta model are more flexible. The general form of a mixture of beta distribution is shown as the Eq. 3.14, where c is the number of components, and w_j is the weighting parameter for each component, $0 < w_j < 1$. The model parameters including α_j, β_j and π_j are described in terms of θ. From the random variables, the parameter shapes of α and β can be described by the terms of μ and σ^2 (Eq. 3.15-3.16).

$$f_{\theta(x)} = \Sigma_{j=1}^{c} w_j \cdot f_{\alpha_j,\beta_j}(x) \quad \text{(3.14)}$$

$$\alpha = \mu\left(\frac{\mu(1-\mu)}{\sigma^2} - 1\right) \quad \text{(3.15)}$$

$$\beta = (1-\mu)\left(\frac{\mu(1-\mu)}{\sigma^2} - 1\right) \quad \text{(3.16)}$$

Given the data have N observations, X = $\{x_{1,\ldots},x_N\}$, the likelihood function for the beta distribution is the Eq. 3.17, and the log-likelihood can take the form as the Eq. 3.18.

$$L(\alpha,\beta|X) = \left(\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\right)^N \prod_{n=1}^{N}(x_n)^{\alpha-1}\prod_{n=1}^{N}(1-x_n)^{\beta-1} \quad \text{(3.17)}$$

$$l(\alpha,\beta|X) = \log[L(\alpha,\beta)|X)] = Nlog\big(\Gamma(\alpha+\beta)\big) - Nlog\big(\Gamma(\alpha)\big) - Nlog\big(\Gamma(\beta)\big) + (\alpha-1)\sum_{n=1}^{N}\log(x_n) + (\beta-1)\sum_{n=1}^{N}\log(1-x_n) \quad \text{(3.18)}$$

To estimate beta parameters using the EM algorithm with the ML method, the procedures in the E-step is similar to a Gaussian mixture model, but the expectation is considered with the probability density function of beta model. We denote X = $\{x_1,\ldots,x_N\}$, Z = $\{z_1,\ldots,z_N\}$, and data have K components. Thus, the expectation of the complete log-likelihood for beta mixture model displays as the Eq. 3.19.

$$E\big[\log\big(P(x,z|\alpha,\beta,w)\big)\big] = \sum_{n=1}^{N}\sum_{k=1}^{K}E[Z_{nk}]\log\left[f(x_n;\alpha_k,\beta_k)w_k\right] \quad \text{(3.19)}$$

The derivative of the expected log-likelihood of beta function is performed with the respect to each parameter. The partial derivatives are set to zero to solve the update parameters (Eq. 3.20-3.22). However, there is no closed form solution to the derivative equation of α and β parameters if any observed data points are equal to 0 or 1.

$$\frac{\partial}{\partial\alpha_k}E\big[\log\big(P(x,z|\alpha,\beta,w)\big)\big] = \sum_{n=1}^{N}E[z_{nk}]\left[\frac{\Gamma'(\alpha+\beta)}{\Gamma(\alpha+\beta)} - \frac{\Gamma'(\alpha)}{\Gamma(\alpha)} + \log(x_n)\right] \quad \text{(3.20)}$$

$$\frac{\partial}{\partial\beta_k}E\big[\log\big(P(x,z|\alpha,\beta,w)\big)\big] = \sum_{n=1}^{N}E[z_{nk}]\left[\frac{\Gamma'(\alpha+\beta)}{\Gamma(\alpha+\beta)} - \frac{\Gamma'(\beta)}{\Gamma(\beta)} + \log(1-x_n)\right] \quad \text{(3.21)}$$

$$\frac{\partial}{\partial w_k}E\big[\log\big(P(x,z|\alpha,\beta,w)\big)\big] = \sum_{n=1}^{N}E[z_{nk}]\cdot\frac{1}{w_k} \quad \text{(3.22)}$$

Since the derivative with the respect to w_k does not rely on the PDF of beta model, thus the estimated w_k can be solved to get closed form same as the formula in Eq. 3.11. The major problem of the log-likelihood function for beta distribution is that it is unable to estimate well for $\alpha \neq 1$ if any observed data points are $x_n = 0$, or for $\beta \neq 1$ if any observed data points are $x_n = 1$. Therefore, the implementations of ML estimators might not be suitable for the best estimation of beta parameters.

As mentioned before, the approach of method of moments is also widely used for parameter estimation, this method is straightforward, and the moment generating functions can get finite forms for solving beta parameters estimation. There are published studies that report the usage of method of moments for beta mixture distributions [170, 174]. With the framework of EM algorithm, the application of method of moments is used instead of ML estimation in the M-step to estimate the update parameters until the estimation get the convergence [170]. In this chapter, the approach of EM algorithm with method of moments was used to develop the model to estimate the parameters of the predicted data from NetMHCpan4.1 and MHCflurry instead of maximum likelihood estimation. The implementation of the EM algorithm with method of moments for beta mixture parameters estimation is entirely described in Section 3.2.5.

3.2 Materials and Methods

3.2.1 Collection of MHC bound peptides derived from mass spectrometry (MS) analysis

Data sets of MHC-bound peptides derived from MS analysis were downloaded from the IEDB (https://www.iedb.org/) [65]. Human peptides identified from MS and bound to HLA-A, -B, and -C were collected. Other eluted peptides from MHC class I "mono-allelic cells", i.e. presented peptides from cells carrying a single HLA allele, were collected from several publications of immunopeptidomics studies [104, 175-177]. Peptides for each HLA allele from those sources were combined, and redundant peptides were removed. Only peptides with lengths of 8, 9, 10, and 11 mers were retained. However, the majority of peptide length in the collected data was found as 9mers peptides for all HLA alleles. HLA alleles that had ≥1000 9mers peptides were collected for onward analysis, totalling 85 HLA alleles covering HLA-A, -B, and -C. Additionally, the "multi-allelic data sets", i.e. a set of peptides presented by cells carrying several alleles, were obtained from the data set contains naturally presented HLA class I ligands derived from chronic myeloid leukemia (CML) patients [178].

3.2.2 Generation of MS-random peptides data sets

To generate the mixture of predicted scores of MS and random peptides, the data set of true binding peptides were combined with random peptides, where the method of random peptide generation was described in Section 2.2.5, in a ratio of 1000 MS to 4000 random peptides. The true binders were sourced from the 85 mono-allelic data sets containing MS identified data sets where peptides presented by genuine MHC alleles were presented. The predicted MHC-peptides binding affinity was performed by NetMHCpan4.1 with a command line using a stand-alone software package.

3.2.3 Generating the data sets from the statistical models

The simulated data sets of normal and beta distributions were generated by the random function of those statistical models using packages in Python. For the normal distribution, the data sets were simulated from the function of numpy.random.normal(μ, σ^2, size), and the function of numpy.random.beta(α, β, size) was used to generate the data sets for the beta distribution. The input parameters for those models were computed from the template data, which are predicted results of MS-random peptides data sets. The statistical values including μ and σ^2 were calculated and were taken to calculate beta parameters using Eq. 3.15 and 3.16.

3.2.4 Similarity measure

Linear regression was performed between the simulated and the scaled predicted score distribution. The resulting R-squared (R^2) statistics were used to evaluate the similarity between the two distributions. Since the simulated beta distributions lie the interval [0, 1], the binding affinity scores ($\log_{10}$ IC_{50}) were scaled to the same interval by diving the predicted scores by the maximum value for each data set.

3.2.5 The modified EM algorithm with the iterated method of moments for the beta mixture model

The parameter estimation algorithm for beta mixture was built by a Python script. The algorithm was proceeded iteratively as in the basis of the EM algorithm. The algorithm consists of four major steps including initialisation, expectation, maximisation, and termination. For each iteration, parameters (θ) including two mixture proportions (π_1, π_2), two means (μ_1, μ_2), and two variance values (σ_1^2, σ_2^2) were estimated for two components. However, in this work, the step of parameter estimation was computed by Pearson's method of moment instead of the maximisation of likelihood, thus, the maximisation step (M-step) was replaced by a method of moments estimation step (MM-step) [170].

Initialisation

As the distribution of predicted scores is a bimodal, thus, two was defined as number of components. The initial mixture proportion (π) of each component *j* was initially set as 0.5. The initial mean and variance (μ_j, σ_j^2) were calculated from the data of each component, and the initial values of α_j and β_j were then computed according to Eq. 3.23 and 3.24.

$$\alpha_j = \left(\frac{1-\mu_j}{\sigma_j^2} - \frac{1}{\mu_j}\right)\mu_j^2 \qquad \text{(3.23)}$$

$$\beta_j = \alpha_j\left(\frac{1}{\mu_j} - 1\right) \qquad \text{(3.24)}$$

Expectation (E-step)

The expected responsibility weight ($W_{i,j}$) of each component *j* and data point x_i was estimated from the probability density function of the current estimates for beta distributions (α_j^t, β_j^t) and the mixture proportion π_j^t (Eq. 3.25).

$$W_{i,j}^t = \frac{\pi_j^t\left(f\left(x_i\,;\alpha_j^t,\beta_j^t\right)\right)}{\sum_{j=1}^{k}\pi_j^t f\left(x_i\,;\alpha_j^t,\beta_j^t\right)}, \textit{where } f(x\,;\alpha,\beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\cdot x^{\alpha-1}\cdot(1-x)^{\beta-1} \qquad \text{(3.25)}$$

Method of moments estimation (MM-step)

For each component *j*, the mixture proportion is updated based on the new values of responsibility weights $W_{i,j}^{t}$ according to Eq. 3.26. Then, component's mean and variance and the beta distribution parameters are updated using the method of moments (Eq. 3.27-3.30).

$$\pi^{t+1}{}_j = \frac{1}{n}\sum_{i=1}^{n} W_{i,j}^{t} \quad \text{(3.26)}$$

$$\mu_j^{t+1} = \frac{\sum_{i=1}^{n} W_{i,j}^{t} \cdot x_i}{\sum_{i=1}^{n} W_{i,j}^{t}} \quad \text{(3.27)}$$

$$\left(\sigma_j^2\right)^{t+1} = \frac{\sum_{i=1}^{n} W_{i,j}^{t} \cdot \left(x_i - \mu_j^{t+1}\right)^2}{\sum_{i=1}^{n} W_{i,j}^{t}} \quad \text{(3.28)}$$

$$\alpha_j^{t+1} = \left(\frac{1-\mu_j^{t+1}}{\left(\sigma_j^2\right)^{t+1}} - \frac{1}{\mu_j^{t+1}}\right)\left(\mu_j^2\right)^{t+1} \quad \text{(3.29)}$$

$$\beta_j^{t+1} = \alpha_j^{t+1}\left(\frac{1}{\mu_j^{t+1}} - 1\right) \quad \text{(3.30)}$$

In this step, the estimated beta parameters for the beta 2 component were further constrained by the ranges of values calculated from the data sets of various sizes (10000, 5000, 1000) of predicted binding affinity scores from random peptides with a length of 8, 9, 10, and 11 mers against 85 HLA alleles. The purpose of this restriction was to ensure that the beta 2 component of the mixture model is certain to capture the false data. Moreover, to ensure the beta 1 component is not fitted to the wrong distribution when presented with all false data, the estimated parameters of the first component are restricted if the estimated $\pi_1 = 0$ and size of the negative set $\neq 0$ (predicted $IC_{50} > 10000$ nM) i.e. indicating that there is only one distribution found, and there are data points in the plausible range for false data. In this case, the ranges of α and β for the first beta component were initially calculated from data points with predicted $IC_{50} \leq 10000$ nM using Eq. 1 and 2, and the range of values are only allowed to deviate 25% from the initial estimates. In practice, these two constraints mean that when the algorithm detects evidence a very large imbalance, in either direction (i.e. all true or all false), the beta 1 or beta 2 is correctly fitted to the appropriate distribution.

Termination

The parameter updates (E-step and MM-step) were repeated until the maximal absolute changes in parameter values, k^t, between step t and $t + 1$ is less than 0.00001 (Eq. 3.31).

$$k^t = max\left(\left\{\frac{\left|\alpha_j^{t+1}-\alpha_j^t\right|}{max\left(\left|\alpha_j^{t+1}\right|,\left|\alpha_j^t\right|\right)}, \frac{\left|\beta_j^{t+1}-\beta_j^t\right|}{max\left(\left|\beta_j^{t+1}\right|,\left|\beta_j^t\right|\right)}, \frac{\left|\pi_j^{t+1}-\pi_j^t\right|}{max\left(\left|\pi_j^{t+1}\right|,\left|\pi_j^t\right|\right)},\ j = 1,2\right\}\right) \quad (3.31)$$

3.2.6 Testing the EM beta mixture model with the predicted data sets

The resulting peptide set of 85 HLA alleles in Section 3.2.2 was run through NetMHCpan4.1 for each the specific HLA allele. The predicted scores of each HLA allele were used as an input data for the estimator model to estimate beta parameters of true (MS) and false (random) data distributions. The correctness of estimation was measured by the relative change between the defined mixture proportions and their estimated values. The similarity between real and simulated data generated by estimated parameters was measured by the linear regression analysis yielding R^2 statistics. Moreover, Kolmogorov-Smirnov (KS) test was used to detect the difference between the real and simulated data sets, the significant threshold was set at p-value < 0.05. Moreover, for testing a robustness of the estimator model, a wider range of data sets with unknown true binding or non-binding peptides were used to test the model, and the accuracy of FDR and PEP values were observed.

3.2.7 Calculation of FDR and PEP for predicted scores

The estimated beta parameters were utilised to calculate values of FDR and PEP of an individual predicted score in the data set using Eq. 3.32 and 3.33, respectively. The number of false and true positive were estimated by the CDF of the beta distribution while density at true and false were estimated by the PDF of the beta distribution.

$$FDR_{x_i} = \frac{F_{\alpha_{false},\beta_{false}(x_i)}}{F_{\alpha_{true},\beta_{true}(x_i)} + F_{\alpha_{false},\beta_{false}(x_i)}} \quad (3.32)$$

$$PEP_{x_i} = \frac{f_{\alpha_{false},\beta_{false}(x_i)}}{f_{\alpha_{true},\beta_{true}(x_i)} + f_{\alpha_{false},\beta_{false}(x_i)}} \quad \text{(3.33)}$$

3.3 Results

3.3.1 The study of the statistical model fitting predicted data distributions

3.3.1.1 Data distribution of MHC-peptide binding predicted scores

The data distribution of the predicted binding affinity scores were represented by the predicted data set from the MS peptides from mono-allelic cells and multi-allelic cells. Data sets of MS peptides from multi-allelic cells were collected from a CML patient who has six alleles of HLA class I including A*03:01, A*68:01, B*07:02, B*44:02, C*07:01, and C*07:02. To compare with data from mono-allelic cells, the same six HLA alleles of MS peptides from mono-allelic cells were selected for representation. The MS peptides from multi- and mono- allelic cells were mixed with the same set of random peptides. MHC-peptide binding affinity values for their specific HLA alleles were then predicted using NetMHCpan4.1. The histogram plots were created from the predicted scores to display the data distribution of peptides identified by MS experiments from mono-allelic cells, multi-allelic cells, and random peptides (Figure 3.4). The distribution shape of the scores for MS peptide binding from mono-allelic cells was almost exclusively a single peak on the left side ($\log_{10}$ (IC_{50}) values < 3 or 3.5 depending on the allele). The overlay of random peptides, which are believed reasonably well model non-binders (or negative results), demonstrated a peak on the right side. Since peptides identified by MS data from mono-allelic cells are highly likely to be genuine binding peptides for a given specific HLA allele, it could imply that the peak on the left with low IC_{50} values is the distribution of binding peptides (positives), whilst the right peak (high IC_{50} values) is the distribution of the non-binding peptides (negatives). The distribution shape of A*03:01, A*68:01, and B*07:01 from multi-allelic cells displayed a bimodal distribution, there are two separated peaks that one located on the left (lower $\log_{10}$ (IC_{50}) values, higher binding affinity) and the other on the right

side (higher $\log_{10}$ (IC_{50}) values, lower binding affinity). This result is expected, since only some of the presented peptides in multi-allelic cell lines are presented by one allele. However, the left peak of B*44:02, C*07:01, and C*07:02 can hardly be observed, most predicted scores located on the side of low binding affinity. This might be cause from biological artifact of different expression level of HLA alleles in a representative sample. The distribution shape of the right-hand peak (low binding affinity peptides) from multi-allelic cells well matches the distribution shape of random peptides, indicating that random peptides also well model peptides not presented by a given HLA allele. The distribution of predicted scores from MS peptides from monoallelic cells mixed with random peptides of 85 HLA alleles were shown in Figure 3.4

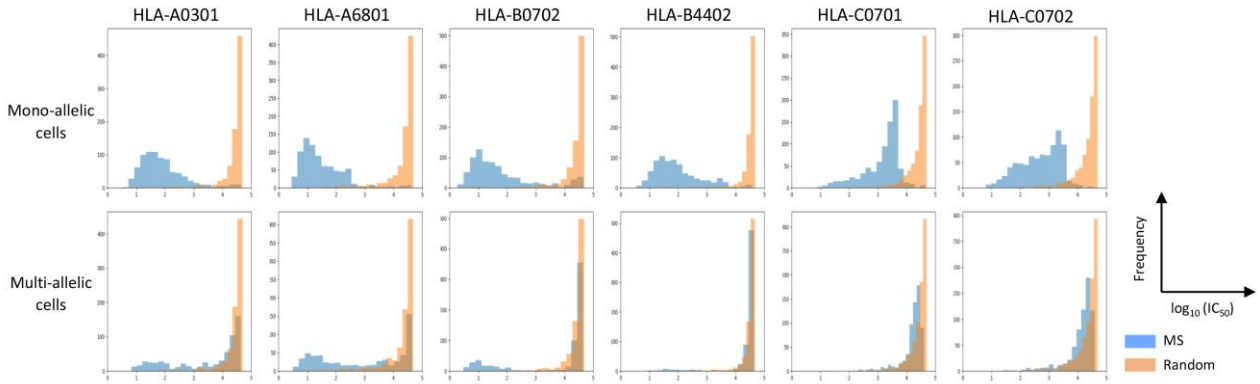

Figure 3.4 The distribution of predicted binding affinity data obtained from MS and random peptides. The distribution of predicted binding affinity of the MS peptides from mono-allelic cells (top) and from multi-allelic cells (bottom) and those MS data sets mixed with random peptides.

3.3.1.2 The mixture of models fitting a bimodal data distribution

From the inspection of predicted binding score distribution in Figure 3.5, the model fitting data distribution was performed. The distribution of MS peptides can fit to beta or Gaussian distributions (represented by HLA-A*0201) were shown in Figure 3.6. The distribution of positive results is generally symmetrical in shape, with an approximate bell shape, which can

be well modelled by Gaussian or beta. The negative distribution is not symmetrical, since it has a hard stop at about 4.69 ($\log_{10}(50000)$) that is the maximum value that the predictor can provide, which cannot be well modelled by a Gaussian distribution. The scatter plots of correlation coefficient values from MS data sets and random data sets of 85 HLA alleles were plotted across all possible combinations of the statistical models for MS and random data sets including mixture of Gaussian-Gaussian (GG), mixture of Gaussian-beta (GB), mixture of beta-Gaussian (BG), mixture of beta-beta (BB). The scatter plots demonstrated that the mixtures of GB and BB have R^2 and slope in range of 0.9 to 1, for most alleles, and the intercept from the mixtures of GB and BB were closer to 0 than the mixtures of GG and BG (Figure 3.7A). It suggested that the true data distribution (left peak) can properly fit both Gaussian and beta models while the beta distribution is the fittest model for false data distribution (right peak). To find statistical models that can properly model the observed bimodal distributions, the values of R^2 from beta and Gaussian model fitting of each HLA allele were compared by a pair t-test. The average R^2 from data sets of 85 HLA alleles from the beta model fitting true data distribution (0.95) was significantly higher than the Gaussian model (0.93) (p-value = 1.77E-0.7) (Figure 3.7B). Therefore, these results can indicate that the beta mixture is the most suitable model to fit the predicted scores of data containing a mixture of binding and non-binding peptides, as would be expected to be observed in practice.

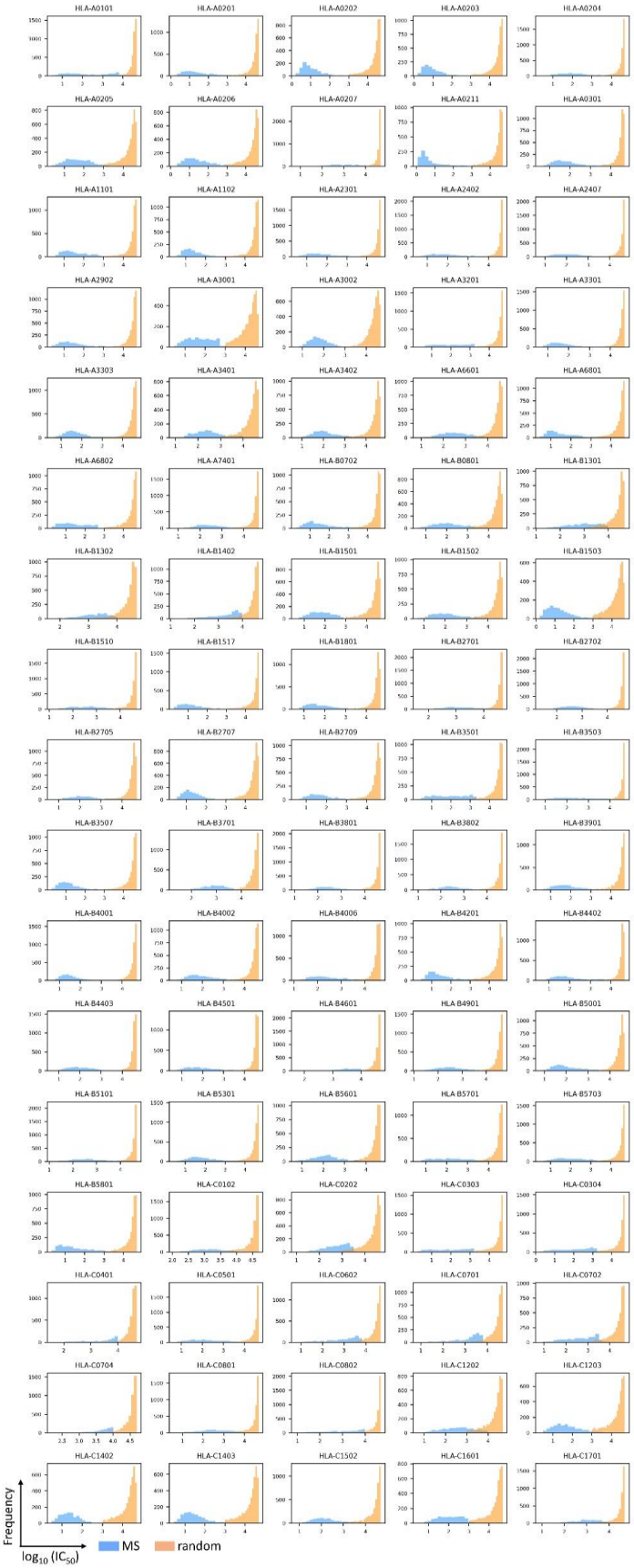

Figure 3.5 Data distribution of the predicted scores (binding affinity in $\log_{10}$ (IC_{50})) of MS and random peptides of 85 HLA alleles.

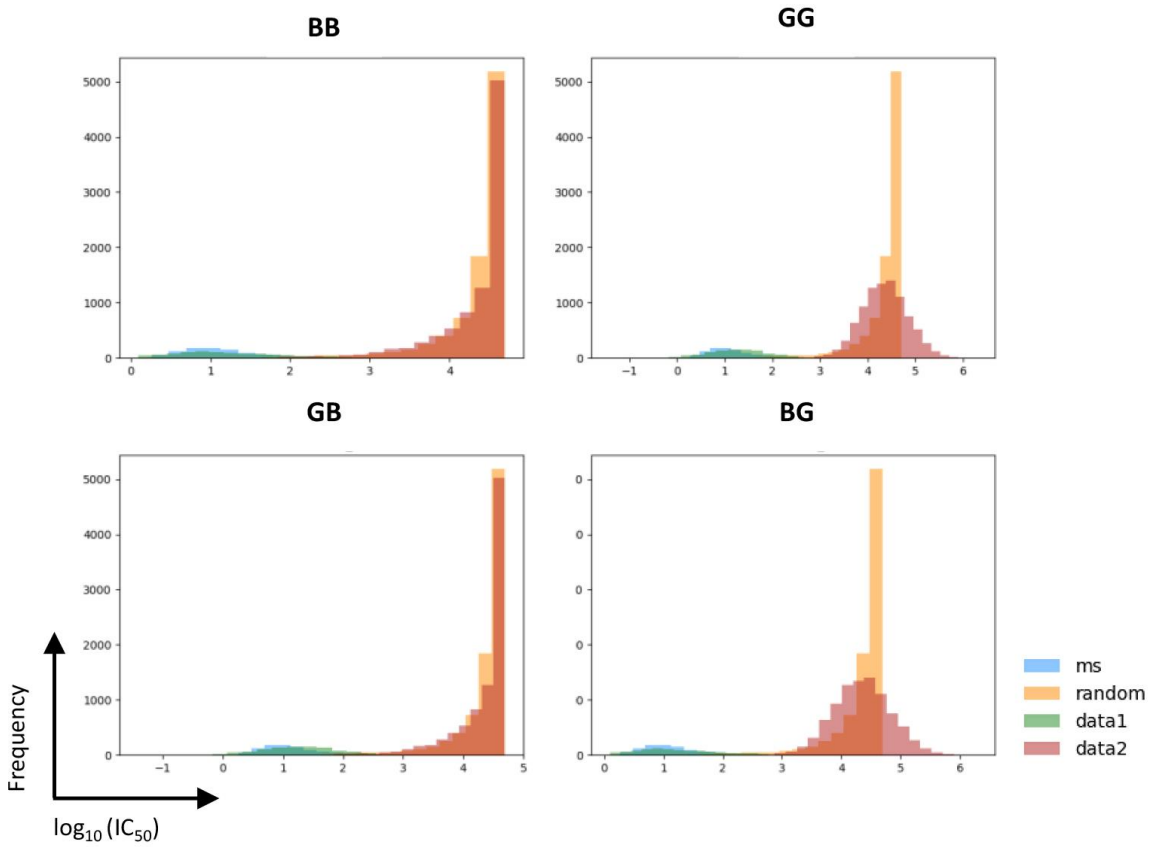

Figure 3.6 The overlaid of distribution between real data and generated data from different statistical distributions. The data sets were generated from the model of mixture of Gaussian-Gaussian (GG), mixture of Gaussian-beta (GB), mixture of beta-Gaussian (BG), mixture of beta-beta (BB).

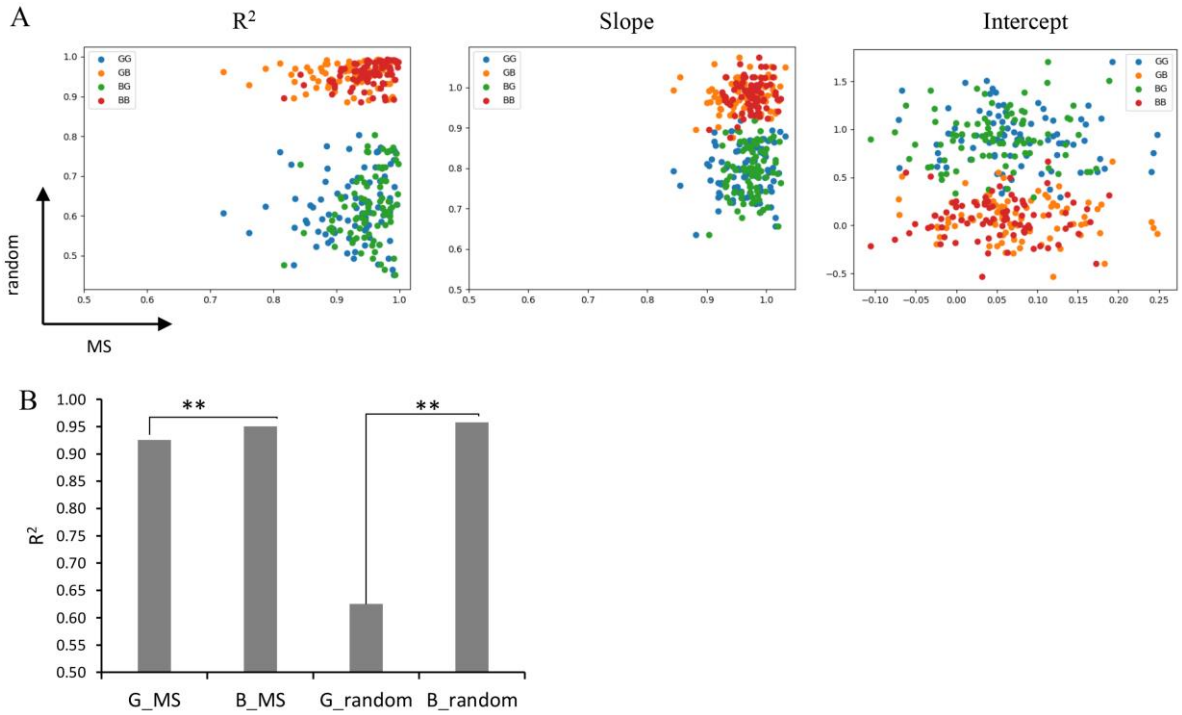

Figure 3.7 The mixture models fitting data distributions of 85 HLA alleles. (A) The scatter plots of correlation coefficient R^2, slope and intercept values from MS data sets (x axis) and random (y axis) data sets of 85 HLA alleles that were fitted by beta or Gaussian models; mixture of Gaussian-Gaussian (GG), mixture of Gaussian-beta (GB), mixture of beta-Gaussian (BG), mixture of beta-beta (BB). (B) The average of R^2 of Gaussian and beta model fitting MS and random data sets from 85 HLA alleles. Each bar represented the mean of R^2 from 85 HLA alleles. (**p-value < 0.01); Gaussian fitting MS data (G_MS), beta fitting MS data (B_MS), Gaussian fitting random data (G_random), beta fitting random data (B_random).

