IN SILICO IDENTIFICATION OF UNIVERSAL HLA STIMULATING B AND T-CELL RESTRICTED MAGE EPITOPES FOR VACCINE DEVELOPMENT

AWINO MAUREIQ EDITH OJWANG'

I56/64134/2010

A thesis submitted to the University of Nairobi in partial fulfillment for the award of Master of Science degree in Bioinformatics

2013

DECLARATION

I declare that this thesis is my own work and to the best of my knowledge has never been submitted or examined for the award of degree in any university.

Awino Maureiq Edith Ojwang'

Registration Number: I56/64134/2010

Signature; ……. Date; **……**

APPROVAL

This thesis has been submitted with our approval as university supervisors

Anton Dormer, MD MS
Associate Professor - Biology Department
Program Coordinator - Bioinformatics Interdisciplinary Program (BIP)
Washington Adventist University
Signature; …….
Date; ……

Dr. Musembi, R. J, PhD
Lecturer - Center of Biotechnology and Bioinformatics
University of Nairobi
Signature; …….
Date; **……**

Prof James Ochanda, PhD
Director – Center of Biotechnology and Bioinformatics (CEBIB)
University of Nairobi
Signature; …….
Date; **……**

Prof Walter Mwanda, PhD
Director – Institute of Tropical and Infectious Diseases (UNITID)
University of Nairobi
Signature; …….
Date; **……**

ACKNOWLEGDEMENTS

I acknowledge my sovereign God-Almighty Father, Lord Jesus Christ and Most Holy Spirit-who has made this work possible. I am sincerely grateful to my father, Dr. Ayoma Ojwang' for devotedly supporting me and paying for all stages of my education. Father, you are a precious gift from God. I am very indebted to my supervisor, Dr. Anton Dormer, for giving me an excellent opportunity to work with him in this amazing field. Dr. Dormer, you have been a model supervisor, serene instructor and a great mentor. I acknowledge Prof James Ochanda for the untiring guidance and critic in every stage of this research work. Prof, thank you for pushing me to finish and graduate in time. I thank you; Dr. Musembi for your invaluable input in computational section of this project, Prof Walter Mwanda, for your guidance through the medical aspects this work and Mr Dan Achinko, for your sincere proof reading and relentless assistance in bioinformatics platforms. Many thanks to the Centre for Biotechnology and Bioinformatics (CEBIB) group; lecturers and non-academic staff, for equipping me with both intellectual and life lessons. I am immensely grateful to my mother, Mrs. Rose Ojwang' and siblings; William Onyango, Lameck Ayoma and Mercy Ojwang' for their unconditional love and unwavering support. A special thanks to my dearest friends; Sr. Tessie Matanguihan Mendoza and Brian Omondi for incessant encouragement and prayers, Gregory Macharia for constant support and to Hillary Omondi and Christabel Ogola Ochieng for hardware assistance. Many thanks to my colleagues: Rosaline Macharia, Peterson Mathenge, Harrison Ndung'u and Eric Maina, with whom I shared joyous and challenging moments in school. Finally, I thank the University of Nairobi for giving me an opportunity to study in the best Bioinformatics program in the country.

Maureiq, 2013

DEDICATION

TO MY FAMILY

DR. AYOMA OJWANG'

For being the definition of a true father and an epitome of excellence

MRS. ROSE OJWANG'

She loved, strengthened and encouraged me all my life

WILLIAM, LAMECK AND MERCY

For being the source of joy, unwavering support and determination to succeed

Table of Contents

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

ACC- Auto cross covariance

AFP- Alpha (α)–fetoprotein

ANN- Artificial neural network

APC- Antigen presenting cell

ARB- Average relative binding

ASCI- Antigen-specific cancer immunotherapeutic

BAGE- B melanoma antigen

BCR- B cell receptor

CAGE - Cancer-associated antigen

CALR- Calreticulin

CD- Cluster of differentiation

cDNA- Complimentary deoxyribonucleic acid

CDR- Complementary differential region

CDS- Coding sequence

CEA- Carcinoembryonic antigen

CLIP- Class II associated-invariant chain peptide

CTL- Cytotoxic T lymphocytes

DC- Dendrite cell

DNA- Deoxyribonucleic acid

ER- Endoplasmic reticulum

GAGE- G antigen

Gp100- Glycoprotein 100

HER-2/neu- Human epidermal receptor-2/neurological

HLA- Human leukocyte antigen

HPV- Human Papilloma virus

IC50- Half maximal (50%) inhibitory concentration (IC)

IEDB- Immune epitope database

MAGE- Melanoma antigen

MART/Melan-A- Melanoma antigen recognized by T cells-1/Melanoma antigen A

MHC- Major histocompatibility complex

MHD- MAGE homology domain

mRNA- Messenger Ribonucleic acid

MSA- Multiple sequence alignment

NCBI- National centre for biotechnology information

NMR- Nuclear magnetic resonance

NY-ESO-1- Newyork esophagus squamous carcinoma 1

PAP- Prostatic acid phosphates

PCR- Polymerase chain reaction

PDB- Protein Data Bank

PSA- Prostate specific antigen

PSMA- Prostate specific membrane antigen

RMSD- Root mean square deviation

RNA- Ribonucleic acid

RNA- Ribonucleic acid

RT-PCR- Reverse transcriptase polymerase chain reaction

SAGE- Sarcoma antigen

siRNA- Small interfering ribosomal nucleic acid

SMM- Stabilized matrix method

SSX- Synovial sarcoma X

SVM- Support vector machine

TA- Tumor Antigen

TAA- Tissue associated antigen

TAP- Transporter associated with antigen processing

TCR- T cell receptor

Th1/Th2- T helper lymphocyte1/2

TM-PAP- Trans-membrane prostatic acid phosphates

TRP-Tyrosinase related protein

TSA- Tissue specific antigen

ABSTRACT

Melanoma antigens (MAGE) are immunogens expressed in malignancies but silenced in somatic tissues except testis. They are grouped into ten subfamilies and at least one subfamily is expressed in a cancer type. In vaccinology, their tumor specific expression eliminates autoimmunity, antitumor effects arrests tumorigenesis and epitopes' promiscuity matches human leukocyte antigen polymorphism. MAGE based vaccines in clinical trials are limited to subfamilies and target single tumors. Many epitopes have been reported but an analysis of universal epitopes that cut across all subfamilies has never been investigated. This *in silico* study was focused on finding conserved epitopes within the subfamilies. A conserved motif analysis of all proteomes was done using ClustalO, Jalview and Cytoscape tools. There was a notable absence of strong conservation explained as a consequence of weak functional constraints during gene evolution. An analysis of the antigens' extracellular topology was achieved using TMHMM server and all antigens except MAGE H1 were found to be extracellular. B and T cell epitopes within the extracellular conserved motifs were then identified via a pipeline of predictive servers (BCPreds, IEDB, ProPred I and II, MHCPred and T-epitope designer) and tested for antigenicity using VaxiJen server. For results, 20 antigenic B cell epitopes (18-20 mer) and >10 T cell epitopes (8-15 mer) binding to human leukocyte antigen alleles; HLA-A*0201, -A*0204, -B*2705, -DRB1*0101, and -DRB1*0401, are given. 8-mer T cell epitopes were found to be the most conserved and are thus considered universal epitopes. However, other epitopes (9-15 mer) are only conserved within subfamilies hence strings of epitopes were formed to get universal polytopes. These findings will inform the design of a multivalent universal MAGE vaccine that targets more than nineteen tumors. This pipeline confirms six reported epitopes (from *in vitro* studies) thus showing the efficacy of using *in silico* tools in epitope prediction.

Keywords

MAGE, HLA, epitope, ClustalO, Jalview, Cytoscape, TMHMM server, BCPreds, IEDB, ProPred I and II, MHCPred, T-epitope designer, VaxiJen

CHAPTER ONE

INTRODUCTION

1.1 General Introduction

Cancer refers to a group of life threatening diseases that are characterised by unabated growth of abnormal cells that can form tumours. Resultant tumours are in many cases immunogenic; they express new antigens that are recognized as foreign by the immune system therefore eliciting a response against the malignancies (Chiarella *et al*., 2009). Melanoma antigens (MAGE) are a type of these new antigens that are expressed in an extensive variety of tumours of different histologies (Oba-Shinjo *et al*., 2008; Kim *et al*., 2001; Sudo *et al*., 1997; Mori *et al*., 1996; Shichijio *et al*., 1995;). Many studies have shown that expression of MAGE antigens in malignancies can trigger the adaptive immune response (Huang *et al*., 1999; Van der Bruggen *et al*., 1991; Van den Eynde *et al*., 1989). The awareness of this immune response has made it possible to target MAGE antigens for cancer immunotherapy (Bettinotti *et al*., 2003; Chianese-Bullock *et al*., 2005; Kim *et al*., 2008a; Toh *et al*., 2009). One way is to develop epitope based vaccines against MAGE antigens. A paradigm of the rational design of this vaccine involves identification of epitopes, *in vitro* analysis of these epitopes, vaccine production, *in vivo* studies in animal models and finally, clinical trials (Chiarella *et al*., 2009).

There are structural and functional experimental methods for epitope prediction. Structural methods (X-ray, Nuclear Magnetic Resonance (NMR), electron microscopy etc.) identify epitopes in complex with Major Histocompatibility Complex (MHC) molecules or paratope of B cell receptor (BCR). On the other hand, functional methods (mass spectrometry, NMR spectroscopy and immunoassays) involve *in vitro* synthesis and testing of epitopes. These experimental methods are tedious and expensive, a better approach would be to use *in silico* tools to predict and select a few candidate epitopes from a vast array of possible epitopes and then subject the few epitopes to experimental validation (Tomaselli *et al*., 2010; Chiarella *et al*., 2009). Several *in silico* prediction tools that offer accurate results have been developed and are significant bioinformatics resources (Guan *et al.,* 2003; Singh & Raghava, 2003; Mayrose *et al*., 2007; Vita *et al*., 2010). Now, the use of *in silico* tools is a standard approach in epitope vaccine development (Chiarella *et al*., 2009) and has been used to identify epitopes in; *Neisseria*

gonorrhea (Barh *et al*, 2010a), *Escherichia coli* 536 (Rai *et al*., 2012a), Ebola glycoproteins (Wu *et al*., 2012), MAGE antigens (Akiyama *et al*., 2012) etc.

The goal of this study was to identify B and T cell epitopes (HLA-epitope complexes) from conserved motifs that cut across all MAGE antigens using *in silico* tools. This multi-epitope approach for targeting B cell, cytotoxic and T helper epitopes is vital to stimulate both humoral and cell-mediated immune responses against endogenous and exogenous MAGE antigens. The proof that such epitopes can be investigated is because they have been successful in therapeutic vaccination trials on cancer patients (see Table 2.2).

1.2 Problem statement

MAGE based vaccines-that are currently in clinical trials-are derived from individual subfamilies and target single tumors (Chianese-Bullock *et al*., 2005; Kavanagh *et al*., 2007; Kim *et al*., 2008b). The implication is that several vaccines are being developed for different cancer types yet a single vaccine can be designed to target many MAGE expressing tumors. Also, not all MAGE antigens are used in developing these vaccines and as a result the full potential of these antigens is not exploited.

1.3 Research question

Can *in silico* tools be used to identify universal MAGE epitopes which can work on several MAGE expressing cell lines and therefore inform the design of a universal vaccine?

1.4 Objectives

1.4.1 General objective

To identify universal B and T cell epitopes derived from MAGE antigens.

1.4.2 Specific objectives

To identify and characterize:

- conserved motifs that cut across all MAGE antigens
- antigenic B cell epitopes
- immunogenic T cell epitopes

1.5 Justification

Although several MAGE vaccines have been successful in clinical trials, they are limited to single tumors and individual MAGE subfamilies (Bettinotti *et al*., 2003; Kavanagh *et al*., 2007; Kim *et al*., 2008a; Takahashi *et al*., 2012). This methodology is designed to identify vaccine candidates that can be used to create one vaccine for several MAGE expressing tumors. The approach in the long run will be more cost effective to make one vaccine vis-a-vi making several vaccines for every MAGE expressing tumor.

In addition, developing a multivalent vaccine will contribute to fighting cancer which now stands as the third leading cause of death worldwide. It is projected that by 2030, there will be ~26 million new cancer cases (Thun *et al.*, 2010). In Kenya, cancer is an economic and health burden ranking as the third main cause of death in the country (Rositch *et al*., 2012). Common cancers e.g. breast, throat, cervical and prostrate cancers, express MAGE antigens (Miyashiro *et al*., 2001; von Boehmer *et al*., 2011; Zhang *et al*., 2010b; Weinert *et al*., 2009). Therefore, the development of a single multivalent cancer vaccine that is effective for the prophylaxis or treatment of many tumors will help alleviate some of the economic burdens that such pathologies place on the Kenya health system

CHAPTER TWO

LITERATURE REVIEW

2.1 Tumor antigens

A tumor antigen (TA) is a protein that is expressed majorly by malignancies and can be recognized as foreign by CD8+ and CD4+ T cells. TAs are classified as tissue specific antigens (TSA) or tissue associated antigens (TAA). TSAs are expressed strictly in tumors as they are gene products of oncogenes, tumor suppressor genes, DNA translocations and isoforms. However, TAAs are expressed in tumors and in some normal cells. They include; oncofetal, tissue lineage, cell cycle regulatory, viral and cancer testis antigens (CTA). Basically, oncofetal antigens are epithelial antigens that are over expressed in tumors and fetal tissues e.g. carcinoembryonic antigen (CEA) and prostate specific membrane antigen (PSMA). Tissue lineage antigens are mostly pigment proteins that are normally expressed in tissues and over expressed in the resultant tumors of the same tissues e.g. prostate specific antigen (PSA) in prostate cancer and melanoma-associated antigens (Melan-A, tyrosinase, Glycoprotein 100 (gp100) and tyrosinase related protein (TRP)) in melanoma. Some products of oncogenes and tumor suppressor genes are over expressed in tumors and in some fetal tissues e.g. human epidermal growth factor 2 (HER-2), p53 and telomerase. These constitute the cell cycle regulatory antigens while viral antigens are products of viral oncogenes.

More importantly, CTAs are expressed in highly proliferating cancer cells but silenced in normal cells except for male germ line and placental cells (Kalejs & Erenpreisa, 2005a). DNA hypermethylation of the gene promoter regions and histones acetylation silence CTA expression in somatic tissues however, gene transcription is induced by hypomethylation during carcinogenesis allowing CTA expression in tumors (Kalejs & Erenpreisa, 2005a). Studies on adaptive immune response on a melanoma patient lead to the discovery of CTLs response on CTAs (Van der Bruggen *et al*., 1991). Since then, CTAs have been studied aggressively for cancer immunotherapy and they include; Melanoma antigens (MAGE), B melanoma antigens (BAGE), cancer associated gene (CAGE), helicase antigen (HAGE), G antigen (GAGE), L antigen (LAGE), New York esophagous-1 (NY-ESO-1), synovial sarcoma on X chromosome (SSX) among others (Kalejs & Erenpreisa, 2005a). MAGE were the initial CTAs to be

identified and are of interest in this study due to their differential expression in an extensive variety of malignancies and success in clinical trials.

2.2 Melanoma antigens

Melanoma antigens are immunogens that are expressed in a variety of tumors of diverse histological types: melanoma, multiple myeloma, ovarian cancer, classical Hodgkins lymphoma etc. (de Carvalho *et al*., 2012; Inaoka *et al*., 2011; Zhang *et al*., 2010a; van der Bruggen *et al*., 1991). They were originally identified and characterized in melanoma, hence the name melanoma antigen and the acronym MAGE. That was in 1991, during studies on the cell-mediated immune response on a melanoma patient MZ-2, that led to the discovery and characterization of MAGE A1, the first member of the MAGE family. Additionally, the first CTL epitope (MAGE A1-HLA-A1 complex) was also identified from this study (van der Bruggen *et al*., 1991). Interestingly, MAGE antigens and some CTAs are expressed in spermatogenic germ cells and are involved in spermatogenesis. MAGE antigens play a role in regulation of transcription and apoptosis process. Other TA's like CAGE and HAGE are involved in recombination DNA repair in the testis and tumors, SSX are repressors of transcription process, OY-TES-1 is involved in acrosome packaging etc. (Kalejs & Erenpreisa, 2005b).

MAGE antigens are encoded by X-linked and non X-linked genes. These genes and their corresponding proteins are classified into two types and total to 10 subfamilies (Katsura & Satta, 2011). Type I MAGE genes are classified into three subfamilies; MAGE A (1-12), B (1-18) and C (1-3). All these genes are considered recent and are localized on a palindrome on the long arm of the X chromosome. Their coding sequence (CDS) encodes ~300 amino acid long immunogenic MAGE proteins in highly proliferating cancerous cells (Katsura & Satta, 2011). Type II MAGE genes are classified as MAGE D through K, L2 and necdin. Unlike type I, type II MAGE genes are considered ancient and are found on the autosomes. They are universally expressed both in tumors and in normal cells and are involved in cell proliferation and apoptosis processes (Katsura & Satta, 2011). MAGE genes have homologs in chicken, fruit fly and zebra fish. While human and mice have multi-gene MAGE families, the mentioned homologous taxa have a single gene each (Katsura & Satta, 2011).

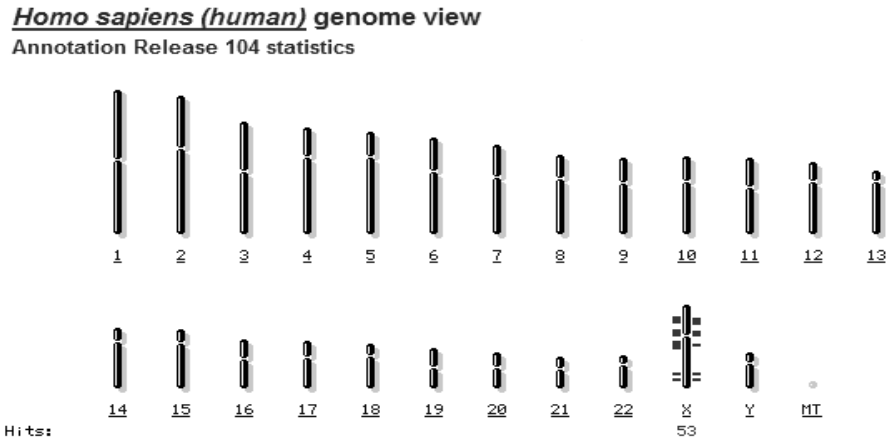


Figure 2.1: Location of type 1 MAGE genes on the X chromosome. This Map Viewer image was obtained from national center for biotechnology information (NCBI) database. The structures represent human chromosomes (1-22, X and Y). The lines on chromosome X show gene, gene prediction and sequence hits of MAGE genes

MAGE genes evolved in four phases from a single ancestral gene (phase I). While all MAGE genes have a one coding exon, MAGE D has fourteen exons. It is speculated that MAGE genes were formed from retro transposition of MAGE D and its processed genes. This retro transposition characterized phase II leading to the formation of eight MAGE members excluding MAGE C. All MAGE genes are only homologous in the CDS region thus evidence based on the similarity beyond the CDS signify that MAGE C members were formed as a duplication of MAGE A members. Phase III was marked with more duplications within each subfamily which led to formation of palindromes in MAGE A. Survival of these palindromes in humans marked phase IV (Katsura & Satta, 2011). All these genes share a 46% conserved region containing ~170 amino acids located at the carboxyl termini of the proteins (Yang *et al*., 2007). This region is called MAGE homology domain (MHD). This MHD is more conserved among sub families with 75% and 70% conservation among MAGE D and A members correspondingly (Yang *et al*., 2007). Studies show that MHD is involved in protein-protein interactions. Using *in vitro* assays, it has been shown that MAGE B18, C2 and G1 bind to E3 RING though the MHD (Yang *et al*., 2007).

MAGE antigens are attractive for cancer immunotherapy since they are genuine targets. Their expression in malignancies, antitumor effects, immunogenicity and even role in tumorigenesis has been verified *in vitro* and *in vivo*. Expression of MAGE in malignancies has been demonstrated by many researchers including Jungbluth and his coworkers who used

monoclonal antibodies to analyze MAGE expression in normal and cancerous tissues (Jungbluth *et al*., 2000). Positive results were noted in the testis, melanomas and carcinomas (head, lung and urinary).

Table 2.1: Expression of MAGE antigens in malignancies of different histologies

Malignancy	**Type**	**Reference**
Carcinoma	Melanoma	(van der Bruggen *et al*., 1991)
	Prostate cancer	(von Boehmer *et al*., 2011)
	Squamous cell lung carcinoma	(Bolli *et al*., 2002)
	Colorectal carcinoma	(Mori *et al*., 1996)
	Esophageal cancer	(Weinert *et al*., 2009)
	Gastric carcinoma	(Kim *et al*., 2001)
	Hepatocellular carcinoma	(Sarobe *et al*., 2004)
	Ovarian cancer	(Zhang *et al*., 2010b)
Sarcoma	Monophasic and biphasic synovial sarcoma	(Antonescu *et al*., 2002)
	Osteosarcoma	(Sudo *et al*., 1997)
Lymphoma and Leukemia	Classical Hogkins lymphoma	(Inaoka *et al*., 2011)
	Lymphocytic leukemia	(Shichijio *et al*., 1995)
	Acute myeloid leukemia	(Greiner *et al*., 2000)
	Multiple Myeloma	(de Carvalho *et al*., 2012)
Central nervous system tumor	Glioma	(Chi *et al*., 1997)
	Medulloblastoma	(Oba-Shinjo *et al*., 2008)
Germ cell tumor	Testicular germ cell tumor	(Aubry *et al*., 2001)

MAGE antigens have been tested for immunogenicity. Huang and his partners showed that MAGE A10-HLA-A2 epitope could be recognized by CTL in a melanoma cell line LB1751-MEL (Huang *et al*., 1999). Evidence of antitumor effect of MAGE antigens can be inferred from the study by Liu and coworkers. They delivered a recombinant adenovirus vector containing calrectin and MAGE A3 genes on glioblastoma cell line U87. The result was suppression of cell proliferation and invasion properties (Liu *et al*., 2012). The role of MAGE in tumorigenesis has also been documented. Expression of MAGE A1 and A3 was shown to be correlated with tumor stage and differentiation in ovarian cancer (Zhang *et al*., 2010b). Also, the application of siRNA suppression of MAGE A and B slows proliferation of cancer cells and induces apoptosis in human mast cell line (Yang *et al*., 2007). In addition, MAGE epitopes are promiscuous; they bind to multiple HLAs hence they serve as good vaccine targets since this promiscuous nature increases their probability of matching HLA polymorphism and thus cover a large population of

patients with different HLA allele variants (Buhler & Sanchez-Mazas, 2011; Parham *et al*., 1988).

2.3 Related work

2.3.1 Cancer immunotherapy

In cancer immunotherapy, tumor antigens are used in developing diagnostic markers, cancer vaccines, monoclonal/polyclonal antibodies and non- specific immunotherapies. Diagnostic markers are used to detect presence of cancer. MAGE A based reverse transcriptase PCR markers have been used to detect presence of mRNA in breast, melanoma and colorectal cancers and MAGE A4 has been shown as a potential diagnostic marker for testicular germ cell tumors (Miyashiro *et al*., 2001; Aubry *et al*., 2001). On the other hand, non-specific immunotherapy refers to synthetic immune components that boost the immune system in a general way. Some MAGE based vaccines in clinical trials are made in conjunction with synthetic cytokines (Chianese-Bullock *et al*., 2005), another example is NCT00020267 (see Table 2.2). Similarly, monoclonocal antibodies are synthetic antibodies that treat cancer by destroying cancer cells. They are being used to detect the presence of MAGE antigens i.e. MA454 against MAGE A1 (Carrel *et al*., 1996), 2B4 against MAGE A4 (Hennard *et al*., 2006), polyvalent 57B targeting MAGE A1, A3, A4, A6 and A12 (Jungbluth *et al*., 2000), among others.

Moving on, therapeutic vaccines now use advanced approaches that boost the immune response against cancer cells. They include; tumor cell and dendrite cell (DC) vaccines both created from a patient's cells and tailored to elicit aggressive immune response, DNA vaccines developed using recombinant DNA that codes for an immunogenic antigen and epitope vaccines constructed using immunogenic epitopes that are guaranteed to trigger CTL response (Chiarella *et al*., 2009). Epitope based vaccines are being investigated as they are more precise, easy to manufacture and have minimal side effects (Chiarella *et al*., 2009). Clinical trials using MAGE epitope based vaccines have been successful although they are limited to individual subfamilies and single tumors. An example is MAGE A3 antigen which part of GlaxoSmithKline's antigen-specific cancer immunotherapeutic (ASCI) pipeline and is currently in phase III trials for melanoma and non-small cell lung cancer and on phase II trial for bladder cancer

(http://www.gsk.com/research/our-product-pipeline.html). Other clinical trials with MAGE antigens are given in the Table 2.2 and Table 7.3 (see appendix).

Table 2.2: clinical trials of MAGE based vaccines obtained from National Cancer Institute's clinical trial database (http://www.clinicaltrials.gov)

Peptides	**Cancer**	**Phase**	**Clinicaltrials.gov Identifier**	**Sponsor**
MAGE A3	Cutaneous melanoma	I	NCT00706238	GlaxoSmithKline
MAGE A3, Melan-A and gp100	Metastatic melanoma	II	NCT00203879	University of Chicago
MAGE A3, gp100	Melanoma	II	NCT00254397	Anderson Cancer Centre
MAGE A12, interleukin-2 and montanide	Lung Cancer Colorectal Cancer Bone Cancer Ovarian sarcoma Melanoma	I	NCT00020267	National Cancer Institute (NCI)
MAGE A1, MAGE A3, survivin, gp100 and tyrosinase	Malignant Melanoma	I/II	NCT00204516	University hospital Tuebingen
MAGE A10 and Melan-A	Melanoma	I	NCT00112216	Ludwig Institute for Cancer Research
MAGE A3	Melanoma	II	NCT00042783	Southwest oncology group

2.3.2 Identification of epitopes

Even though many MAGE epitopes have been reported, they are derived from single subfamily peptides and target sole HLA alleles (see table 7.4 in appendices). An example is the recent work by Akiyama and his colleagues who used *in silico* peptide docking assay to identify MAGE A6 and A12 epitopes binding to HLA*A24 allele (Akiyama *et al*., 2012). An analysis of universal epitopes has only been attempted for MAGE A family by Graff-Dubois and his colleagues; however, an analysis of universal epitopes which cut across all MAGE subfamilies has never been fully investigated. Graff-Dubois and his co-workers did *in vivo* studies of CTL responses against given heterolytic epitopes in transgenic mice. They concluded that YLEYRQVPG and YLEYRQVPD are acceptable universal epitopes for HLA*0201 allele targeting MAGE A1, A2, A3, A4, A6, A10, A12 (Graff-Dubois *et al*., 2002). On a slightly similar note, Miyashiro and his co-workers developed a universal MAGE A reverse transcription PCR assay to detect mRNAs in multiple cancers expressing MAGE A antigens. They designed

universal MAGE A primers and probes using MAGE A1, A6, A3, A5 and A12 sequences (Miyashiro *et al*., 2001).

There is a need to identify consensus epitopes across MAGE members that can be applied in developing a universal vaccine targeting all MAGE expressing tumors. An epitope is a section of an antigen that can trigger an immune response. In the context of B lymphocytes, the paratope of the BCR simply binds to epitopes on exogenous MAGE antigens. However for T lymphocytes, MAGE epitopes are generated from intracellular proteosomal cleavage of endogenous MAGE antigens. These epitopes are presented by human leukocyte antigens (HLA) molecules on the cell surface.

2.3 Human leukocyte antigens

Human leukocyte antigens (HLA) synonymous to Major Histocompatibility Complex (MHC) are a group of related transmembrane glycoproteins that are expressed on the surface of almost every cell in the human body. Their primary function is to bind and present antigens to T lymphocytes in addition to acting as self-markers that the immune system uses to distinguish self-cells from non-self-cells in an individual (Parham *et al*., 1988). HLAs are encoded by over 220 genes localized on chromosome six of the human genome. These genes and their associated proteins are categorized into classes I, II and III. HLA class I genes are distinguished into three main types; HLA-A, HLA-B and HLA-C, and seven minor types; HLA (E-L), each gene coding for an α polypeptide chain of the HLA molecules which are expressed on the surface of platelets and all nucleated cells excluding the central nervous system cells. HLA class I molecules are assembled in the endoplasmic reticulum as heterodimers comprising of a three domain α (1-3) heavy chain and a lighter β2-microglobulin protein (Parham *et al*., 1988). The α1 and α2 domains bind to the TCR while α3 domain binds to the CD8 molecules on the T lymphocyte. An 8-10 amino acid peptide binding groove exists between α1 and α2 domains. In the cytosol, standard proteasome or immunoproteosome degrades endogenous proteins into short antigenic peptides. These peptides are transported by transporter associated with antigen processing (TAP) to the ER where they bind to the molecules (Parham *et al*., 1988). This complex is then taken to the cell surface through the transport vesicles to the scrutiny of CD8+ T cells. Presentation of foreign antigens results in the binding of the TCR to the complex and the subsequent production of cytotoxins and cytokines by the CTLs leading destruction of the foreign bearing cells.

HLA class II genes are divided into six main types: HLA-DPA1, HLA-DPB1, HLA-DQA1, HLA-DQB1, HLA-DRA, and HLA-DRB1 (Parham *et al*., 1988). These genes encode HLA antigens that are solely present on the surfaces of antigen presenting cells (APC) including macrophages, DCs, thymic epithelial cells, B and activated T lymphocytes. HLA class II molecules are multimers consisting of a two domain α and β heavy chains. Peptides allied with HLA class II molecules are 12-18 amino acids long and they bind to the peptide binding groove located between the α1 and β1 domains. Exogenous antigens are engulfed by the APC's and transported to the endosomes where they are degraded by vesicular proteases into antigenic peptides. The endosome is then transported into the interior of the cell where it fuses with vesicles containing MHC class II molecules. Unlike class I molecules, class II molecules are multimers in complex with an invariant chain protein (li or CD74) (Parham *et al*., 1988). This invariant chain is cleaved in the endocytic pathway and 'CLIP', the remaining portion in the binding groove is exchanged for the antigenic peptides. This is facilitated by the non-classical molecule HLA-DM. The complex is then transported to the cell surface to the scrutiny of CD4+ T cells. Presentation of foreign antigens results in the attraction of other APC's to the attack site in context of Th1 or for Th2, stimulation of the B lymphocytes which undergo clonal expansion and differentiation into plasma cells that produce antibodies against the foreign antigen (Parham *et al*., 1988). It is important to note that, B cells can be activated in two ways; when the paratope of the BCR binds to the epitope or when B cells act as APCs. The other HLA gene type; class III, encode proteins whose functions differ from the classes I and II. These proteins are involved in inflammation activities among other activities (Parham *et al*., 1988). The CTAs expressed by the male germ line and placental cells are not recognized by CTLs because these normal cells do not express HLA molecules on their surfaces hence they cannot present antigens to T lymphocytes (Kalejs & Erenpreisa, 2005b).

HLAs are highly polymorphic (Parham *et al*., 1988) especially in the peptide binding groove domains; they differ just slightly from each other in their amino acid sequences and in the subsequent micro-pockets of the binding groove. Currently, there are over 6000 HLA class I allele and over 1000 HLA class II allele variants (http://www.ebi.ac.uk/imgt/hla/stats.html). HLAs are population specific (Buhler & Sanchez-Mazas, 2011). The HLAs common in different ethnic groups are: A*0201, A*0204, B*2705, DRB1*0101, and DRB1*0401 (Barh *et al*., 2010a). Identification of MAGE epitopes in complex with these common alleles is a prerequisite

for this study. This MAGE peptide–HLA complex is crucial in epitope vaccine development as the T cells activated by vaccination are MHC restricted; they can only recognise these complexes.

2.4 Bioinformatics

Biologically, epitopes must affix to the binding pockets of HLA molecules for transportation to the cell surface and subsequent presentation to the scrutiny of CD4+ and CD8+ T cells. Many *in silico* tools and servers that mimic and predict this process have been developed (Niesel *et al*., 2003). Classical tools were built on the assumption that each amino acid of the epitope binds to the HLA pockets with a binding energy. Prediction relied on the overall binding energy given as the sum of individual amino acid energies (Niesel *et al*., 2003). However, newer prediction tools and servers use 'models' that are generally based on sequence mining techniques. These models are trained and tested using authenticated experimental data to develop specific prediction tools. Models are derived from: machine learning techniques e.g. artificial neural networks and support vector machines, mathematical functions e.g. hidden Markov model and regression functions, matrices e.g. stabilized matrix model and quantitative matrices, statistical methods e.g. partial least squares multivariate method, structural based methods e.g. structure prediction by energy minimization and by 3D structure inference etc. It is important to note that a peptide sequence must be encoded first by methods e.g. sparse encoding where each amino acid sequence is represented as a 20-digit binary number consisting of a sole 1 and 19 zeros or as a BLOSUM50 matrix score in BLOSUM encoding. Binary vector encoding is similar to sparse encoding but representation is in vector format (Niesel *et al*., 2003).

Some prediction servers used in this work are based on matrices. An example is ProPred I which uses quantitative matrices (multiplication and addition) to predict 9-mer epitopes binding to 47 HLA alleles (Singh & Raghava, 2003). These matrices were trained on scores calculated from binders and non-binders of experimentally determined 9-mer epitopes. The scores are derived from physicochemical properties, amino acid side chain interactions etc. Multiplication matrices compute the total score for a peptide by multiplying scores of each amino acid at each position while addition matrices add the scores to get the total. The server follows a simple pathway; given a sequence of length n, n+1-9 9-mer overlapping peptides are generated, total score for each is calculated using quantitative matrix of the chosen HLA allele, probable epitopes

are then picked above given thresholds. This same process is used in ProPred II servers in addition to an amino acid position coefficient table (Singh & Raghava, 2001). Other matrices include the stabilized matrix model (SMM) used in IEDB server for class I prediction (Peters & Sette, 2005). Here, a program implementing the SMM algorithm uses quantitative models derived from experimental IC50 values of binding affinities. Given a peptide sequence as input, the sequence data is encoded into a binary vector. The vector is then transposed and incorporated into a matrix which also contains constant offset values as the first column and pair coefficient values as the last columns. The matrix is multiplied with a weighted vector and the result vector contain the scoring matrix (gives the contribution of each amino acid at every position) and values which show the impact of pair coefficient. The scoring matrix is then used to predict the epitopes based on the quantitative models. The SMM algorithm has been converted by Nielsen and his colleagues into a method that integrates alignment and Gibbs samplers method called SMM-align for class II epitope prediction (Nielsen *et al*., 2007). Similarly, the model is trained on HLA-epitope binding data from IEDB databases; sequence data is encoded using sparse, BLOSUM and position specific scoring matrix and epitope binding is generated from the average of calculations from the different encoding techniques.

Statistical methods e.g. quantitative structure activity relationship (QSAR) has been proven to be effective in predicting HLA-peptide binding. MHCPred, one of the servers used in this study, is statistically based (Guan *et al*., 2003). This server uses HLA specific QSAR models generated using partial least square (PLS) method. These models are trained from experimentally determined IC50 values from radioligand assays. These IC50 values are transformed to appropriate log values e.g. log (1/IC50) and are correlated to the free energy of binding. The new values; pIC50, are then used as the dependent variables in QSAR regression. The models predict 9-mer epitopes and consider; a) effects of every amino acid side chain in each position b) interactions between nearest neighbors. For prediction, a peptide's pIC50 value is calculated and compared to the trained data set and a pIC50>6.3 imply recognition by the TCR.

Machine learning methods e.g. ANN is also efficient for prediction. An artificial neural network is a technique that mimics the neurological transmission process, to make predictions after training and testing on known data. An example of an ANN that is used for prediction was developed by Niesel and his colleagues (Niesel *et al*., 2003). Their feed forward ANN is not only based on the binding affinities, but also on chemical similarities, evolutionary information,

competition for binding space, correlations amid amino acids positioned at anchor locations etc. This ANN consists of an input layer of 180 neurons, one hidden layer of 2-10 neurons and one output layer. The input layer is given a peptide sequence and a HMM trained to recognize allele binding affinities. The hidden layer is fed with a mutual information matrix containing data on correlations amid amino acids positioned at anchor locations while the output layer is a linear combination of the given network. Prediction relies on output values that are assigned according to binding affinities (Niesel *et al*., 2003).

Epitopes can be predicted using structure based methods. An example is the method used in T-epitope designer (Kangueane & Sakharkar, 2005b). This server uses a statistically and structurally derived model. The model is trained on 29 experimentally determined X-ray crystal structures of peptide-HLA complexes obtained from PDB database. It uses virtual binding pockets that are derived from 8 unique HLA molecules. As mentioned earlier, sequence data must be encoded. Here, the sequence data is encoded to position specific weight matrix. This matrix-containing peptide residue anchor information-is mapped to the virtual pockets based on the experimental data. Finally, compatibility scores between the peptide and the binding pockets are calculated using the Q matrix method (Kangueane & Sakharkar, 2005b). These same authors also state that other structure based methods use energy minimization techniques and protein-ligand calculations which is computationally intensive.