3.3.2 The development of parameter estimating model using the EM algorithm

3.3.2.1 Parameter estimation using the EM for beta mixture model

The true and false distributions were estimated by the beta parameters estimation model with the framework of the EM algorithm as described in Section 3.2.5. To observe the feasibility of the EM model, the predicted binding affinity scores of the mixture of 1000 MS and 4000 random peptides from 85 HLA alleles were then estimated. The overall distribution of the observed data in Figure 3.5 is captured well by a two-component beta mixture model, with the first component representing low IC_{50} values (true data) and the second component for high

IC_{50} values (false data). To estimate the sizes of true and false data from the predicted results, the parameters of beta mixture distribution including two mixture proportions (π_{true}, π_{false}), $\alpha_{true}, \alpha_{false}$, and $\beta_{true}, \beta_{false}$ were estimated from the predicted data sets for 85 HLA alleles using the EM algorithm with a method of moments estimation for the beta mixtures. As the data sets are scores of MHC-peptide binding affinity prediction, the real parameter shapes of α and β of the data were not exactly known, but the ratio of MS and random size was defined as 0.2 and 0.8, respectively. Therefore, the relative change between real and estimated values were computed to evaluate the correctness of estimated parameters. The bar graph of relative change values demonstrated very low relative changes for almost data sets, though, some HLA alleles data showed a high difference (a relative change ≥ 0.5) between real and estimated values, which are found in a few alleles in B locus e.g. B*13:01, B*14:02 and most in HLA-C e.g. C*04:01, C*07:02, C*12:02 (Figure 3.8). Taken together, the analysis from invented data sets and predicted scores from 85 HLA alleles indicated that feasibility of the current version of the EM for beta mixture model might not generally robust for any beta mixture distributions.

3.3.2.2 The EM model with constraining of false parameters

To improve the performance of the EM model to give more sensible estimation, it is considered that the estimate numbers of parameter shapes should be reasonable for data distribution of a specific allele. Since the second component of the data distribution is a set of scores of non-binding peptides, the distribution shape of any predicted scores of non-binding peptides with the same HLA allele should be similar. Therefore, the estimated beta parameters for the second component were then constrained by the range of values calculated from the data sets of various sizes (1000, 5000, and 10000) of predicted binding affinity scores from random peptides with a length of 8, 9, 10, and 11 mers against 85 HLA alleles. The purpose of this restriction is to ensure that one component of the beta mixture model is certainly captured as the false data. It should be noted that since some random peptides may be true binders, random peptides with

IC_{50} < 1000 nM (~2.5% of all generated random peptides on average) were exclude from consideration (Figure 3.9). The values of α_{false} and β_{false} from different data sets of each HLA allele were explored to test for a variation of beta model parameter ranges dependent on data set sizes and peptide lengths. The calculated values of α and β from random data sets have a small variation across different data sizes for most HLA alleles (Figure 3.10), from which it can be inferred that the calculated values of α and β can be utilisied to apply to any false data sets for the same specific HLA allele. The ranges of calculated values of α_{false} and β_{false} were used to constrain the estimated α_{false} and β_{false} in the MM-step The data sets of predicted scores from 1000 MS and 4000 random peptides with 9mers for 85 HLA alleles (same data sets in Section 3.3.2.1) were used to test the feasibility of the modified EM model. The relative change of real and estimated values from the non-constrained model was compared to the constrained model. The relative changes of several HLA alleles, especially data sets in HLA-C were dramatically reduced with using the constrained model (Figure 3.11) indicating that restriction of estimated values for false data in the sensible ranges can improve the performance of the EM model. The R^2 between the real and simulated data set for all 85 alleles are greater than 0.99 (Figure 12A). However, the R^2 can only describe a similarity of distribution shape but not scaling between two data sets. To ensure that simulated data can represent a given observed data, we also considered other values in the linear equation including the slope and intercept, and they are also close to 1 and 0, respectively (Figure 3.12B and 3.12C). Furthermore, the difference between two distributions of real and simulated data for 85 HLA alleles was tested by the KS analysis. The p-values from KS test are higher than 0.05 for almost all alleles indicating that distributions of real and simulated data are not significantly different, although there are few alleles that have p-value less than 0.05, which are A*01:01, C*04:01, and C*07:01 (Figure 3.12D). The overlaying of data distributions of each HLA allele between the predicted scores and simulated data set were shown in the Figure 3.13. In most HLA alleles,

the distribution of real and simulated data for both the left and right peak showed a good alignment (81 of 85 alleles, $R^2 \geq 0.995$). However, there are four data sets, that have right skew distribution of MS data including B*14:02, C*04:01, C*06:02, and C*07:02, displayed less good alignment between real MS data and their simulated data (those alleles have $R^2 < 0.995$). Nevertheless, the distribution of false data was well captured by the simulated data indicating that the ratio of false positive to true positive in that area of the MS data should be still correct.

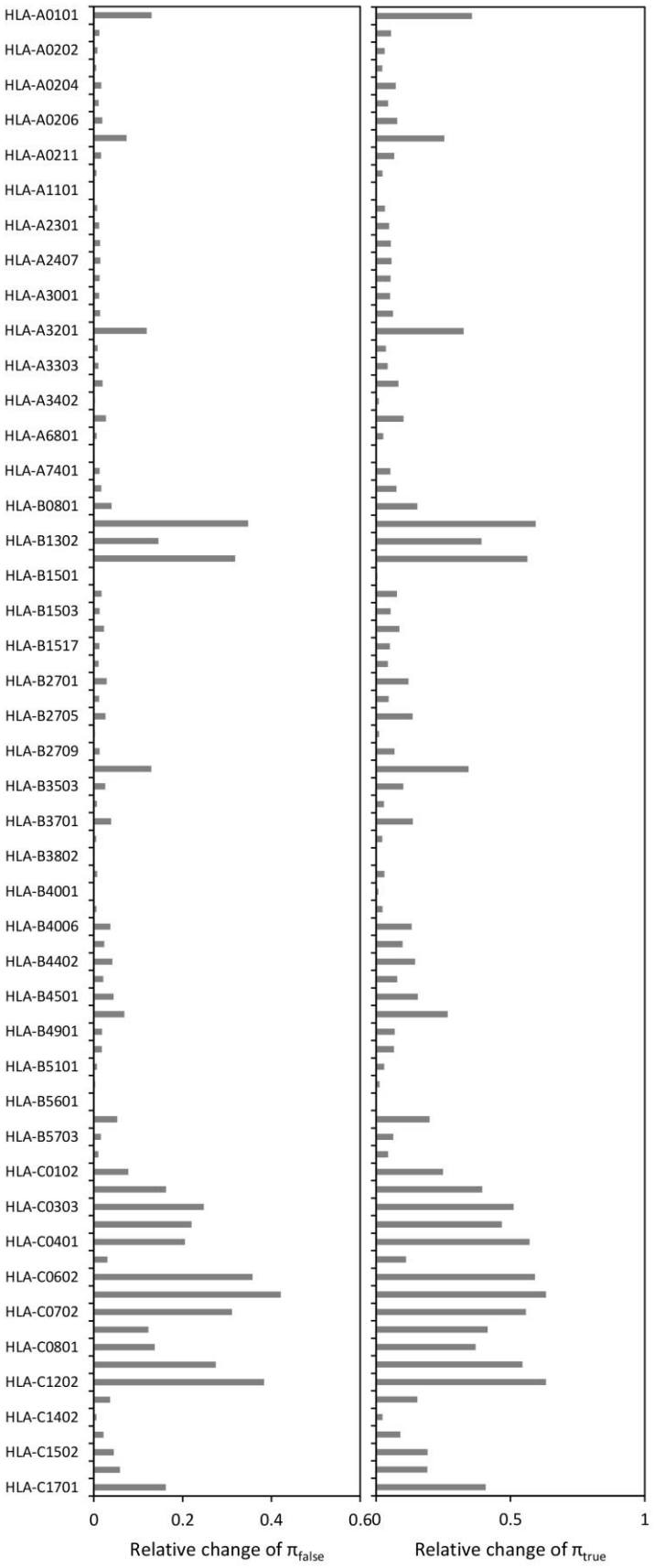

Figure 3.8 The relative change between designated and estimated proportion mixture of MS (π_{true}) and random (π_{false}) for data sets of 85 HLA alleles.

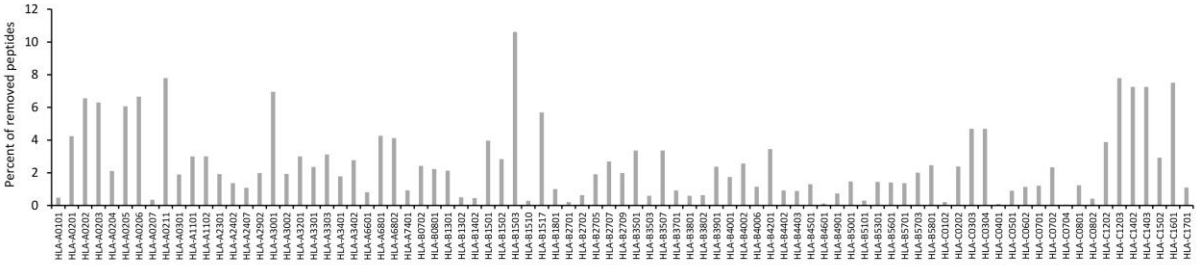

Figure 3.9 The percentage of removed data points (predicted IC_{50} < 1000 nM) from random peptides for 85 HLA alleles.

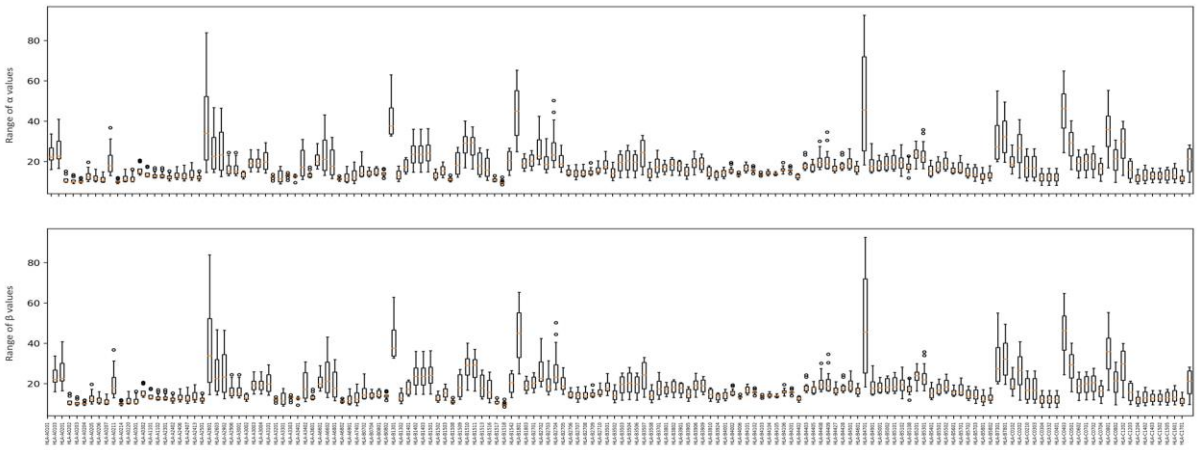

Figure 3.10 The box plots of calculated parameter shapes (α and β) of beta distribution. The two parameters were calculated from data sets with various sizes (1000, 5000, and 10,000 peptides) and lengths (8 to 11 mers) of predicted binding scores of random peptides against 85 HLA alleles.

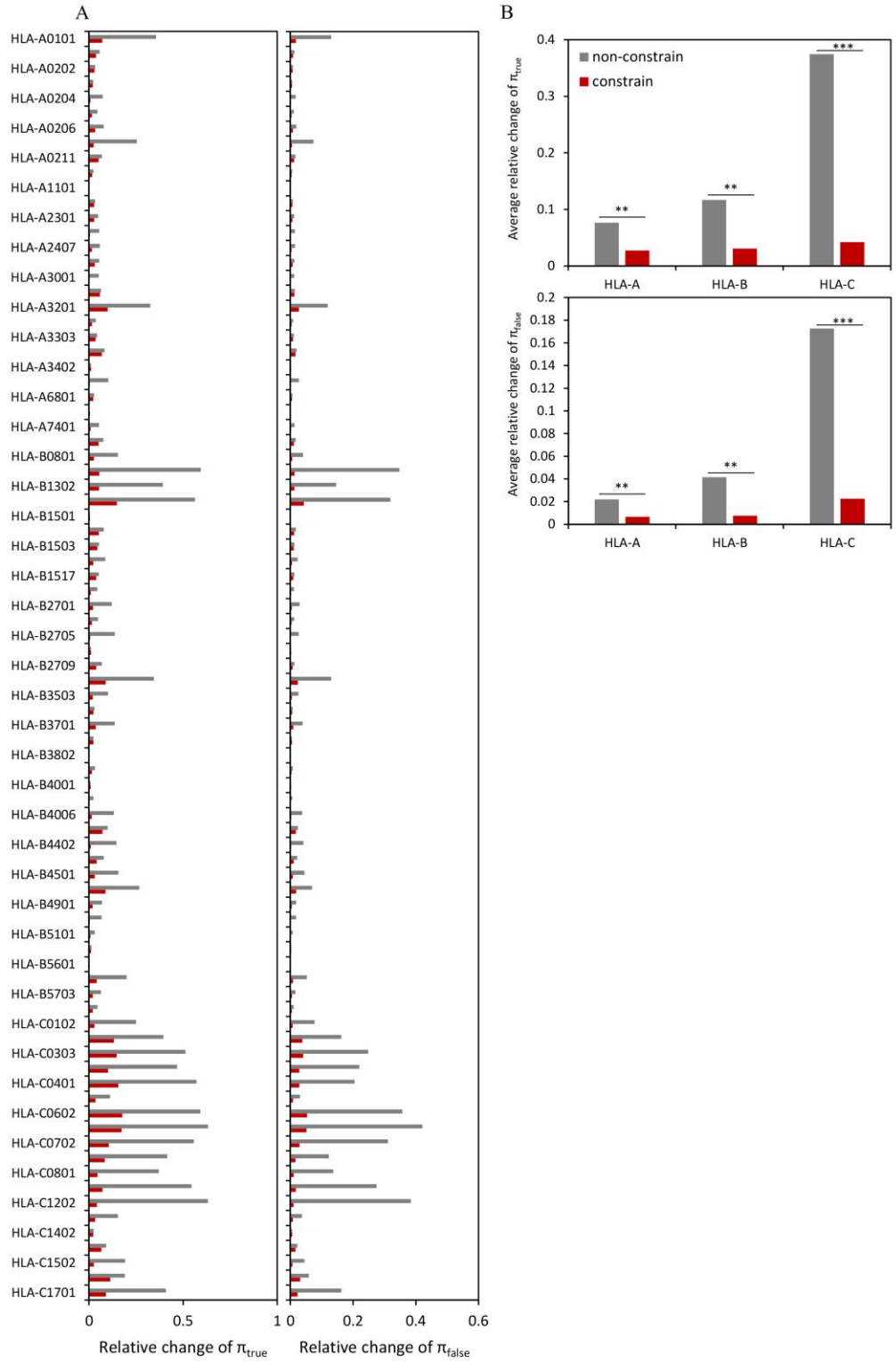

Figure 3.11 The performance of beta parameter estimation models with non-constrained and constrained estimated false parameters. (A) The relative change between designated and estimated proportion mixture of MS (π_{true}) and random (π_{false}) for data sets of 85 HLA alleles. (B) The average relative change of π_{true} and π_{false} from HLA-A, B and C (**p-value < 0.01, *** < 0.001).

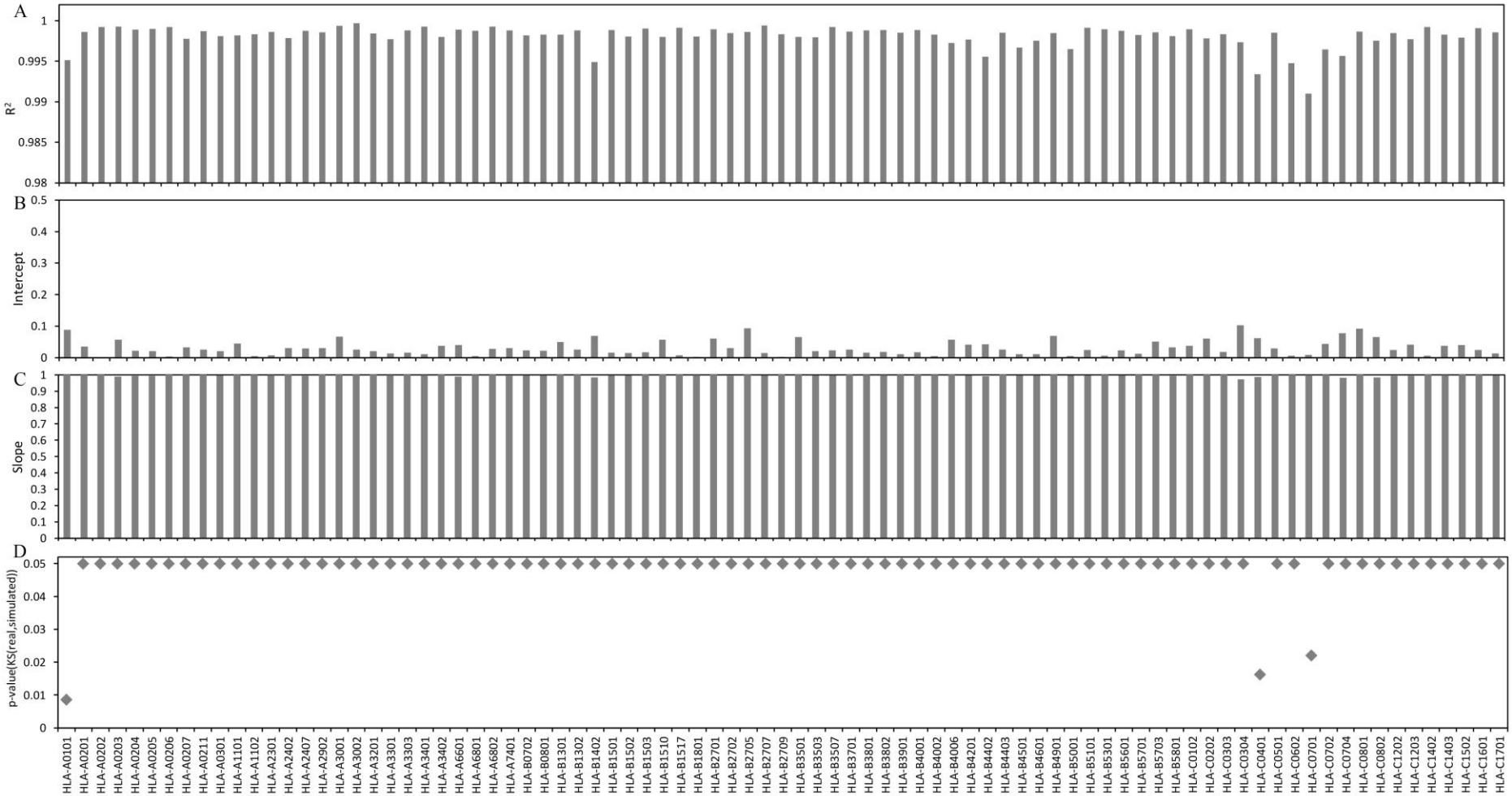

Figure 3.12 The analysis of the parameter estimation model for beta mixture testing with the predicted data sets of 9mers peptides of 85 HLA alleles. The correlation between the real and simulated data set calculated from the linear regression model, R^2 (A), Intercept (B), and Slope (C). The p-values from KS test between the real and simulated data (D).

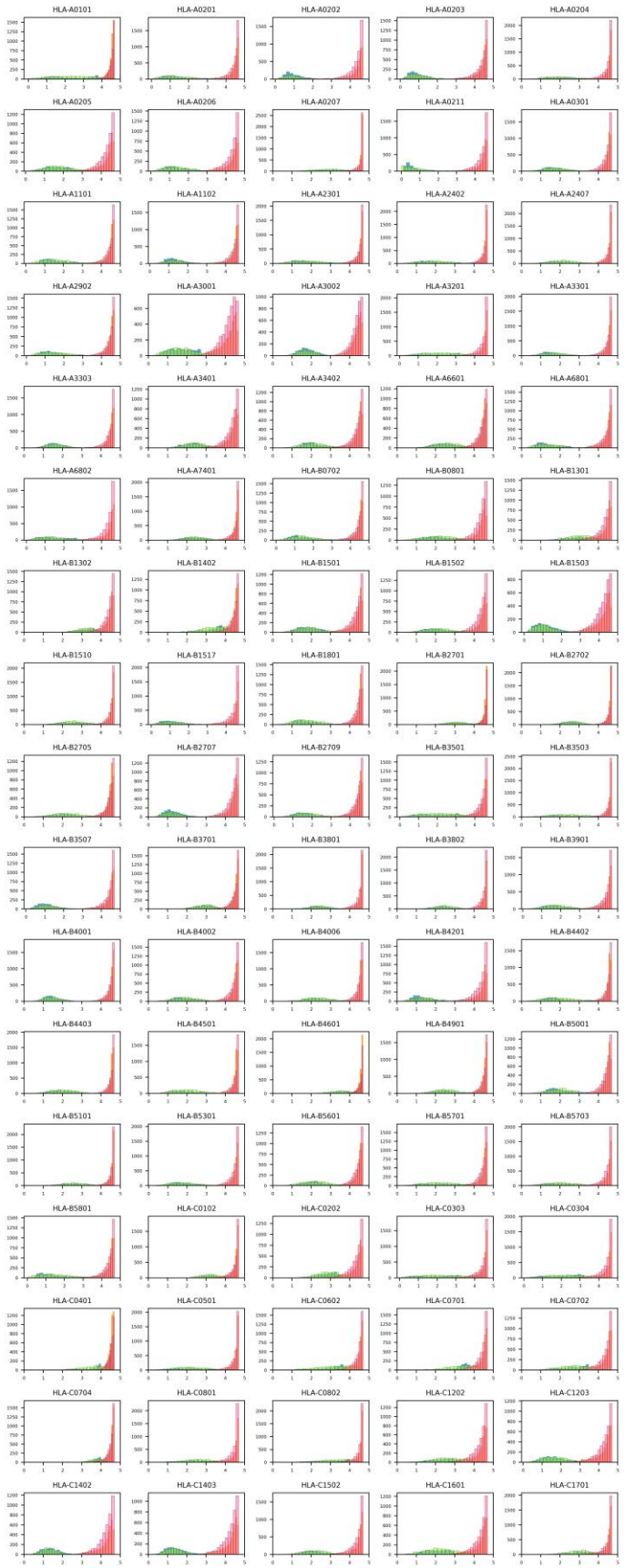

Figure 3.13 The overlaying of data distribution between the real (predicted IC_{50} scores) and simulated data sets.

3.3.2.3 Beta parameter estimation for massive imbalance data

To test the robustness of the estimator model with a large imbalance in true and false data, the final implementation of the estimator model was not only constrained with estimated parameters of the beta 2 component, but also restricted estimated parameters of the beta 1 component. However, the estimated beta parameters of true data are only restricted if the estimated $\pi_1 = 0$ and size of the negative set $\neq 0$ (predicted IC_{50} > 10000 nM) i.e. indicating that there is only one distribution found, and there are data points in the plausible range for false data. In this case, the ranges of α and β for the first beta component were initially calculated from data points with predicted $IC_{50} \leq 10000$ nM using Eq. 3.23 and 3.24, and the range of values are only allowed to deviate 25% from the initial estimates. In this analysis, the parameter estimation analysis was also performed with data sets with a larger imbalance ratio containing 1000 MS peptides and 8000 random peptides, and the result showed that the similarity between real and simulated data sets for 85 HLA alleles are close to 1 ($R^2 > 0.995$), and they are similar to those from 4000 random peptide (Figure 3.14). Furthermore, highly imbalanced distributions (i.e. almost all true, or almost all false), where selected MS data sets and random data sets were used to test with the model separately (Figure 3.15) and the predicted IC_{50} of peptides derived from MHC I multi-allelic cells (Figure 3.16). The result of similarity measure from those data sets revealed a high similarity between the real and simulated data indicating that the model can work well with data sets that are not in our sets of data used to learn and train the model and provide sensible estimated parameters for data distributions with a large imbalance between true and false data.

3.3.2.4 Beta parameter estimation for multi-lengths peptides

Since the common lengths of peptides for MHC class I are 8 to 11 mers, hence, the mixture of MS and random data sets with different peptide lengths were generated for more thoroughly

testing the performance of the constrained EM model. There are 16 HLA alleles, with MS peptides available for all lengths (8, 9, 10, and 11 mers), which were used to test the estimation performance of the model, 800 MS peptides (200 per length) and 3,200 random peptides (800 per length). It was found that the R^2 values of the real and simulated data sets for 16 HLA alleles are highly close to 1 (Figure 3.17A). The value of R^2 suggested that the parameter estimation model functions well for the data sets with multi-lengths of peptides, which are shown by a good alignment between the real and simulated data sets created by the estimated parameters (Figure 3.17B). Altogether, the results of the R^2 values and the overlaying of data distributions indicated that the framework of EM algorithm with a modified MM step for constraining estimated parameters can provide the sensible estimated parameters that can be further used for generating a data set for resembling the real predicted data set. The estimated true and false data of the predicted results can be further used to calculate the values of FDR and PEP for an individual predicted score.

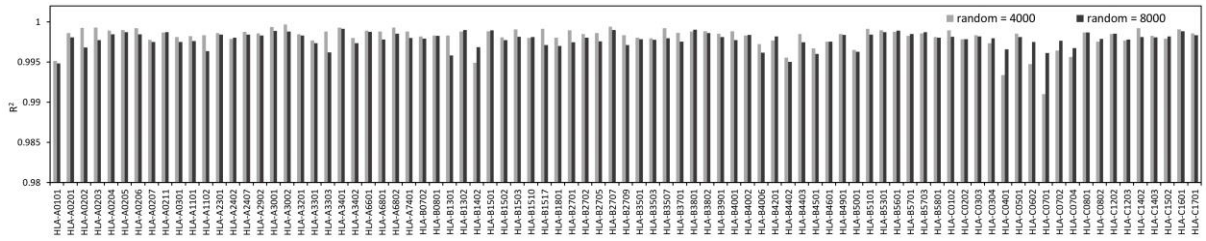

Figure 3.14 The R^2 between real and simulated data sets from data with 4000 and 8000 random peptides.

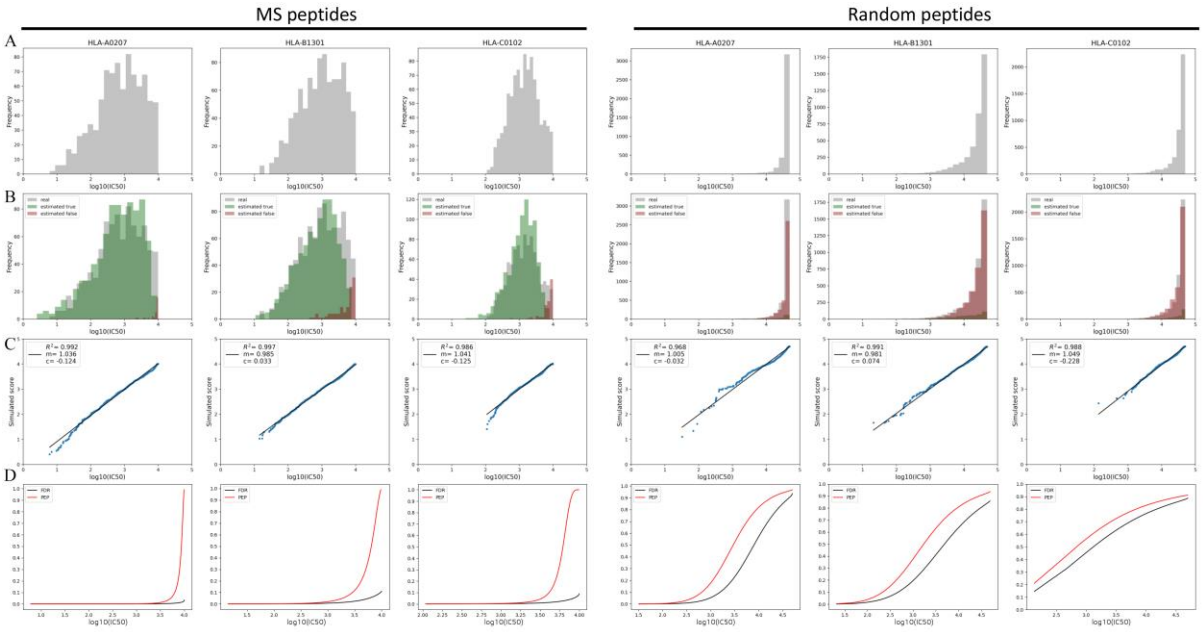

Figure 3.15 The estimation results from the beta mixture model on data set with a very large imbalance ratios between two components. (A) The data distribution of predicted binding affinity ($\log_{10}(IC_{50})$) of a specific HLA allele. (B) The beta models can fit to two components of data distribution to estimate parameters for true and false data. (C) The similarity measure from the linear regression model fitting correlation of the real and simulated data set, m = slope, c = y-intercept. (D) FDRs and PEPs calculated from estimated true and false data of each predicted IC_{50}.

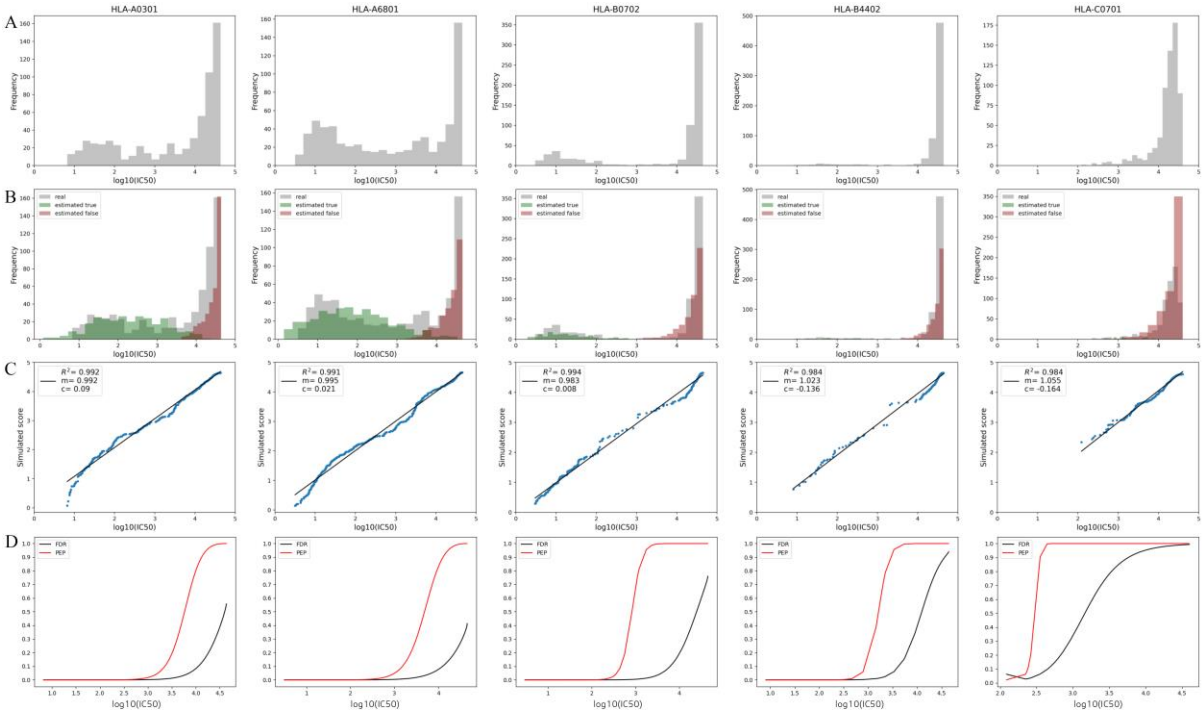

Figure 3.16 The estimation results from the beta mixture model on predicted IC_{50} of peptides from multi-allelic cells. (A) The data distribution of predicted binding affinity ($log_{10}(IC_{50})$) of each HLA allele expressed by multi-allelic cells. (B) The beta models can fit to two components of the data distribution to estimate parameters for true and false data. (C) The similarity measure from the linear regression model fitting correlation of the real and simulated data set, m = slope, c = y-intercept. (D) FDRs and PEPs calculated from estimated true and false data of each predicted IC50.

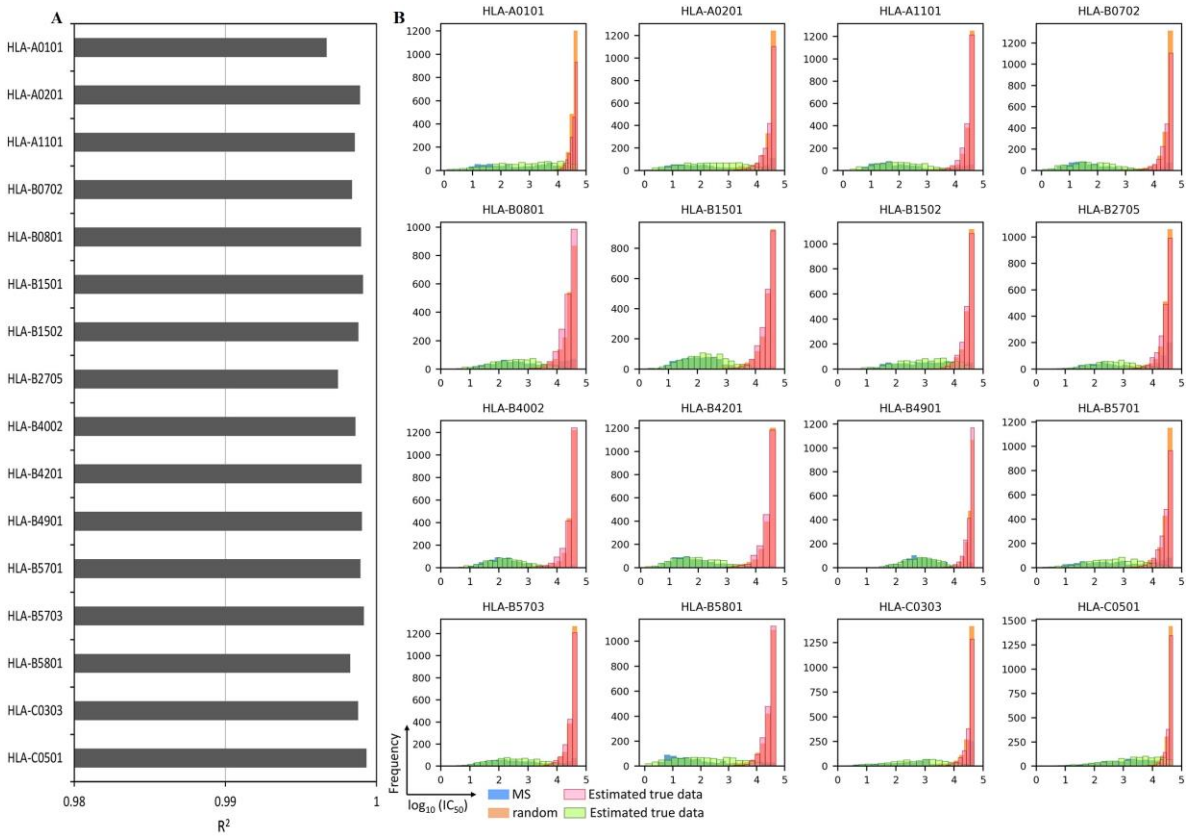

Figure 3.17 The performance of parameter estimation model for beta mixture testing with data sets with multi-lengths peptides (8-11 mers). (A) The R^2 between the real and the simulated data sets. (B) The overlaying of data distributions between the real and simulated data sets.

3.3.3. The estimation of FDR and PEP from simulated data sets generated by estimate parameters for the predicted scores

The values of FDR and PEP of an individual predicted score were calculated from the estimated parameters by beta distribution functions. The FDR was estimated using a CDF (Eq. 3.32), while the PEP was computed based on a PDF (Eq. 3.33) given by estimated beta parameters derived from the EM model, which are $\alpha_{true}, \beta_{true}, \alpha_{flase}$, and β_{false}. From NetMHCpan's documentation, the 2% rank is recommended to use as a threshold for binding peptide selection. The % rank scores were estimated from number of random peptides that have IC_{50} scores located in the range of predicted scores of a set of naturally presented MHC ligands. Hence, the role of % rank score is assumed to estimate the FPR of true predicted data. Here, this analysis performed the estimation of statistical confidence measure of FDR and PEP for peptide

binding prediction from the test data sets of a mixture of MS and random peptides. The results in Figure 3.18 showed that the accumulated FDR value at the 2% rank score of most alleles are less than 0.1, but about 25% of the representative data sets (21 of 85 alleles) have the FDR at the 2% rank score reach up to 0.26 e.g. B*15:10, C*01:02, and C*07:04, i.e. 26% of peptides passing the threshold are predicted to be false positives (Figure 3.18A). At the 2% rank threshold, the FDR of HLA-C (0.13) is the highest on average followed by HLA-B (0.07), and the average of HLA-A (0.03) is the lowest (Figure 3.19).

To assess the confidence of each peptide's predicted score, the PEP was computed for each peptide in the data set. The analysis demonstrated that 48 of 85 data sets have the PEP at the 2% rank over 0.5, i.e. peptides close the threshold have only a 50% chance of being a true positive (Figure 3.18B). Moreover, the PEP at the 2% rank of HLA-B and HLA-C on average are greater than 0.5 (0.64 and 0.63, respectively) whilst the average PEP of HLA-A is 0.38 (Figure 3.19).The overlaying of PEP values on the data distribution of predicted IC_{50} scores from 85 HLA alleles in Figure 3.18C showed that the $\log_{10}(IC_{50}) < 2$ or > 4 have a high certainty for being true or false binding peptides, their PEP values close to 0 and 1. For the scores in the range of 2 to 4 have less certainty to determine whether they should be true or false binding peptides, especially for less well separated data sets of some alleles. Several data sets have PEP values close to 1 for peptides with the % rank ~ 2% e.g. A*02:07, A*29:02, B*15:10, B*27:02, C*08:01, and C*14:03, in contrast, some data have very low PEP values, even if those scores have the % rank $\geq$ 2%, scores are determined as non-binding peptides, e.g. A*02:11, B*15:17, B*27:05. These results suggest that PEP values provide considerable added value over the use of the % rank for estimation of confidence in an individual data point. Beyond the MS:random data sets generated for 85 alleles data, the FDRs and PEPs from data containing almost all true or all false data (Figure 3.15D) and multi-allelic data (Figure 3.16D) were calculated from estimated beta parameters using Eq. 3.32 and 3.33. The results

demonstrated that the values of FDR and PEP correspond well with expected true and false data, giving confidence that the model will perform well when presented with genuine data sets.

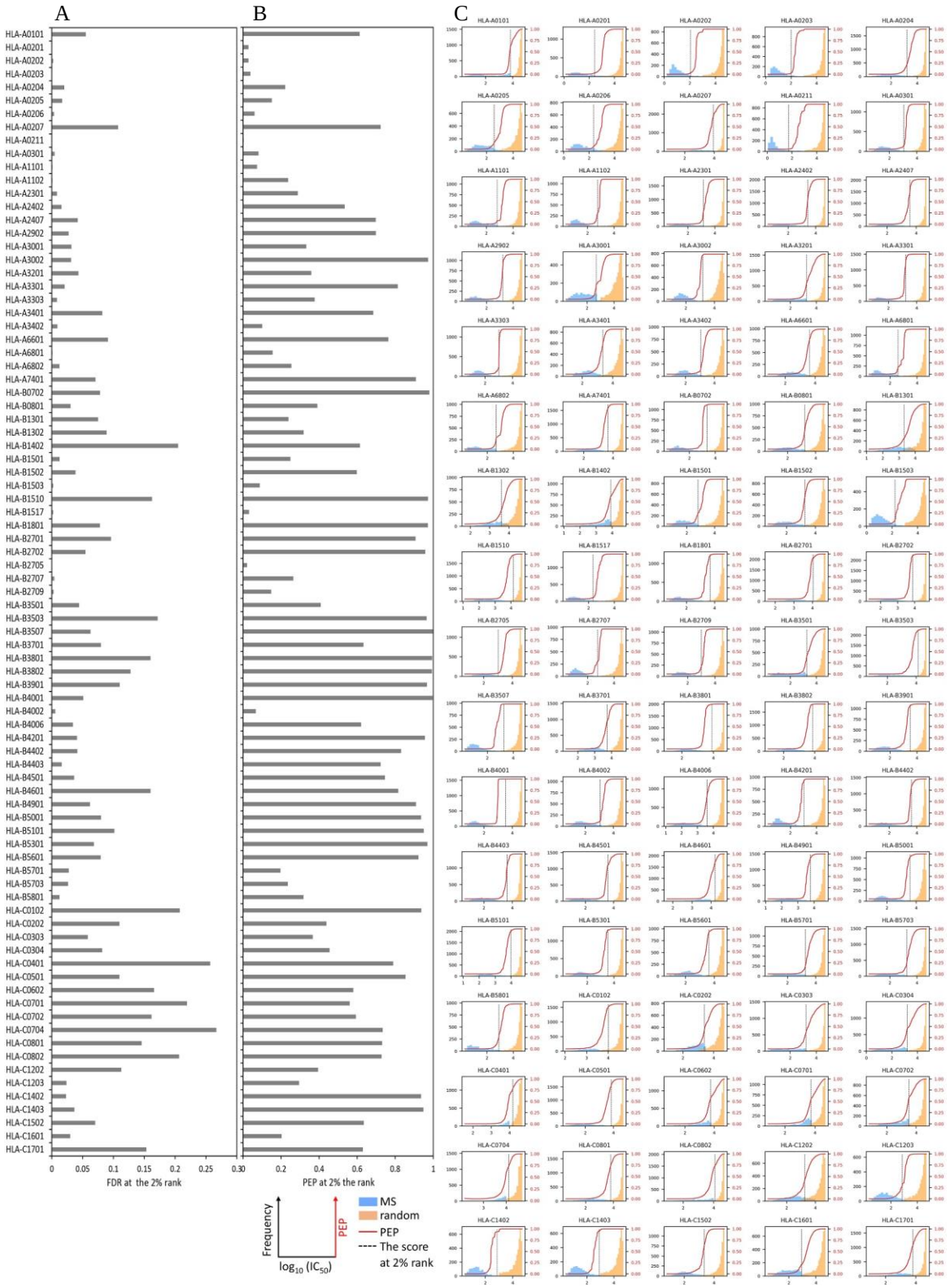

Figure 3.18 Estimation of FDR and PEP for predicted scores of 85 HLA alleles. The values of accumulated global FDR (A) and PEP (B) at the 2% rank. (C) The overlaying of PEP values on the data distribution of predicted scores of 85 HLA alleles, the dashed black line was marked at the score with 2% rank.

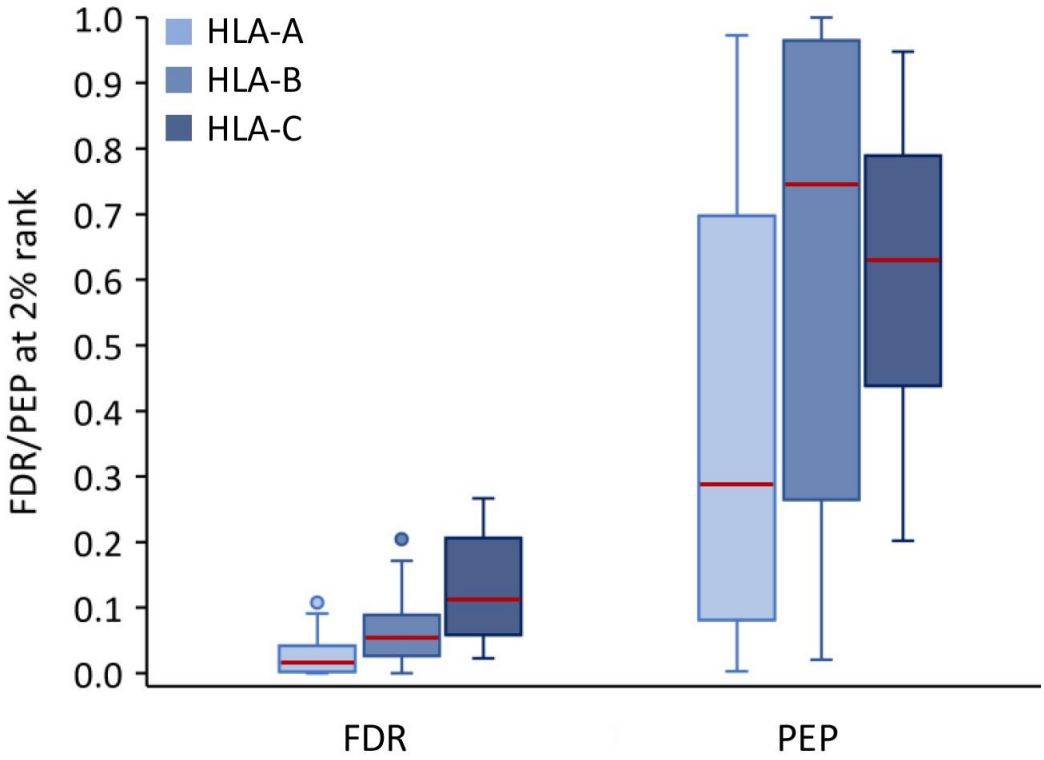

Figure 3.19 FDR and PEP at the 2% rank score of HLA-A, HLA-B, and HLA-C. The red line in the box represents median of FDR or PEP.

3.3.4 Extensibility for MHCflurry prediction

The previous analysis has been primarily tested with NetMHCpan, though, past benchmarking results suggest that MHCflurry gives similarly strong performance for peptide binding prediction, the approach in this study was thus extended for predicted results coming from MHCflurry2.0. MHCflurry also reports predicted IC_{50} and % rank, however, the MHCflurry's documentation does not suggest the cut off threshold of the % rank. Therefore, the 2% rank was assumed as a possible threshold for distinguishing binders and non-binders, as for NetMHCpan. There are 79 HLA alleles supported by MHCflurry, and there are 55 alleles of 9mers MS-random peptides in this study, which are available for those supported alleles. To estimate parameters from data distributions from MHCflurry prediction results, the parameter ranges that are calculated from MHCflurry predicted scores of random peptides in various data sizes (1000, 5000, and 10000) were applied to constrain the EM model instead of parameters

ranges calculated from NetMHCpan predicted scores. The R^2 between the real (predicted IC_{50} scores) and simulated data set ranged from 0.995 to 0.999 for most alleles (Figure 3.20A), and the overlay between the real and simulated data is shown in Figure 3.21. Thus, those results indicate that the approach of EM algorithm with method of moments also works well for predicted data coming from MHCflurry. The analysis of FDR and PEP estimation showed that if using a 2% rank threshold, over 10% global FDR occurs for 18 alleles, and PEP is higher than 50% for 27 of 55 alleles – indicating that as for MHCflurry, 2% rank is not an ideal threshold for controlling FDR for many alleles (Figure 3.20B and 3.20C).

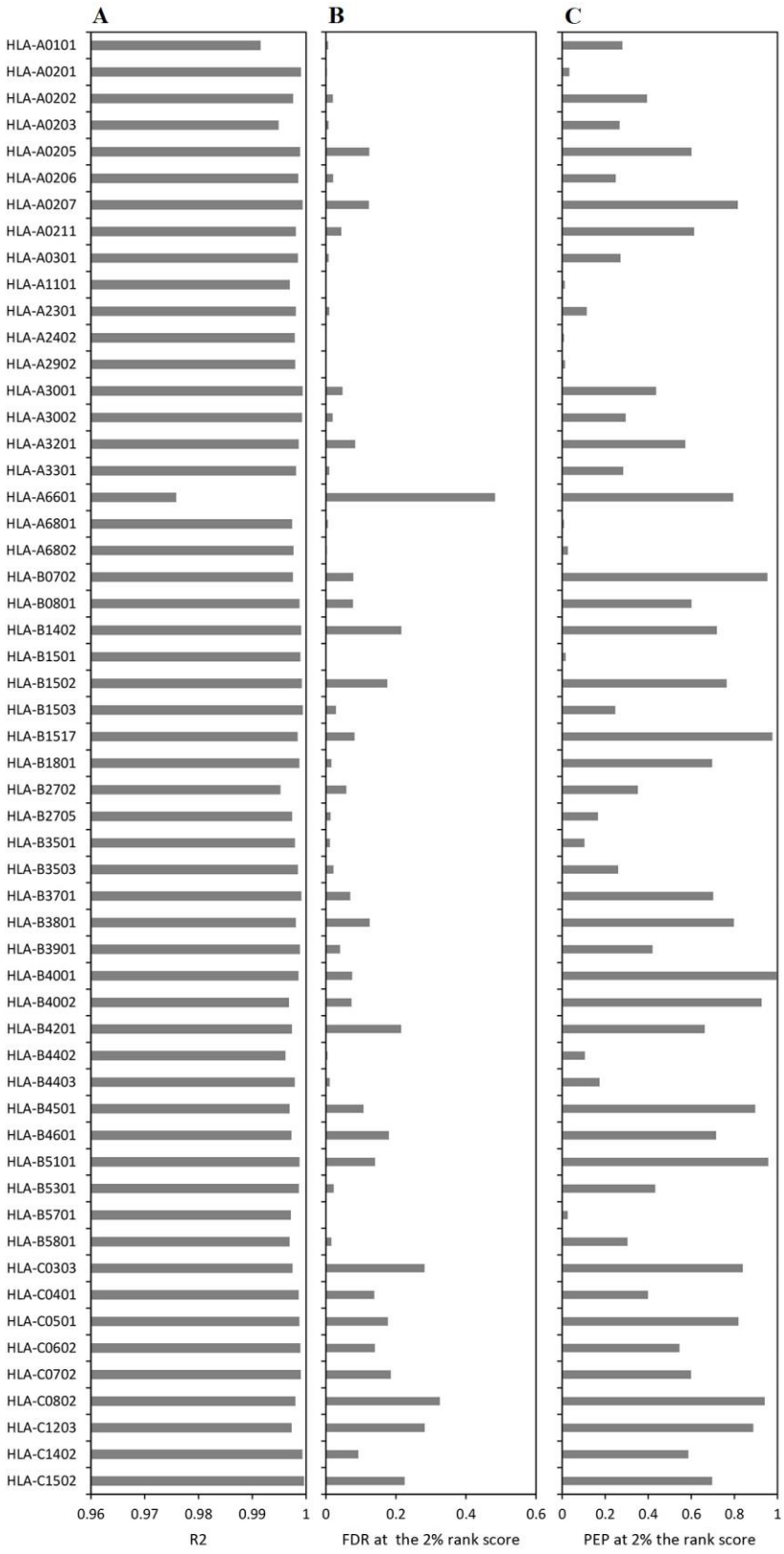

Figure 3.20 The analysis of the model with predicted results from MHCflurry. (A) The R^2 between the real and simulated data sets. The values of estimated FDR (B) and PEP (C) for predicted scores of 55 HLA alleles at the 2% rank score.

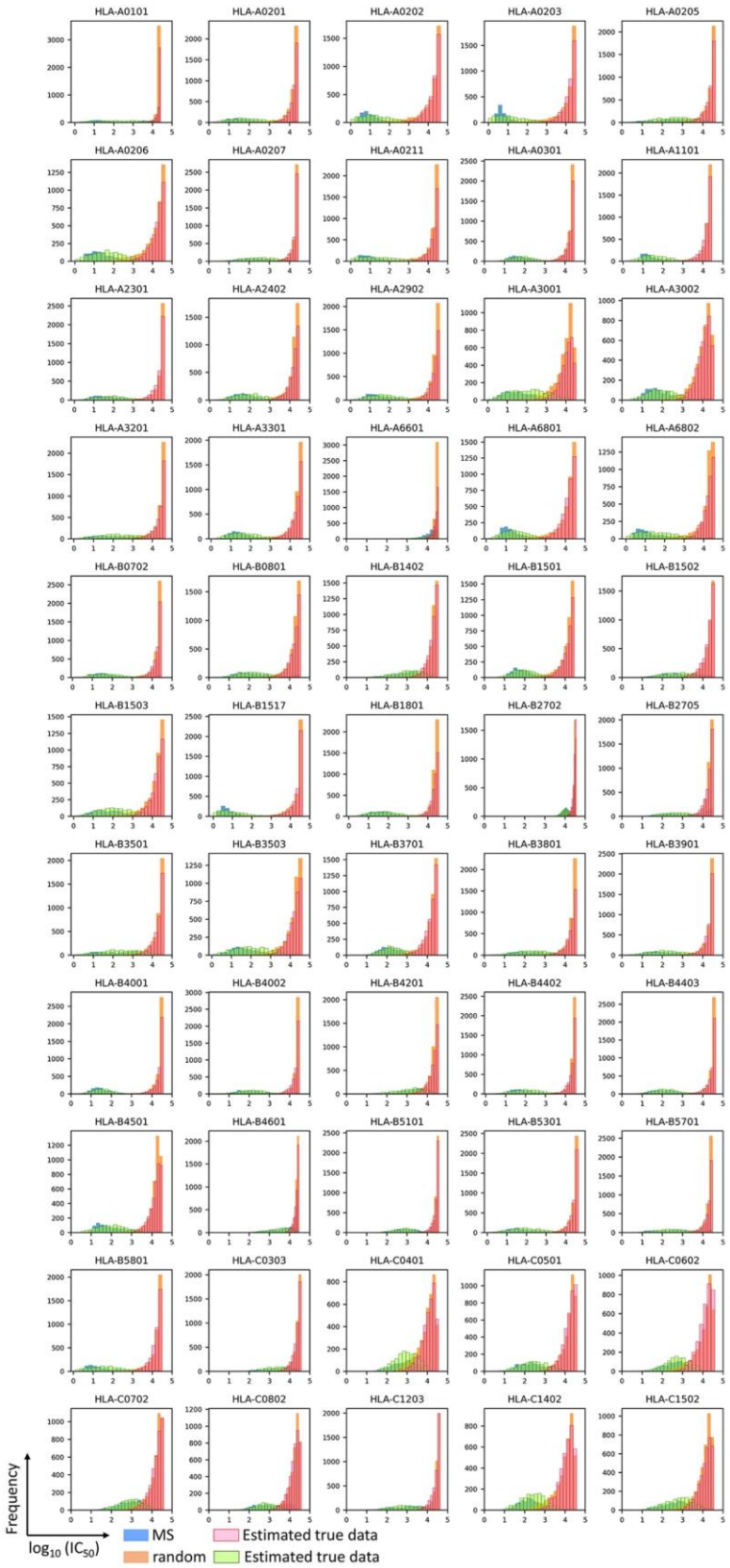

Figure 3.21 The overlaying of data distributions between the predicted scores from MHCflurry and their simulated data sets.

3.4 Discussion

MHC-peptide binding affinity prediction is widely used in the immunology research e.g. designing immunogenic peptides for vaccine development as shown in the previous chapter. NetMHCpan (from version 4.0 onward) produces a % rank score for each peptide predicted, estimated as the rank position of a given score within a list of scores from a set of 125,000 of 8-12 mers random natural peptides (25,000 of each length), assumed to represent the distribution of false results (non-binders). Rank scores < 2% is commonly used to distinguish binding and non-binding peptides because it reduces the known bias of binding preference across different MHC molecules [106]. If random peptides are assumed as false results, thus, the predicted % rank for each predicted IC_{50} score is approximated as the false positive rate. However, using only the FPR might not be sufficient to quantitatively evaluate whether a predicted binding score for a peptide is a true or false positive, hence, the statistical measurement that can control the false positive rate might help to increase the accuracy for binding peptide selection. In this chapter, the model to estimate statistical values including FDR and PEP of an individual predicted score was developed.

The distribution of predicted binding affinity scores (IC_{50}) coming from NetMHCpan4.1 was firstly observed. The distribution of predicted scores of MS peptides from multi-allelic cells to an MHC molecule displayed a bimodal distribution that contains two separated peaks. From a bimodal distribution, it can infer that the left peak contains peptides that will truly bind to an HLA allele (true positives). The right peak contains non-binding peptides to a given HLA allele. The distribution of the mixture of MS-random peptides for 85 HLA alleles in Figure 3.5 also demonstrated a bimodal shape with clear separation between MS and random distribution. Although, there are a small number of alleles that do not follow the expected distribution shape. First, for some alleles e.g. A*34:01, B*15:02, C*12:02, a small number of MS peptides overlap to the random peptide distribution suggesting that they could be incorrect identifications of

peptide sequences from MS data. Second, almost alleles displayed symmetrical shape for binding peptides and right skewed distribution for random peptides, however, there are some alleles, mostly HLA-C, where their distribution of MS peptides showed asymmetrical shapes and do not have particularly clear separation of assumed "true" and "false" positive distributions e.g. B*14:02, C*04:01, and C*07:02. From the overall inspection of 85 HLA alleles and the analysis of statistical models fitting data distribution, the beta mixture distribution was finally selected to model a bimodal distribution of predicted scores from the mixture of binding and non-binding peptides. The usage of beta model fitting MHC I predicted scores is agreed by the study of Zeng's group that they used beta distribution to model the data distribution of MHC-peptide binding affinity for MHC class I [179]. Moreover, the beta model is the most flexible distribution shapes depending on different combinations of the parameters of α and β [170]. As the beta mixture was used to model the data distribution of predicted scores, thus, the EM algorithm with a method of moments was implemented for estimating parameters of a beta mixture distribution.