The T cell epitopes predicted in this work are derived from B cell epitopes. Tools that predict the later epitopes are classified as residue or epitope based (El-Manzalawy *et al.*, 2008b). In residue based, a peptide input is encoded via binary method, the sequence is scanned using a sliding window and residues are assigned some calculated values, finally a set of neighbor residues with positive scores are chosen as the predicted B cell epitope. The process differs somewhat in epitope based methods. Here a sequence and the desired epitope length are given as the input. A sliding window of the chosen length is used to extract peptides which are then subjected to classifiers (e.g. ANN, SVM) trained using experimental data (El-Manzalawy *et al*., 2008b). BCPreds server used in this study uses SVM classifiers to predict B cell epitopes (El-Manzalawy *et al*., 2008b). These classifiers are trained implementing 5 kernel methods (subsequence, string, spectrum, mismatch and local alignment). Training data sets are experimentally determined 701 epitopes and 701 non-epitopes (El-Manzalawy *et al.*, 2008b). Training SVM classifiers involves finding a hyperplane that linearly separates positive and

negative data sets. Sometimes the data is not linearly separable and in this situation, a kernel function is used to transform this non-linearly separable data into a separable feature space.

Antigenicity is a vital issue as epitopes must be capable of triggering an immune response. VaxiJen server was used for this purpose. It contains three prediction models (bacteria, virus and tumor) derived through ACC pre-processing of amino acid properties (hydrophobicity, molecular size and polarity). These properties were the principal components for variance; z1, z2 and z3 respectively used in the PCA analysis of 29 experimental datasets. Training data sets were 100 known antigens and 100 non-antigens for each of the three models. VaxiJen accepts a protein sequence as input, computes antigen probability and ACC calculations, predicts antigenicity based on thresholds from the trained models (Doytchinova & Flower, 2007)

Every MAGE antigen was subjected to 3D structure prediction based on both homology modeling and de novo techniques. Homology modeling is a process where novel protein is constructed based on a known structure while de novo constructs structures from the sequence data alone based on amino acid properties. Phyre2 server was used for this purpose because it constructs 3D structures using both methods (Kelley & Sternberg, 2009). It searches for homologous proteins in PDB database, computes multiple sequence alignments, models matching motifs using homology modeling and dissimilar motifs using de novo methods. Having the 3D structures, epitope analysis was achieved by mapping the predicted epitopes to these structures using Pepitope server (Mayrose *et al*., 2007). Pepitope uses three algorithms; pepsurf, mapitope, and a combination of both, that mimic the genuine epitopes in terms of spatial and physiochemical properties. The pepsurf algorithm maps each epitope onto the surface of an antigen. The algorithm searches for all 3D paths on the antigen structure that are similar to the epitope sequence and returns the best alignment between an epitope and the antigen. The resulting alignments are then clustered and epitope location is inferred. On the other hand, mapitope algorithm searches for shared residue pairs among a set of epitope sequences. It then proceeds to search for a cluster of the shared pairs on the antigen structure and returns the epitope location.

CHAPTER THREE

MATERIALS AND METHODS

This chapter describes the data sets, *in silico* pipeline of epitope predictive servers and sequence analysis bioinformatics tools that were used in this work.

3.1 Data sets

3.1.1 Full protein data

A set of 39 *Homo sapiens* MAGE amino acid sequences was retrieved from UniProt Knowledgebase (UniProtKB) database (www.UniProt.org); MAGE A (A1,A2,A3,A4,A5,A6,A8,A9,AA,AB,AC,A13p), B (B1,B2,B3,B4,B5,B6,B10,B16,B17,B18), C (C1,C2,C3,C4), D (D1,D2,D3,D4), E (E1,E2), F1, G1, H1, L2, N and necdin. UniProtKB database is the principal core of protein data (Magrane & Consortium, 2011). All these sequences have been reviewed by both UniProtKB and Swiss-Prot (Bairoch and Apweiler, 2000) databases and are experimentally authenticated valuable protein sequences. All sequences except (A5, AC, A9, A10, A8, B10, B3, E2, H1 and B5) have evidence at protein level; the proteins were experimentally determined by sequencing, mass spectrometry or NMR. However, the mentioned have evidence at transcript level; their transcripts were determined from expression profiles e.g. RT-PCR or cDNAs. Both type I (MAGE A, B and C) and II (the rest) MAGE proteins were considered for analysis.

3.1.2 MAGE homology domain sequences

A set of 39 domain sequences were derived from the full length sequences, by pruning each of the 39 MAGE sequences at the C and N terminal ends, according to the MHD annotation obtained from UniProtKB database. Each MHD sequence contains 200 amino acid residues except for MAGE C1 and E1 which contain 2 MHDs and MAGE A5 which has 123 amino acid residues.

3.1.3 Protein structures

A set of 39 3D protein structures were derived by modeling each full length MAGE sequence using Phyre version 2.0 web-server (Kelley & Sternberg, 2009). This server uses both *ab initio*

methods and homology modeling using experimentally determined templates from protein data bank (PDB). The best 3D structures were chosen considering acceptable energy minimization values and RMSD scores.

3.2 Conserved motif analysis

Identification of universal epitopes involves analysis of conserved regions that cut across all MAGE sequences. To do this, a multiple sequence alignment (MSA) of all thirty nine sequences was done using CLUSTAL Omega (ClustalO). ClustalO is specifically designed to compute proteins' MSAs; it produces superior MSAs and is time efficient (Sievers *et al.,* 2011). A first alignment was done using ClustalO in UniProtKB where the results were used to retrieve MAGE homology domain (MHD) annotation position from every MAGE sequence. A second alignment involved using ClustalO hosted at www.ebi.ac.uk/Tools/msa/clustalo. The results were downloaded and analyzed using Jalview software (Waterhouse *et al*., 2009). Jalview alignment editor can generate superior alignment figures and provides interactive editing (Waterhouse *et al*., 2009). Conservation, quality and consensus values were the key points in selecting conserved motifs in Jalview. Using the second alignment result, mobyle@pasteur phylip package (http://mobyle.pasteur.fr) was used to create a phylogenetic tree based on calculated protein distances. The resultant tree was converted into a network using Cytoscape software to show possible monopolic clusters (Cline *et al*., 2007; Smoot *et al*., 2011). In addition, Cytoscape was used to create MHD relation networks of MAGE antigens. Cytoscape is a java based open source platform that provides a powerful graphical presentation and visualization of complex networks (Cline *et al*., 2007; Smoot *et al*., 2011).

Another crucial step was to determine whether the conserved motifs identified from the previous steps were located within extracellular regions of MAGE sequences, since it is a prerequisite in epitope design that epitopes be exposed externally. A topology analysis of full length sequences was therefore carried out using TMHMM server to identify the extracellular, transmembrane and intracellular helices of every MAGE sequence (Krogh *et al*., 2001). This server is based on hidden Markov model (HMM) approach and predicts helices with a high level of accuracy (97-98%) (Krogh *et al*., 2001). Only fully extracellular peptides or portions that were found to be extracellular were used for subsequent steps.

3.3 B cell epitope prediction

Potential epitopes should trigger both humoral and cell-mediated immune response (An, L.L., 1997). Full length MAGE sequences were queried into BCPreds server for B cell epitope prediction. BCPreds is based on Support Vector Machine (SVM) technique which uses five kernels to predict potential epitopes (El-Manzalawy *et al.*, 2008b). Epitopes were selected above predefined cut off values i.e.

1. >0.8 for BCPreds (AUC~0.8)
2. >0.8 for AAP (AUC~0.8)
3. >0.5 for VaxiJen (AUC~0.5)

Chosen epitopes were compared to conserved motifs and THMMM server results in order to categorize extracellular conserved epitopes. Selected epitopes were then subjected to an antigenicity analysis using VaxiJen server. VaxiJen is a bioinformatics resource used to predict bacterial, viral and tumor antigens (Doytchinova & Flower, 2007).

3.4 T cell epitope prediction

HLA alleles common in different ethnic groups are: A*0201, A*0204, B*2705, DRB1*0101, and DRB1*0401 (Barh *et al*., 2010a). Identification MAGE peptide–HLA complex is crucial in epitope vaccine design as T cells activated by vaccination are MHC restricted; they can only recognise peptide–HLA complexes. As mentioned earlier, a viable vaccine should be able to induce both humoral and cell mediated response. A pipeline of predictive servers was used to predict T cell epitopes from antigenic B cell epitopes identified from the previous step.

3.4.1 HLA class I epitope prediction using IEDB server

A variety of prediction algorithms that are available are based on artificial neural networks (ANN), stabilized matrix method (SMM), average relative binding (ARB) among others. ANN prediction method uses a combination of several neural networks and hidden Markov model (HMM) trained to recognize epitopes (Nielsen *et al*., 2003), ARB approach is based on the matrix and linear coefficient method (Bui *et al*., 2005) and SMM method uses an algorithm implemented in C++ code for prediction (Peters & Sette, 2005). Retaining the default and recommended settings, 8-11 mer epitopes were chosen based on

1. affinity for HLA-A*0201, -A*0204, and -B*2705
2. percentage rank score of <30 (low percentile rank=good binders)
3. default score of >0.5 for VaxiJen

3.4.2 HLA class II epitope prediction using IEDB server

Available prediction methods use algorithms such as SMM-align, Sturniolo, among others (Nielsen *et al*., 2007). Selected B cell epitopes were used to predict 15 mer class II epitopes. Retaining default and recommended settings, epitopes were chosen based on

1. affinity for HLA-DRB1*0101 and -DRB1*0401
2. percentage rank score of <30 (low percentile rank=good binders)
3. default score of >0.5 for VaxiJen

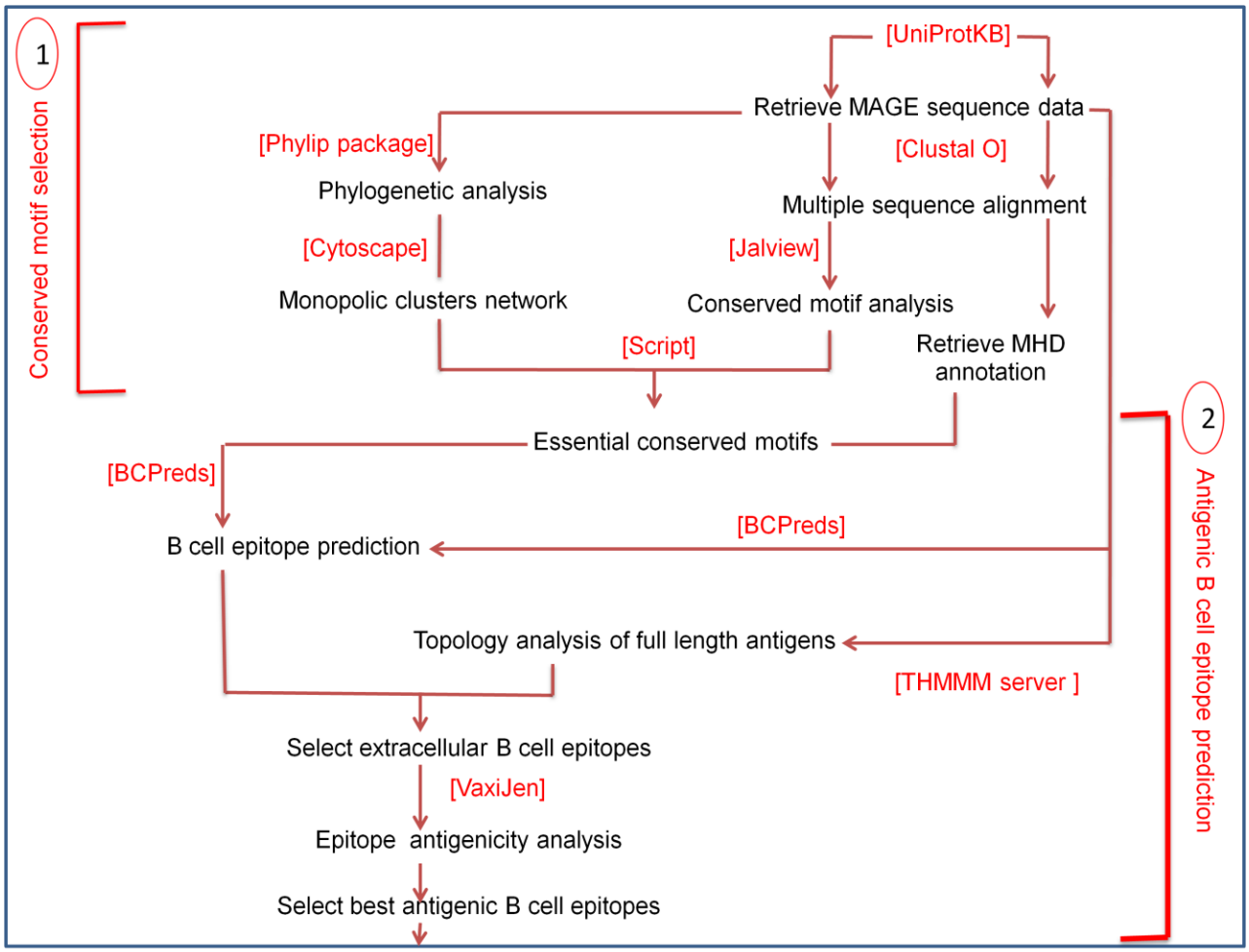
3.4.3 T cell prediction using ProPreds, MHCPred and T-epitope designer

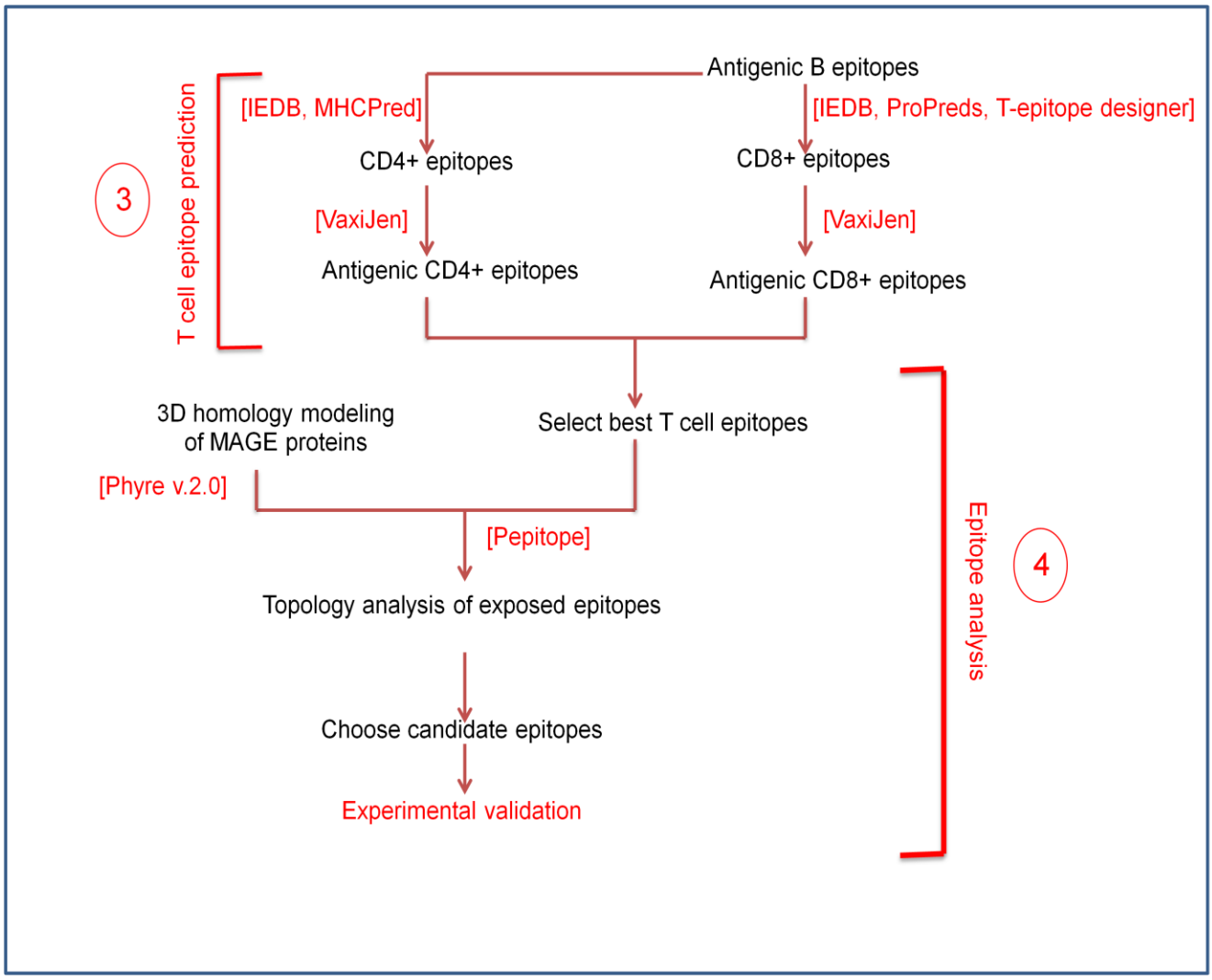
This method was adopted from *Neisseria gonorrhoeae* epitope design (Barh *et al*., 2010b). Chosen B cell epitopes were queried into ProPred I (Singh & Raghava, 2003) and ProPred II (Singh & Raghava, 2001) servers for HLA class I and II epitope prediction respectively. These servers use quantitative matrices models-derived from experimental amino acid position data-for prediction. The number of times an epitope binds to an HLA allele in these servers was counted. Epitopes that bind more than fifteen alleles and are antigenic according to VaxiJen, were queried into MHCPred server (Guan *et al*., 2003) for IC50 value prediction. Only epitopes with IC50<100 were considered. The next step involved validation of chosen epitopes using T-Epitope designer (Kangueane & Sakharkar, 2005a) for class I alleles and MHCPred server for class II alleles. T-Epitope designer predicts epitope-HLA complexes using a model trained on experimental X-ray crystal structures while MHCPred calculates half maximal (50%) inhibitory concentration and predicts epitopes above a certain threshold.

3.5 Epitope analysis

An epitope analysis was carried out using Pepitope server (Mayrose *et al*., 2007). A set of candidate epitopes and a single 3D MAGE protein structure (at a time) were input into Pepitope server for an analysis of epitopes' topology and clusters on protein structures. Those epitopes that were surface exposed and had acceptable cluster scores were chosen as candidate epitopes.

3.6 Methodology Pipeline




Antigenic B epitopes
[IEDB, MHCPred]
[IEDB, ProPreds, T-epitope designer]
CD4+ epitopes
CD8+ epitopes
3
T cell epitope prediction
[VaxiJen]
[VaxiJen]
Antigenic CD4+ epitopes
Antigenic CD8+ epitopes
3D homology modeling
of MAGE proteins
Select best T cell epitopes
[Phyre v.2.0]
[Pepitope]
Topology analysis of exposed epitopes
Choose candidate epitopes
Experimental validation
Epitope analysis
4

CHAPTER FOUR

RESULTS

4.1 Conserved motif analysis

In the first analysis, MAGE homology domain data set was used to evaluate similar amino acid residues among MHDs of MAGE A, B and C as shown in Figure 4.1. MHDs of MAGE A and B form clusters with respect to their subfamilies, while MAGE C defy this norm as C3 and C1 lie in MAGE A cluster while C3 in B cluster. This significant cluster groups highlight the dissimilarities in MAGE domains unlike expectations of co-evolution, therefore implying an impossibility of finding strong conservations across all domains.

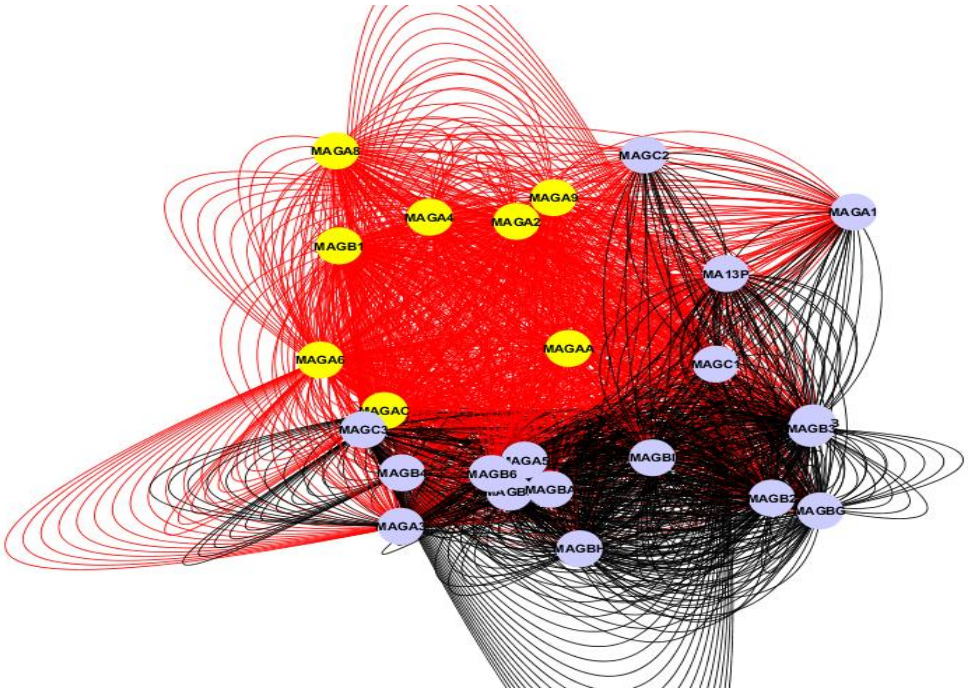


Figure 4.1: Cytoscape network showing monopolic clusters formed by MAGE A, B and C MHDs. Highlighted nodes in yellow represent MAGE A family while the edges in red show relations between amino acids in the MHDs. The edges between nodes represent identical amino acid residues. MHD has ~200 amino acids hence the numerous nested edges

This impossibility of finding strong conserved motifs was further shown in the MSA analysis results. The MSA showed a 1.5% identity across all the MAGE peptides used while a 26.6% pair wise identity for amino acid percentage identity was observed. Comparing the general alignment percentage identities and pair wise identities for the various MAGE classes,

the following was observed 19.0% against 60.6% for MAGE class A, 10.2% against 45.7% for MAGE class B, 6.9% against 20.7% for MAGE class C, 10.1% against 24.0% for MAGE class D and 2.5% against 20.2% for the other MAGE variants. MAGE class A showed a better clustering percentage compared to the other classes. The MAGE Homology domain (Figure according to UniProt) located close to the C-terminal was observed across all the MAGE protein sequences with sequence length between 180 – 200 amino acids. This domain conservation showed a clustering of the various MAGE peptide sub-classes as observed with the polytomy analyses. An analysis beyond MHD reported no consistent conserved domains for all sequences neither at N nor C terminals. Small domains were found conserved within each subfamily i.e. MAGE A (6), B (1), and C (1) in N terminal and a single domain for each family on the C terminal. Interestingly, an analysis within MHD reported an absence of strong sequence similarity in the domain. A lone conserved motif was revealed from the results; YEFLWGPRA (Figure 4.2). The dearth of strong sequence similarity is a consequence of weak functional constraints during gene evolution (Katsura & Satta, 2011).

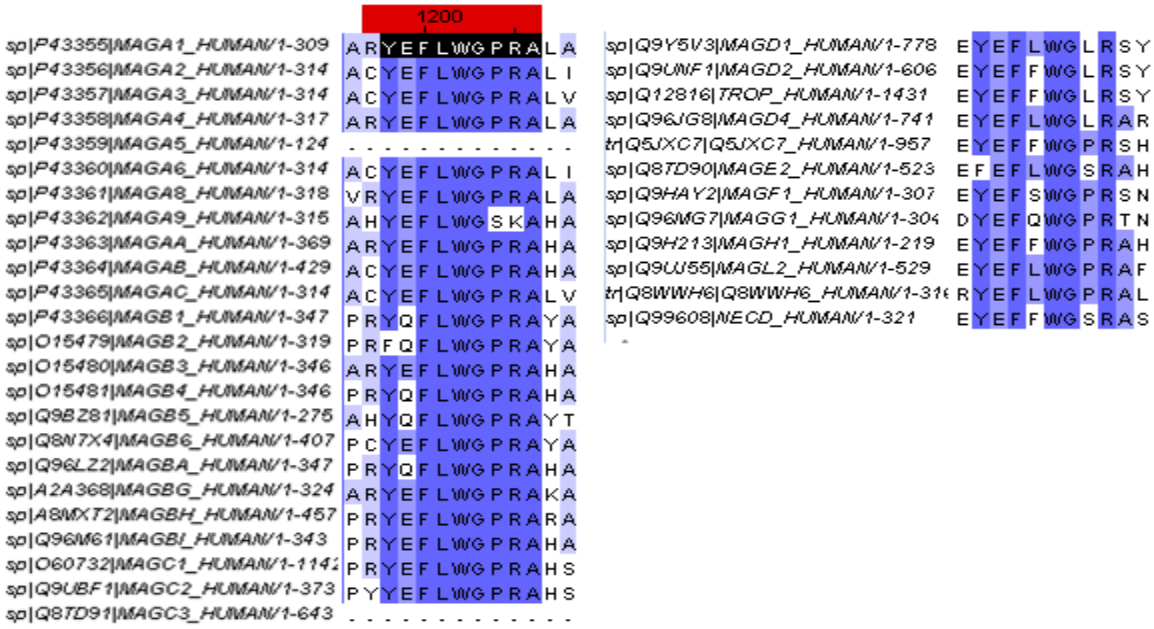


Figure 4.2: A snapshot from MSA results showing YEFLWGPRA conservation. YEFLWGPRA is present in ten MAGE A, five B (B3, B6, and BG-BI), C1, C2, L2 and N

A phylogenetic tree constructed from the MSA results show distinct monopolies formed by MAGE A, B and C (see appendix). The clustering pattern of MAGE A reveals that the family is closely related to N and C families. The monopoly clusters are better visualized using

Cytoscape networks (Figure 4.3 and Figure 7-see appendix). The network in Figure 4.3, analogous to Figure 4.1, shows that MAGE A family forms a more distinct monopoly cluster than the rest explained by the existence of palindromes and a strong sequence conservation in this family (Katsura & Satta, 2011). Again, MAGE C and N are shown to be closer than A.

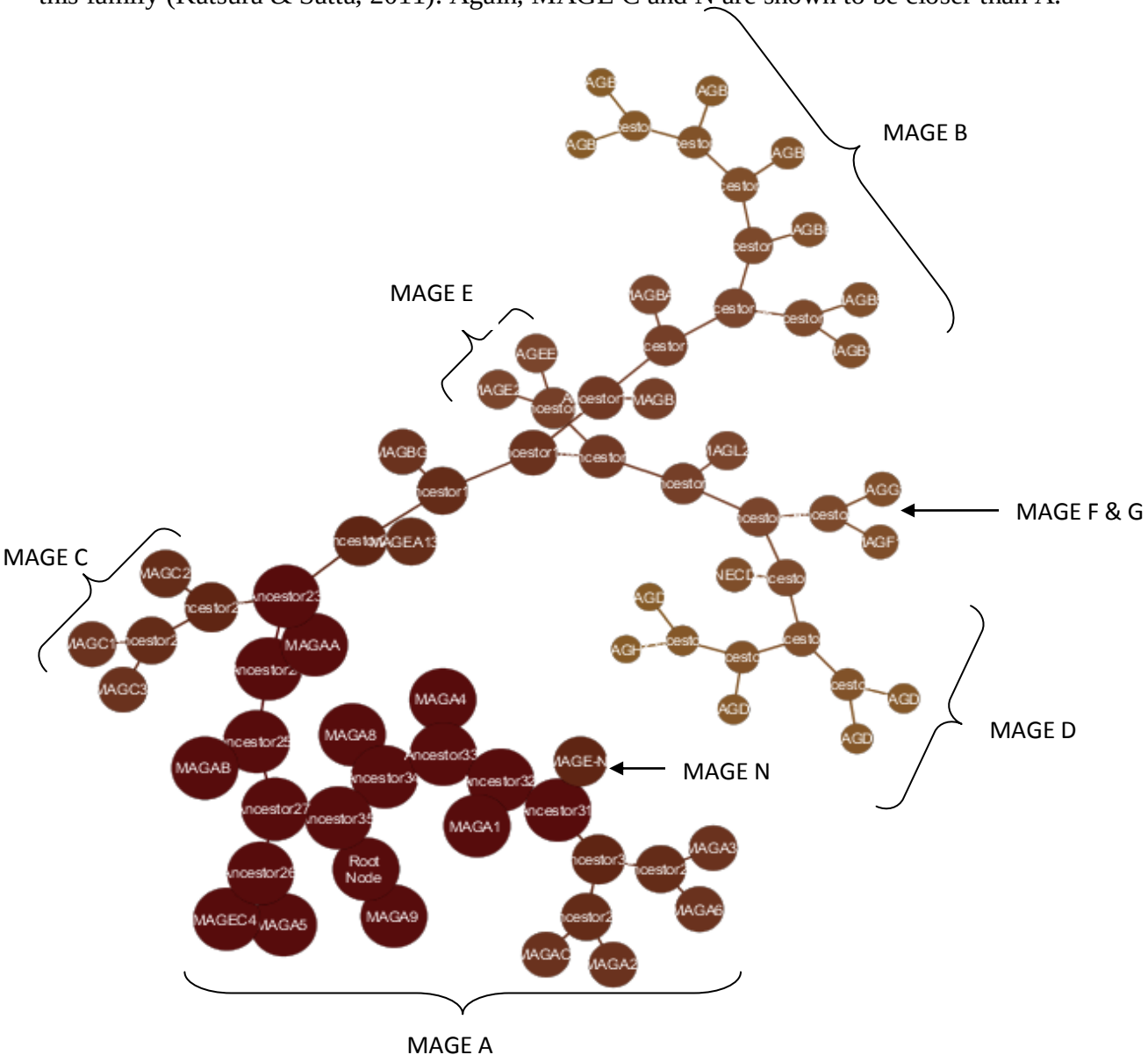


Figure 4.3: Network representation of monopolic clusters formed by individual MAGE subfamilies. Node sizes and colors were created using protein distances obtained from MAGE phylogenetic tree. The size (largeness) of node sizes is proportional to closeness within subfamilies.

Topology analysis results from TMHMM server reveal that all MAGE proteomes except MAGE H1 are extracellular; they are exposed outside the cell membrane. MAGE H1 contains an

internal (position 1-89) and trans-membrane (position 90-112) helices which were excluded from analysis. Only the extracellular portion (position 113-219) was considered (see appendix).

Antigenicity results of MAGE proteins from VaxiJen server range from 0.3221-0.6335, with MAGE B4 and C4 having the highest (0.6335) and lowest (0.3221) scores respectively. For the MAGE class A peptides only two peptides were above the cutoff threshold of 0.5 used for the predictions and these were; MAGE A5 (0.5544) and MAGE A10 (0.5122). The rest were below the cutoff value. For the class B peptides only one was below the cutoff value, MAGE B16 (0.4253) while all the rest were above the cutoff value making them Probable antigens. For the MAGE class C peptides only one was above the cutoff value, MAGE C2 (0.5134) while the rest were below the cutoff value. For the MAGE class D peptides, all were above the cutoff value. For the MAGE class E peptides one was above the cutoff value, MAGE E2 (0.5261) and the other was below the cutoff value, MAGE E1 (0.4425). For the other subfamilies (Table 4.1) only MAGE G1 was above the cutoff value (0.5048). The rest were below. Given the scores, it was observed that MAGE class B & D peptides were more antigenic compared to the rest of the classes. As seen above MAGE A, N and C were poor in their predicted antigenicity further bringing these classes together.

In summary, the results reported in Table 4.1 contain an analysis of each MAGE antigen's sequence length, presence or absence of trans-membrane helices (0=absence and 1=presence) (column 3) and antigenicity scores from VaxiJen (column 4). All extracellular proteins, antigenic and non-antigenic, were considered for the subsequent steps.

Table 4.1: TMHMM server results and VaxiJen scores of full length MAGE sequences

MAGE FAMILY	Sequence length	Predicted TMHs	VaxiJen Score
MAGE A			
A1	309	0	0.4049 (Probable NON-ANTIGEN)
A2	314	0	0.3599 (Probable NON-ANTIGEN)
A3	314	0	0.3735 (Probable NON-ANTIGEN)
A4	317	0	0.4380 (Probable NON-ANTIGEN)
A5	125	0	0.5544 (Probable ANTIGEN))
A6	314	0	0.3688 (Probable NON-ANTIGEN)
A8	318	0	0.4698 (Probable NON-ANTIGEN)
A9	315	0	0.4482 (Probable NON-ANTIGEN)
AA/A10	369	0	0.5122 (Probable ANTIGEN)
AB/A11	489	0	0.4529 (Probable NON-ANTIGEN)
AC/A12	314	0	0.3660 (Probable NON-ANTIGEN)
A13p	341	0	0.4715 (Probable NON-ANTIGEN)
MAGE B			
B1	347	0	0.6118 (Probable ANTIGEN)
B2	319	0	0.6293 (Probable ANTIGEN)
B3	346	0	0.6334 (Probable ANTIGEN)
B4	346	0	0.6335 (Probable ANTIGEN)
B5	275	0	0.5326 (Probable ANTIGEN)
B6	407	0	0.6166 (Probable ANTIGEN)
B10	347	0	0.6321 (Probable ANTIGEN)
B16	324	0	0.4253 (Probable NON-ANTIGEN)
B17	457	0	0.5281 (Probable ANTIGEN)
B18	343	0	0.5884 (Probable ANTIGEN)
MAGE C			
C1	1142	0	0.4161 (Probable NON-ANTIGEN)
C2	373	0	0.5134 (Probable ANTIGEN)
C3	643	0	0.4760 (Probable NON-ANTIGEN)
C4	90	0	0.3221 (Probable NON-ANTIGEN)
MAGE D			
D1	778	0	0.5190 (Probable ANTIGEN)
D2	606	0	0.6208 (Probable ANTIGEN)
D3	1431	0	0.5281 (Probable ANTIGEN)
D4	741	0	0.5344 (Probable ANTIGEN)
MAGE E			
E1	957	0	0.4425 (Probable NON-ANTIGEN)
E2	523	0	0.5261 (Probable ANTIGEN)
OTHER SUBFAMILIES			
F1	307	0	0.4387 (Probable NON-ANTIGEN)
H1	219	1	0.3876 (Probable NON-ANTIGEN)
G1	304	0	0.5048 (Probable ANTIGEN)
L2	529	0	0.3962 (Probable NON-ANTIGEN)
N	316	0	0.3902 (Probable NON-ANTIGEN)
Necdin	321	0	0.4527 (Probable NON-ANTIGEN)

4.2 B cell epitope prediction

B cell epitopes were predicted for each MAGE peptide class with a target of 20 mers. 12 to 18 mer epitopes were equally considered in case where 20 mer epitopes for each MAGE peptide was not predicted. Based on the cutoff values for B epitopes (Table 4.2), MAGE class A predicted 7 epitopes majority of which were 20-mer in length. The second epitope as per the table was 14 mer in length and a fragment of the first. Epitopes five and six were 18 mers in length. Based on the VaxiJen antigenic cutoff score of 0.5 (Table 4.2), the second (0.1867) and seventh (0.1851) epitopes showed poor antigenic scores. Within the MAGE A peptide class, MAGE A2, A3 and A6 were majorly observed across the various epitopes and it was observed

that the first and fifth epitopes were made up only of these 3 classes with the expected 20 mer length, good epitope scores and good antigenic scores. MAGE A6 was common across 6 of the 7 epitopes predicted. The epitopes which cut across majority of the MAGE A peptides (second and seventh) showed poor antigenic scores though the good epitope scores. For the MAGE class B peptides, two epitopes were predicted having 20 mer and 18 mer amino acid lengths. Both of them showed good epitope binding scores, and good antigenic scores. MAGE class B1 was of the expected epitope size while majority of the other class B peptides grouped into the 18 mer epitope length. For the MAGE class C peptides, a 20 mer epitope length was predicted for classes C2, C3 and C4. C1 had no epitope predicted for the considered length. The epitope had a good binding score and epitope score. For MAGE class D, a 20 mer epitope length was predicted for class D1 with a good binding score and good antigenic score. For the rest of the MAGE peptide classes a 20 mer epitope length was predicted except for class E2 which showed just a 16 mer epitope length. The various epitopes all showed good binding and antigenic scores. Looking at the variation of epitope peptide length distribution in the various classes (Figure 4.3), it was observed that a 20 mer epitope length was more favored for binding by the B-cell antibody then followed by an 18 mer length epitope. The variation in epitope sequence length relates to the class variations observed above given that common epitopes were not observed for the various classes. The similarity in epitope sequence and length for MAGE class A and N further confirms the sequence similarity between these two groups as seen above. It was also observed that most of the B-epitopes were predicted from the N-terminal end of the MAGE peptides. Only two MAGE class A peptide (VHFLLLKYRAREPVTKAEML, LEYRQVPGSDPACYEFLW, one class B (FLWGPRAHAETSKMKVLE) and one class C (SVIFIKGNCASEEVIWEVLN) were predicted within the MAGE domain. The high variation in B-cell epitope resulted from the site of prediction of these epitopes. The results reported in Table 4.2 summarize the predicted 21 B cell epitopes from every individual family with corresponding definition scores (size comparisons with their scores (BCPreds and AAP) are given)