The study of parameter estimation using the EM algorithm with method of moments for beta mixture model has been previously reported for the application in the field of molecular biology [170]. The performance of the non-constrained model for the predicted scores of 85 alleles (Figure 3.8) is not accurate for some data sets, specifically, those data sets have been described as having unusual distributions in Figure 3.5 e.g. B*14:02, C*04:01, and C*07:02. These data sets do not have markedly clear separation of presumed true and false distributions indicating that the parameters from indistinct separate data is not well estimated by the typical EM algorithm. Thus, it is important to improve the model for unclear separate data because the predicted scores in the overlapped area are very uncertain if they should belong to true or false data. The EM algorithm was then modified by constraining the estimated parameters of false distribution with the ranges of α_2 and β_2 calculated from predicted scores of random data with

different sizes and peptide lengths for 85 alleles. The restriction causes the false data to be well captured that might consequently forces the true data to be correctly modelled. That assumption is supported by the results in Figure 3.11 demonstrating that the constrained model has obvious improvement for the unclear separate data sets that are not correctly estimated by the non-constrained model. In practice, these two constraints mean that when the algorithm detects evidence a very large imbalance, in either direction (i.e. all true or all false), the beta 1 or beta 2 is correctly fitted to the appropriate distribution. Since this developed model was built by relied on the predicted scores from NetMHCpan4.1, thus the application of this model is available for MHC types supported in NetMHCpan4.1, which cover for 2,915 alleles for HLA-A, -B, and -C [67].

The global FDR can describe the error rate that accumulates in the selected binding peptides from the prediction across the whole data set, while PEP values can describe a local false probability of an individual peptide in the data set. The results demonstrated that some data sets might get over 10% FDR when using the 2% rank as a threshold, which might be too high risk to control false positives (Figure 3.18A). In practice, the FDR observed is dependent upon the allele selected, as well as the actual (unknown) count of true positives in the data, relative to false positives. Moreover, there is variability in PEP values close to the 2% rank score. In some data sets the predicted scores $\leq$ 2% rank can have PEP values very close to 1, but in other data sets the predicted scores $\geq$ 2% rank have a PEP less than 0.1. Furthermore, the analysis of predicted results for 55 alleles coming from MHCflurry discovered similar trends as for NetMHCpan. This finding indicates that using only the predicted % rank for thresholding might wrongly accept false binding peptides or miss some true binding peptides in different cases, which cannot normally be differentiated straightforwardly. The final implementation of parameter estimation model and FDR/PEP calculation were built by a Python script, the software was named MHCVision, available at https://github.com/PGB-LIV/MHCVision. In

brief, MHCVision performs parameter estimation using the EM framework for a two-component beta mixture model, representing the distribution of true and false scores of the predicted data set. The estimated parameters are further used to calculate FDR/PEP of an individual peptide's predicted score. The input requires a column of predicted IC_{50} scores, and the output will return the estimate statistical values including FDR and true posterior probability (a converse PEP, 1-PEP) for every predicted peptide in each data set for a specific HLA allele. Moreover, the approach of the model was also extended to MHCflurry because the performance of this tool has been reported as good as NetMHCpan [68], users can opt to run with MHCflurry, the supported alleles are limited to 79. Finally, for different downstream uses of peptide binding data, rather using solely the fixed threshold as the predicted % rank to classify or prioritise binding peptides, this study would recommend using MHCVision for calculation of FDR and PEP and selection of appropriate threshold to reduce a risk of getting false positive and gain confidence for those peptides that their scores might be determined as non-binders via 2% rank threshold.

3.5 Conclusions

This chapter reports on the successful development of a parameter estimation model for beta mixtures for predicted binding affinity scores, and tested with data from 85 HLA alleles. The statistical values including FDR and PEP of an individual predicted score can be computed from the best estimated parameters derived from the beta parameter estimation model. The converse PEP value, true probability, of an individual predicted score is promising for prioritisation of peptides. The software was implemented and deposited at https://github.com/PGB-LIV/MHCVision.

Chapter 4

The development of an immunogenicity prediction model for distinguishing immunogenic and non-immunogenic peptides using Random Forest

4.1 Introduction

The model from the previous chapter can estimate global and local FDR for predicted MHC-peptide binging affinity scores. Using FDRs (either local or global, depending on the context) as one of criteria to select binding peptides would help to avoid selecting false binding peptides, and thus reduce the risk for getting non-functional neoantigen because non-binding peptides cannot be epitopes. However, as the context described in Section 1.6, all epitopes must be MHC binding peptides, but some MHC presented peptides can be non-immunogenic peptides i.e. those that do not generate an immune response. Therefore, to increase the chances to obtain genuine neoantigens, prediction of immunogenicity is also required. The process of antigen processing and MHC presentation allows T cells to detect antigens presented by MHC molecules. The interaction of TCR and antigen involves a strong binding between TCR, MHC molecule and presented peptide. Due to sophisticated steps for T cell recognition and extremely high variety of T cell receptors, resulting in enormous variety of preference patterns of TCR-peptide binding (Section 1.1.1), the characterisation of the specificity for TCR-peptide interaction using prediction algorithms is very challenging. The recent immunogenicity prediction approaches consider the peptide sequence as the starting point, because TCR-epitope interaction is governed by physicochemical principles like other protein-protein interactions, and the concept of "foreignness" since a host's T cells will be stimulated by non-self-antigens. In this chapter, a machine learning model for immunogenicity prediction was developed. The model was built using the Random Forest (RF) algorithm and aimed to classify peptides to two categories, those that are immunogenic and non-immunogenic peptides. The training data were collected from a data set of MHC class I presented peptides including immunogenic and non-immunogenic peptides derived from previously published immunogenicity experiments. A set of features related to physicochemical properties and divergence from the nearest human homolog, were exploited to create the classifier model, reported in Section 4.3.1. In Section

4.3.2, the benchmarking analysis was performed, and predicted probability scores obtained from the classifier model were studied and calibrated to real probability scores relying on the probability density of the data distribution, described in Section 4.3.3. For the final part, the Random Forest model was further investigated to understand how the model makes decisions, summarised in Section 4.3.4.

4.2 Methods

4.2.1 Data collection

This aim of this chapter was to classify immunogenic and non-immunogenic peptides on MHC class I molecules. These peptides were obtained from 'tcell_full_v3.csv' downloaded from IEDB [138] with the following inclusion criteria: linear epitope, 9mers in length, MHC class I, non-human parent peptides, and any host species (Table 4.1). In this study, the determination of true neoantigens of human cancer is emphasised, thus, only non-human parent peptides, as shown in Table 4.1, were retained for training data to avoid the bias from matching self-proteins obtained from blasting peptides against to human proteome. Epitope and non-epitope peptides were labelled as "positive" and "negative", respectively.

Table 4.1 Summary of parent species, host species, and experimental assays of 9mers peptides specific MHC class I collected from IEDB

	Positive (n = 2127)	Negative (n=5042)
Parent Species		
Homo sapiens	430	326
Non-Homo sapiens	1697	4716
Host Name		
Homo sapiens	1514	1708
Mus musculus	563	3497
Pan troglodytes	12	12
Sus scrofa	13	7
Equus caballus	5	0
Macaca mulatta	15	139
Gallus gallus	2	14
Rattus norvegicus	2	25
Oryctolagus cuniculus	1	0
Assay group		
qualitative binding	589	219
cytotoxicity	386	423
IFN-γ release	1058	4704
proliferation	23	35
dissociation constant KD	20	0
granzyme B release	27	1
TNF release	2	3
activation	17	12
pathogen burden after challenge	2	2
T cell help	1	0
CCL4/MIP-1b release	2	0
disease exacerbation	0	1
degranulation	0	2

4.2.2 Generation of data sets with matching binding affinity scores

We first analysed the MHC binding affinity of peptides within the training set, *a priori* classified as immunogenic and non-immunogenic. Data sets containing immunogenic peptides are biased towards containing peptides that are also strong MHC binders (Figure 4.1, the top panel) i.e. median $\log_{10} IC_{50}$ = 1.93 (positive) vs 2.67 (negative). Given that there are already reliable predictors for MHC-peptide binding, and statistics developed via MHCVision (covered

in Chapter 3), in this work we aimed to train a peptide immunogenicity model that is statistically independent of whether a peptide is predicted to be bound by MHC. As such, to prevent the bias being introduced by features that predict binding and non-binding properties, the data distributions of predicted MHC-peptide binding affinity between epitopes and non-epitopes data were divided into 20 bins, and the distribution of both data sets was standardised for every bin by sub-sampling predicted binding affinity scores within the same range for both epitopes and non-epitopes data (Figure 4.1, the bottom panel). The final data for training and testing consisted of 1,146 immunogenic peptides and 1,356 non-immunogenic peptides, with near identical distributions of peptide binding affinity, as predicted by NetMHCpan. These peptide sets are biased towards those with high-binding affinity e.g. ~70% of both sets (71% for positive and 72% for negative) have IC_{50} < 500 nM ($\log_{10} IC_{50}$ < 2.7), indicative of being strong binders, but ~30% of the data are relatively weak binders or non-binders. The rationale for this approach is to learn features that will be useful for determining immunogenicity, unrelated to peptide affinity and MHC binding, working under the assumption only peptides will be tested for immunogenicity if they are a reasonable peptide binding affinity from another tool.

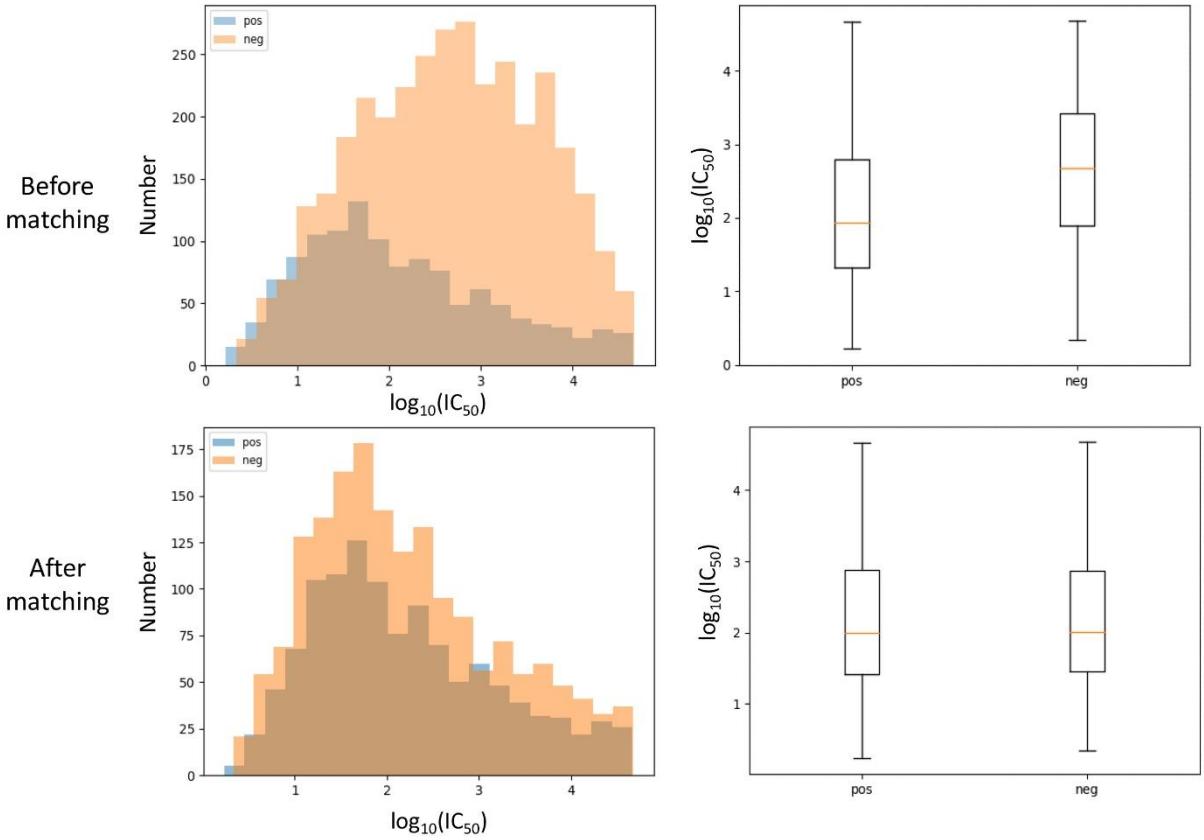

Figure 4.1 Matching data distributions of predicted MHC-peptide binding affinity of immunogenic and non-immunogenic peptides. The upper panel shows histogram (left) and boxplot (right) for $\log_{10}$ IC_{50} values derived from NetMHCpan for all peptides in the set. The bottom panel shows the same after sub-sampling from score bins to match score distributions between the positive and negative training set.

4.2.3 Construction of physicochemical properties for immunogenicity features

The physicochemical properties were selected based on the properties have been studied in immunodominant peptides and reported by the previous studies. Those properties include molecular weight, bulkiness, entropy, hydrophobicity, polarity and charge, and other properties related to binding interaction e.g. side chain orientation, bonded energy per residue [117, 146]. The scores of each property for 20 amino acids were obtained from the AAindex database [180]. The redundant physicochemical properties in the AAindex database were defined by their strong correlation (Pearson's correlation test) with the absolute correlation coefficient ≥ 0.9. There are in total 18 selected physicochemical properties, shown in Table 4.2. Each

property consists of ten features, which are nine features from nine residues and one feature from summation of all residues in a peptide, and two features from similarity features including BLAST score and mismatched number(s) (Table 4.3).

Table 4.2 The physicochemical properties obtained from the AAindex database

Physicochemical property	Index	Description
Entropy	HUTJ700103	Entropy of formation (Hutchens, 1970)
	KRIW790102	Fraction of site occupied by water (Krigbaum-Komoriya, 1979)
Hydrophobicity	EISD860102	Atom-based hydrophobic moment (Eisenberg-McLachlan, 1986)
	EISD840101	Consensus normalized hydrophobicity scale (Eisenberg, 1984)
	EISD860103	Direction of hydrophobic moment (Eisenberg-McLachlan, 1986)
	GOLD730101	Hydrophobicity factor (Goldsack-Chalifoux, 1973)
	BLAS910101	Scaled side chain hydrophobicity values (Black-Mould, 1991)
	PRAM900101	Hydrophobicity (Prabhakaran, 1990)
	FAUJ880103	Normalized van der Waals volume (Fauchere et al., 1988)
Binding interaction	OOBM770102	Short and medium range non-bonded energy per atom (Oobatake-Ooi, 1977)
	KRIW710101	Side chain interaction parameter (Krigbaum-Rubin, 1971)
	OOBM770103	Long range non-bonded energy per atom (Oobatake-Ooi, 1977)
Polarity	ZIMJ680104	Isoelectric point (Zimmerman et al., 1968)
	GRAR740102	Polarity (Grantham, 1974)
	ZIMJ680103	Polarity (Zimmerman et al., 1968)
Size	FASG760101	Molecular weight (Fasman, 1976)
	DAWD720101	Size (Dawson, 1972)
	CHAM830103	The number of atoms in the side chain labelled 1+1 (Charton-Charton, 1983)

Table 4.3 The initial set of features (182 features)

Property index	Description	Amino acid position									Summation of the entire peptide
		p1	p2	p3	p4	p5	p6	p7	p8	p9	
HUTJ700103	Entropy_1	p1_HUTJ700103	p2_HUTJ700103	p3_HUTJ700103	p4_HUTJ700103	p5_HUTJ700103	p6_HUTJ700103	p7_HUTJ700103	p8_HUTJ700103	p9_HUTJ700103	sum_HUTJ700103
KRIW790102	Entropy_2	p1_KRIW790102	p2_KRIW790102	p3_KRIW790102	p4_KRIW790102	p5_KRIW790102	p6_KRIW790102	p7_KRIW790102	p8_KRIW790102	p9_KRIW790102	sum_KRIW790102
EISD860102	Hydrophobicity_1	p1_EISD860102	p2_EISD860102	p3_EISD860102	p4_EISD860102	p5_EISD860102	p6_EISD860102	p7_EISD860102	p8_EISD860102	p9_EISD860102	sum_EISD860102
EISD840101	Hydrophobicity_2	p1_EISD840101	p2_EISD840101	p3_EISD840101	p4_EISD840101	p5_EISD840101	p6_EISD840101	p7_EISD840101	p8_EISD840101	p9_EISD840101	sum_EISD840101
EISD860103	Hydrophobicity_3	p1_EISD860103	p2_EISD860103	p3_EISD860103	p4_EISD860103	p5_EISD860103	p6_EISD860103	p7_EISD860103	p8_EISD860103	p9_EISD860103	sum_EISD860103
GOLD730101	Hydrophobicity_4	p1_GOLD730101	p2_GOLD730101	p3_GOLD730101	p4_GOLD730101	p5_GOLD730101	p6_GOLD730101	p7_GOLD730101	p8_GOLD730101	p9_GOLD730101	sum_GOLD730101
BLAS910101	Hydrophobicity_5	p1_BLAS910101	p2_BLAS910101	p3_BLAS910101	p4_BLAS910101	p5_BLAS910101	p6_BLAS910101	p7_BLAS910101	p8_BLAS910101	p9_BLAS910101	sum_BLAS910101
PRAM900101	Hydrophobicity_6	p1_PRAM900101	p2_PRAM900101	p3_PRAM900101	p4_PRAM900101	p5_PRAM900101	p6_PRAM900101	p7_PRAM900101	p8_PRAM900101	p9_PRAM900101	sum_PRAM900101
FAUJ880103	Hydrophobicity_7	p1_FAUJ880103	p2_FAUJ880103	p3_FAUJ880103	p4_FAUJ880103	p5_FAUJ880103	p6_FAUJ880103	p7_FAUJ880103	p8_FAUJ880103	p9_FAUJ880103	sum_FAUJ880103
OOBM770102	Binding_1	p1_OOBM770102	p2_OOBM770102	p3_OOBM770102	p4_OOBM770102	p5_OOBM770102	p6_OOBM770102	p7_OOBM770102	p8_OOBM770102	p9_OOBM770102	sum_OOBM770102
KRIW710101	Binding_2	p1_KRIW710101	p2_KRIW710101	p3_KRIW710101	p4_KRIW710101	p5_KRIW710101	p6_KRIW710101	p7_KRIW710101	p8_KRIW710101	p9_KRIW710101	sum_KRIW710101
OOBM770103	Binding_3	p1_OOBM770103	p2_OOBM770103	p3_OOBM770103	p4_OOBM770103	p5_OOBM770103	p6_OOBM770103	p7_OOBM770103	p8_OOBM770103	p9_OOBM770103	sum_OOBM770103
ZIMJ680104	Polarity_1	p1_ZIMJ680104	p2_ZIMJ680104	p3_ZIMJ680104	p4_ZIMJ680104	p5_ZIMJ680104	p6_ZIMJ680104	p7_ZIMJ680104	p8_ZIMJ680104	p9_ZIMJ680104	sum_ZIMJ680104
GRAR740102	Polarity_2	p1_GRAR740102	p2_GRAR740102	p3_GRAR740102	p4_GRAR740102	p5_GRAR740102	p6_GRAR740102	p7_GRAR740102	p8_GRAR740102	p9_GRAR740102	sum_GRAR740102
ZIMJ680103	Polarity_3	p1_ZIMJ680103	p2_ZIMJ680103	p3_ZIMJ680103	p4_ZIMJ680103	p5_ZIMJ680103	p6_ZIMJ680103	p7_ZIMJ680103	p8_ZIMJ680103	p9_ZIMJ680103	sum_ZIMJ680103
FASG760101	Size_1	p1_FASG760101	p2_FASG760101	p3_FASG760101	p4_FASG760101	p5_FASG760101	p6_FASG760101	p7_FASG760101	p8_FASG760101	p9_FASG760101	sum_FASG760101
DAWD720101	Size_2	p1_DAWD720101	p2_DAWD720101	p3_DAWD720101	p4_DAWD720101	p5_DAWD720101	p6_DAWD720101	p7_DAWD720101	p8_DAWD720101	p9_DAWD720101	sum_DAWD720101
CHAM830103	Size_3	p1_CHAM830103	p2_CHAM830103	p3_CHAM830103	p4_CHAM830103	p5_CHAM830103	p6_CHAM830103	p7_CHAM830103	p8_CHAM830103	p9_CHAM830103	sum_CHAM830103
blast_score	Similarity of peptides and top hit from host's proteome										
mismatch	Number of mismatch residue(s) between a peptide and a host's top hit counterpart										

4.2.4 Similarity properties of peptides

"Foreignness" is a crucial factor to trigger the host immune system since only non-self-peptides can be recognised and stimulate host's T cells. Therefore, the similarity between immunogenic peptides and the host's proteome was created as one of features for epitope and non-epitope classification. Peptides for training the model were searched against their host proteome using Basic Local Alignment Search Tool (BLAST) with a stand-alone version 2.7.1. The optimal searching parameters are shown in Table 4.4. The best matched peptide was defined by the highest similarity score with 9mers in length and no gap. The similarity score and number of mismatches of the best matched peptide were selected as similarity features for the model.

Table 4.4 Input parameters for BLAST search

Option	Parameter	Description
program	blastp	Compare a protein query to a protein database
task name	blastp-short	Optimized for queries shorter than 30 residues
evalue	10^6	Expect value (E) for saving hits
word_size	2 (default)	Length of initial exact match
matrix	PAM30 (default)	A scoring matrix

4.2.5 The Random Forest classification model and model evaluation

For creating the Random Forest classification model, *RandomForestClassifier* from Scikit-learn packages was implemented in Python 3.7 to build the model from training data. The Random Forest machine learning package in Scikit-learn provides automatic iterative selection of optimal parameters; n_estimator = 100, and other parameters were set as default. The performance of the model was evaluated by the area under a curve (AUC) of a receiver operating characteristic (ROC) curve. The function named *roc_curve* was used to generate ROC, and the AUC score was computed by the *auc* function in Scikit-learn packages. The cross validation was performed by *cross_val_score* function with a parameter *cv=10*. Furthermore,

the F_1 scores that contributes to a weighted average of the precision and recall was also reported, where the best value is 1 and the worst value is 0.

$$F_1 = 2 \times \frac{precision \times recall}{precision + recall} = \frac{TP}{TP + \frac{1}{2}(FP + FN)}$$

TP = number of true positives

FP = number of false positives

FN = number of false negatives

4.2.6 Feature selection

Feature selection is the process of reducing the number of input variables for predictive model development since fewer input variables can help to reduce the complexity of the algorithm and make it more understandable. The process of selection involves evaluating the relationship between each input variable and the target variable using statistical methods to select those input variables that have the strong relationship with the target variable. In this work, the feature screening results were generated in Python 3.7 using the function called *SelectFromModel* in the Scikit-learn packages. In this work, the estimator for this function is the *RandomForestClassifier* algorithm, and the threshold for feature selection was set as default, which is the mean of all feature importance values. The feature selection was performed iteratively, for the first iteration, features whose importance value is greater than or equal to the mean importance of all features are kept for the next iteration. Then, selected features from the second iteration were combined to a set from the first iteration, these steps were repeatedly performed until the number of features are less than 30 and the stationary phase is reached (getting the same number of features > 10 iterations).

4.2.7 Calibrating predicted probability and constant value estimation

The predicted probability of each class was obtained from *predict_proba* function. To calibrate the probability scores produced by the Random Forest model, the logistic regression function (Eq.4.1) was used to transform pseudo-probability to calibrated probability.

Given pseudo-score data X = $\{x_1, x_2,\ldots x_i\}$ and calibrated probability data Y = $\{y_1, y_2,\ldots y_i\}$

$$y_i = n + \frac{1}{1+e^{-ax_i+b}} \qquad \text{(4.1)}$$

The constant values including a, b, and n were estimated by *curve_fit(sigmoid, X,Y)* function from Scipy packages in Python 3.7. The constant values estimation was performed from different sizes of testing data (20%, 30%, and 40%).

4.2.8 Decision tree interpretation

Interpreting the basis for prediction a model is important to check the reliability, i.e. does the combination of features make sense, and allows decomposing successful prediction to understand any bias and feature contribution. The feature contribution result was generated by *treeinterpreter*, which is a library that computes contribution values of each feature on prediction for tree-based models of Scikit-learn including *RandomForestClassifier*.

4.3 Results

4.3.1 Immunogenicity classification prediction model

The initial set of features was created from physicochemical properties in Table 4.3 as well as the similarity features. Numerical values of a property from the AAindex database were applied to each amino acid in a 9mers peptide. Thus, a 9mers peptide can generate 182 features including 180 features from 18 physicochemical properties and two features from similarity properties. The first model was built from all 182 features using *RandomForestClassifier* model with 70% training and 30% testing. The model performance was evaluated by AUC

score from ROC curve with 10-fold cross-validation. With 182 features, the average AUC is 0.726 which means there is ~73% chance that the model will be able to distinguish between immunogenic and non-immunogenic peptides. Even though the AUC score is not near to 1, it showed that the model can classify two classes of peptide with fairly high discrimination capacity. Moreover, the average F_1 scores for positive and negative classification from 10 runs are 0.709 and 0.601, respectively suggesting that the model has a slightly better accuracy for classifying immunogenic class than non-immunogenic class. The high importance values were mostly found in summation of all residues features and similarity features (Figure 4.3). However, importance values from the model with 182 features were very low per feature, indicating that most features are only making a small contribution to the model performance. This makes for a model that is hard to interpret and difficult to know if it will work well beyond the source data used for training.

To reduce the number of input variables that might not contribute to the model decision and sculpt more interpretable decision trees, feature selection was therefore performed to retain a small number of key features that contribute more highly to model performance. The feature selection was performed as described in Section 4.2.6. The algorithm was terminated when AUC scores were substantially declining. It was found that decreasing feature numbers does not significantly improve AUC scores (Figure 4.4). From inspection, the AUC score started falling from the model with 70 features (AUC = 0.723), and the AUC score of the final model (17 features) is 0.715 (Figure 4.4A). The objective for this analysis is to determine a set of features that should not decrease the performance model's predictability, therefore, a set of features that is as small as possible and does not substantially drop the AUC score was selected. From the result in Figure 4.4B, the models trained with 42 and 17 features were assessed, as follows. The AUC scores from 10-fold cross validation of 42 features model is not significantly different from the original model (182 features), but the AUC score from 17 features model is

substantially decreased from the original model (Figure 4.5). Therefore, a set of 42 features was determined as the optimal set of features for training the Random Forest model, the list of 17 and 42 features were shown in Table 4.5 and 4.6, respectively. Although, a set of features from feature selection analysis cannot considerably improve the performance of the Random Forest model for immunogenicity classification, a small set of feature number does make the model more understandable and can reduce inconsistency from irrelevant features contributing to a tree decision.

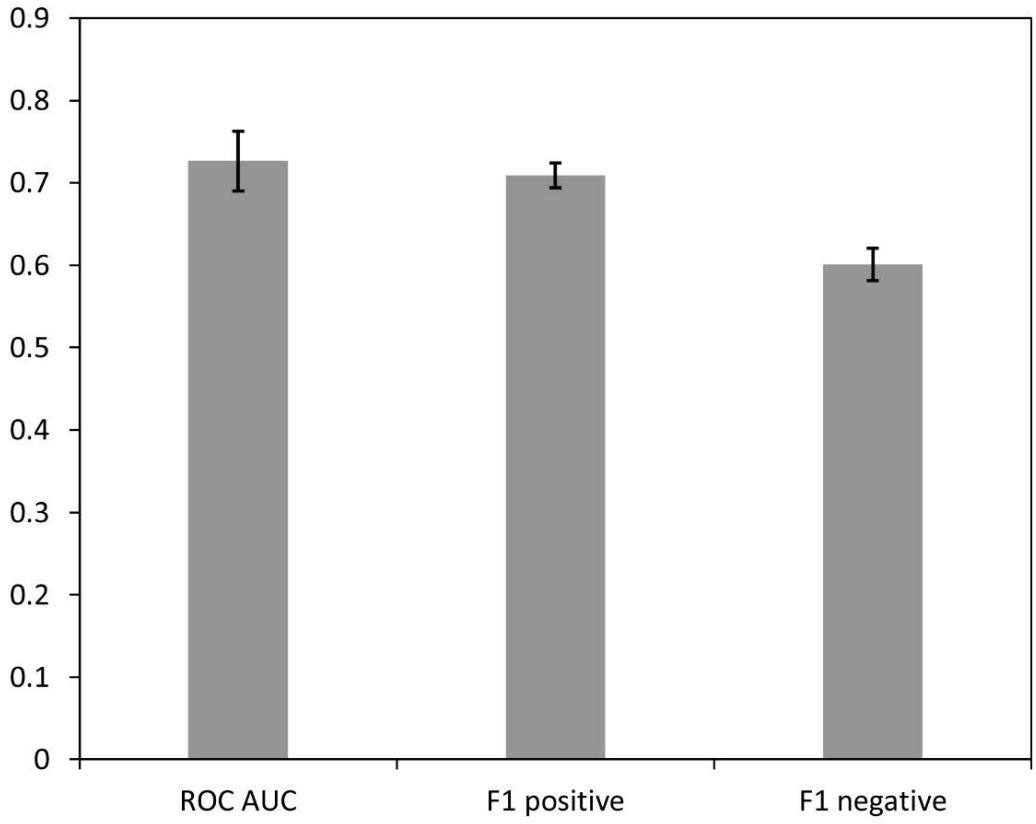

Figure 4.2 The reported performance of the Random Forest model with 182 features. The average of AUC scores, F_1 scores for positive and negative classes were obtained from 10 runs.