Table 4.2: B cell epitopes

MAGE family	MAGE sequences having predicted epitope	Number	B cell epitopes	Length	BCPreds server results	
					BCPreds scores	AAP scores
A	A2,A3,A6,N	**1**	LEQRSQHCKPEEGLEARGEA	20	0.982	-
	A1,A2,A6,AB,AC,N	**1**	LEQRSQHCKPEEGL	14	0.739	_
	A2,A3,N	**2**	LGLVGAQAPATEEQEAASSS	20	0.986	1
	A3,A6,N	**3**	PQSPQGASSLPTTMNYPLWS	20	0.986	1
	A2,A3,A6	**4**	VHFLLLKYRAREPVTKAEML	20	_	1
	A1,A2,A6,AC	**4**	LEYRQVPGSDPACYEFLW	18	0.888	_
	A2,A3,A6,A9,AA,AC	**7**	VQENYLEYRQVPGSDP	18	0.906	_
B	B1	**1**	MPRGQKSKLRAREKRRKARE	20	0.983	_
	B3,B4,BA,BI,AB	**2**	FLWGPRAHAETSKMKVLE	18	0.822	_
C	C2,C3,C4	**1**	SVIFIKGNCASEEVIWEVLN	20	0.881	_
D	D4	**1**	LPPRNVTLLQERANKLVKYL	20	0.881	_
E	E2	**1**	DQFPEILRRASAHLDQ	16	_	0.979
F	F1	**1**	AEGEKDGGHDGETRAPTASQ	20	1	_
G	G1	**1**	MLQKPRNRGRSGGQAERDRD	20	0.932	_
H	H1	**1**	PRGRKSRRRRNARAAEENRN	20	0.994	_
L2	L2	**1**	R R S G K A T R K K K H L E A Q E D S R	20	0.976	_
N	N, A2,A3,A6	**1**	LEQRSQHCKPEEGLEARGEA	20	0.982	_
	N, A3,A6	**2**	LGLVGAQAPATEEQEAASSS	20	0.986	_
	N, A3,A6	**3**	PQSPQGASSLPTTMNYPLWS	20	0.986	_
	N, A3,A6	**4**	DSSNQEEEGPSTFPDLESEF	20	1	_
	N,A1	**5**	MEGGHAPEEEIWEELSVMEV	20	0.954	_

Table 4.2.1 B cell epitope score variations relative to size compared to Vaxijen predictions

Number	Length	VaxiJen score
1	20	0.5320 (Probable ANTIGEN)
1	14	0.1867 (Probable NON-ANTIGEN)
2	20	0.5041 (Probable ANTIGEN)
3	20	0.5041 (Probable ANTIGEN)
4	20	0.6506 (Probable ANTIGEN)
4	18	0.5800 (Probable ANTIGEN)
7	18	0.1851 (Probable NON-ANTIGEN)
1	20	1.4092 (Probable ANTIGEN)
2	18	0.5472(Probable ANTIGEN)
1	20	0.6749 (Probable ANTIGEN)
1	20	0.7955 (Probable ANTIGEN)
1	16	0.7864 (Probable ANTIGEN)
1	20	1.2950 (Probable ANTIGEN)
1	20	1.4041 (Probable ANTIGEN)
1	20	1.1039 (Probable ANTIGEN)
1	20	1.2627 (Probable ANTIGEN)
1	20	0.5320 (Probable ANTIGEN))
2	20	0.5041 (Probable ANTIGEN))
3	20	0.5041 (Probable ANTIGEN))
4	20	0.5531 (Probable ANTIGEN)
5	20	0.5343 (Probable ANTIGEN)

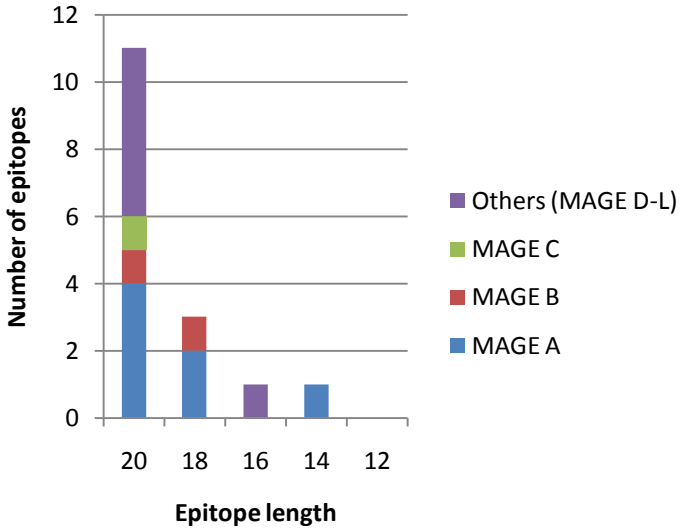


Figure 4.3 Length distributions of B cell epitopes derived from MAGE families

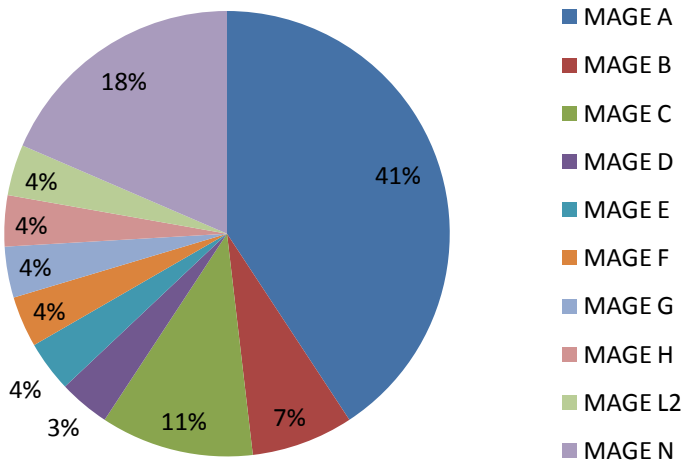


Figure 4.4b Frequency distributions of B cell epitopes derived from MAGE families

4.3 T cell epitope prediction

The predicted B-cell epitopes above were used to predict T-cell epitopes for MHC class I and class II HLA variants. Five variants were considered for this analysis; HLA-A*02:01, HLA-A*02:04 and HLA-A*27:05 for MHC class I, and DRB1*0101, and DRB1*0401 for class II. Binding epitopes were predicted for eight to eleven monomers. The best considered binding scores relative to IEDB for this analysis ranged from 0 to 21. All epitopes were validated for antigenicity using VaxiJen. The nine monomer length showed the highest occurrence of 19 for the various MAGE peptide classes, seconded by the eight epitope monomer length of 9, and this was based on the range of IEDB score groupings. The nine monomer epitope length showed a wider coverage for the various MAGE peptide classes (A, B, C, D & E) and seconded by the ten epitope peptide length (A, B, C & D) then eight monomer epitope length (A, B & C) and finally the eleven monomer epitope length (C). This occurrence and coverage observed for the nine monomer epitope length goes further to confirm them as the best peptide lengths for recognition by T-cell epitopes. Basing on the B-epitopes and derived MAGE peptide class predicted across the various epitope monomer lengths, C1 (SVIFIKGNCASEEVIWEVLN) was present for all the various lengths followed by A1 (LEQRSQHCKPEEGLEARGEA) making these two peptide classes essential for both B and T-cell epitopes given that they were 20 mers in length. C1 was predicted within the MAGE domain while A1 was at the N terminal of the peptide sequence. T cells showed a better preference for epitopes within the domain while B-cell epitopes showed preference out of the domain.

Analysis of the most conserved peptide, YEFLWGPRA, by IEDB and T epitope designer servers reveal that it binds all the select alleles (Tables 4.3 and 4.4). Similarly, using ProPreds servers, this peptide binds to 21 MHC alleles (Table 4.3) which is above the desired count number of 15.

Table 4.3: Analysis of YEFLWGPRA using ProPreds and IEDB servers

MAGE family	**B cell epitope**		**Predicted epitope** (ProPreds MHCPred+ IEDB server)	**IEDB server**		**VaxiJen scores**	**IC50 value of epitopes for DRB1*0101** (MHCPred)	**Number of MHC Class I binding alleles** (ProPred1)		**Number of MHC Class II binding alleles** (ProPred)		**Total number of MHC binding alleles**	
				Alleles	**Scores**			**>threshold**	**<threshold**	**>threshold**	**<threshold**	**>threshold**	**<threshold**
Total=18: A1-A8, AA-AC, B3, B6,BG-BI,C1,C2,L2, N	YRQVPGSDPAC**YEFLWGPRA**-(0.969)	**0.2461** (Probable NON-ANTIGEN))	YEFLWGPRA	**HLA-A*02:01**	17.80	**-0.4486** (Probable NON-ANTIGEN)	27.73	19	26	2	20	21	46
				HLA-A*02:04	19								
				HLA-A*02:05	19								
				HLA-B*27:05	16.90								

Table 4.4: Analysis of YEFLWGPRA using T-Epitope designer

MAGE family	**Epitope**	**T-Epitope Designer**				**MHCPred (IC50 Value)**	
		A*0201	**A*0204**	**A*0205**	**B*2705**	**DRB1*0101**	**DRB1*0401**
All	YEFLWGPRA	860.52	1186.12	973.10	1654.01	27.73	937.56

4.3.1 IEDB results

As mentioned earlier, HLA alleles common in different ethnic groups are A*0201, A*0204, B*2705, DRB1*0101, and DRB1*0401 (Barh *et al*., 2010a). The best epitopes that bind these alleles were chosen and are presented in six tables. Each table contains epitopes of varying lengths; 8-15 amino acids. Selected epitopes were subjected to epitope analysis by Phyre and Pepitope servers.

The results reported in Table 4.5 summarize 18 8-mer epitopes predicted from individual MAGE subfamily members. In terms of numbers, MAGE A family again outperforms the rest, having 10 epitopes. The results show that many epitopes are not only conserved and shared among different members, but also promiscuous to the select four HLA alleles. This observation gives proof that 8-mer epitopes are our desired promiscuous universal epitopes. For antigenicity analysis, results of the epitopes range from 0.5035-1.7858, with MAGE A and C having the lowest and highest scores respectively. The results reported in Table 4.6 summarize 15 9-mer epitopes, Table 4.7 summarizes 7 10-mer epitopes, Table 4.8 summarizes 6 11-mer epitopes and Table 4.9 summarizes 11 15-mer epitopes. **For each epitope, IEDB binding scores of each HLA allele and VaxiJen antigenicity scores are given. Also, the numbers of MAGE members containing the epitope are listed.**

4.3.2 ProPreds, MHCPred and T-epitope designer results

As mentioned earlier, 9-mer epitopes have been shown to fit well in the binding groove of HLA class I molecules and are industrial favored (Chakraborty *et al*., 2010). The epitope prediction pipeline used by Barh and his colleagues (Barh *et al*., 2010b) was used to identify 9-mer epitopes under a more stringent analysis. Moreover, this second method as described in materials and methods, allows for prediction of 9-mer class II epitopes (Madden, 1995), an impossibility using IEDB server. There was no major difference from IEDB predictions, implying that these tools can be used to validate each other. The results reported in Table 4.10 summarize 31 9-mer epitopes predicted from individual MAGE subfamily members.

Table 4.5: 8-mer T cell epitopes

MAGE family	B cell epitope	MHC class I alleles and epitopes			IEDB score	VaxiJen score
		Allele	**Epitope**	**MAGE members containing epitope**		
MAGE A	**1**	**HLA- A*02:01**	RSQHCKPE	A2, A3,A6,AB,AC,N	18.10	0.7187 (Probable ANTIGEN)
		HLA- A*02:04	GLEARGEA	A2,A3,A6	29	1.4522 (Probable ANTIGEN)
		HLA- A*02:05				
		HLA- A*02:01	SQHCKPEE	A2-A6,AB,AC,N	86	0.5240 (Probable ANTIGEN)
		HLA- A*02:04			60	
		HLA- A*02:05			51	
		HLA-B*27:05			24.40	
	4	**HLA- A*02:01**	LKYRAREP	A1,A2, A3,A6, AC,N	22.40	1.0087 (Probable ANTIGEN)
		HLA-B*27:05	VHFLLLKY	A2,A3,A6,AC,N	7.0	0.8499 (Probable ANTIGEN)
		HLA-B*27:05	HFLLLKYR	A1,A2, A3,A6, AC,N	25.90	0.5580 (Probable ANTIGEN)
	5	**HLA- A*02:01**	CLGLSYDG	A1-A8,AC,N	3.10	1.1055 (Probable ANTIGEN)
		HLA-B*27:05			10.0	
		HLA- A*02:05			15	
		HLA- A*02:04			8.3	
		HLA- A*02:01	LSYDGLLG		9.0	0.8264 (Probable ANTIGEN)
		HLA- A*02:04			37	
		HLA- A*02:05			36	
		HLA-B*27:05			33.80	
	6	**HLA- A*02:01**	VQENYLEY	A1-AA,AC	64.0	0.5035 (Probable ANTIGEN)
		HLA- A*02:04			37.0	
		HLA- A*02:05			44	
		HLA-B*27:05			84.7	
		HLA- A*02:01	CYEFLWGP	A1,A4,A8,AA,B6	30.6	0.5135 (Probable ANTIGEN)
		HLA- A*02:04			71.0	
		HLA- A*02:05			70.0	
		HLA-B*27:05			60.0	
		HLA- A*02:01	RYEFLWGP	A2,A3,A6,AB,AC, B3,BG,BH,BI,CI,N	34.0	0.5135 (Probable ANTIGEN)
		HLA- A*02:04			64.0	
		HLA- A*02:05			70.0	
		HLA-B*27:05			25.0	
MAGE B	**2**	**HLA-A*02:01**	ETSKMKVL	B2-B5,BA-BI, AB,D1-D3,G1	34.00	0.7627 (Probable ANTIGEN)
		HLA- A*02:04			49.0	
		HLA- A*02:05			26.0	
		HLA-B*27:05			93.5	
		HLA-A*02:01	SKMKVLEF	B2-B4,BA-BI	7.70	1.5328 (Probable ANTIGEN)
		HLA- A*02:04			34.0	
		HLA- A*02:05			27.0	
		HLA-B*27:05			18.10	
MAGE C	**1**	**HLA- A*02:01**	IKGNCASE	C2,C3,C4	6.9	1.1447 (Probable ANTIGEN)
	1	**HLA- A*02:05**	FIKGNCAS	C2,C3,C4	14	1.6066 (Probable ANTIGEN)

		HLA- A*02:04			29	
		HLA- A*02:01	VIFIKGNC	C2,C3,C4	20	1.7858 (Probable ANTIGEN)
		HLA- A*02:04			18	
		HLA- A*02:05			29	
		HLA-B*27:05			40.30	
		HLA- A*02:05	IFIKGNCA	C2,C3,C4	19	1.1170 (Probable ANTIGEN))
MAGE D	**1**	**HLA- A*02:05**	LQERANKL	D1,D3,D4	18.0	1.7812 (Probable ANTIGEN)
		HLA- A*02:01			78.9	
		HLA- A*02:04			25	
		HLA-B*27:05			59.7	

Table 4.6: 9-mer T cell epitopes

MAGE family	B cell epitope	MHC class I alleles and epitopes					
		Allele	Epitope	MAGE members having epitope		IEDB score	VaxiJen score
				In Subfamilies	Out Subfamilies		
MAGE A	1	HLA-B*27:05	LEQRSQHCK	A2, A3,A6,AB,AC	N	27.80	**0.6135** (Probable ANTIGEN)
	2	HLA- A*02:04	LLLKYRARE	A1,A2, A3,A6, AC	N	25.0	**0.8609** (Probable ANTIGEN)
		HLA- A*02:05	LKYRAREPV	A1,A2,A3,A6	_	31	**0.6338** (Probable ANTIGEN)
	1	HLA-B*27:05	QRSQHCKPE	A2, A3,A6,AB,AC	N	17.80	**0.8311** (Probable ANTIGEN))
	5	HLA-B*27:05	CLGLSYDGL	A1-A8,AC	N	10.0	**0.6048** (Probable ANTIGEN)
		HLA- A*02:01				9.9	
		HLA- A*02:05				15	
		HLA- A*02:04				8.3	
		HLA- A*02:01	LGLSYDGLL			20.7	**0.6074** (Probable ANTIGEN)
		HLA- A*02:05				33	
		HLA- A*02:04				24	
		HLA-B*27:05				35.70	
		HLA- A*02:01	LSYDGLLGD			29.60	**0.6595** (Probable ANTIGEN)
	6	HLA- A*02:01	RQVPGSDPA	A2,A3,A6,A9,AA,AC	_	12.00	**0.5390** (Probable ANTIGEN)
		HLA-B*27:05				7.00	
		HLA- A*02:04				8.5	
		HLA- A*02:05				4.2	
MAGE B	2	HLA-A*02:01	TSKMKVLEF	B2-B4,BA-BI	_	68.7	**1.4262** (Probable ANTIGEN)
		HLA- A*02:04				60	
		HLA- A*02:05				51	
		HLA-B*27:05				34.70	
	1	HLA-B*27:05	SKLRAREKR	B1,B2,B4,BA,BI	_	6.20	**1.7147** (Probable ANTIGEN)
		HLA-B*27:05	MPRGQKSKL			47.95	**1.0360** (Probable ANTIGEN)
		HLA-B*27:05	GQKSKLRAR			9.10	**1.3405** (Probable ANTIGEN)
		HLA-B*27:05	RGQKSKLRA			15.50	**1.6363** (Probable ANTIGEN)
MAGE C	1	HLA-A*02:01	VIFIKGNCA	C2-C4	-	12	**1.1892** (Probable ANTIGEN)
		HLA- A*02:04				16.80	
		HLA- A*02:05				19	
		HLA-B*27:05				61.0	
		HLA-A*02:01	KGNCASEEV			15.30	**0.8621** (Probable ANTIGEN)
		HLA- A*02:04				19	
		HLA- A*02:05				15	
		HLA-B*27:05				43.10	
MAGE D	1	HLA-A*02:01	LQERANKLV			20.40	**1.1716** (Probable ANTIGEN)
		HLA- A*02:04				20	
		HLA- A*02:05				21.20	
		HLA-B*27:05				22	
MAGE E	1	HLA-A*02:01	QFPEILRRA	E1,E2	_	29.9	**1.1853** (Probable ANTIGEN)
		HLA- A*02:04				23	
		HLA- A*02:05				19	
		HLA-B*27:05				67.2	

Table 4.7: 10-mer T cell epitopes

MAGE family	B cell epitope	MHC class I alleles and epitopes					
		Allele	Epitope	MAGE members having epitope		IEDB score	VaxiJen score
				In Subfamilies	Out Subfamilies		
MAGE A	1	HLA-B*27:05	QRSQHCKPEE	A2, A3,A6,AB,AC	N	8.75	**0.7396** (Probable ANTIGEN)
	2	HLA- A*02:01	CLGLSYDGLL	A1-A8,AC	N	7.20	**0.7010** (Probable ANTIGEN)
		HLA- A*02:04				7.9	
		HLA- A*02:01	TCLGLSYDGL			21	**0.9305** (Probable ANTIGEN)
		HLA- A*02:04				22.40	
		HLA- A*02:05				27	
		HLA-B*27:05				39.35	
MAGE B	1	HLA-B*27:05	PRAHAETSKM	B1,B2,B4,BA,BI	_	5.40	**0.6613 (** Probable ANTIGEN)
		HLA-B*27:05	RAHAETSKMK			10.95	**0.8207 (** Probable ANTIGEN)
MAGE C	1	HLA- A*02:05	SVIFIKGNCA	C2,C3,C4	_	9.7	**0.6648** (Probable ANTIGEN)
MAGE D	1	HLA-B*27:05	LQERANKLVY	D1,D3,D4	_	12.20	**1.0671** (Probable ANTIGEN)

Table 4.8: 11-mer T cell epitopes

MAGE family	B cell epitope	MHC class I alleles and epitopes					
		Allele	Epitope	MAGE members having epitope		IEDB score	VaxiJen score
				In Subfamilies	Out Subfamilies		
MAGE A	None	HLA-B*27:05	TCLGLSYDGLL	A1-A8,AC	N	15.30	**0.9353** (Probable ANTIGEN)
	None	HLA-B*27:05	LSYDGLLGDNQ	A1-A8	-	9.60	**0.6911** (Probable ANTIGEN)
MAGE B	7	HLA-A*02:01	AETSKMKVLEF	B2-B4,BA-BI	_	26.20	**1.2259** (Probable ANTIGEN)
		HLA- A*02:04				64.0	
		HLA- A*02:05				61.0	
		HLA-B*27:05				79.6	
MAGE C	1	HLA- A*02:05	FIKGNCASEEV	C2,C3,C4	_	0.2	**1.1161** (Probable ANTIGEN)
		HLA-A*02:01				2.8	
		HLA- A*02:04				29.4	
		HLA-B*27:05				67.70	
		HLA-B*27:05	VIFIKGNCASE			29.60	**1.0653** (Probable ANTIGEN)
		HLA-A*02:01	ASEEVIWEVLN	C2,C3,C4	_	14.70	**0.6055** (Probable ANTIGEN)

Table 4.9: 15-mer T cell epitopes

MAGE family	Allele	Epitope	MAGE having epitope		IEDB score	VaxiJen score
			In Subfamilies	Out Subfamilies		
MAGE A	**HLA-DRB1*01:01**	SQHCKPEEGLEARGE	A2,A3,A6	_	71.94	**0.4792** (Probable NON-ANTIGEN))
	HLA-DRB1*04:01				86.72	
	HLA-DRB1*01:01	FLLLKYRAREPVTKA	A2,A3,A6	_	18.86	**0.5937** (Probable ANTIGEN)
	HLA-DRB1*04:01				8.85	
	HLA-DRB1*01:01	TCLGLSYDGLLGDNQ	A1-A4, AC	D3	16.03	**0.7239** (Probable ANTIGEN)
	HLA-DRB1*04:01					
MAGE B	**HLA-DRB1*01:01**	PRAHAETSKMKVLEF	B3,B4,BA,BI	AB	74.49	0.9487 (Probable ANTIGEN)
	HLA-DRB1*04:01				39.99	
	HLA-DRB1*01:01	MPRGQKSKLRAREKR	B1,B2,B4,BA,BI	_	51.65	**1.3450** (Probable ANTIGEN)
	HLA-DRB1*04:01				72.56	
MAGE C	**HLA-DRB1*01:01**	SVIFIKGNCASEEVI	C2,C3,C4	_	26.22	**0.5430** (Probable ANTIGEN)
		VIFIKGNCASEEVIW	C2,C3,C4	_	26.97	**0.7193** (Probable ANTIGEN)
	HLA-DRB1*04:01	SVIFIKGNCASEEVI	C2,C3,C4	_	18.04	**0.5430** (Probable ANTIGEN)
		VIFIKGNCASEEVIW	C2,C3,C4	_	19.37	**0.7193** (Probable ANTIGEN)
MAGE D	**HLA-DRB1*01:01**	VTLLQERANKLVKYL	D4	_		**0.8903** (Probable ANTIGEN)
	HLA-DRB1*04:01					
MAGE E	**HLA-DRB1*01:01**	QFPEILRRASAHLDQ	E2	_	6.94	**0.7417** (Probable ANTIGEN)
	HLA-DRB1*04:01				7.96	

Table 4.10: 9-mer T cell epitopes predicted using ProPred I, ProPred II and MHCPred servers. For each epitope, VaxiJen and MHCPred scores are given. The numbers of MAGE members containing the epitope are listed and the number of times an epitope binds to any HLA allele in ProPred I and ProPred II servers is given

MAGE family	B cell epitope	Predicted Epitopes (ProPreds + MHCPred)	MAGE having epitope		VaxiJen scores	IC50 value of epitopes for DRB1*0101 (MHCPred)	Number of MHC Class I binding alleles (ProPred1)		Number of MHC Class II binding alleles (ProPred)		Total number of MHC Binding alleles	
			Within Subfamilies	Shared			>threshold	<threshold	>threshold	<threshold	>threshold	<threshold
A	1	LEQRSQHCK	A2, A3,A6,AB,AC	N	**0.6135** (Probable ANTIGEN)	98.17	7	5	0	48	7	53
	4	LLLKYRARE	A1,A2, A3,A6, AC	N	**0.8609** (Probable ANTIGEN)	75.51	3	1	12	5	15	6
		LLKYRAREP	A1,A2, A3,A6, AC	N	**1.0927** (Probable ANTIGEN)	8.89	1	1	1	3	2	4
	5	LSYDGLLGD	A1-A8	N	**0.6595** (Probable ANTIGEN)	46.99	5	15	0	16	5	31
		LGLSYDGLL	A1-A8,AC	N	**0.6074** (Probable ANTIGEN)	207.49	18	8	0	0	18	8
		CLGLSYDGL	A1-A8,AC	N	**0.6048** (Probable ANTIGEN))	755.09	30	10	0	0	30	10
		RQVPGSDPA	A2,A3,A6,A9,AA,AC	_	**0.5390** (Probable ANTIGEN)	17.99	11	15	0	0	11	15
B	1	MPRGQKSKL	B1,B2,B4,BA,BI	_	**1.0360** (Probable ANTIGEN)	94.84	24	15	0	10	24	35
		GQKSKLRAR	B1,B2,B4,BA,BI	_	**1.3405** (Probable ANTIGEN)	11.78	13	10	0	0	13	10
		RGQKSKLRA	B1,B2,B4,BA,BI	_	**1.6363** (Probable ANTIGEN)	56.49	11	26	0	0	11	26
	2	WGPRAHAET	B3,B4,BA,BI	AB	**0.5833** (Probable ANTIGEN)	88.31	7	12	3	6	10	18
		AETSKMKVL	B2-B4,BA-BI	AB	**0.8782** (Probable ANTIGEN)	73.11	17	10	0	0	17	10
		TSKMKVLEF			**1.4262** (Probable ANTIGEN)	6.61	14	9	0	0	14	9
		RAHAETSKM	B3,B4,BA,BI	AB	**0.5426** (Probable ANTIGEN)	20.46	20	11	0		20	11
C	1	FIKGNCASE	C2,C3,C4	_	**1.1670** (Probable ANTIGEN)	4.04	4	2	0	4	4	6
		SVIFIKGNC			**0.9398** (Probable ANTIGEN)	40.18	4	2	0		4	2
		IKGNCASEE			**0.9351** (Probable ANTIGEN)	69.66	1	0	0	4	1	0
		IFIKGNCAS			**1.3998** (Probable ANTIGEN)	358.92	1	1	25	20	26	21
		VIFIKGNCA			**1.1892** (Probable ANTIGEN)	177.01	11	9	12	25	23	34
D	1	LLQERANKL	D1,D4	–	**1.8833** (Probable ANTIGEN)	2.3	27	1	2	22	28	23
		ERANKLVKY	D1,D3,D4		**1.4243** (Probable ANTIGEN)	23.93	8	3	0		8	3
		QERANKLVK			**1.5610** (Probable ANTIGEN)	26.73	5	2	0		5	2
		LQERANKLV			**1.1716** (Probable ANTIGEN)	275.42	9	4	0		9	4
E	1	QFPEILRRA	E1,E2	_	**1.1853** (Probable ANTIGEN)	4.75	14	33	0		14	33

F1	**1**	ETRAPTASQ	F1	_	**0.7990** (Probable ANTIGEN)	0.78	7	8	0		7	8
		GETRAPTAS			**0.7988** (Probable ANTIGEN)	0.67	6	4	0		6	4
		DGETRAPTA			**0.6993** (Probable ANTIGEN))	0.78	5	15	0		5	15
G1	**1**	MLQKPRNRG	G1	_	**1.6284** (Probable ANTIGEN)	61.80	3	12	14	25	17	37
H1	**1**	RRRRNARAA	H1	_	**1.0127** (Probable ANTIGEN)	7.52	4	11	0		4	11
L2	**1**	RRSGKATRK	L2	_	**1.8865** (Probable ANTIGEN)	18.0	7	1	0	47	7	48
N	**1**	EQRSQHCKP	N	A2, A3,A6,AB,AC	**0.7740** (Probable ANTIGEN)	27.54	1	2	0		1	2
		RSQHCKPEE			**0.6344** (Probable ANTIGEN)	78.16	1	5	0		1	5

Table 4.11: T cell epitopes validated using T-epitope designer and MHCPred server

MAGE family	**Epitopes**	**T-Epitope Designer**				**MHCPred (IC50 Value)**	
		A*0201	**A*0204**	**A*0205**	**B*2705**	**DRB1*0101**	**DRB1*0401**
MAGE A	LEQRSQHCK	392.30	-306.05	395.83	846.90	98.17	2023.02
	LLLKYRARE	894.99	753.01	607.88	815.51	75.51	0
	LLKYRAREP	1420.28	423.25	1241.42	1863.02	8.89	3767.04
	LSYDGLLGD	-681.19	-498.45	-632.43	184.71	46.99	1081.43
	LGLSYDGLL	-420.99	-408.49	-591.23	-648.33	207.49	820.35
	CLGLSYDGL	-1156.41	-1333.98	-989.5	-289.86	755.09	0
	RQVPGSDPA	-536.10	-873.27	-539.76	1215.97	17.99	272.90
MAGE B	MPRGQKSKL	1174.00	507.80	1739.85	1449.93	94.84	2157.74
	GQKSKLRAR	692.77	-320.66	700.03	307.06	11.78	615.19
	RGQKSKLRA	67.23	-172.82	382.10	-334.04	56.49	801.68
	WGPRAHAET	318.20	277.83	227.01	857.07	88.31	272.90
	AETSKMKVL	-156.92	-266.65	199.36	580.54	73.11	1432.19
	TSKMKVLEF	189.13	119.98	693.96	-547.40	6.61	133.97
	RAHAETSKM	460.67	388.27	464.13	765.82	20.46	663.74
MAGE C	FIKGNCASE	-235.89	-591.12	-233.13	744.29	4.04	417.83
	SVIFIKGNC	271.54	621.56	-101.02	-13.53	40.18	583.45
	IKGNCASEE	-819.73	-922.83	-1049.27	1230.11	69.66	2851.02
	IFIKGNCAS	-383.85	-360.95	-816.5	55.77	358.92	1713
	VIFIKGNCA	-119.48	-317.14	-340.36	487.57	177.01	431.52
MAGE D	LLQERANKL	860.12	504.29	962.68	1380.68	2.3	0
	ERANKLVKY	201.18	256.61	-16.45	798.35	23.93	217.27
	QERANKLVK	739.34	-127.95	1038.89	347.19	26.73	386.37
	LQERANKLV	-104.11	-623.37	-319.82	828.41	275.42	468.81
MAGE E	QFPEILRRA	244.57	355.62	7.82	547.81	4.75	1811.34
MAGE G1	MLQKPRNRG	722.41	-145.96	689.21	1512.93	61.80	772.68
MAGE H1	RRRRNARAA	722.41	-145.96	689.21	1512.93	7.52	522.40
MAGE L2	RRSGKATRK	155.40	-517.14	78.56	1276.24	18.62	2285.60
MAGE N	EQRSQHCKP	576.40	122.27	843.19	539.30	27.54	851.14
	RSQHCKPEE	361.81	-219.62	361.81	927.58	78.16	219.28

4.4 Epitope analysis

A fold level analysis of chosen epitopes was performed by mapping them onto 3D protein structures for extracellular verification. Each MAGE sequence was queried into Phyre v2.0 server for structure prediction. Table 7.2 (see appendix) summarizes protein structures obtained from PDB used as templates for homology modeling of thirty nine MAGE structures. Almost all structures were based on crystal structures of MAGE A4 and G1. Modeled structures and selected epitopes were input into Pepitope server for mapping analysis. Figure 4.4 illustrates images retrieved from Pepitope server showing location of epitopes on MAGE A2. Most epitopes are surface exposed and occur in acceptable cluster scores (see appendix-Table 7.1). These results confirm that the chosen epitopes are viable candidates for vaccine development.

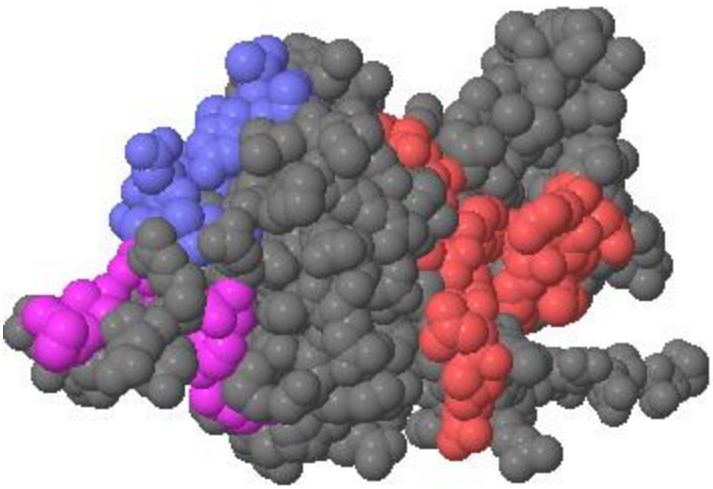


Figure 4.5 (a): A snap shot from Pepitope view of clusters containing epitopes on MAGE A2. It shows a bubble view of MAGE A2 protein, illustrating the surface exposure of the epitopes. The three colors (red, blue and pink) represent different clusters containing epitopes on the protein (grey).

Figure 4.5 (b): The second image shows detailed existence of the epitope clusters within an alpha helix (blue), beta sheet (red and pink) and in the 1D structure (red and pink).

CHAPTER FIVE

DISCUSSION, CONCLUSION AND RECOMMENDATION

5.1 DISCUSSION

5.1.1 Conserved motif analysis

Assessments of alignment sections within and beyond MHD show great sequence dissimilarities and small conserved motifs that cut across all MAGE members. Failure to get larger conserved motifs is a consequence of weak functional constraints during gene evolution (Katsura & Satta, 2011). The same authors state that phase II of gene evolution characterized by retro transposition lead to emergence of eight MAGE members with different processed genes as ancestors for each subfamily. Therefore, massive sequence dissimilarities are expected. This is further shown in monopolic clusters formed in the phylogenetic tree and Cytoscape networks. The close clustering patterns of MAGE A, B and C families differ from other families (D through L) because the former genes' families diverged more recently than the others (Katsura & Satta, 2011). The closeness of MAGE A to MAGE C genes was discussed in the analysis of MAGE gene family by Katsura and Satta who concluded that the latter were duplicated from MAGE A therefore their protein products are more likely to be similar (Katsura & Satta, 2011) Also, MAGE A family forms a more distinct monopoly cluster than the rest explained by the strong 60 % sequence similarity reported in results section, existence of palindromes and strong sequence conservation in this family (Katsura & Satta, 2011). This explains why MAGE A family produced many conserved epitopes than other families. Cytoscape platform was the networking tool of choice as it has been used to create a network of novel cancer genes (Östlund *et al*., 2010) and a network of cancer genes and their homologs (D'Antonio *et al*., 2011).

Results from MSA of MAGE members reveal only one complete conserved motif; YEFLWGPRA. As mentioned earlier, lack of sequence similarity is as a result of weak functional constraints during gene evolution (Katsura & Satta, 2011). Using IEDB server, YEFLWGPRA binds to all select alleles. Using ProPreds servers, YEFLWGPRA binds to 21 MHC class I and class II alleles and through analysis using the T epitope designer, YEFLWGPRA performs well to all the selected alleles. This *in silico* approach selects YEFLWGPRA as a probable candidate, however YEFLWGPRA is non-antigenic from VaxiJen

results, therefore it is discarded. This epitope was identified and analyzed by Graff-Dubois and his colleagues both *in vitro* and *in vivo* (Graff-Dubois *et al*., 2002).This clearly shows the efficacy of using *in silico* tools as a first approach. From their study, they concluded that although YEFLWGPRA was the most conserved domain, E residue at anchor p2 renders the epitope less likely to be naturally processed and thus cannot be a viable epitope. This prompted the study to shift to analyzing individual sub families' conserved motifs. Some of the reported epitopes are conserved in a few number of sub family members. This data is included because complete epitope sequence homology is not a must, it is the homology of the sequences which bind to the TCR that is important (Graff-Dubois *et al*., 2002).

In epitope design, it is required that only extracellular peptides be used. For this reason, each of the thirty nine sequences was queried into THMMM server for topology analysis and only extracellular peptides predicted from this server were used for B cell epitope prediction. TMHMM server has been used to identify transmembrane prostatic acid phosphatases (TM-PAP) a target for vaccine therapy in prostate cancer *(*Quintero *et al.,* 2007*)* and to identify transmembrane lung cancer related proteins in the blood *(*Xiao *et al.,* 2005*)* etc.

Antigenicity is a vital issue as epitopes must be capable of triggering an immune response. Although only MAGE class B & D were shown to be antigenic, MAGE A, N and C were considered for further analysis to determine whether this non-antigenicity affected their performance during epitope prediction. VaxiJen is a powerful antigenicity prediction tool that has been applied to determine antigenicity of; *Escheria coli* 536 (Rai *et al*., 2012b), *Neisseria gonorrhea* epitopes (Barh *et al*., 2010a) and *Neisseria meningitidis* B epitopes (Chandra *et al*., 2010).