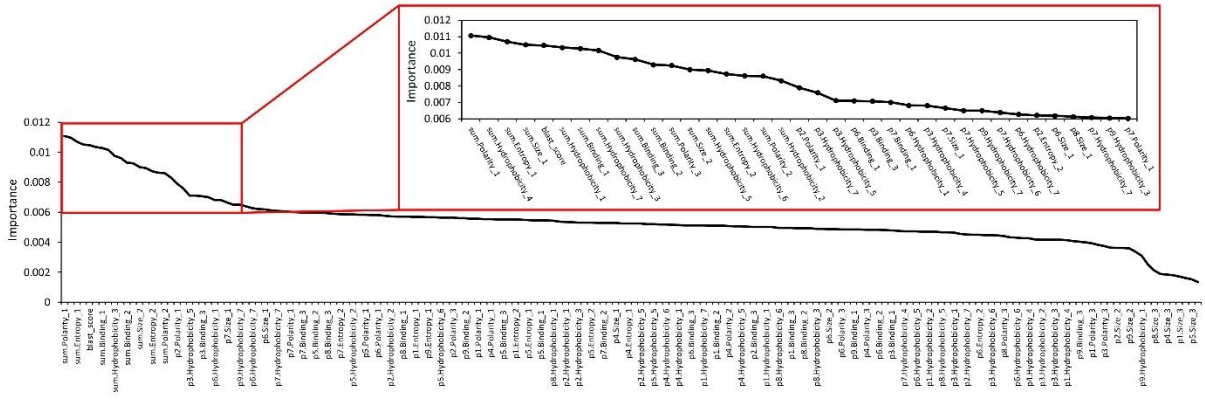

Figure 4.3 The importance values for 182 features. The values of each feature was computed from the Random Forest algorithm.

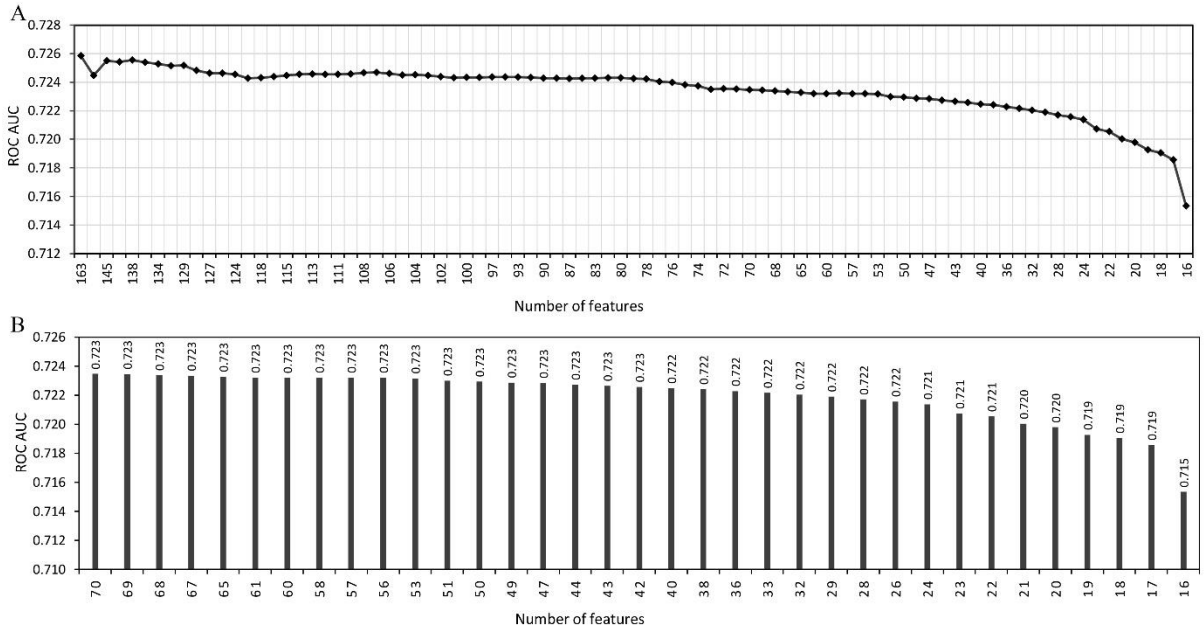

Figure 4.4 Feature selection analysis. (A) The AUC scores from models with different number of features during feature selection analysis. (B) The AUC scores of models with number of features less than 70.

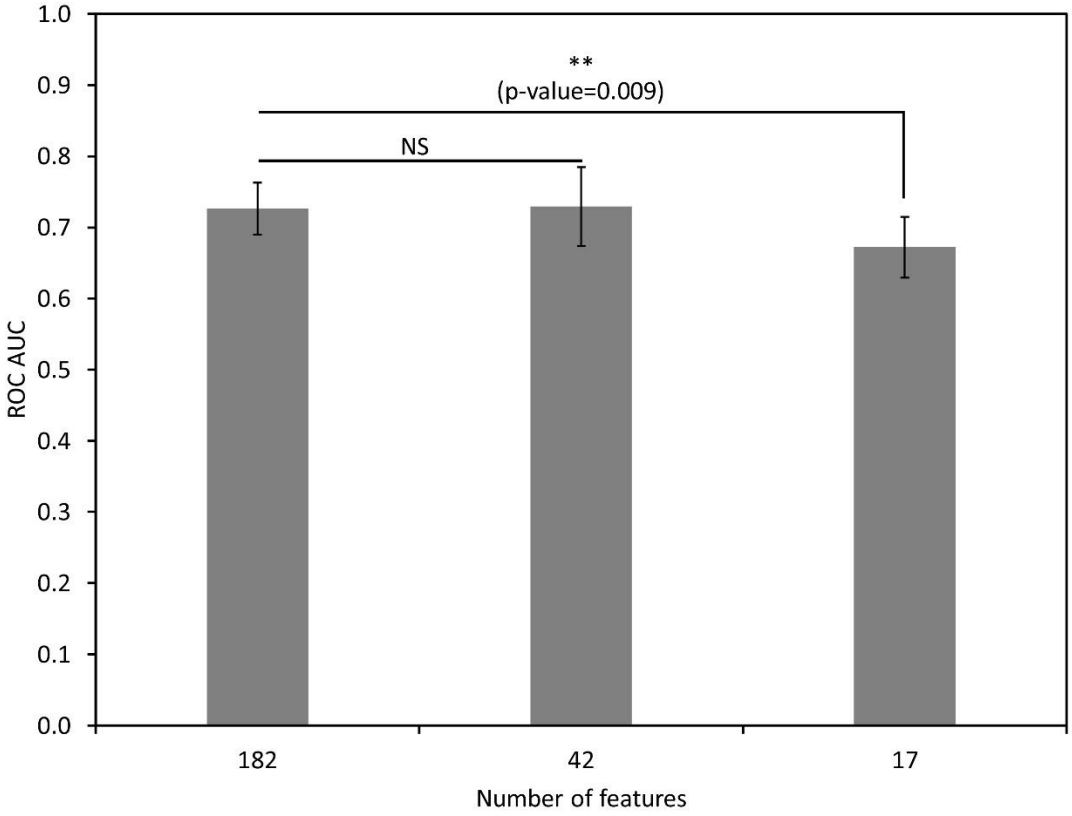

Figure 4.5 The AUC scores from 10-fold cross validation from models with 182, 42, and 17 features. Each bar represents mean±sd. The p-values were obtained from Student's t-test analysis, NS (non-significant), **(p-value, 0.01).

Table 4.5 The set of 17 features yielded from the feature selection analysis

Features	Description
p2_ZIMJ680104	Isoelectric point of position 2
sum_ZIMJ680104	Isoelectric point of a peptide
sum_GRAR740102	Polarity (Grantham, 1974) of a peptide
p3_FAUJ880103	Normalized van der Waals volume of positon 3
p9_FAUJ880103	Normalized van der Waals volume of positon 9
sum_FAUJ880103	Normalized van der Waals volume of a peptide
sum_HUTJ700103	Entropy of a peptide
sum_OOBM770102	Short and medium non-bonded energy per atom of a peptide
sum_BLAS910101	side chain hydrophobicity of a peptide
blast_score	Similarity of peptides and host's proteome
sum_EISD860102	Atom-based hydrophobic moment of a peptide
sum_OOBM770103	Long range non-bonded energy per atom of a peptide
sum_GOLD730101	Hydrophobicity factor of a peptide
sum_EISD860103	Direction of hydrophobic moment of a peptide
sum_FASG760101	Molecular weight of a peptide
sum_ZIMJ680103	Polarity (Zimmerman et al., 1968) of a peptide
sum_KRIW790102	Fraction of site occupied by water of a peptide

Table 4.6 The set of 42 features yielded from the feature selection analysis

Features	Description
p2_ZIMJ680104	Isoelectric point of position 2
p9_ZIMJ680104	Isoelectric point of position 9
sum_ZIMJ680104	Isoelectric point of a peptide
p2_GRAR740102	Polarity (Grantham, 1974) of position 2
sum_GRAR740102	Polarity (Grantham, 1974) of a peptide
p3_FAUJ880103	Normalized van der Waals volume of positon 3
p9_FAUJ880103	Normalized van der Waals volume of positon 9
sum_FAUJ880103	Normalized van der Waals volume of a peptide
p6_FAUJ880103	Normalized van der Waals volume of positon 6
sum_HUTJ700103	Entropy of a peptide
p1_HUTJ700103	Entropy of position 1
p8_HUTJ700103	Entropy of position8
p2_OOBM770102	Short and medium non-bonded energy per atom of position 2
p7_OOBM770102	Short and medium non-bonded energy per atom of position 7
p6_OOBM770102	Short and medium non-bonded energy per atom of position 6
sum_OOBM770102	Short and medium non-bonded energy per atom of a peptide
p3_BLAS910101	side chain hydrophobicity of position 3
p7_BLAS910101	side chain hydrophobicity of position 7
sum_BLAS910101	side chain hydrophobicity of a peptide
blast_score	Similarity of peptides and host's proteome
sum_EISD860102	Atom-based hydrophobic moment of a peptide
p6_EISD860102	Atom-based hydrophobic moment of position 6
sum_OOBM770103	Long range non-bonded energy per atom of a peptide
p8_OOBM770103	Long range non-bonded energy per atom of position 8
p3_OOBM770103	Long range non-bonded energy per atom of position 3
p1_PRAM900101	Hydrophobicity of position 1
p5_PRAM900101	Hydrophobicity of position 5
p9_PRAM900101	Hydrophobicity of position 9
p7_PRAM900101	Hydrophobicity of position 7
sum_PRAM900101	Hydrophobicity of a peptide
p5_KRIW710101	Side chain interaction parameter of position 5
p4_KRIW710101	Side chain interaction parameter of position 4
sum_KRIW710101	Side chain interaction parameter of a peptide
sum_GOLD730101	Hydrophobicity factor of a peptide
sum_EISD860103	Direction of hydrophobic moment of a peptide
p9_FASG760101	Molecular weight of position 9
sum_FASG760101	Molecular weight of a peptide
sum_ZIMJ680103	Polarity (Zimmerman et al., 1968) of a peptide
p2_KRIW790102	Fraction of site occupied by water of position 2
sum_KRIW790102	Fraction of site occupied by water of a peptide
sum_EISD840101	Consensus normalized hydrophobicity scale of a peptide
sum_DAWD720101	Size of a peptide

4.3.2 Benchmarking analysis

Form the previous analysis, the optimised model was created by a set of 42 features in Table 4.6. To evaluate the model performance compared to the existing tools, the model was compared to existing MHC class I immunogenicity prediction tools which are Immunogenicity [134] and INeo-Epp [146], these models were built by sequence-based learning and trained with physicochemical properties. The data set for benchmarking was split from the whole data set for 10%, hence, the 10% validating data was used to test with the Random Forest model and those two published models and had not been used to train our model or select features. The AUC scores from the model in this study and those two models were calculated from the prediction of the same data set (10% validating data). It revealed that the performance of the Random Forest model in this work (AUC=0.729) outperforms Immunogenicity (AUC=0.516) and INeo-Epp (AUC=0.699) with respect to F_1 scores of the Random Forest model, Immunogenicity, and INeo-Epp (0.649, 0.510, and 0.578, respectively) (Figure 4.6A and B).

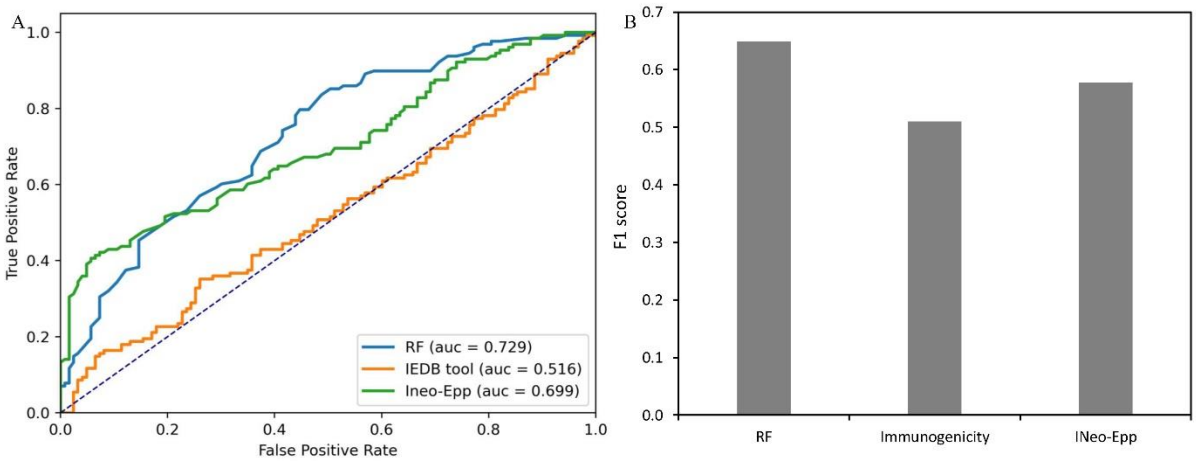

Figure 4.6 The benchmarking analysis of the Random Forest (RF) model and the existing tools. (A) The ROC plots representing AUC scores for all tools. (B) The bar plot of F1 scores obtaining from all tools.

4.3.3 Predicted probability calibration

The probability scores from the Random Forest model were calculated from the average probabilities over the number of trees in the forest, although it may not be a real probability for other unseen data that might not be different in data size or a ratio between positives and negatives data. Therefore, the predicted probability scores produced from the Random Forest model should be calibrated to real probabilities that can be further applied for any data prediction. The distribution of pseudo-probability scores corresponding to immunogenic class was observed for known true and false data set (Figure 4.7A). Then, the distributions were fitted to two beta components, and true posterior probability of each predicted probability (a reverse posterior error probability (1-PEP)) was calculated from the PDF of beta distribution. The plot of pseudo-probability against to true posterior probability displayed a non-linear relationship between pseudo-probability and true posterior probability, especially scores in range of 0.8 to 1, but it seems to be a sigmoid curve-like (Figure 4.7B). Therefore, the logistic regression function in Eq. 4.1 was selected to model the data and fit to that sigmoid curve to estimate those constant values that could form the best equation for transforming pseudo-probability to calibrated probability. The estimated constant values from different training data sizes showed small variation (Table 4.7). Kullback–Leibler (KL) divergence was used to evaluate similarity between each set of fit data and three observed true posterior probability data with 20%, 30%, and 40% testing data. A set of constant values from 30% testing data has the lowest average KL value for all three data sets (Table 4.7) indicating that transforming data by the logistic regression with these constant yields the best fitted data (Figure 4.8). Hence, the formula in Eq. 4.2 was further used to transform pseudo-probability values from the Random Forest model to calibrated probability scores for the immunogenicity prediction.

$$y_i = 0.048 + \frac{1}{1+e^{-5.87x_i+2.89}} \qquad (4.2)$$

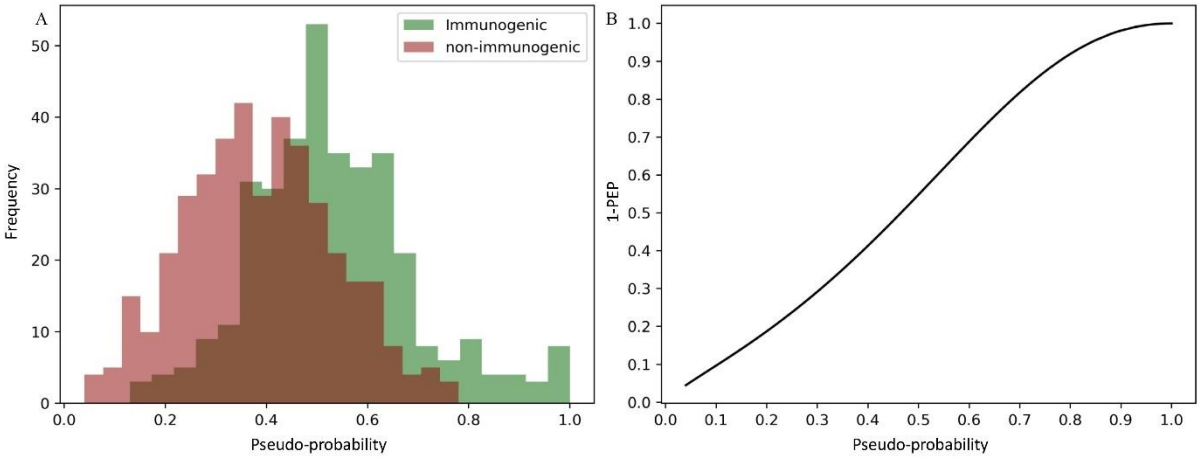

Figure 4.7 The pseudo-probability scores and true posterior error probability (1-PEP). (A) The distribution of pseudo-probability of immunogenic and non-immunogenic data sets. (B) The plot between each pseudo-probability score against its true posterior probability (1-PEP), which is estimated from the density of the data distribution.

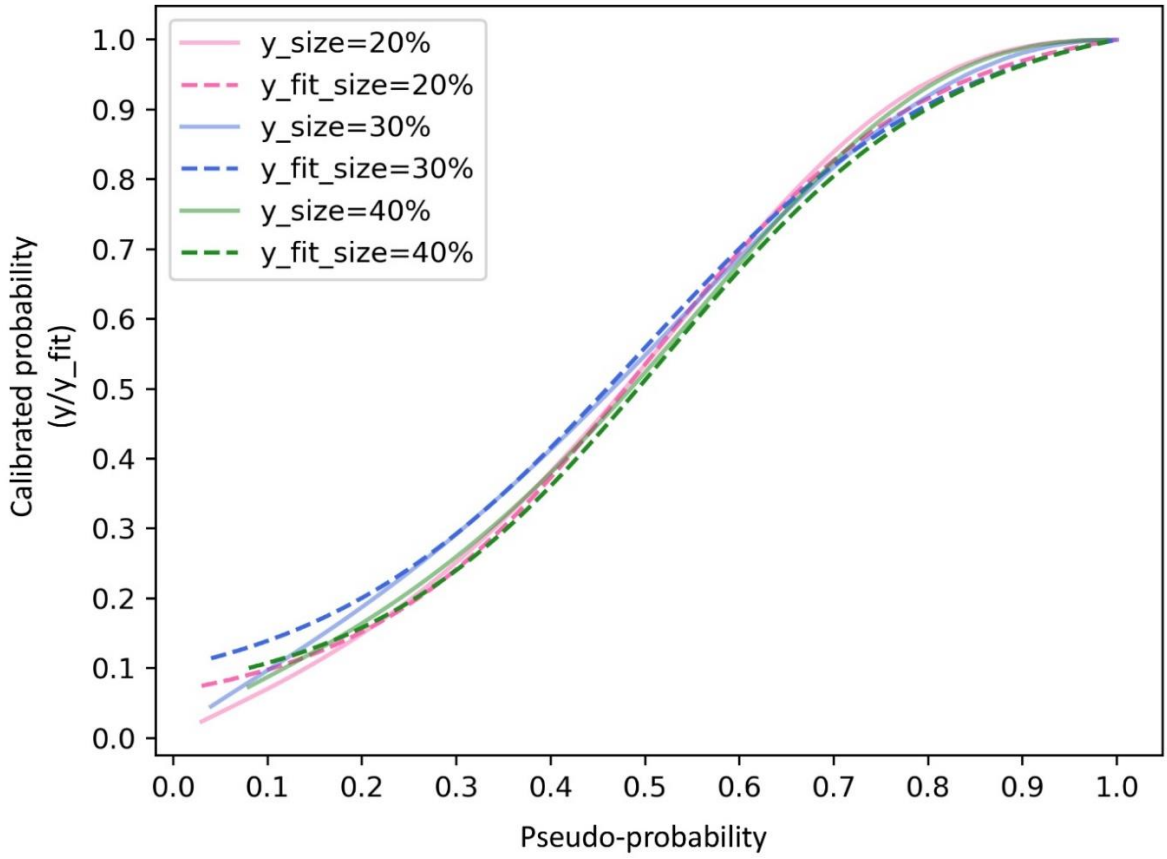

Figure 4.8 The plot of pseudo-probability scores against to 1-PEP (y) and calibrated probability scores (y_fit). The calibrated probability scores of different testing data sizes are transformed by the logistic regression function.

Table 4.7 The estimated constant values from the logistic regression fit and Kullback–Leibler (KL) divergence values

Observed data (% testing size)	Constants			KL divergence value			
	a	b	n	y fit_20%	y fit_30%	y fit_40%	Average
20%	6.713	3.351	0.033	1.528	2.004	1.790	1.774
30%	5.870	2.892	0.048	1.168	1.538	1.373	1.360
40%	6.415	3.333	0.044	1.307	1.721	1.531	1.520

4.3.4 The model interpretation

The *treeinterpreter* function provides a contribution value of each feature for the prediction of each class. The average prediction value is yielded from the average of all possible predictions in data from the path going through an individual node in a tree. Each node in the decision tree represents some feature and makes a decision based on the feature value in the sample. For the Random Forest where there are multiple trees, the final prediction is computed from an average of all trees. The contribution value of each feature in Table 4.6 was assessed using the *treeinterpreter* function to reveal the features that influence the immunogenicity classification model. The result from the *treeinterpreter* returns contributed values of every feature for every data point in a training data set. For each data point, the feature thar has the highest contributed value was determined. The percent frequency of each feature was computed by counting from a set of highest contributed features across all data points. A larger number means the feature has been frequently found in a set of highest contributed features indicating that the feature has more influence on the model decision. The percent frequency for all features is shown in Table 4.8. Among those features, the properties related to polarity highly contribute to the model, which are *isoelectric point* (22.4%) *and polarity* (11.1%). The second highest influence was found in properties involved in *hydrophobicity*, the summation of those feature has a frequency of 30.9%. Moreover, features that relate to a strength of binding interaction associated with non-covalent intermolecular interaction have a contributed frequency of 15.3%, which are

Short and medium non-bonded energy per atom and *Long range non-bonded energy per atom*. Furthermore, *entropy* has a moderate impact (9.3%) on the model, this property is also involved in a strength of binding affinity. *Molecular weight* (2.2%) of amino acid and the similarity feature (*blast score*, 2.3%) contribute less to the model compared to other features (Figure 4.9). Overall, the result from the model interpretation revealed properties related to the strength of binding interaction mostly contribute to the decision of the model to classify immunogenic and non-immunogenic peptides suggesting that those properties might be favourable for interaction between T cell receptors and peptides.

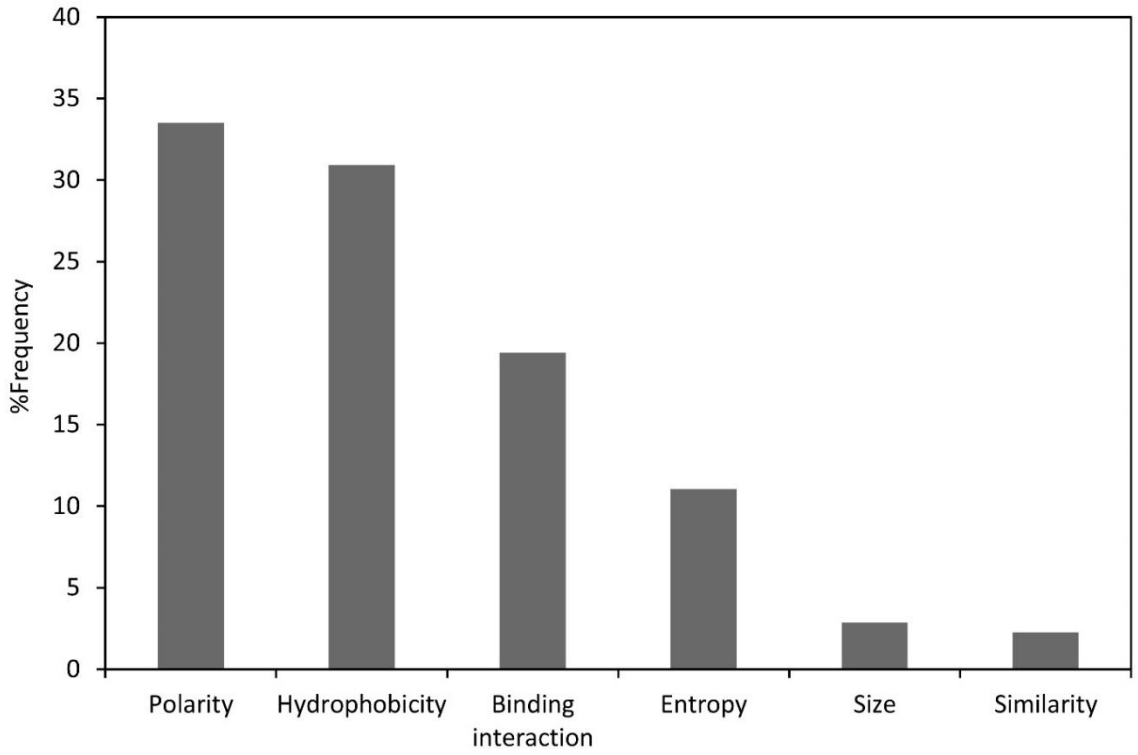

Figure 4.9 The contribution of important features to the prediction model. The bar plot displays the percent frequency computed by counting from a set of highest contributed features across all data points.

Table 4.8 The percent frequency counting from number of found highest contributed values of 42 features

Features	Description	% Frequency
sum_HUTJ700103	Entropy of formation (Hutchens, 1970)	4.06
p1_HUTJ700103		3.19
p8_HUTJ700103		2.06
p2_KRIW790102	Fraction of site occupied by water (Krigbaum-Komoriya, 1979)	1.06
sum_KRIW790102		0.66
p1_PRAM900101	Hydrophobicity (Prabhakaran, 1990)	1.86
p5_PRAM900101		1.6
p9_PRAM900101		1.46
p7_PRAM900101		1.26
sum_PRAM900101		0.33
p3_FAUJ880103	Normalized van der Waals volume (Fauchere et al., 1988)	7.65
p9_FAUJ880103		3.19
sum_FAUJ880103		0.6
p6_FAUJ880103		0.73
sum_EISD860103	Direction of hydrophobic moment (Eisenberg-McLachlan, 1986)	1.6
sum_EISD840101	Consensus normalized hydrophobicity scale (Eisenberg, 1984)	0.73
p3_BLAS910101	Scaled side chain hydrophobicity values (Black-Mould, 1991)	3.32
p7_BLAS910101		0.66
sum_BLAS910101		0.6
sum_GOLD730101	Hydrophobicity factor (Goldsack-Chalifoux, 1973)	1.66
sum_EISD860102	Atom-based hydrophobic moment (Eisenberg-McLachlan, 1986)	2.13
p6_EISD860102		1.53
p2_ZIMJ680104	Isoelectric point (Zimmerman et al., 1968)	10.64
p9_ZIMJ680104		7.38
sum_ZIMJ680104		4.39
p2_GRAR740102	Polarity (Grantham, 1974)	9.11
sum_GRAR740102		0.8
sum_ZIMJ680103	Polarity (Zimmerman et al., 1968)	1.2
p2_OOBM770102	Short and medium range non-bonded energy per atom (Oobatake-Ooi, 1977)	3.59
p7_OOBM770102		2.99
p6_OOBM770102		2.73
sum_OOBM770102		1.93
sum_OOBM770103	Long range non-bonded energy per atom (Oobatake-Ooi, 1977)	1.93
p8_OOBM770103		1.26
p3_OOBM770103		0.86
p5_KRIW710101	Side chain interaction parameter (Krigbaum-Rubin, 1971)	1.86
p4_KRIW710101		1.4
sum_KRIW710101		0.86
p9_FASG760101	Molecular weight (Fasman, 1976)	1.53
sum_FASG760101		0.66
sum_DAWD720101	Size (Dawson, 1972)	0.66
blast_score	Similarity of peptides and host's proteome	2.26

4.4 Discussion

The determination of immunogenicity for MHC presented peptides aims to identify short peptides that can activate T cell response, either CD4+ or CD8+ T cells. Identification of immunogenic peptides is of great interest for immunology research such as understanding disease etiology, monitoring of immune response, or designing epitope-based vaccine. For neoantigen-based cancer vaccine development, the identification of immunogenicity is essential for the selection of true neoantigens to reduce the risk of getting a false positive and thus help to increase the success rate of neoantigen-based cancer vaccine therapeutic. Nevertheless, the process of antigen presenting and TCR recognition is highly complicated, the precise mechanism of binding interaction between TCR and an MHC presented peptide has not been clearly revealed. Previous studies have been studied importance characters of amino acids and positions in immunogenic peptides. Those studies reported that the physicochemical property of amino acids corresponding to size, hydrophobicity, entropy, polarity, and binding interaction are associated with the preference of TCRs [117, 134, 146]. Those properties were encoded to numerical data using AAindex database and applied for amino acids in a peptide sequence to create a set of features for building a prediction model with sequence-based learning. The other protein encoding sequence method, called Context-Free Encoding Scheme (CFreeEnS), has been reported for use in the development of antigenicity prediction for a various of influenza A viruses. CFreeEnS takes advantage of rich information about the physiochemical and structural properties of amino acids. This encoding scheme keeps the information about conserved properties of amino acids, which makes it possible for learning methods (e.g. Random Forest) to capture the antigenic pattern of influenza viruses. That study reported that CFreeEnS can improve the prediction of antigenicity and outperforms existing models [181]. The other existing models showed the capability to distinguish immunogenic peptides from non-immunogenic peptides with moderate performance, the reported ROC score

is approximately 0.75 (0.65 and 0.78 for Immunogenicity and INeo-Epp, respectively) [134, 146]. An antigen or a peptide that can elicit an immune response must be a foreign substance to the host immune or can be recognised as non-self by the host's immune system. Since during T cell development, those T cells who have a strong binding to self-peptides are eliminated, the negative selection results in a population of T cells that promptly bind to non-self antigens [2]. Thus, the properties of foreignness to self of input peptides were utilised as features for training the immunogenicity classification model.

For developing an immunogenicity predictor, an alternative would be to search only for simple sequence-based motifs, or for example to use logistic regression over numerical properties, derived for amino acid positions. However, the interaction between TCR and epitopes is very complicated and involved by many factors from TCR and MHC molecules. Thus, we required a method that would be easy to interpret how the prediction model works and be able to assess the contribution of different feature. Therefore, the RF algorithm was selected to develop a predictive model in this study. RF has a general property of high performance for creating predictive models from complex numerical data, with relatively simple interpretation (compared to neutral networks for example), and there is very little pre-processing that needs to be done [182]. This algorithm is an ensemble learning method for classification, regression, and other tasks that operates by constructing a multitude of decision trees at the training phase. RF eradicates some of the limitations of a decision tree algorithm. It reduces the overfitting of datasets and increases precision. In this chapter, target variables in the training data can be classified into true or false, thus, we used the RF classification as a method to develop the predictive model. The training data is fed to train various decision trees. This dataset consists of observations and features that will be selected randomly during the splitting of nodes [183]. Moreover, RF generates predictions without requiring many configurations in packages such as scikit-learn that was used in this study. Nevertheless, many algorithms can work for binary

classification, for example, Naive Bayes, Logistic regression, Support Vector Machine, Voting Classification, as well as more complex methods such as deep learning. All models have pros and cons, in terms of performance, generalisability, or interpretability. RF tends to occupy a middle ground of good performance, simple to train and relatively interpretable. However, in future work, it may be worth trying out several different modelling approaches to determine the optimal performing approach.

Initially, 182 features were used to build the model, the feature importance analysis was found that the summation physicochemical property of all amino acids in a 9mers peptide and a BLAST score feature have the top ranks compared to others. In this study, we considered the importance of each feature and each position individually, but the side chain group of amino acids in a peptide might affect their neighbour which might be a key residue interacting with TCR. It is also possible that there could be interactions between different positions. However, including features derived from all pairs (or triplets) of positions vastly increased the number of features, potentially making feature selection more challenging. RF does have the inherent ability to find interactions between features, as there can be paths on decision trees that are followed only when certain numerical thresholds are reached for certain variables working together.

Overall, the individual importance values were very low indicating a small contribution from many features, likely highly correlated with each other, leading to a model that is difficult to interpret. The approach of feature selection is commonly used for high-dimensional data analysis to improve the model predictability by removing irrelevant and redundant features resulting to improvement of learning accuracy, reducing learning time, and generating understandable learning results [184]. Although, the analysis from the feature selection experiment demonstrated that a subset from the original feature set does not improve the AUC score, a small set of features might reduce variation from irrelevant features and can simplify

the model prediction that can further elaborate which features mostly contribute to the model classification.

The result from benchmarking analysis showed that the Random Forest classification model in this study outperforms Immunogenicity from Calis *et al*., this might be due to update of training data, and the model from Immunogenicity masked the positions corresponding to anchor residues this might miss signals to differentiate epitope and non-epitope peptides [134]. Even though the model in this study did not mask position related to MHC anchored residues, the training data set was cleaned to match a distribution of predicted IC_{50} between positive and negative data to prevent bias from binding affinity property. INeo-Epp is a current immunogenicity prediction tools trained by only human peptides presented by HLA supertypes, and the AUC from external validation from this study showed about 0.779. It is interesting that the AUC score came from the model which removed peptides that have predicted % rank > 2. Moreover, the most importance feature contributing to the model was found as % rank from MHC-peptide binding prediction [146]. Therefore, it is likely that the reported prediction statistics from INeo-Epp might be mostly contributed from distinguishing binding and non-binding peptides but might not genuinely classify immunogenic and non-immunogenic peptides. This also might explain why INeo-Epp yielded poorer performance in our benchmark, with the data sets matched by MHC-peptide binding affinity for both positive and negative data. Overall, even the model performance from this work does not reach very high accuracy (e.g. >90%), it is still outperforms existing tools.

Finally, this developed model returns probability scores, which are computed from the average probabilities over the number of trees in the forest [185], instead of predicted class of immunogenic or non-immunogenic. Probability scores provide several applications such as ranking, thresholding with uncertainty predicted scores, and deciding how to interpret the predicted result. Moreover, to prevent an inconsistency from probability estimation for

different input data sets in the future, the pseudo-probability scores produced from the Random Forest model were transformed to calibrated probability. The calibrated probability score allows for better comparison across results from different prediction runs, and as will be shown in Chapter 5, can be combined straightforwardly with probability of peptide binding to develop a full prediction pipeline.

4.5 Conclusions

In this chapter, the immunogenicity prediction model was developed using the Random Forest algorithm, and the model was trained by physicochemical properties and foreignness features corresponding to immunogenic and non-immunogenic peptides derived from immunogenicity experiments. The developed model in this work exhibits performance improvement over existing tools. Moreover, the predictability of this model is independent of MHC-peptide binding affinity. Thus, this predictor should truly contribute to distinguish epitopes and non-epitopes relying on characters of T cell preference. To apply the immunogenicity prediction for neoantigen selection, the integration of ability of MHC-peptide binding obtaining from Chapter 3 should be combined to immunogenic probability scores, which is discussed in the following chapter.

Chapter 5

A pipeline for ranking predicted neoantigens using the estimation of local FDR and immunogenicity prediction

5.1 Introduction

The previous chapters described the development of models for scoring MHC-peptide binding prediction and for immunogenicity prediction. The global/local FDR estimation model, MHCVision, in Chapter 3 can help to improve criteria selection for MHC binding peptides, whilst the Random Forest model in Chapter 4 can predicted immunogenicity of peptide sequences based on T cell preferences of chemical properties of amino acids. The outcomes from those two models contribute to a potent of neoantigen properties i.e. peptides that have a strong binding with MHC molecules and can stimulate T cell activity, thus, those models should be integrated to be a pipeline that produce a final probability of MHC binding and T cell recognition. The final probability can be used for neoantigen selection or prioritisation. In this chapter, MHCVision and the Random Forest models were combined to be a pipeline, so called MHCVision-RF, and the final probability was computed from true MHC binding probability (1-PEP) and immunogenicity probability. The final probability produced by MHCVision-RF can be served as ranking scores for neoantigen selection. The mathematical operation for final probability calculation was reported in Section 5.3.1, then the workflow of a pipeline and code implementation were described in Section 5.3.2. In Section 5.3.3, the pipeline was applied to data with experimental validation from previous published studies to explore if the ranking score from the pipeline can separate neoantigen and non-neoantigen data.

5.2 Materials and Methods

5.2.1 Software implementation and client software requirement

MHCVision-RF was built by integrating MHCVision and the immunogenicity prediction model, and the program was implemented using Python version 3.7. The software is non-graphical user interface and can be run on Unix operating system using the command line. The following programs and packages in Python are required for optimal processing.

a.) Python with version 3.7 onward and the following Python packages (Table 5.1)

Table 5.1 Python packages and their versions for client requirement

Package	Version
numpy	≥ 1.19.1
pandas	≥ 1.1.2
scipy	≥ 1.5.2
scikit-learn	≥ 0.23.2

b.) Standalone BLAST for Unix (version 2.7.1)

The installation process for Mac OSX, Windows, and Linux can be found in Tao T., 2008 [186].

5.2.2 Observation of the relationship between true MHC binding probability and immunogenicity probability

The relationship between true MHC binding probability and immunogenicity probability was observed using a date set of 1000 9mers peptides generated from human proteome. MHC-peptide binding affinity prediction was made by NetMHCpan 4.1 against to HLA-A*02:01, and true MHC binding probability were estimated using MHCVision. The immunogenicity of each peptide was predicted using the Random Forest model described in Chapter 4. The linear regression model was used to evaluate the correlation between those two probability scores.

5.2.3 Generation of validating data from published neoantigen data

The data sets of peptides with the experimental validation of T cell reactivity towards predicted neoantigens from two previous published studies were used to validate the ranking score produced by MHCVision-RF. In the study from Patrick Ott *et al.,* 2020 [187], candidate neoantigens were selected based on bioinformatic analysis and MHC-peptide binding predictions, and IFN-γ ELISpot assay was used to validate the immunogenicity of peptides. From this study, peptides with 9-11mers of two melanoma patients (M1 and M3) and a lung

cancer patient (L7) were applied to MHCVision-RF. Only data from these three patients were selected because their HLA alleles are reported in *Figure S4* of the original paper [187]. The raw data of peptides from M1, M3, and L7 patients can be found in the supplementary data[1] in the original paper of Patrick Ott *et al.*, 2020.

Furthermore, data sets of patients from the study of Yong Fang *et al.*, 2020 [188] including P001 (Melanoma), P003 (Adrenal Sebaceous Adenocarcinoma), P004 (Small Cell Lung Cancer), P011 (Ovarian Cancer), and P016 (Non-Small Cell Lung Cancer) were selected to test the pipeline. The raw data of peptides from P001, P003, P004, P011 and P016 patients can be found in the supplementary tables[2] in the original paper of Yong Fang *et al.*, 2020. In this study, they reported designed peptides for synthesis which are long peptides (25-30 amino acids) and used IFN-γ ELISpot assay to test T cell reactivity of candidate peptides. To imitate the step of short peptide preparation for neoantigen prediction, those long peptides were cut to 9mers via a sliding window method. A set of 9mers peptides of each patient were then applied to MHCVision-RF.

[1] https://ars.els-cdn.com/content/image/1-s2.0-S0092867420311417-mmc2.xlsx

[2] https://clincancerres.aacrjournals.org/highwire/filestream/182017/field_highwire_adjunct_files/0/228188_3_supp_6278638_q9vzyf.xlsx

5.3 Results

5.3.1 Generation of the final probability of MHC binding and T cell recognition

The final probability is the combination of true MHC binding probability (1-PEP) from MHCVision and immunogenicity probability from the Random Forest model, and this score will be used as the ranking score for candidate neoantigen selection. To find the method of mathematical operation for final probability calculation, the correlation between true MHC binding probability and immunogenicity probability was observed using a scatter plot. The correlation coefficient between two data was evaluated by R^2 calculated from the linear regression model. It was found that there is no correlation between those two scores (R^2 = 0.019) suggesting that the production of those probability scores is independent (Figure 5.1A). Moreover, there are no correlation between those two probabilities when true binding probability was constrained for binding (1-PEP $\geq$ 0.8), or non-binding peptides (1-PEP $\leq$ 0.1) shown in Figure 5.1B and 5.1C, respectively. As mentioned before in Chapter 4, immunogenic peptides tend to be biased for MHC-binding peptides, but the data training for the RF model in this research was standardised the MHC binding ability between positive and negative classes to avoid the bias from binding and non-binding classification. Therefore, a lack of correlation between high binding probability and immunogenicity probability can indicate that the RF model in Chapter 4 purely distinguish immunogenic and non-immunogenic peptides based on T cell preference's properties. Thus, the final probability produced from MHCVision-RF was calculated from a multiplication between true MHC binding probability and immunogenicity probability with an equal weight value of each factor.

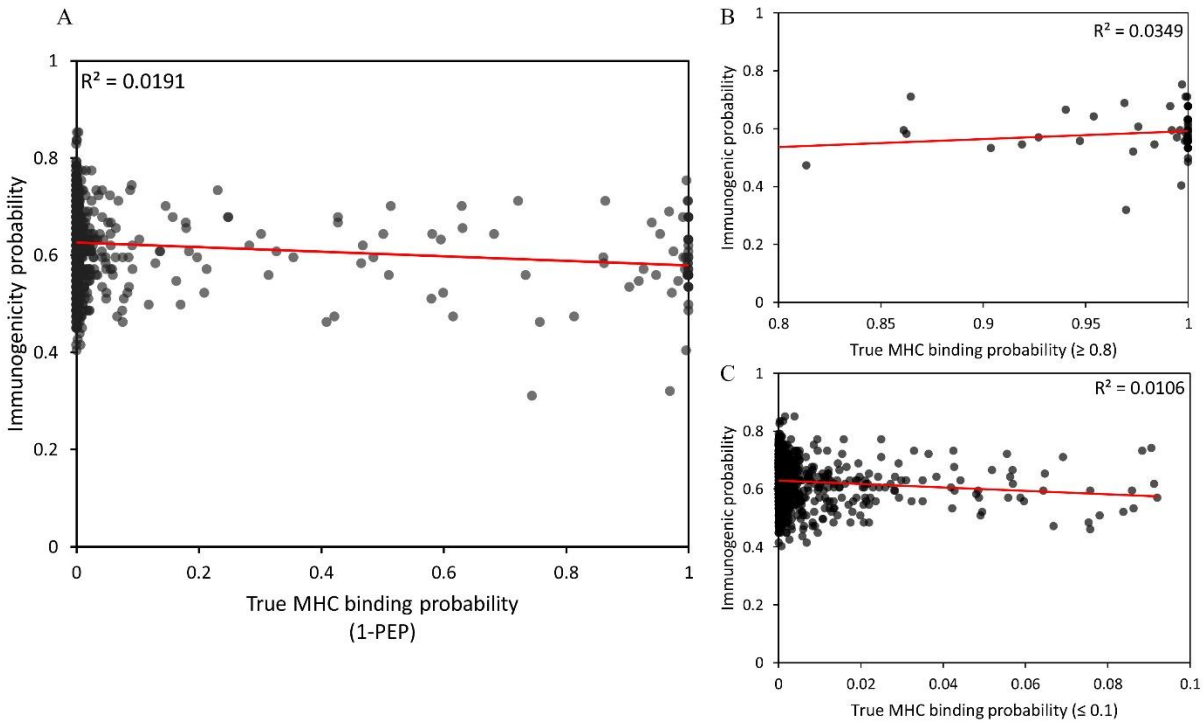

Figure 5.1 The relationship between true MHC binding probability from MHCVision and immunogenicity probability from the Random Forest model. The correlation of immunogenic probability and non-restricted MHC binding probability (A), immunogenic probability and high MHC binding probability (B), and immunogenic probability and low MHC binding probability (C)

5.3.2 The overall workflow of MHCVision-RF pipeline

The pipeline for ranking MHC class I neoantigen was built by integrating of MHCVision which provides true MHC binding probability and the Random Forest model of immunogenicity prediction. Users can opt to use either MHCVision alone for other works that do not need immunogenicity scores, or they can use the whole pipeline that produces the final probability scores which can further apply to the process of candidate neoantigen selection. The algorithm and implementation of MHCVision were fully described in Chapter 3, and the algorithm of the Random Forest model was reported in Chapter 4. Here, the workflow MHCVision-RF was described (Figure 5.2).

a.) Prediction of MHC-peptide binding affinity

MHC-peptide binding predictions are made between peptides and HLA types. The current version of MHCVision is available for only MHC class I that are supported in NetMHCpan 4.1 or MHCflurry. Users can opt to use either NetMHCpan or MHCflurry with the option that provides predicted binding affinity (IC_{50}) in nM unit because this score is used for FDR/PEP estimation.

b.) Input data preparation

MHCVision will run for an individual HLA allele at a time since the algorithm was built with restriction of HLA-specific estimated values. Before running, the output file from NetMHCpan or MHCflurry need to be formatted in comma delimited format (CSV) with one HLA allele for a file, the input table must contain columns of peptides and their predicted IC_{50} in nM unit (Figure 5.3A).

c.) Estimation of true MHC binding probability by MHCVision

Global and local FDRs will be estimated from the distribution of predicted IC_{50} by MHCVision. The algorithm will annotate FDRs, PEPs, and reversed PEPs, so called true MHC binding probability for each peptide.

d.) Immunogenicity probability prediction by the Random Forest model

Peptides from an input data will be translated to numerical matrix and taken as input to the Random Forest model. The model will predict a probability of each peptide, and that score will be transformed to the real probability, which is annotated as immunogenicity probability, by the logistic regression model.

e.) Generation of the final probability of MHC binding and T cell recognition

The final probability is computed from a multiplication between true MHC binding probability and immunogenicity score, thus final scores range from 0 to 1 where 1 is the best score for being a true neoantigen that means the peptide has a strong binding to MHC molecule and high potent for T cell recognition.

The final output will return extra columns of statistical information from MHCVision, immunogenicity probability, and final probability MHC binding and T cell recognition for each peptide (Figure 5.3B). The information from MHCVision-RF gives users the ability to make an informed selection of neoantigens with high potential of being true neoantigens.

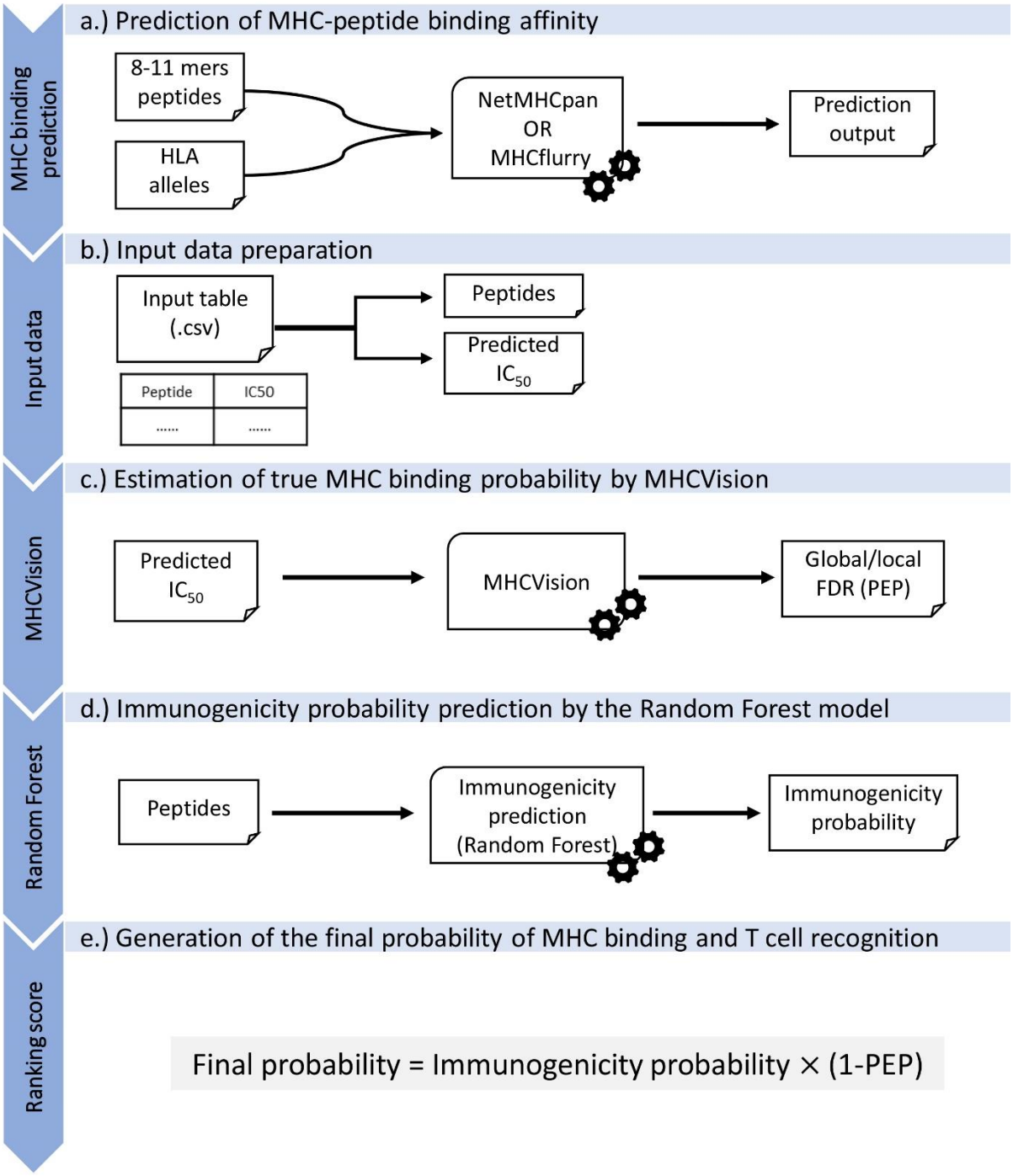

Figure 5.2 The workflow of MHCVision-RF and the calculation of final probability.

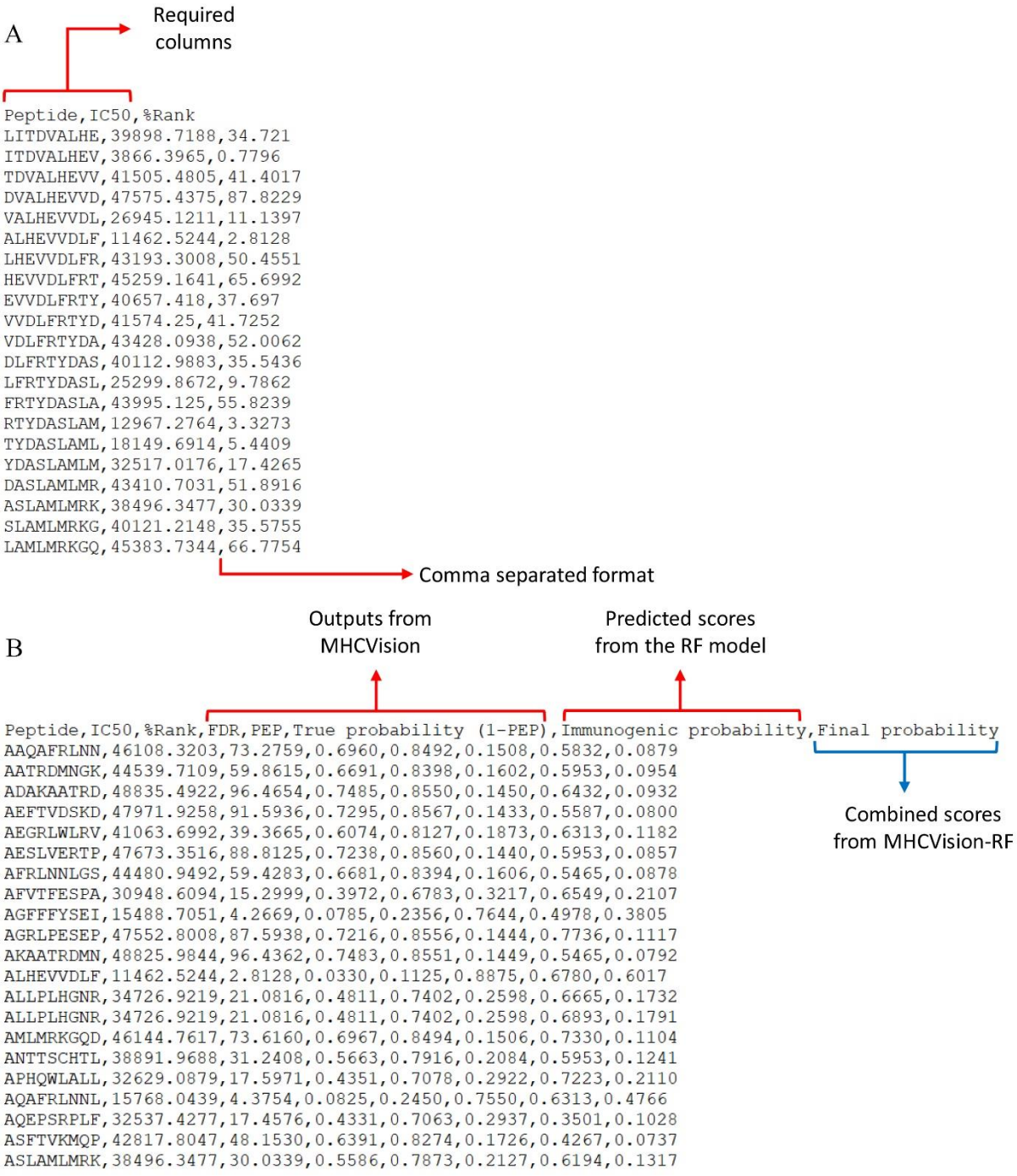

Figure 5.3 The example of input and output files of MHCVision-RF. The example of an input file format (A) and the columns written in an output file (B).

5.3.3 Assessment of the final probability of MHC binding and T cell recognition with data sets from published studies

To assess the separation ability of MHCVision-RF for neoantigen selection, the assessment analysis was performed by using data from two previously published studies. The list of 9 to 11 mers peptides of three patients with available HLA data in the experiment from Patrick Ott *et al.*, 2020 was used to apply with MHCVision-RF. Peptides obtained from each patient consist of both immunogenic and non-immunogenic peptides. The distribution of final probability scores between the immunogenic and non-immunogenic groups of M1 and L7 patients is not clearly different (Figure 5.4A and 5.4D, respectively). While the final probability of the immunogenic group showed higher median than that from non-immunogenic peptides in an M3 patient for both HLA-A*11:01 and A*68:01 (Figure 5.4B and 5.4C, respectively). Especially in A*11:01 of M3, the result displayed a significant difference of mean between immunogenic and non-immunogenic probabilities (student's t-test, p-value = 0.0015; Figure 5.4B). A limitation of this analysis is that it was performed against only one or two HLA alleles of a patient that is available in the publication (more detail in Section 5.2.3), but a person can express at least three HLA alleles and potentially up to six alleles. As such, it is possible that some low probabilities found in immunogenic peptides might be due to the peptide binding to other HLA types carried by the patients that we have not been able to predict.

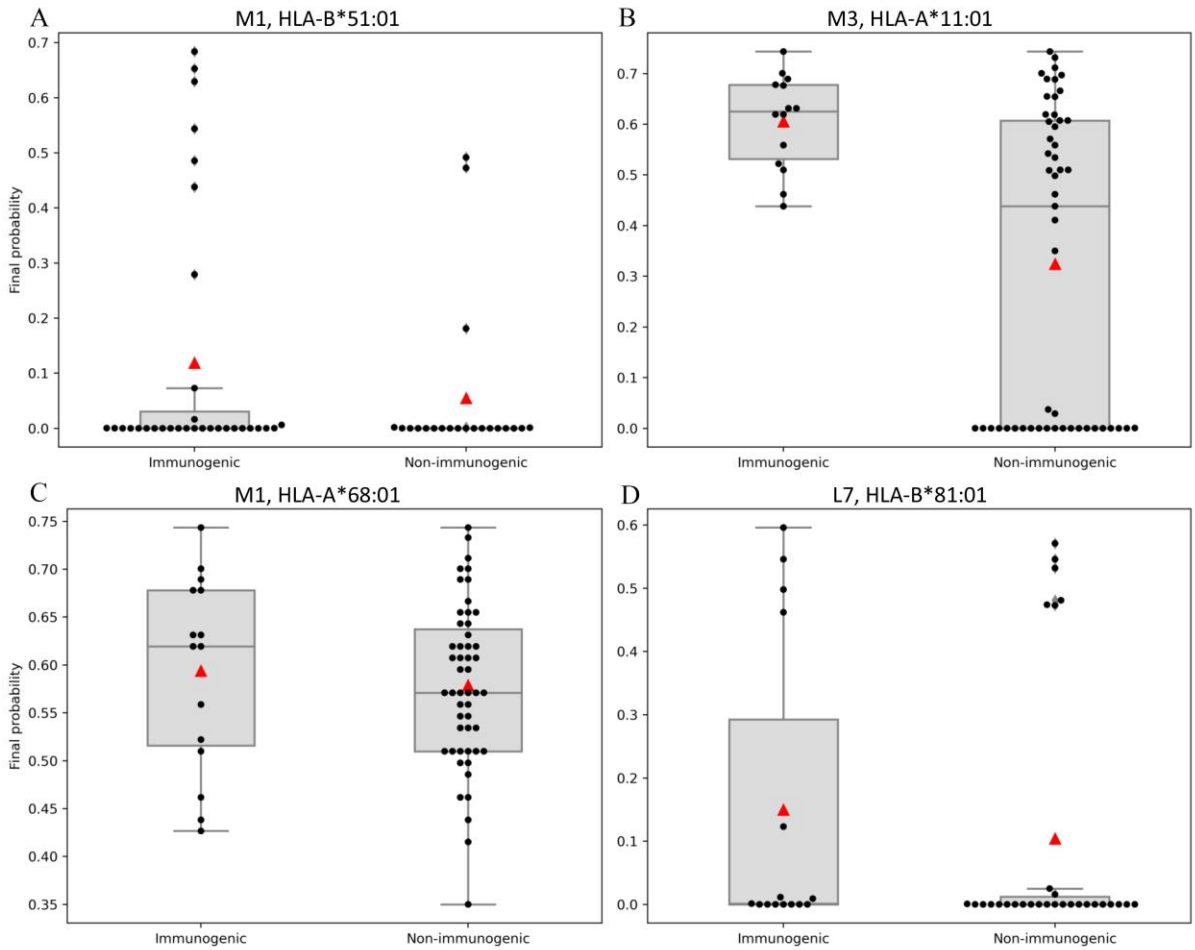

Figure 5.4 The final probability of immunogenic and non-immunogenic peptides from melanoma patients (M1 and M3) and a lung cancer patient (L7) (Patrick Ott *et al.*, 2020). Each box plot represents the final probability scores, mean (red triangle), and median derived from immunogenic and non-immunogenic peptides. The analysis was performed against to HLA-B*51:01 for M1 (A), HLA-A*11:01 (B) and A*68:01 (C) for M3, and HLA-B*81:01 for L7 (D) [187].