5.1.2 B cell epitopes

BCPreds server was chosen for B cell epitope prediction tool as it has been applied in identification of B cell epitopes of nucleocapsid protein region from Japanese Encephalitis virus (Ingale, 2010) and from *Neisseria gonorrhea* protein sequence (Barh *et al*., 2010a) among other studies. The wide occurrence of 20-mers across all MAGE classes makes them the favorable epitope length in this work. This was also stated by EL-Manzalawy and his co-workers during their effort to determine appropriate lengths for B cell epitopes (EL-Manzalawy *et al*, 2008). 14-18 mers were considered since linear epitopes vary from 12-20-mer (EL-Manzalawy *et al*, 2008)

with most being ~15-mer in length (Kringelum *et al.,* 2013). The epitopes predicted for MAGE A family were mostly shared between A2, A3 and A6. This is due to their great sequence similarity and clusterization of MAGE A family. Epitopes like LEQR**SQHCKPEE**GLEARGEA and LEY**RQVPGSDPA**CYEFLW have the regions**-in bold**-which are antigenic, shared among 8 and 6 classes respectively. These epitopes are present in A2. A3 and A6 but can be used to represent 8 class A members because complete epitope sequence homology is not a must, it is the homology of the sequences which bind to the paratope that is important (Graff-Dubois *et al*., 2002). Epitope FLWGPRAHA**ETSKMKVL**E from MAGE B family, though shared among 5 members, can be argued in the same way. The region**-in bold**-which is antigenic, cuts across all B classes. Epitope SVIFIKGNCASEEVIWEVLN from MAGE C is universally present in 3 classes and is thus a good epitope. The mentioned epitopes are ideal based on their occurrence within the C-terminal of MHD which implies protection by co-evolution conservation. The other epitopes that lie beyond MHD are conserved based on motif sequence conservation.

5.1.3 Immunogenic T cell epitopes

B cell epitopes from the above step were used to predict T cell epitopes because these epitopes are required to induce both humoral and cell-mediated immune response. Two different approaches were used in this study. This is because one approach is limited to 9-mer class I epitopes and allows for prediction of 9-mer class II epitopes, while the other approach predicts 8-11 mer class I and 15 mer class II epitopes. As mentioned in the literature review, MAGE epitopes are generated from intracellular proteosomal cleavage of endogenous MAGE antigens and are presented by HLA molecules on the cell surface. These epitopes vary in length i.e. 8-12 mer for HLA class I and 12-15 mer for HLA class II. And more importantly, promiscuous epitopes are preferred as they can bind to many alleles of the human population. IEDB server was used for the first screening to predict 8-15-mer epitopes. It is a reservoir of experimentally determined epitope data from tumor, viral and self-antigens. This server has been used to predict epitopes from HPV-16 E6- and E7 (Riemer *et al*., 2010)**.**

Different epitope lengths are acceptable in vaccine design as shown by Gilleland and his co-workers for examining *Pseudomonas aeruginosa's* 5-12-mer epitopes (Gilleland *et al*., 1997). To begin, 8-mer epitopes can trigger an immune response (Gregoriadis *et al*., 1998). This is supported by Mott and his colleagues for showing that an 8-mer epitope against Herpes virus

could trigger both CD4+ and CD8+ response (Mott *et al*., 2009). Also, in the effort of Duan's group to study the role of ubiquitin–proteasome system in presenting the MUT1 epitope to the CTLs, it was shown that MUT1 8-mer epitope could trigger a strong CTL response (Duan *et al*., 2006). Moreover, Harboe and his team successfully generated anti-peptide antibodies against ESAT-6 Protein of *Mycobacterium tuberculosis* by immunization with an 8-mer epitope (Harboe *et al*., 1998). Therefore, 8-mer epitopes were analyzed in this study and were found to be the most conserved (Table 4.3). One universal epitope here is **R**YEFLWGP present in eleven MAGE members. The "YEFLWGP" region is shared among 19 members and as mentioned earlier, complete epitope sequence homology is not a must, it is the homology of the sequences which bind to the TCR that is important (Graff-Dubois *et al*., 2002). This epitope performs well for HLA- A*02:01 and HLA-B*27:05 and is antigenic with 0.501 VaxiJen score. Analogous to this is **C**YEFLWGP- HLA- A*02:01 epitope which is also antigenic and present in six members. Another epitope to note is ETSKMKVL as it's the most conserved appearing in 13 MAGE members. Other MAGE members differ only in a few sequences i.e. B1; ET**T**KMKVL, MAGE A family shares; ETS**YV**KVL. ETSKMKVL epitope is antigenic with a high VaxiJen score of 0.7627 and binds well to HLA- A*02:01 although it performs poorly for the other HLA alleles of interest.

Now onto individual sub-families, for MAGE A, from the eleven 8 mer epitopes, CLGLSYDG and LSYDGLLG (shared with N) are the best candidates. CLGLSYDG is promiscuous with the best IEDB percentile rank scores <15 for all the selected HLA alleles and a very high antigenicity score of 1.1055. Although LSYDGLLG is also antigenic with a score of 0.8264, it performs very well only for HLA- A*02:01 having IEDB percentile rank score of 0.9. Other epitopes like VQENYLEY and SQHCKPEE (shared with MAGE N) can be considered since they are present in many members, even though they perform fairly well. The other potential epitopes CYEFLWGP (present in MAGE B6), GLEARGEA, RSQHCKPE, LKYRAREP, VHFLLLKY and HFLLLKYR (the last 4 are present in MAGE N) can be excluded since they are present in only five members. On the other hand, MAGE B's potential epitopes are ETSKMKVL and SKMKVLEF. ETSKMKVL has been discussed in the previous paragragh. SKMKVLEF is a perfect as it has a very high antigenicity score of 1.5328, very low percentile rank scores of <10 for all the HLA alleles of interest, and is present in seven MAGE B members. For MAGE C family, VIFIKGNC binds to all select alleles with acceptable scores and

IKGNCASE- HLA- A*02:01 can be considered based on the strong antigenicity of 1.1447 and IEDB score of 6.9. These epitopes from individual MAGE subfamilies can be used to create strings of epitopes that form universal polytope. Therefore, CLGLSYDG-SKMKVLEF-VIFIKGNC is a very good promiscuous antigenic universal tri-epitope targeting MAGE A, B and C members. Also, HLA-B*27:05-RYEFLWGP/ HLA- A*02:01-ETSKMKVL universal di-epitope targets MAGE A1, A4, A8, B3, BG-BI, CI, B2-B5,BA-BI, AB,D1-D3,G1.

T cell epitopes that are 9-mer are the most preferred and have been shown to fit well in the binding groove of HLA class I molecules (Chakraborty *et al*., 2010). Here, MAGE A has six potential 9-mer epitopes with the best being CLGLSYDGL, LGLSYDGLL and RQVPGSDPA. Just like its 8-mer analog, CLGLSYDGL epitope is not only antigenic (VaxiJen score 0.6048) but also promiscuous with low percentile rank scores <15 for all the selected alleles. The other epitope "LGLSYDGLL" is also promiscuous with IEDB scores between 20-35 and VaxiJen score of 0.6078. The third epitope "RQVPGSDPA" has a lower antigenicity score of 0.5390 but performs excellently in IEDB since it has scores of <12 for all the selected alleles. All these epitopes are present in six to seven MAGE A members. Other epitopes that perform well but are limited to singe alleles are; LEQRSQHCK and LLLKYRARE. They too can be considered since they are shared by MAGE N. On another note, MAGE B epitopes proved to be the strongest antigens with antigenicity scores all >1, however, they are limited to HLA-B*27:05. Epitopes like SKLRAREKR, GQKSKLRAR and RGQKSKLRA have very strong antigenic results (1.7147, 1.3405 and 1.6363) and good IEDB scores for this allele. They are unfortunately only present in five MAGE B members and are not promiscuous. Another strong antigenic (score of 1.4262) epitope, TSKMKVLEF which is shared among seven members also only performs well for HLA-B*27:05. MAGE C family also has strong antigenic epitopes like VIFIKGNCA and KGNCASEEV with VaxiJen scores of 1.1892 and 0.8621 respectively. Both epitopes are promiscuous with IEDB scores of <20 for HLA-A alleles but perform poorly for HLA-B*27:05 allele with IEDB scores of >43. Nevertheless, CLGLSYDGL-TSKMKVLEF-VIFIKGNCA can be a promiscuous antigenic universal tri-epitope targeting MAGE A, B and C members.

A 10-mer epitope RTLNAWVKVV in HIV Gag presented by HLA-A*0201 could trigger CTL response (Costanzo, 2012). Here, CLGLSYDGLL was identified as a potential epitope for MAGE A family; however this epitope has already identified and analyzed by (Graff-Dubois *et al*., 2002). And as mentioned earlier, this again proves the accuracy and efficiency of

using *in silico* tools. Another very antigenic (VaxiJen score of 0.9305) epitope is TCLGLSYDGL which is akin to the above epitope, and binds to the four alleles with good IEDB scores. One more is QRSQHCKPEE that performs well only for HLA-B*27:05 in MAGE A and N. MAGE B family failed to produce promiscuous epitopes. PRAHAETSKM and RAHAETSKMK are very antigenic but are limited again to HLA-B*27:05. The same can be said for MAGE C which has SVIFIKGNCA as a potential epitope limited again to HLA-B*27:05. Therefore, TCLGLSYDGL- SVIFIKGNCA- PRAHAETSKM can be a promiscuous antigenic universal tri-epitope targeting HLA-B*27:05.

Using proteasome/TAP assays, it has been shown that 11-mer epitopes are also generated by the proteosome then transported by TAP to the ER (Knuehl *et al*., 2001). Four potential 11-mer epitopes which are very antigenic, with VaxiJen scores between 0.6055-1.2259, were identified in this study. For MAGE A, TCLGLSYDGLL and LSYDGLLGDNQ restricted to HLA-B*27:05 are potential epitopes. MAGE B's AETSKMKVLEF is promiscuous, binding to all the select alleles, and very antigenic with a score of 1.2259 and thus is a good candidate. FIKGNCASEEV epitope for MAGE C also performs well for all alleles except HLA-B*27:05. And so, TCLGLSYDGLL-AETSKMKVLEF-FIKGNCASEEV can form a universal tri-epitope.

A 12-mer epitope of Heat-Shock Protein 60 of *Yersinia enterocolitica* has been found to trigger both CTL and CD4+ immune response and a 15-mer epitope has been used to study HIV1-vaccine induced CD8+ response (Mertz *et al*., 2000; Li *et al*., 2011). However, finding both 12 and 15-mer consensus epitopes was difficult. Reported epitopes bind to a few members. For MAGE A and N, FLLLKYRAREPVTKA and TCLGLSYDGLLGDNQ are candidate epitopes for both class II alleles; HLA-DRB1*01:01 and HLA-DRB1*04:01. FLLLKYRAREPVTKA is not very antigenic based on the VaxiJen score of 0.5937 but it binds well to the select alleles with IEDB scores <20. Similarly the other epitope TCLGLSYDGLLGDNQ performs well for the select alleles with IEDB scores <20 and is very antigenic according to the 0.7239 VaxiJen score. For MAGE B, PRAHAETSKMKVLEF (present in MAGE AB) and MPRGQKSKLRAREKR are the designated epitopes. Although the former epitope is very antigenic with a score of 0.9487, it performs rather poorly in IEDB with scores of 74.49 and 39.99 for HLA-DRB1*01:01 and HLA-DRB1*04:01 respectively. Similarly, MPRGQKSKLRAREKR is a very strong antigen (VaxiJen score of 1.3450) but performs poorly for HLA-DRB1*01:01 (IEDB score of 51.1) and fairly for HLA-DRB1*04:01. However, MAGE

C's VIFIKGNCASEEVIW is not only antigenic (score of 0.7193) but has IEDB good scores for the selected alleles all <26. These epitopes can be combined to a universal multi-epitope e.g. TCLGLSYDGLLGDNQ-VIFIKGNCASEEVIW- PRAHAETSKMKVLEF.

As have been mentioned earlier, 9-mer epitopes are preferred and have been shown to fit well in the binding groove of HLA class I molecules (Chakraborty *et al*., 2010). This prompted a second screening method for potential T cell epitopes using a method adopted from *Neisseria gonorrhoeae* epitope design (Barh *et al*., 2010b). Using B cell epitopes as templates, T cell epitopes were predicted using ProPred I and II servers, MHCPred and T-epitope designer. ProPred servers have been used to identify SSX-2 epitope that can be recognized by the CTL in metastatic melanoma (Ayyoub *et al*., 2002), MHCPred was one of the pipeline tools used to select epitopes from *Escherichia coli* 536 (Rai *et al*., 2012c) and T-epitope designer was applied *Neisseria gonorrhoeae* epitope design (Barh *et al*., 2010b).

This method predicted epitopes akin to IEDB server. Interestingly, an epitope like TSKMKVLEF only binds to HLA B-2705 in IEDB server, yet this epitope binds to all the three alleles except HLA B-2705 in this analysis. The first step using ProPred servers predicted many epitopes; seven for MAGE A shared with N, seven for B, three for C and F, four for D, and one for each the remaining families (E, G1, H and L2). The epitopes had to be screened a second time and minimized using T-epitope designer for class I and MHCPred for class II. Based on positive results from the minimization tools, LEQRSQHCK, LLLKYRARE and LLKYRAREP are the candidate epitopes for MAGE A. For MAGE B, MPRGQKSKL and new epitopes WGPRAHAET and RAHAETSKM are candidates. All epitopes identified for MAGE C have negative values for individual alleles. Similarly, all for MAGE D have negative values except one; LLQERANKL. Other candidate epitopes include; QFPEILRRA for MAGE E, ETRAPTASQ for F, RRRRNARAA for H1 and RRSGKATRK for L2. One universal quadra-epitope is; LLLKYRARE-MPRGQKSKL-SVIFIKGNC-LLQERANKL. HLA B-2705-LSYDGLLGD-TSKMKVLEF-SVIFIKGNC is another good tri-epitope.

5.1.4 Epitope analysis

All MAGE antigens were modeled using Phyre2 server; based on seven known structures with the prototype being a crystal structure of MAGE A4. Phyre2 was one of the modeling servers applied in the 3D structure prediction of CCRL1, a key protein in masked immune

diseases and in Human SIRT1 3D protein structure prediction (Behjati *et al*., 2012; Peck *et al*., 2010). Having 3D structures, epitope analysis was achieved by mapping predicted epitopes to these structures using Pepitope server. This server was applied in mapping *Neisseria gonorrhea* epitopes to the 3D structure of their associated protein (Barh *et al*., 2010a). From the results, most of the epitopes were found within acceptable clusters in the proteins. This was the final analysis to confirm the possibility of the epitopes as viable vaccine candidates. Table 5.1 summarizes the final candidate epitopes.

Table 5.1 Candidate epitopes

T cell epitope	MAGE A	MAGE B	MAGE C
8-mer	CLGLSYDG	SKMKVLEF	VIFIKGNC
8-mer	HLA-B*27:05-RYEFLWGP/ HLA- A*02:01-ETSKMKVL		
9-mer	CLGLSYDGL/LGLSYDGLL	HLA B-2705-TSKMKVLEF	VIFIKGNCA
9-mer (2^{nd})	LLLKYRARE	MPRGQKSKL	SVIFIKGNC
	HLA B-2705-LSYDGLLGD	TSKMKVLEF	SVIFIKGNC
10-mer	TCLGLSYDGL	HLA-B*27:05-RAHAETSKMK	SVIFIKGNCA
11-mer	TCLGLSYDGLL	AETSKMKVLEF	FIKGNCASEEV
15-mer	TCLGLSYDGLLGDNQ	PRAHAETSKMKVLEF	VIFIKGNCASEEVIW

5.2 Conclusion and Recommendation

The goal of this study was to identify conserved epitopes that cut across all MAGE members. MAGE epitopes are attractive for cancer immunotherapy for many reasons. First, MAGE antigens are ideal candidates as they are genuine targets. Their expression in malignancies and benign tumors, antitumor effects, immunogenicity and even role in tumorigenesis has been verified *in vitro* and *in vivo*. Therefore universal epitopes encourage designing a vaccine that not only target many cancers but also triggers strong immune response and antitumor effects thus arresting tumorigenesis. Secondly, the strict tumor specific and differential expression makes MAGE a safe target for cancer immunotherapy because likelihood of autoimmunity is prevented by lack of expression in normal tissues. Thirdly, the testis is protected by blood testis barrier and absent HLA expression. Fourthly, MAGE epitopes are promiscuous; they bind to multiple HLAs hence are good targets since this promiscuous nature increases the probability of matching HLA polymorphism and thus cover a large population of patients with different HLA allele variants.

Even though many MAGE epitopes have been reported, an analysis of universal epitopes that cut across all MAGE sub-families has never been investigated. This study has revealed twenty three unique epitopes and six epitopes that have been reported from *in vitro* analysis. This clearly indicates the efficacy and reliability of using *in silico* tools as a first step in epitope design. Chosen candidate epitopes possess the following properties; 1) **promiscuisity** to alleles that are common in different ethnic groups therefore, implying wide coverage, 2) **varying lengths** (8-15 mer), which can be used independently or as strings of epitopes (polytope), 3) **conservation** across families implying designing vaccine that cuts across MAGE families, 4) **locality in MHD** implies conservation based on co-evolution. The epitopes that lay beyond MHD but possessed the mentioned properties can also be considered since they occur in conserved motifs that cut across families. Majority of these out of domain motifs were mostly B cell specific and could be used to check evolutionary variations within the MAGE peptide family. Moreover, epitopes that were predicted from families other than MAGE A, B, and C-, allow for design of future vaccines should the other families become tumorigenic. Success of these epitopes implementation however depends on the outcome of an experimental validation.

It is recommended that these epitopes be experimentally validated. A paradigm of *in vitro* tests involves MHC Class I and II Binding assay, T-cell proliferation assay and Elispots. *In vivo* studies will be conducted in mice and/or primate and finally, human clinical trials. The final goal will be to attach these epitopes to a polytope liposomal construct for a therapeutic cancer vaccine.

REFERENCES

Akiyama, Y., Komiyama, M., Nakamura, Y., Iizuka, A., Oshita, C., Kume, A., … Yamaguchi,K. (2012). Identification of novel MAGE-A6- and MAGE-A12-derived HLA-A24-restricted cytotoxic T lymphocyte epitopes using an *in silico* peptide-docking assay. *Cancer immunology, immunotherapy: CII*. doi:10.1007/s00262-012-1298-1.

Antonescu, C. R., Busam, K. J., Iversen, K., Kolb, D., Coplan, K., Spagnoli, G. C., … Jungbluth, A. A. (2002). MAGE antigen expression in monophasic and biphasic synovial sarcoma. *Human pathology*, *33*(2), 225–229.