Due to the limitation of HLA information, the experimental data from Yong Fang *et al.*, 2020 was also used to evaluate the pipeline because this work provided all HLA types of each patient. In that study, they reported peptides in a long sequence format that are designed by adding amino acid sequences to short mutated peptides from the prediction methods described in the paper. To reverse the process of neoantigen prediction, long peptides were needed to chop into 9mers which is the common length for MHC class I ligands. Since they pooled two or three peptides for T cell activity assay, it is a limitation to specifically identify which peptide in the

pool can trigger T cell activity. Therefore, among all patients, four patients who has all positive pools (P003 = Pos1 and P004 = Pos2) or all negative pools (P011 = Neg1 and P016 = Neg2) were selected to perform the analysis. The predicted results from all HLA alleles of each patient were combined, and the top 20 scores were selected to observe the difference between positive and negative class. The analysis result was found that the highest 20 ranking scores of Pos2 and Neg1 are obviously different, and those from positive group have substantially greater than negative group (Figure 5.5). However, the ranking scores from Pos1 are not different from Neg1 and Neg2, and the scores of Neg2 has high outliers. In other words, we do not observe that patients with positive reactions generally are predicted to have peptides with higher immunogenicity. This result might be due to variation from external factors relating to different cancer types, general overall disease burden or a personal genetic background. Therefore, the comparison among patients with different cancer types might not be appropriate to represent a difference between ranking scores of immunogenic and non-immunogenic peptides because the ability of immune response from individuals are different. It might be possible that the experimental result from Neg2 might not be due to wrong selection neoantigen but might be affected by immunodeficiency of the patient.

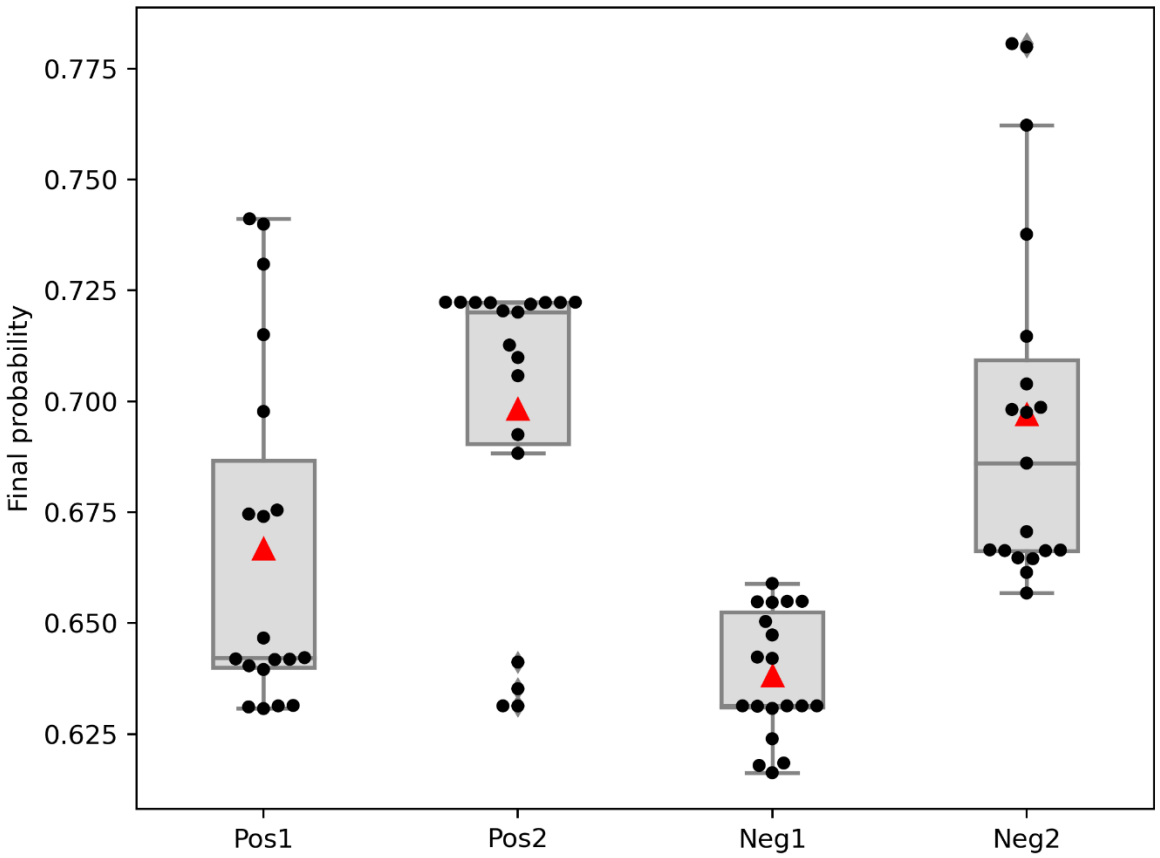

Figure 5.5 The top 20 final probability scores of data obtained from positive and negative pooled peptides (Yong Fang *et al.*, 2020). A box plot represents data of the highest final probability scores, mean (red triangle), and median of positive and negative pooled peptides from four cancer patients (Pos1: Adrenal Sebaceous Adenocarcinoma (P003), Pos2: Small Cell Lung Cancer (P004), Neg1: Ovarian Cancer (P011), and Neg2: Non-Small Cell Lung Cancer (P016)) [188].

To avoid the factor of immunogenetics across patients, a comparison within the same patient was performed using the data set from a patient who has both positive and negative pools (P001). The result showed that the ranking scores from positive pools are much higher than those from negative pools (student's t-test, p-value < 0.00001; Figure 5.6A). Moreover, the scores with control 10% FDR, that means all peptides in both positive and negative group are potentially binding peptides, still showed higher scores in positives than negatives (student's t-test, p-value = 0.0001; Figure 5.6B) indicating that the differentiation is not only classified by MHC-binding affinity prediction, but also from immunogenicity prediction. In summary, even

if there are some limitations because of lacking full information for neoantigen prediction analysis, results in the assessment analysis indicate ranking scores can distinguish neoepitopes from non-neoepitopes suggesting that the ranking score produced by the pipeline can assist with true neoantigen selection.

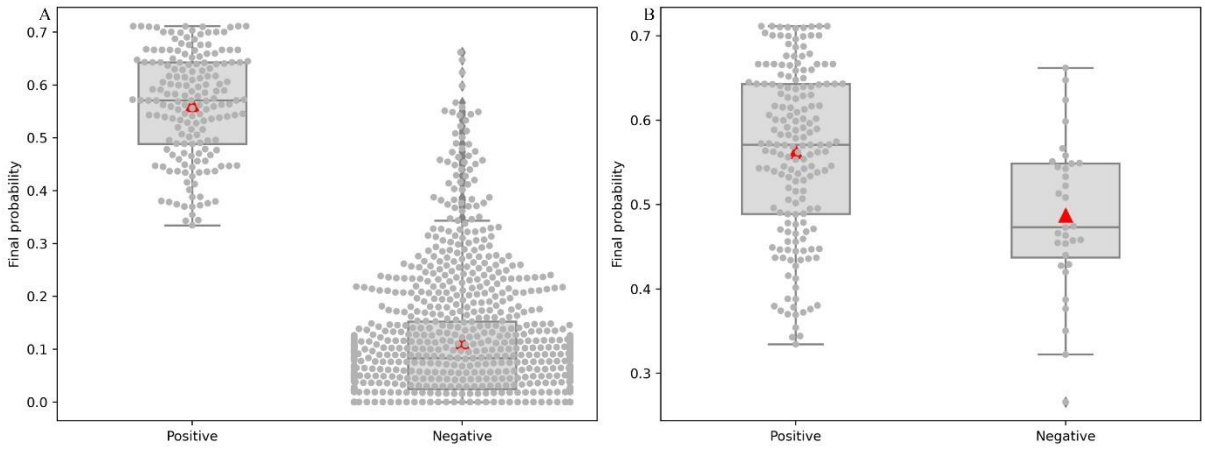

Figure 5.6 The final probability of positive and negative pooled peptides from one cancer patient (Yong Fang *et al*., 2020). The box plots represent the data of final probability scores, mean (red triangle), and median of positive and negative pooled peptides from a melanoma patient (P001) with non-control FDR (A) and within 10% FDR (B) [188].

5.4 Discussion

Identification of neoantigen from NGS data utilising the approach of bioinformatics is a complicated task involving several processes of biological sample preparation, bioinformatic analysis of NGS data, computational prediction, and candidate neoantigen selection. Most efforts of neoantigen prediction focus on a strength of MHC-peptide binding affinity using MHC-peptide binding prediction tools to exclude non-binding peptides. Some predictions also incorporate the biological processes of antigen processing including proteasomal cleavage e.g. NetChop [189] and peptide transports efficiency e.g. NetCTL [190] , or a stability between peptides and MHC molecules e.g. NetMHCStab [142], those information are expected to help to rule out false binding peptides. It has been reported that about thousands of somatic mutations are identified in most neoantigen studies, and hundreds of peptides are predicted as

MHC binding peptides, however, only a handful are found to elicit T cell response [191]. Therefore, there is a high risk to get false positive neoantigen if the selection method relies on MHC-peptide binding. Even MHC binding and antigen processing are necessary process for being neoantigens, they might not be sufficient for determining true neoantigen because an ability of immunisation is obligatory for being an epitope. The final probability from the pipeline in this study is the combination between the same weight of scores from true MHC binding probability and immunogenicity probability, which can be served as ranking scores that can help users to rank candidate neoantigens and make a short list from the top rank scores. Although the final probability from the multiplicative function might have a risk to be awry in case some input data might not fit well to beta mixtures distribution in MHCVision or not accurately predict from the immunogenicity predication, users can alternatively consider those probabilities separately provided in the output table (Figure 5.3B). The multiplication without weight consideration from two variables of MHCVision-RF is a much simpler mathematic operation than the method for calculating priority scores from MuPeXI which uses the negative logistic function contributed by six values from MHC-peptide binding affinity, RNA level expression, variant allele frequency, and similarity between self and mutated peptides [116].

The expression of genes that neoepitopes originate from is the most essential for neoantigen based cancer vaccine therapy in practice because it is meaningless to inject non-expressed peptides into cancer patients. However, RNA expression is not included in the final probability produced by this current model because gene expression levels are personalised data that is specific for an individual. The major limitation for building a model with the prediction of gene expression is RNA level is dynamic and highly tissue specific, which means that different types of cancer or different stages of the same cancer types might have different sets of gene expression. For the best result with using MHCVision-RF for neoantigen identification, users can apply gene expression levels, e.g. TPM > 1 is typically used as a threshold for gene

expression, to rule out non-expressed peptides that have high final probability scores so that could reduce risk for getting false neoantigens. Moreover, the study of neoantigen expression in breast cancer revealed that the high expression in RNA level where neoantigens originate is correlated with improved patient survival in a cohort of breast cancer patients [192]. Therefore, the RNA expression level could not only reduce a risk of getting false positives but also increase the chance of getting immunogenic neoantigens.

The benchmarking analysis with existing tools that provide prioritising scores for neoantigen identification have not done yet in this chapter because the complete of WES and RNA sequencing data as well as the experimental validation are needed for the comparison analysis. A benchmark analysis would require using the raw FASTQ data to control pre-processing steps i.e. sequence alignment, variant calling, quantification of transcripts, and HLA genotyping to ensure that the quality of an input data is same for any software. Besides genomic and transcriptomic data, the experiments of T cell reactivity are necessary to validate the immunogenicity of selected candidate neoantigens. The clinical study of neoantigen based cancer vaccine is ongoing at Dr. Trairak Pisitkun's research centre (CUSB), Chulalongkorn University. Therefore, once the complete data from experimental could be accessed, the comparison analysis between MHCVision-RF and exiting pipelines will be performed and published, if results are encouraging.

5.5 Conclusion

In this chapter, the MHCVision-RF pipeline was built from the integration of MHCVision and the immunogenicity prediction model. Scores from those two models were multiplied to produce a ranking score that can contributes to MHC-peptide binding and immunogenicity of each predicted peptides. The capability of ranking scores produced by this software was validated with data from published studies, and those ranking scores can differentiate data

between positive and negative class. Finally, the source code of MHCVision-RF was implemented and available at https://github.com/PGB-LIV/MHCVision-RF.

Chapter 6

General discussion, conclusion, and future work

The aim of this chapter is to summarise the content of this thesis, extend the general discussion of the models performed in each chapter and describe a perspective of future work related to this current research. Finally, the general conclusions of this work are summarised at the end.

6.1 Summary of thesis

The work carried out in this thesis can be summarised in three main components: first, the development of the model for estimating global and local FDR for MHC-peptide binding affinity predicted data; second, the development of a prediction algorithm for determining a probability of immunogenicity, and third, the integration of those two models to produce a pipeline that provides a probability of true MHC binding probability and T cell recognition probability. The following presents a summary of the conclusions from each chapter.

Chapter 1

This chapter is the general introduction that provides information related to the background of the adaptive immune system, cancer immunotherapies, neoantigen based cancer vaccines, and the methodology of neoantigen prediction. The MHC I-peptide binding prediction algorithms summarised in this chapter were used to produce the data for further analysis in the following chapters. Moreover, the basis of T cell preferences described in this chapter was used for developing the immunogenicity prediction model.

Chapter 2

This chapter demonstrated the practicability of neoantigen identification using the approach of bioinformatics and prediction algorithms. Moreover, the concept of protein structural analysis using the MD simulation technique was performed, and the limitations of this method were reported. The analysis of the random background and false positive rate of the outputs produced by MHC-peptide binding prediction algorithms (NetMHCpan and MHCflurry) were performed to gain a better understanding of the behaviour of those prediction tools. The analysis

summarised in this chapter is the rationale for the study of the improvement of criteria for neoantigen selection based on predicted MHC-peptide binding affinity and immunogenicity.

Chapter 3

This chapter initially described the background of statistical data distribution models and the mathematical context of the EM algorithm. This chapter described the best fit of the beta mixture distribution model for predicted data produced by NetMHCpan. The modification of the EM algorithm with the method of moments for beta parameter estimation was also included in this chapter. The application of this developed model, named as MHCVision, for global and local FDR estimation was demonstrated at the end of this chapter, and these results have been published in *Bioinformatics* [193].

Chapter 4

In this chapter, the development of the model for immunogenicity prediction was described. This chapter included the explanation of features related to T cell preference properties and the Random Forest algorithm that is used for building the immunogenicity prediction model. The performance and comparison analysis of this model against existing tools was demonstrated at the end of this chapter.

Chapter 5

The assembly of MHCVision from Chapter 3 and the immunogenicity prediction model developed in Chapter 4 was performed in this chapter, called MHCVision-RF. A generation of the final probability from a combination of true MHC binding probability and immunogenicity probability was explained. In addition, the code implementation and workflow of this pipeline were also described. Finally, the assessment of the separation ability of the model to distinguish neoantigen and non-neoantigen from published data sets was demonstrated in the final part of the chapter.

6.2 General Discussion

The selection of candidate neoantigens is a crucial step to enable the usage of neoantigen based cancer vaccines in clinical practice. There are several factors related to biological events including antigen processing, MHC-peptide binding, and T cell recognition that must be carefully considered to determine whether a candidate peptide could be a neoantigen. At present, computational methods typically exploit the prediction of binding affinity between MHC molecules and peptides as the primary judge to distinguishing putative neoantigens from non-neoantigens. The precision and accuracy of MHC-peptide binding affinity prediction is therefore important for neoantigen identification. Although the current benchmark of HLA class I binding prediction results showed the best performance of 90% sensitivity and 98% specificity, there remains a high risk of getting false positives if an inappropriate threshold is used, and prior to this work, we are now aware of a straightforward method for quantifying this phenomenon. FDR is an acceptable statistical value to control false positive rate in predicted results. The study in Chapter 3 is a novel perspective in the field of MHC-peptide binding affinity prediction. Instead of focussing on the improvement of MHC-peptide binding prediction algorithms, the model developed in Chapter 3, MHCVision, emphasised providing statistical values, global and local FDR, for predicted scores coming from MHC-peptide binding prediction tools. The statistical values produced by MHCVision can serve as additional information to facilitate users to define binding peptides. Moreover, the performance of MHCVision is independent of the accuracy of prediction tools because this model estimates global and local FDRs from the data distribution of the predicted result. Apart from neoantigen selection, MHCVision can be further applied for any kind of work related to the prediction of MHC-peptide binding affinity. It is worth stating potential limitations and caveats to this work. First, we have tested the algorithm under a range of scenarios, where we believe we have good control of the ground truth i.e. through mixing true (MS identified peptides) and false (random)

data points in different ratios. However, within both sets, there is potential for imperfections. Within MS data sets, as we comment in Chapter 3, it is possible that some incorrect peptides have been identified, which would not be true positives. Similarly, some proportion of random peptides are indeed true binders. Nevertheless, MHCVision was not trained on the labels *per se*, these were used to generate the types of distribution shapes we expect will be encountered in real data sets. It is of course possible that certain peptide set – allele combinations could produce completely unexpected data distributions which we have never seen before, e.g. multi-modal, which might cause some inaccuracies for MHCVision prediction. We have tested the performance of MHCVision with NetMHCpan and MHCflurry, two of the best performing and most popular binding algorithms, but we cannot guarantee performance with other predictors, and the model would likely need retraining for MHC Class II prediction, which is a more complex problem. Further detail on how this could be done is given below in Future Work.

As mentioned above, there are several factors involved in biological events that are normally used to consider for neoantigen identification. Beyond MHC binding ability, ability for being T cell epitopes is a critical property to be a neoantigen. Determining T cell epitopes is very challenging because of the extreme diversity of TCRs and the limitation of T cell epitope data. T cell epitopes obtained from T cell assays experiments deposited in databases do not fully cover all types of TCRs diversity. Thus, the immunogenicity prediction model developed in Chapter 4 is not a completely novel framework, since the database of physicochemical properties of amino acids for generating a set of features and data for training the model have been also used in previous studies. However, a new aspect of this developed model is concerning the bias from binding and non-binding classification. T cell epitopes must be MHC binding peptides, thus T cell epitopes derived from experiments mostly have a good MHC binding affinity. In contrast, non-epitopes that have a negative result from T cell reactivity assays might not be able to bind to MHC molecules, meaning they are not presented to T cells.

If epitope and non-epitope data are used as positive and negative classes to train the model without calibrating the MHC binding affinity, it might be possible that the model will learn binding and non-binding properties from amino acids in a peptide instead of properties for T cell recognition. The standardisation of positive and negative data for model training in Chapter 4 can help to rule out that bias, and build on the fact that very extensive work has already been done to build excellent classifiers for MHC-peptide binding. Even though the performance of this developed model might not reach 90% accuracy, it still outperforms the existing tools. There are several ways in which to improve this immunogenicity prediction model in the future. This model was built from training data that does not consider HLA allele specific peptides and trained only for peptides with nine amino acids in length. In this thesis, the model was built to be assembled with MHCVision to produce a probability that describe the ability for a peptide to be a strong MHC binder and immunogenic.

The integration of MHCVision and the immunogenicity prediction model was implemented to build a pipeline named MHCVision-RF in Chapter 5. The final probability produced from this pipeline is from the combination of true MHC binding probability and immunogenicity probability. As described above, MHC-peptide binding scores of the training data for the immunogenicity prediction model were standardised between epitopes and non-epitopes. Hence, it can ensure that a combined score is generated from the independent scores from MHC binding and immunogenicity probability. Although a probability provided from MHCVision-RF does not describe whether the peptide can be expressed in protein level, users can manually apply the RNA expression levels to consider in neoantigen selection or prioritisation to get the best results with a low risk of getting false neoantigens. At the moment we do not have any sufficiently large training data to know how to calibrate or combine the immunogenicity/binding probability with gene/protein abundance data, in terms of the importance of contribution from each. This must be an area for future focus.

6.3 Future work

The following topics are projects that could extend from the current works in this thesis and should be performed in the future.

6.3.1 Extensibility of MHCVision model

The current version of MHCVision is available for predicted data produced by NetMHCpan and MHCflurry. Extending the application of MHCVision for other MHC class I prediction tools should be further performed to give more flexibility for users. The shape of data distribution produced from other prediction tools must be explored, if they can fit well to beta mixture distributions, it could be possible to use the current version with those tools. However, if their data distribution is not well modelled by a beta mixture distribution, the data distribution model and constrained values might need to be modified. Furthermore, the development of the MHCVision for MHC class II binding prediction also should be performed for the next version. To extend the MHCVision algorithm to MHC class II, the data sets of natural MHC II presented peptides from MS analysis or other biological experiments must be collected for the learning phase, and the core concept of the algorithm of the current version with minor modification could apply to the distribution of predicted data produced from MHC class II-peptide binding affinity prediction tools.

6.3.2 The development of the automated software for neoantigen identification by assembling a package of bioinformatic software to MHCVision-RF

To make MHCVision-RF more practical for clinical research or application, automated software with the upstream steps of data pre-processing and downstream for neoantigen prioritisation should be assembled. The steps of pre-processing data include WES data analysis, variant calling, HLA-genotyping, and short mutated peptides extraction, whereby users could opt to provide the input file either a raw FASTQ format or a variant calling file format. The

main prediction part will take a list of short mutated peptides and patients' HLA alleles to the MHC-peptide binding prediction tool, then MHCVision-RF will compute the final probability of true MHC binding and T cell recognition for each peptide. If the RNA sequencing data is available, users could provide a raw FASTQ file or a level of transcripts file for the input parameters, the expression level of genes that mutated peptide originate from would be considered together with the final probability from MHCVision-RF to prioritise or select candidate neoantigens. Finally, the output would return data in a tabular format containing the multiple scores of each peptide, which are predicted binding affinity scores, true MHC binding probability, immunogenicity probability, and a final probability of MHC binding and T cell recognition as well as gene expression level (if applicable). The installation of the automated software pipeline is planned as a step of neoantigen identification in future projects on the development of neoantigen based cancer vaccine at Chulalongkorn University to test the workability of this software in the clinical level.

6.4 General conclusion

Vaccines are a type of immunotherapy, which normally protects people from diseases. They are generally made from weakened or innocuous versions of pathogens. When people get vaccinated, their immune system will be stimulated, and naïve T cells will be active and develop to memory T cells specific to those pathogens. With the same concept of immunostimulants but unlike general vaccines that are used for protection, cancer vaccines are designed for people who already have cancer. Cancer vaccines are typically designed from a part of protein particularly expressed in cancer cells but not expressed in normal cells, i.e. neoantigens. Once cancer patients get a vaccine formulated from neoantigens, the immune system will recognise those neoantigens to attack and destroy the cancer cells that carry those neoantigens. The approach of personalised neoantigen-based cancer vaccines might be feasible for various types of cancer compared to other cancer immunotherapies, such as monoclonal

antibodies or CAR T cells because the generation of neoantigens relies on the individual genetic background. Moreover, the potential of this approach on neoantigen-specific T cells activation provides the development and proliferation of memory T cells that might achieve long-term protection against disease recurrence.

In general, the processes for obtaining a list of neoantigens based on NGS data primarily exploit software packages in bioinformatics and computational methods. One of the limitations of peptide-based cancer vaccines is that a small handful of peptides are practically selected for the step of peptide manufacture and vaccine production due to cost and time effectiveness. The current criteria for neoantigen selection utilising the information provided by the prediction algorithms might yield too many numbers of peptides to proceed in vaccine production. This thesis was mainly focused on the development of the models to facilitate the improvement of the criteria for selecting and prioritising potent neoantigens to make a shortlist of candidate neoantigens with a low risk of getting false positives. In this thesis, the two models that contribute true MHC binding probability and immunogenicity probability were successfully developed, and the integration of those two models provided a pipeline producing the final probability that describes the ability of MHC binding and a potent for being T cell epitopes. Finally, the pipeline developed in this thesis can provide a probability score that can describ a potent for being real neoantigen. The software and the source code of MHCVision and MHCVision-RF are freely available at https://github.com/PGB-LIV/MHCVision and https://github.com/PGB-LIV/MHCVision-RF, respectively.

References

1. Roberts, K., et al., *Molecular biology of the cell.* New York: Garland Science, 2002.
2. Janeway, C.A., et al., *Immunobiology*. 2001: Taylor & Francis Group UK: Garland Science.
3. Krangel, M.S., *Mechanics of T cell receptor gene rearrangement.* Current opinion in immunology, 2009. **21**(2): p. 133-139.
4. Turner, S.J., et al., *Structural determinants of T-cell receptor bias in immunity.* Nature reviews immunology, 2006. **6**(12): p. 883-894.
5. Rock, K.L., E. Reits, and J. Neefjes, *Present yourself! By MHC class I and MHC class II molecules.* Trends in immunology, 2016. **37**(11): p. 724-737.
6. Choo, S.Y., *The HLA system: genetics, immunology, clinical testing, and clinical implications.* Yonsei medical journal, 2007. **48**(1): p. 11-23.
7. Unanue, E.R., *From antigen processing to peptide-MHC binding.* Nature immunology, 2006. **7**(12): p. 1277-1279.
8. Robinson, J., et al., *IPD-IMGT/HLA Database.* Nucleic acids research, 2020. **48**(D1): p. D948-D955.
9. Mack, S.J., et al., *Common and well-documented HLA alleles: 2012 update to the CWD catalogue.* Tissue antigens, 2013. **81**(4): p. 194-203.
10. Hurley, C.K., *Naming HLA diversity: a review of HLA nomenclature.* Human immunology, 2020.
11. Marsh, S.G., et al., *An update to HLA nomenclature, 2010.* Bone marrow transplantation, 2010. **45**(5): p. 846-848.
12. Michalek, M.T., et al., *A role for the ubiquitin-dependent proteolytic pathway in MHC class I-restricted antigen presentation.* Nature, 1993. **363**(6429): p. 552-554.
13. Rock, K.L., et al., *Inhibitors of the proteasome block the degradation of most cell proteins and the generation of peptides presented on MHC class I molecules.* Cell, 1994. **78**(5): p. 761-771.
14. Reits, E., et al., *Peptide diffusion, protection, and degradation in nuclear and cytoplasmic compartments before antigen presentation by MHC class I.* Immunity, 2003. **18**(1): p. 97-108.
15. Cresswell, P., et al., *The nature of the MHC class I peptide loading complex.* Immunological reviews, 1999. **172**(1): p. 21-28.
16. Stern, L.J., et al., *Crystal structure of the human class II MHC protein HLA-DR1 complexed with an influenza virus peptide.* Nature, 1994. **368**(6468): p. 215-221.
17. Suri, A., S.B. Lovitch, and E.R. Unanue, *The wide diversity and complexity of peptides bound to class II MHC molecules.* Current opinion in immunology, 2006. **18**(1): p. 70-77.
18. Unanue, E.R., V. Turk, and J. Neefjes, *Variations in MHC class II antigen processing and presentation in health and disease.* Annual review of immunology, 2016. **34**: p. 265-297.
19. Ghosh, P., et al., *The structure of an intermediate in class II MHC maturation: CLIP bound to HLA-DR3.* Nature, 1995. **378**(6556): p. 457-462.
20. Denzin, L.K. and P. Cresswell, *HLA-DM induces CLIP dissociation from MHC class II αβ dimers and facilitates peptide loading.* Cell, 1995. **82**(1): p. 155-165.
21. Pos, W., et al., *Crystal structure of the HLA-DM–HLA-DR1 complex defines mechanisms for rapid peptide selection.* Cell, 2012. **151**(7): p. 1557-1568.
22. Borg, N.A., et al., *The CDR3 regions of an immunodominant T cell receptor dictate the'energetic landscape'of peptide-MHC recognition.* Nature immunology, 2005. **6**(2): p. 171-180.
23. Zoete, V., et al., *Structure-based, rational design of T cell receptors.* Frontiers in immunology, 2013. **4**: p. 268.
24. Schuster, M., A. Nechansky, and R. Kircheis, *Cancer immunotherapy.* Biotechnology Journal: Healthcare Nutrition Technology, 2006. **1**(2): p. 138-147.
25. Tan, S., D. Li, and X. Zhu, *Cancer immunotherapy: Pros, cons and beyond.* Biomedicine & Pharmacotherapy, 2020. **124**: p. 109821.
26. Galon, J. and D. Bruni, *Approaches to treat immune hot, altered and cold tumours with combination immunotherapies.* Nature reviews Drug discovery, 2019. **18**(3): p. 197-218.

27. Swann, J.B. and M.J. Smyth, *Immune surveillance of tumors.* The Journal of clinical investigation, 2007. **117**(5): p. 1137-1146.
28. Soothill, J., *Immunological Surveillance.* Journal of Clinical Pathology, 1971. **24**(4): p. 372.
29. Dunn, G.P., et al., *Cancer immunoediting: from immunosurveillance to tumor escape.* Nature immunology, 2002. **3**(11): p. 991-998.
30. Vukusic, P., et al., *Now you see it—now you don't.* Nature, 2001. **410**(6824): p. 36-36.
31. Diefenbach, A., et al., *Rae1 and H60 ligands of the NKG2D receptor stimulate tumour immunity.* Nature, 2001. **413**(6852): p. 165-171.
32. Bromberg, J.F., et al., *Transcriptionally active Stat1 is required for the antiproliferative effects of both interferon alpha and interferon gamma.* Proceedings of the national academy of sciences, 1996. **93**(15): p. 7673-7678.
33. Pardoll, D.M., *The blockade of immune checkpoints in cancer immunotherapy.* Nature Reviews Cancer, 2012. **12**(4): p. 252.
34. Cameron, F., G. Whiteside, and C. Perry, *Ipilimumab.* Drugs, 2011. **71**(8): p. 1093-1104.
35. Remon, J. and B. Besse, *Immune checkpoint inhibitors in first-line therapy of advanced non-small cell lung cancer.* Current opinion in oncology, 2017. **29**(2): p. 97-104.
36. Rosenberg, S.A. and N.P. Restifo, *Adoptive cell transfer as personalized immunotherapy for human cancer.* Science, 2015. **348**(6230): p. 62-68.
37. Kochenderfer, J.N., et al., *Adoptive transfer of syngeneic T cells transduced with a chimeric antigen receptor that recognizes murine CD19 can eradicate lymphoma and normal B cells.* Blood, 2010. **116**(19): p. 3875-3886.
38. Tran, E., P.F. Robbins, and S.A. Rosenberg, *'Final common pathway'of human cancer immunotherapy: targeting random somatic mutations.* Nature immunology, 2017. **18**(3): p. 255-262.
39. Verdegaal, E.M., et al., *Neoantigen landscape dynamics during human melanoma–T cell interactions.* Nature, 2016. **536**(7614): p. 91-95.
40. Perumal, D., et al., *Mutation-derived neoantigen-specific T-cell responses in multiple myeloma.* Clinical Cancer Research, 2020. **26**(2): p. 450-464.
41. Peng, M., et al., *Neoantigen vaccine: an emerging tumor immunotherapy.* Molecular cancer, 2019. **18**(1): p. 1-14.
42. Li, L., S. Goedegebuure, and W.E. Gillanders, *Preclinical and clinical development of neoantigen vaccines.* Annals of Oncology, 2017. **28**: p. xii11-xii17.
43. Ott, P.A., et al., *An immunogenic personal neoantigen vaccine for patients with melanoma.* Nature, 2017. **547**(7662): p. 217.
44. Sahin, U., et al., *Personalized RNA mutanome vaccines mobilize poly-specific therapeutic immunity against cancer.* Nature, 2017. **547**(7662): p. 222.
45. Franzese, O., et al., *Drug-induced xenogenization of tumors: A possible role in the immune control of malignant cell growth in the brain?* Pharmacological Research, 2018. **131**: p. 1-6.
46. Wirth, T.C. and F. Kühnel, *Neoantigen targeting—dawn of a new era in cancer immunotherapy?* Frontiers in immunology, 2017. **8**: p. 1848.
47. Perez, C.R. and M. De Palma, *Engineering dendritic cell vaccines to improve cancer immunotherapy.* Nature communications, 2019. **10**(1): p. 1-10.
48. Khong, H. and W.W. Overwijk, *Adjuvants for peptide-based cancer vaccines.* Journal for immunotherapy of Cancer, 2016. **4**(1): p. 1-11.
49. Terbuch, A. and J. Lopez, *Next Generation Cancer Vaccines—Make It Personal!* Vaccines, 2018. **6**(3): p. 52.
50. Jahanafrooz, Z., et al., *Comparison of DNA and mRNA vaccines against cancer.* Drug discovery today, 2020. **25**(3): p. 552-560.
51. Castle, J.C., et al., *Exploiting the mutanome for tumor vaccination.* Cancer research, 2012. **72**(5): p. 1081-1091.

52. Yadav, M., et al., *Predicting immunogenic tumour mutations by combining mass spectrometry and exome sequencing.* Nature, 2014. **515**(7528): p. 572-576.
53. Carreno, B.M., et al., *A dendritic cell vaccine increases the breadth and diversity of melanoma neoantigen-specific T cells.* Science, 2015. **348**(6236): p. 803-808.
54. Rutledge, W.C., et al., *Tumor-infiltrating lymphocytes in glioblastoma are associated with specific genomic alterations and related to transcriptional class.* Clinical Cancer Research, 2013. **19**(18): p. 4951-4960.
55. Hilf, N., et al., *Actively personalized vaccination trial for newly diagnosed glioblastoma.* Nature, 2019. **565**(7738): p. 240-245.
56. Keskin, D.B., et al., *Neoantigen vaccine generates intratumoral T cell responses in phase Ib glioblastoma trial.* Nature, 2019. **565**(7738): p. 234-239.
57. Bassani-Sternberg, M., *Mass spectrometry based immunopeptidomics for the discovery of cancer neoantigens*, in *Peptidomics*. 2018, Springer. p. 209-221.
58. Bassani-Sternberg, M., et al., *Direct identification of clinically relevant neoepitopes presented on native human melanoma tissue by mass spectrometry.* Nature communications, 2016. **7**(1): p. 1-16.
59. Zhang, X., et al., *Application of mass spectrometry-based MHC immunopeptidome profiling in neoantigen identification for tumor immunotherapy.* Biomedicine & Pharmacotherapy, 2019. **120**: p. 109542.
60. Garcia-Garijo, A., C.A. Fajardo, and A. Gros, *Determinants for neoantigen identification.* Frontiers in immunology, 2019. **10**: p. 1392.
61. Kote, S., et al., *Mass Spectrometry-based identification of MHC-associated peptides.* Cancers, 2020. **12**(3): p. 535.
62. Richters, M.M., et al., *Best practices for bioinformatic characterization of neoantigens for clinical utility.* Genome medicine, 2019. **11**(1): p. 1-21.
63. El Achi, H., J.D. Khoury, and S. Loghavi, *Liquid biopsy by next-generation sequencing: a multimodality test for management of cancer.* Current hematologic malignancy reports, 2019. **14**(5): p. 358-367.
64. De Mattos-Arruda, L. and G. Siravegna, *How to use liquid biopsies to treat patients with cancer.* ESMO open, 2021. **6**(2): p. 100060.
65. Fleri, W., et al., *The immune epitope database and analysis resource in epitope discovery and synthetic vaccine design.* Frontiers in immunology, 2017. **8**: p. 278.
66. Zhang, H., C. Lundegaard, and M. Nielsen, *Pan-specific MHC class I predictors: a benchmark of HLA class I pan-specific prediction methods.* Bioinformatics, 2009. **25**(1): p. 83-89.
67. Reynisson, B., et al., *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, 2020.
68. O'Donnell, T.J., et al., *MHCflurry: open-source class I MHC binding affinity prediction.* Cell systems, 2018. **7**(1): p. 129-132. e4.
69. Boegel, S., et al., *Bioinformatic methods for cancer neoantigen prediction.* Progress in molecular biology and translational science, 2019. **164**: p. 25-60.
70. Fang, H., et al., *Reducing INDEL calling errors in whole genome and exome sequencing data.* Genome medicine, 2014. **6**(10): p. 1-17.
71. Belkadi, A., et al., *Whole-genome sequencing is more powerful than whole-exome sequencing for detecting exome variants.* Proceedings of the National Academy of Sciences, 2015. **112**(17): p. 5473-5478.
72. Nam, J.-Y., et al., *Evaluation of somatic copy number estimation tools for whole-exome sequencing data.* Briefings in bioinformatics, 2016. **17**(2): p. 185-192.
73. Griffith, M., et al., *Optimizing cancer genome sequencing and analysis.* Cell systems, 2015. **1**(3): p. 210-223.

74. Li, H. and R. Durbin, *Fast and accurate long-read alignment with Burrows–Wheeler transform.* Bioinformatics, 2010. **26**(5): p. 589-595.
75. Van der Auwera, G.A., et al., *From FastQ data to high-confidence variant calls: the genome analysis toolkit best practices pipeline.* Current protocols in bioinformatics, 2013. **43**(1): p. 11.10. 1-11.10. 33.
76. Schneider, V.A., et al., *Evaluation of GRCh38 and de novo haploid genome assemblies demonstrates the enduring quality of the reference assembly.* Genome research, 2017. **27**(5): p. 849-864.
77. Li, H., et al., *The sequence alignment/map format and SAMtools.* Bioinformatics, 2009. **25**(16): p. 2078-2079.
78. Institute, B., *Picard Tools. http://broadinstitute.github.io/picard.* Accessed 1 Feb 2021.
79. Tarasov, A., et al., *Sambamba: fast processing of NGS alignment formats.* Bioinformatics, 2015. **31**(12): p. 2032-2034.
80. Cibulskis, K., et al., *Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples.* Nature biotechnology, 2013. **31**(3): p. 213-219.
81. Saunders, C.T., et al., *Strelka: accurate somatic small-variant calling from sequenced tumor–normal sample pairs.* Bioinformatics, 2012. **28**(14): p. 1811-1817.
82. Koboldt, D.C., et al., *VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing.* Genome research, 2012. **22**(3): p. 568-576.
83. Callari, M., et al., *Intersect-then-combine approach: improving the performance of somatic variant calling in whole exome sequencing data using multiple aligners and callers.* Genome medicine, 2017. **9**(1): p. 1-11.
84. Fang, L.T., et al., *An ensemble approach to accurately detect somatic mutations using SomaticSeq.* Genome biology, 2015. **16**(1): p. 1-13.
85. Danecek, P., et al., *The variant call format and VCFtools.* Bioinformatics, 2011. **27**(15): p. 2156-2158.
86. McLaren, W., et al., *The ensembl variant effect predictor.* Genome biology, 2016. **17**(1): p. 1-14.
87. Mortazavi, A., et al., *Mapping and quantifying mammalian transcriptomes by RNA-Seq.* Nature methods, 2008. **5**(7): p. 621-628.
88. Liao, Y., G.K. Smyth, and W. Shi, *featureCounts: an efficient general purpose program for assigning sequence reads to genomic features.* Bioinformatics, 2014. **30**(7): p. 923-930.
89. Trapnell, C., et al., *Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks.* Nature protocols, 2012. **7**(3): p. 562.
90. Pertea, M., et al., *Transcript-level expression analysis of RNA-seq experiments with HISAT, StringTie and Ballgown.* Nature protocols, 2016. **11**(9): p. 1650.
91. Trapnell, C., et al., *Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation.* Nature biotechnology, 2010. **28**(5): p. 511-515.
92. Patro, R., S.M. Mount, and C. Kingsford, *Sailfish enables alignment-free isoform quantification from RNA-seq reads using lightweight algorithms.* Nature biotechnology, 2014. **32**(5): p. 462-464.
93. Bray, N.L., et al., *Near-optimal probabilistic RNA-seq quantification.* Nature biotechnology, 2016. **34**(5): p. 525-527.
94. Patro, R., et al., *Salmon provides fast and bias-aware quantification of transcript expression.* Nature methods, 2017. **14**(4): p. 417-419.
95. Rapaport, F., et al., *Comprehensive evaluation of differential gene expression analysis methods for RNA-seq data.* Genome biology, 2013. **14**(9): p. 1-13.
96. Teng, M., et al., *A benchmark for RNA-seq quantification pipelines.* Genome biology, 2016. **17**(1): p. 1-12.

97. Baruzzo, G., et al., *Simulation-based comprehensive benchmarking of RNA-seq aligners.* Nature methods, 2017. **14**(2): p. 135-139.
98. Bunce, M. and B. Passey, *HLA typing by sequence-specific primers*, in *Transplantation Immunology*. 2013, Springer. p. 147-159.
99. Cereb, N., et al., *Advances in DNA sequencing technologies for high resolution HLA typing.* Human immunology, 2015. **76**(12): p. 923-927.
100. Bauer, D.C., et al., *Evaluation of computational programs to predict HLA genotypes from genomic sequencing data.* Briefings in bioinformatics, 2018. **19**(2): p. 179-187.
101. Kiyotani, K., T.H. Mai, and Y. Nakamura, *Comparison of exome-based HLA class I genotyping tools: identification of platform-specific genotyping errors.* Journal of human genetics, 2017. **62**(3): p. 397-405.
102. Szolek, A., et al., *OptiType: precision HLA typing from next-generation sequencing data.* Bioinformatics, 2014. **30**(23): p. 3310-3316.
103. Bai, Y., D. Wang, and W. Fury, *PHLAT: inference of high-resolution HLA types from RNA and whole exome sequencing*, in *HLA Typing*. 2018, Springer. p. 193-201.
104. Abelin, J.G., et al., *Mass spectrometry profiling of HLA-associated peptidomes in mono-allelic cells enables more accurate epitope prediction.* Immunity, 2017. **46**(2): p. 315-326.
105. Bulik-Sullivan, B., et al., *Deep learning using tumor HLA peptide mass spectrometry datasets improves neoantigen identification.* Nature biotechnology, 2019. **37**(1): p. 55-63.
106. Jurtz, V., et al., *NetMHCpan-4.0: Improved peptide–MHC class I interaction predictions integrating eluted ligand and peptide binding affinity data.* The Journal of Immunology, 2017. **199**(9): p. 3360-3368.
107. Zhao, W. and X. Sher, *Systematically benchmarking peptide-MHC binding predictors: From synthetic to naturally processed epitopes.* PLoS computational biology, 2018. **14**(11): p. e1006457.
108. Sette, A., et al., *The relationship between class I binding affinity and immunogenicity of potential cytotoxic T cell epitopes.* The Journal of Immunology, 1994. **153**(12): p. 5586-5592.
109. Andreatta, M. and M. Nielsen, *Gapped sequence alignment using artificial neural networks: application to the MHC class I system.* Bioinformatics, 2016. **32**(4): p. 511-517.
110. Karosiene, E., et al., *NetMHCcons: a consensus method for the major histocompatibility complex class I predictions.* Immunogenetics, 2012. **64**(3): p. 177-186.
111. O'Donnell, T.J., A. Rubinsteyn, and U. Laserson, *MHCflurry 2.0: Improved Pan-Allele Prediction of MHC Class I-Presented Peptides by Incorporating Antigen Processing.* Cell Systems, 2020. **11**(1): p. 42-48. e7.
112. Shao, X.M., et al., *High-throughput prediction of MHC class i and ii neoantigens with MHCnuggets.* Cancer Immunology Research, 2020. **8**(3): p. 396-408.
113. Phloyphisut, P., et al., *MHCSeqNet: a deep neural network model for universal MHC binding prediction.* BMC bioinformatics, 2019. **20**(1): p. 270.
114. Hundal, J., et al., *pVAC-Seq: A genome-guided in silico approach to identifying tumor neoantigens.* Genome medicine, 2016. **8**(1): p. 1-11.
115. Hundal, J., et al., *pVACtools: a computational toolkit to identify and visualize cancer neoantigens.* Cancer immunology research, 2020. **8**(3): p. 409-420.
116. Bjerregaard, A.-M., et al., *MuPeXI: prediction of neo-epitopes from tumor sequencing data.* Cancer Immunology, Immunotherapy, 2017. **66**(9): p. 1123-1130.
117. Kim, S., et al., *Neopepsee: accurate genome-level prediction of neoantigens by harnessing sequence and amino acid immunogenicity information.* Annals of Oncology, 2018. **29**(4): p. 1030-1036.
118. Stranzl, T., et al., *NetCTLpan: pan-specific MHC class I pathway epitope predictions.* Immunogenetics, 2010. **62**(6): p. 357-368.
119. Li, Y., et al., *ProGeo-neo: a customized proteogenomic workflow for neoantigen prediction and selection.* BMC medical genomics, 2020. **13**: p. 1-11.

120. Coelho, A., et al., *neoANT-HILL: an integrated tool for identification of potential neoantigens.* 2020.
121. Zhou, C., et al., *pTuneos: p rioritizing tu mor neo antigens from next-generation s equencing data.* Genome medicine, 2019. **11**(1): p. 1-17.
122. Wood, M.A., et al., *neoepiscope improves neoepitope prediction with multivariant phasing.* Bioinformatics, 2020. **36**(3): p. 713-720.
123. Wang, T.-Y., et al., *ScanNeo: identifying indel-derived neoantigens using RNA-Seq data.* Bioinformatics, 2019. **35**(20): p. 4159-4161.
124. Richman, L.P., R.H. Vonderheide, and A.J. Rech, *Neoantigen dissimilarity to the self-proteome predicts immunogenicity and response to immune checkpoint blockade.* Cell systems, 2019. **9**(4): p. 375-382. e4.
125. Schenck, R.O., et al., *NeoPredPipe: high-throughput neoantigen prediction and recognition potential pipeline.* BMC bioinformatics, 2019. **20**(1): p. 1-6.
126. Smart, A.C., et al., *Intron retention is a source of neoepitopes in cancer.* Nature biotechnology, 2018. **36**(11): p. 1056-1058.
127. Rubinsteyn, A., et al., *Vaxrank: A computational tool for designing personalized cancer vaccines.* bioRxiv, 2017: p. 142919.
128. Zhou, Z., et al., *TSNAD: an integrated software for cancer somatic mutation and tumour-specific neoantigen detection.* Royal Society open science, 2017. **4**(4): p. 170050.
129. Chang, T.-C., et al., *The neoepitope landscape in pediatric cancers.* Genome medicine, 2017. **9**(1): p. 1-12.
130. Bais, P., et al., *CloudNeo: a cloud pipeline for identifying patient-specific tumor neoantigens.* Bioinformatics, 2017. **33**(19): p. 3110-3112.
131. Zhang, W., et al. *Predicting immunogenic T-cell epitopes by combining various sequence-derived features*. in *2013 IEEE International Conference on Bioinformatics and Biomedicine*. 2013. IEEE.
132. Kotturi, M.F., et al., *Naive precursor frequencies and MHC binding rather than the degree of epitope diversity shape CD8+ T cell immunodominance.* The Journal of Immunology, 2008. **181**(3): p. 2124-2133.
133. Assarsson, E., et al., *A quantitative analysis of the variables affecting the repertoire of T cell specificities recognized after vaccinia virus infection.* The Journal of Immunology, 2007. **178**(12): p. 7890-7901.
134. Calis, J.J., et al., *Properties of MHC class I presented peptides that enhance immunogenicity.* PLoS Comput Biol, 2013. **9**(10): p. e1003266.
135. Wucherpfennig, K.W., et al., *Structural alterations in peptide–MHC recognition by self-reactive T cell receptors.* Current opinion in immunology, 2009. **21**(6): p. 590-595.
136. Hoof, I., et al., *Interdisciplinary analysis of HIV-specific CD8+ T cell responses against variant epitopes reveals restricted TCR promiscuity.* The Journal of Immunology, 2010. **184**(9): p. 5383-5391.
137. Rammensee, H.-G., et al., *SYFPEITHI: database for MHC ligands and peptide motifs.* Immunogenetics, 1999. **50**(3): p. 213-219.
138. Vita, R., et al., *The immune epitope database (IEDB): 2018 update.* Nucleic acids research, 2019. **47**(D1): p. D339-D343.
139. Vita, R., et al., *The immune epitope database (IEDB) 3.0.* Nucleic acids research, 2015. **43**(D1): p. D405-D412.
140. Vita R, M.S., Overton JA, Dhanda SK, Martini S, Cantrell JR, Wheeler DK, Sette A, Peters B., *www.iedb.org.* accessed 20 April 2021, 2021.
141. Sidney, J., B. Peters, and A. Sette. *Epitope prediction and identification-adaptive T cell responses in humans*. in *Seminars in Immunology*. 2020. Elsevier.
142. Jørgensen, K.W., et al., *Net MHC stab–predicting stability of peptide–MHC-I complexes; impacts for cytotoxic T lymphocyte epitope discovery.* Immunology, 2014. **141**(1): p. 18-26.

143. Trolle, T. and M. Nielsen, *NetTepi: an integrated method for the prediction of T cell epitopes.* Immunogenetics, 2014. **66**(7): p. 449-456.
144. De Neuter, N., et al., *On the feasibility of mining CD8+ T cell receptor patterns underlying immunogenic peptide recognition.* Immunogenetics, 2018. **70**(3): p. 159-168.
145. Ogishi, M. and H. Yotsuyanagi, *Quantitative prediction of the landscape of T cell epitope immunogenicity in sequence space.* Frontiers in immunology, 2019. **10**: p. 827.
146. Wang, G., et al., *INeo-Epp: A novel T-cell HLA class-I Immunogenicity or neoantigenic epitope prediction method based on sequence-related amino acid features.* BioMed research international, 2020. **2020**.
147. Hsiue, E.H.-C., et al., *Targeting a neoantigen derived from a common TP53 mutation.* Science, 2021. **371**(6533): p. eabc8697.
148. Martin, M., *Cutadapt removes adapter sequences from high-throughput sequencing reads.* EMBnet. journal, 2011. **17**(1): p. 10-12.
149. Andrews, S., *FastQC: a quality control tool for high throughput sequence data*. 2010, Babraham Bioinformatics, Babraham Institute, Cambridge, United Kingdom.
150. McKenna, A., et al., *The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data.* Genome research, 2010. **20**(9): p. 1297-1303.
151. McLaren, W., et al., *The ensembl variant effect predictor.* Genome biology, 2016. **17**(1): p. 122.
152. Liu, C., et al., *ATHLATES: accurate typing of human leukocyte antigen through exome sequencing.* Nucleic acids research, 2013. **41**(14): p. e142-e142.
153. Nielsen, M. and M. Andreatta, *NetMHCpan-3.0; improved prediction of binding to MHC class I molecules integrating information from multiple receptor and peptide length datasets.* Genome medicine, 2016. **8**(1): p. 1-9.
154. Borbulevych, O.Y., et al., *TCRs used in cancer gene therapy cross-react with MART-1/Melan-A tumor antigens via distinct mechanisms.* The Journal of Immunology, 2011. **187**(5): p. 2453-2463.
155. BIOVIA, D.S., *Discovery Studio Modeling Environment.* Release 2017.
156. Olsson, M.H., et al., *PROPKA3: consistent treatment of internal and surface residues in empirical p K a predictions.* Journal of chemical theory and computation, 2011. **7**(2): p. 525-537.
157. D.A. Case, V.B., J.T. Berryman, R.M. Betz, Q. Cai, D.S. Cerutti, T.E. Cheatham, III, T.A. Darden, R.E., et al., *AMBER 14.* 2014.
158. Tsui, V. and D.A. Case, *Theory and applications of the generalized Born solvation model in macromolecular simulations.* Biopolymers: Original Research on Biomolecules, 2000. **56**(4): p. 275-291.
159. *UniProt: the universal protein knowledgebase.* Nucleic acids research, 2017. **45**(D1): p. D158-D169.
160. Huang, D., et al., *Mutations of key driver genes in colorectal cancer progression and metastasis.* Cancer and Metastasis Reviews, 2018. **37**(1): p. 173-187.
161. Paul, S., et al., *Benchmarking predictions of MHC class I restricted T cell epitopes in a comprehensively studied model system.* PLOS Computational Biology, 2020. **16**(5): p. e1007757.
162. Weinstein, J.N., et al., *The cancer genome atlas pan-cancer analysis project.* Nature genetics, 2013. **45**(10): p. 1113.
163. Castle, J.C., *Mutation-derived neoantigens for cancer immunotherapy.* Frontiers in immunology, 2019. **10**: p. 1856.
164. Dempster, A.P., N.M. Laird, and D.B. Rubin, *Maximum likelihood from incomplete data via the EM algorithm.* Journal of the Royal Statistical Society: Series B (Methodological), 1977. **39**(1): p. 1-22.
165. Chen, S., *The application of the expectation-maximization algorithm to the identification of biological models*. 2006, Virginia Tech.

166. Zhang, Z., *Parameter estimation techniques: A tutorial with application to conic fitting.* Image and vision Computing, 1997. **15**(1): p. 59-76.
167. Do, C.B. and S. Batzoglou, *What is the expectation maximization algorithm?* Nature biotechnology, 2008. **26**(8): p. 897-899.
168. Pearson, K., *Contributions to the mathematical theory of evolution.* Philosophical Transactions of the Royal Society of London. A, 1894. **185**: p. 71-110.
169. Mishra, S. and A. Datta-Gupta, *Applied statistical modeling and data analytics: A practical guide for the petroleum geosciences*. 2017: Elsevier.
170. Schröder, C. and S. Rahmann, *A hybrid parameter estimation algorithm for beta mixtures and applications to methylation state classification.* Algorithms for Molecular Biology, 2017. **12**(1): p. 21.
171. Casella, G. and R.L. Berger, *Statistical inference*. 2021: Cengage Learning.
172. Lyon, A., *Why are normal distributions normal?* The British Journal for the Philosophy of Science, 2014. **65**(3): p. 621-649.
173. Ji, Y., et al., *Applications of beta-mixture models in bioinformatics.* Bioinformatics, 2005. **21**(9): p. 2118-2122.
174. Owen, C.E.B., *Parameter estimation for the beta distribution.* 2008.
175. Sarkizova, S., et al., *A large peptidome dataset improves HLA class I epitope prediction across most of the human population.* Nature Biotechnology, 2020. **38**(2): p. 199-209.
176. Schittenhelm, R.B., et al., *A comprehensive analysis of constitutive naturally processed and presented HLA-C* 04: 01 (Cw4)–specific peptides.* Tissue antigens, 2014. **83**(3): p. 174-179.
177. Solleder, M., et al., *Mass spectrometry based immunopeptidomics leads to robust predictions of phosphorylated HLA class I ligands.* Molecular & Cellular Proteomics, 2020. **19**(2): p. 390-404.
178. Stickel, J.S., et al., *HLA ligandome analysis of chronic myeloid leukemia (CML), revealed novel tumor associated antigens for peptide based immunotherapy*. 2013, American Society of Hematology Washington, DC.
179. Zeng, H. and D.K. Gifford, *Quantification of uncertainty in peptide-MHC binding prediction improves high-affinity peptide Selection for therapeutic design.* Cell systems, 2019. **9**(2): p. 159-166. e3.
180. Kawashima, S. and M. Kanehisa, *AAindex: amino acid index database.* Nucleic acids research, 2000. **28**(1): p. 374-374.
181. Zhou, X., et al., *A context-free encoding scheme of protein sequences for predicting antigenicity of diverse influenza A viruses.* BMC genomics, 2018. **19**(10): p. 145-154.
182. Speiser, J.L., et al., *A comparison of random forest variable selection methods for classification prediction modeling.* Expert systems with applications, 2019. **134**: p. 93-101.
183. Palczewska, A., et al., *Interpreting random forest classification models using a feature contribution method*, in *Integration of reusable systems*. 2014, Springer. p. 193-218.
184. Cai, J., et al., *Feature selection in machine learning: A new perspective.* Neurocomputing, 2018. **300**: p. 70-79.
185. Breiman, L., *Random forests.* Machine learning, 2001. **45**(1): p. 5-32.
186. Tao, T., *Standalone BLAST setup for Unix.* BLAST® Help [Internet]. Bethesda (MD): National Center for Biotechnology Information (US), 2008.
187. Ott, P.A., et al., *A phase Ib trial of personalized neoantigen therapy plus anti-PD-1 in patients with advanced melanoma, non-small cell lung cancer, or bladder cancer.* Cell, 2020. **183**(2): p. 347-362. e24.
188. Fang, Y., et al., *A pan-cancer clinical study of personalized neoantigen vaccine monotherapy in treating patients with various types of advanced solid tumors.* Clinical Cancer Research, 2020. **26**(17): p. 4511-4520.
189. Saxová, P., et al., *Predicting proteasomal cleavage sites: a comparison of available methods.* International immunology, 2003. **15**(7): p. 781-787.

190. Larsen, M.V., et al., *An integrative approach to CTL epitope prediction: a combined algorithm integrating MHC class I binding, TAP transport efficiency, and proteasomal cleavage predictions.* European journal of immunology, 2005. **35**(8): p. 2295-2303.
191. Editorial, N., *The problem with neoantigen prediction.* Nat. Biotechnol, 2017. **35**: p. 97-97.
192. Li, W., et al., *Impact of Neoantigen Expression and T-Cell Activation on Breast Cancer Survival.* Cancers, 2021. **13**(12): p. 2879.
193. Pearngam, P., et al., *MHCVision: estimation of global and local false discovery rate for MHC class I peptide binding prediction.* Bioinformatics, 2021.