Aubry, F., Satie, A.-P., Rioux-Leclercq, N., Rajpert-De Meyts, E., Spagnoli, G. C., Chomez, P., … Samson, M. (2001). MAGE-A4, a germ cell specific marker, is expressed differentially in testicular tumors. *Cancer*, *92*(11), 2778–2785. doi:10.1002/1097-0142(20011201)92:11<2778::AID-CNCR10125>3.0.CO;2-S

Ayyoub, M., Stevanovic, S., Sahin, U., Guillaume, P., Servis, C., Rimoldi, D., … Lévy, F. (2002). Proteasome-assisted identification of a SSX-2-derived epitope recognized by tumor-reactive CTL infiltrating metastatic melanoma. *Journal of immunology (Baltimore, Md.: 1950)*, *168*(4), 1717–1722.

Barh, D., Misra, A. N., Kumar, A., & Vasco, A. (2010a). A novel strategy of epitope design in Neisseria gonorrhoeae. *Bioinformation*, *5*(2), 77–85.

Behjati, M., Torktaz, I., Mohammadpour, M., Ahmadian, G., & Easton, A. J. (2012). Comparative modeling of CCRL1, a key protein in masked immune diseases and virtual screening for finding inhibitor of this protein. *Bioinformation*, *8*(7), 336–340. doi:10.6026/97320630008336

Bettinotti, M. P., Panelli, M. C., Ruppe, E., Mocellin, S., Phan, G. Q., White, D. E., & Marincola, F. M. (2003). Clinical and immunological evaluation of patients with metastatic melanoma undergoing immunization with the HLA-Cw*0702-associated epitope MAGE-A12:170-178. *International journal of cancer. Journal international du cancer*, *105*(2), 210–216. doi:10.1002/ijc.11045

Bolli, M., Kocher, T., Adamina, M., Guller, U., Dalquen, P., Haas, P., … Spagnoli, G. C. (2002). Tissue microarray evaluation of Melanoma antigen E (MAGE) tumor-associated antigen expression: potential indications for specific immunotherapy and prognostic relevance in squamous cell lung carcinoma. *Annals of surgery*, *236*(6), 785–793; discussion 793. doi:10.1097/01.SLA.0000036266.09823.6C

Buhler, S., & Sanchez-Mazas, A. (2011). HLA DNA Sequence Variation among Human Populations: Molecular Signatures of Demographic and Selective Events. *PLoS ONE*, *6*(2), e14643. doi:10.1371/journal.pone.0014643

Bui, H.-H., Sidney, J., Peters, B., Sathiamurthy, M., Sinichi, A., Purton, K.-A., … Sette, A. (2005). Automated generation and evaluation of specific MHC binding predictive tools: ARB matrix applications. *Immunogenetics*, *57*(5), 304–314. doi:10.1007/s00251-005-0798-y

Carrel, S., Schreyer, M., Spagnoli, G., Cerottini, J.-C., & Rimoldi, D. (1996). Monoclonal antibodies against recombinant-MAGE-1 protein identify a cross-reacting 72-kDa antigen which is co-expressed with MAGE-1 protein in melanoma cells. *International Journal of Cancer*, *67*(3), 417–422. doi:10.1002/(SICI)1097-0215(19960729)67:3<417::AID-IJC17>3.0.CO;2-4

Chakraborty, S., Chakravorty, R., Ahmed, M., Rahman, A., Waise, T. Z., Hassan, F., … Shamsuzzaman, S. (2010). A computational approach for identification of epitopes in dengue virus envelope protein: a step towards designing a universal dengue vaccine targeting endemic regions. *In silico biology*, *10*(5), 235–246. doi:10.3233/ISB-2010-0435

Chandra, S., Singh, D., & Singh, T. R. (2010). Prediction and characterization of T-cell epitopes for epitope vaccine design from outer membrane protein of Neisseria meningitidis serogroup B. *Bioinformation*, *5*(4), 155–161.

Chi, D. D., Merchant, R. E., Rand, R., Conrad, A. J., Garrison, D., Turner, R., … Hoon, D. S. (1997). Molecular detection of tumor-associated antigens shared by human cutaneous melanomas and gliomas. *The American journal of pathology*, *150*(6), 2143–2152.

Chianese-Bullock, K. A., Pressley, J., Garbee, C., Hibbitts, S., Murphy, C., Yamshchikov, G., … Slingluff, C. L., Jr. (2005). MAGE-A1-, MAGE-A10-, and gp100-derived peptides are immunogenic when combined with granulocyte-macrophage colony-stimulating factor and montanide ISA-51 adjuvant and administered as part of a multipeptide vaccine for melanoma. *Journal of immunology (Baltimore, Md.: 1950)*, *174*(5), 3080–3086.

Chiarella, P., Massi, E., De Robertis, M., Fazio, V. M., & Signori, E. (2009). Recent advances in epitope design for immunotherapy of cancer. *Recent patents on anti-cancer drug discovery*, *4*(3), 227–240.

Cline, M. S., Smoot, M., Cerami, E., Kuchinsky, A., Landys, N., Workman, C., … Bader, G. D. (2007). Integration of biological networks and gene expression data using Cytoscape. *Nature protocols*, *2*(10), 2366–2382. doi:10.1038/nprot.2007.324

Costanzo, M. C. (2012). Mobilizing subdominant HIV-specific CTLS as a novel vaccine strategy. *ETD Collection for University of Texas, El Paso*, 1–77.

D'Antonio, M., Pendino, V., Sinha, S., & Ciccarelli, F. D. (2011). Network of Cancer Genes (NCG 3.0): integration and analysis of genetic and network properties of cancer genes. *Nucleic Acids Research*, *40*(D1), D978–D983. doi:10.1093/nar/gkr952

de Carvalho, F., Vettore, A. L., & Colleoni, G. W. B. (2012). Cancer/Testis Antigen MAGE-C1/CT7: new target for multiple myeloma therapy. *Clinical & developmental immunology*, *2012*, 257695. doi:10.1155/2012/257695

Doytchinova, I. A., & Flower, D. R. (2007). VaxiJen: a server for prediction of protective antigens, tumour antigens and subunit vaccines. *BMC Bioinformatics*, *8*(1), 4. doi:10.1186/1471-2105-8-4

El-Manzalawy, Y., Dobbs, D., & Honavar, V. (2008b). Predicting linear B-cell epitopes using string kernels. *Journal of molecular recognition: JMR*, *21*(4), 243–255. doi:10.1002/jmr.893

Gilleland, H. E., Jr, Gilleland, L. B., Staczek, J., Harty, R. N., Garcia-Sastre, A., Engelhardt, O. G., & Palese, P. (1997). Chimeric influenza viruses incorporating epitopes of outer membrane protein F as a vaccine against pulmonary infection with Pseudomonas aeruginosa. *Behring Institute Mitteilungen*, (98), 291–301.

Graff-Dubois, S., Faure, O., Gross, D.-A., Alves, P., Scardino, A., Chouaib, S., … Kosmatopoulos, K. (2002). Generation of CTL Recognizing an HLA-A*0201-Restricted Epitope Shared by MAGE-A1, -A2, -A3, -A4, -A6, -A10, and -A12 Tumor Antigens: Implication in a Broad-Spectrum Tumor Immunotherapy. *The Journal of Immunology*, *169*(1), 575–580.

Gregoriadis, G., McCormack, B., & Allison, A. C. (1998). *Vaccine Design: The Role of Cytokine Networks*. Springer.

Greiner, J., Ringhoffer, M., Simikopinko, O., Szmaragowska, A., Huebsch, S., Maurer, U., … Schmitt, M. (2000). Simultaneous expression of different immunogenic antigens in acute myeloid leukemia. *Experimental Hematology*, *28*(12), 1413–1422. doi:10.1016/S0301-472X(00)00550-6

Guan, P., Doytchinova, I. A., Zygouri, C., & Flower, D. R. (2003). MHCPred: a server for quantitative prediction of peptide–MHC binding. *Nucleic Acids Research*, *31*(13), 3621–3624.

Harboe, M., Malin, A. S., Dockrell, H. S., Gotten Wiker, H., Ulvund, G., Holm, A., … Andersen, P. (1998). B-Cell Epitopes and Quantification of the ESAT-6 Protein of Mycobacterium tuberculosis. *Infection and Immunity*, *66*(2), 717–723.

Hennard, C., Pfuhl, T., Buettner, M., Becker, K.-F., Knöfel, T., Middeldorp, J., … Grässer, F. (2006). The antibody 2B4 directed against the Epstein-Barr virus (EBV)-encoded nuclear antigen 1 (EBNA1) detects MAGE-4: implications for studies on the EBV association of human cancers. *The Journal of pathology*, *209*(4), 430–435. doi:10.1002/path.1996

Huang, L. Q., Brasseur, F., Serrano, A., De Plaen, E., van der Bruggen, P., Boon, T., & Van Pel, A. (1999). Cytolytic T lymphocytes recognize an antigen encoded by MAGE-A10 on a human melanoma. *Journal of immunology (Baltimore, Md.: 1950)*, *162*(11), 6849–6854.

Inaoka, R. J., Jungbluth, A. A., Baiocchi, O. C., Assis, M. C., Hanson, N. C., Frosina, D., … Colleoni, G. W. (2011). An overview of cancer/testis antigens expression in classical Hodgkin's lymphoma (cHL) identifies MAGE-A family and MAGE-C1 as the most frequently expressed antigens in a set of Brazilian cHL patients. *BMC cancer*, *11*(1), 416. doi:10.1186/1471-2407-11-416

Ingale, A. . (2010). *In silico* Homology Modeling and Epitope Prediction of Nucleocapsid Protein region from Japanese Encephalitis Virus. *Journal of Computer Science & Systems Biology*, *03*(02), 053–058. doi:10.4172/jcsb.1000056

Jungbluth, A A, Busam, K. J., Kolb, D., Iversen, K., Coplan, K., Chen, Y. T., … Old, L. J. (2000). Expression of MAGE-antigens in normal tissues and cancer. *International journal of cancer. Journal international du cancer*, *85*(4), 460–465.

Kalejs, M., & Erenpreisa, J. (2005a). Cancer/testis antigens and gametogenesis: a review and. *Cancer Cell International*, *5*(1), 4. doi:10.1186/1475-2867-5-4

Kangueane, P., & Sakharkar, M. K. (2005a). T-Epitope Designer: A HLA-peptide binding prediction server. *Bioinformation*, *1*(1), 21–24.

Katsura, Y., & Satta, Y. (2011). Evolutionary History of the Cancer Immunity Antigen MAGE Gene Family. *PLoS ONE*, *6*(6). doi:10.1371/journal.pone.0020365

Kavanagh, B., Ko, A., Venook, A., Margolin, K., Zeh, H., Lotze, M., … Fong, L. (2007). Vaccination of metastatic colorectal cancer patients with matured dendritic cells loaded with multiple major histocompatibility complex class I peptides. *Journal of immunotherapy (Hagerstown, Md.: 1997)*, *30*(7), 762–772. doi:10.1097/CJI.0b013e318133451c

Kelley, L. A., & Sternberg, M. J. E. (2009). Protein structure prediction on the Web: a case study using the Phyre server. *Nature protocols*, *4*(3), 363–371. doi:10.1038/nprot.2009.2

Kim, S. H., Castro, F., Gonzalez, D., Maciag, P. C., Paterson, Y., & Gravekamp, C. (2008a). Mage-b vaccine delivered by recombinant Listeria monocytogenes is highly effective against breast cancer metastases. *British Journal of Cancer*, *99*(5), 741–749. doi:10.1038/sj.bjc.6604526

Kim, Y. M., Lee, Y. H., Shin, S. H., Kim, E. H., Choi, Y. W., Lee, K. M., … Kim, M. C. (2001). Expression of MAGE-1, -2, and -3 genes in gastric carcinomas and cancer cell lines derived from Korean patients. *Journal of Korean medical science*, *16*(1), 62–68.

Knuehl, C., Spee, P., Ruppert, T., Kuckelkorn, U., Henklein, P., Neefjes, J., & Kloetzel, P. M. (2001). The murine cytomegalovirus pp89 immunodominant H-2Ld epitope is generated and translocated into the endoplasmic reticulum as an 11-mer precursor peptide. *Journal of immunology (Baltimore, Md.: 1950)*, *167*(3), 1515–1521.

Krogh, A., Larsson, B., von Heijne, G., & Sonnhammer, E. L. (2001). Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. *Journal of molecular biology*, *305*(3), 567–580. doi:10.1006/jmbi.2000.4315

Lin, H. H., Zhang, G. L., Tongchusak, S., Reinherz, E. L., & Brusic, V. (2008). Evaluation of MHC-II peptide binding prediction servers: applications for vaccine research. *BMC bioinformatics*, *9 Suppl 12*, S22. doi:10.1186/1471-2105-9-S12-S22

Liu, X.-L., Zhao, D., Sun, D.-P., Wang, Y., Li, Y., Qiu, F.-Q., & Ma, P. (2012). Adenovirus-mediated delivery of CALR and MAGE-A3 inhibits invasion and angiogenesis of glioblastoma cell line U87. *Journal of experimental & clinical cancer research: CR*, *31*, 8. doi:10.1186/1756-9966-31-8

Magrane, M., & Consortium, U. (2011). UniProt Knowledgebase: a hub of integrated protein data. *Database*, *2011*(0), bar009–bar009. doi:10.1093/database/bar009

Mayrose, I., Penn, O., Erez, E., Rubinstein, N. D., Shlomi, T., Freund, N. T., … Pupko, T. (2007). Pepitope: epitope mapping from affinity-selected peptides. *Bioinformatics*, *23*(23), 3244–3246. doi:10.1093/bioinformatics/btm493

Mertz, A. K. H., Wu, P., Sturniolo, T., Stoll, D., Rudwaleit, M., Lauster, R., … Sieper, J. (2000). Multispecific CD4+ T Cell Response to a Single 12-mer Epitope of the Immunodominant Heat-Shock Protein 60 of Yersinia enterocolitica in Yersinia-Triggered Reactive Arthritis: Overlap with the B27-Restricted CD8 Epitope, Functional Properties, and Epitope Presentation by Multiple DR Alleles. *The Journal of Immunology*, *164*(3), 1529–1537.

Miyashiro, I., Kuo, C., Huynh, K., Iida, A., Morton, D., Bilchik, A., … Hoon, D. S. B. (2001). Molecular Strategy for Detecting Metastatic Cancers with Use of Multiple Tumor-specific MAGE-A Genes. *Clinical Chemistry*, *47*(3), 505–512.

Mori, M., Inoue, H., Mimori, K., Shibuta, K., Baba, K., Nakashima, H., … Akiyoshi, T. (1996). Expression of MAGE genes in human colorectal carcinoma. *Annals of surgery*, *224*(2), 183–188.

Mott, K. R., Chentoufi, A. A., Carpenter, D., BenMohamed, L., Wechsler, S. L., & Ghiasi, H. (2009). The Role of a Glycoprotein K (gK) CD8+ T-Cell Epitope of Herpes Simplex Virus on Virus Replication and Pathogenicity. *Investigative Ophthalmology & Visual Science*, *50*(6), 2903–2912. doi:10.1167/iovs.08-2957

Nielsen, M., Lundegaard, C., & Lund, O. (2007). Prediction of MHC class II binding affinity using SMM-align, a novel stabilization matrix alignment method. *BMC bioinformatics*, *8*, 238. doi:10.1186/1471-2105-8-238

Nielsen, M., Lundegaard, C., Worning, P., Lauemøller, S. L., Lamberth, K., Buus, S., … Lund, O. (2003). Reliable prediction of T-cell epitopes using neural networks with novel sequence representations. *Protein science: a publication of the Protein Society*, *12*(5), 1007–1017. doi:10.1110/ps.0239403

Oba-Shinjo, S. M., Caballero, O. L., Jungbluth, A. A., Rosemberg, S., Old, L. J., Simpson, A. J. G., & Marie, S. K. N. (2008). Cancer-testis (CT) antigen expression in medulloblastoma. *Cancer immunity*, *8*, 7.

Östlund, G., Lindskog, M., & Sonnhammer, E. L. L. (2010). Network-based Identification of Novel Cancer Genes. *Molecular & Cellular Proteomics*, *9*(4), 648–655. doi:10.1074/mcp.M900227-MCP200

Parham, P., Lomen, C. E., Lawlor, D. A., Ways, J. P., Holmes, N., Coppin, H. L., … Ennis, P. D. (1988). Nature of polymorphism in HLA-A, -B, and -C molecules. *Proceedings of the National Academy of Sciences of the United States of America*, *85*(11), 4005–4009.

Park, M.-S., Park, J.-W., Jeon, C.-H., Lee, K.-D., & Chang, H.-K. (2002). Expression of melanoma antigen-encoding genes (MAGE) by common primers for MAGE-A1 to -A6 in colorectal carcinomas among Koreans. *Journal of Korean medical science*, *17*(4), 497–501.

Peck, B., Chen, C.-Y., Ho, K.-K., Fruscia, P. D., Myatt, S. S., Coombes, R. C., … Lam, E. W.-F. (2010). SIRT Inhibitors Induce Cell Death and p53 Acetylation through Targeting Both SIRT1 and SIRT2. *Molecular Cancer Therapeutics*, *9*(4), 844–855. doi:10.1158/1535-7163.MCT-09-0971

Peters, B., & Sette, A. (2005). Generating quantitative models describing the sequence specificity of biological processes with the stabilized matrix method. *BMC bioinformatics*, *6*, 132. doi:10.1186/1471-2105-6-132

, I. B., Araujo, C. L., Pulkka, A. E., Wirkkala, R. S., Herrala, A. M., Eskelinen, E.-L., … Vihko, P. T. (2007). Prostatic Acid Phosphatase Is Not a Prostate Specific Target. *Cancer Research*, *67*(14), 6549–6554. doi:10.1158/0008-5472.CAN-07-1651

Rai, J., Lok, K. I., Mok, C. Y., Mann, H., Noor, M., Patel, P., & Flower, D. R. (2012a). Immunoinformatic evaluation of multiple epitope ensembles as vaccine candidates: E coli 536. *Bioinformation*, *8*(6), 272–275. doi:10.6026/97320630008272

Riemer, A. B., Keskin, D. B., Zhang, G., Handley, M., Anderson, K. S., Brusic, V., … Reinherz, E. L. (2010). A Conserved E7-derived Cytotoxic T Lymphocyte Epitope Expressed on Human Papillomavirus 16-transformed HLA-A2+ Epithelial Cancers. *Journal of Biological Chemistry*, *285*(38), 29608–29622. doi:10.1074/jbc.M110.126722

Shichijio, S., Tsunosue, R., Masuoka, K., Natori, H., Tamai, M., Miyajima, J., … Itoh, K. (1995). Expression of the MAGE gene family in human lymphocytic leukemia. *Cancer Immunology, Immunotherapy*, *41*(2), 95–103. doi:10.1007/BF01527405

Singh, H., & Raghava, G. P. S. (2001). ProPred: prediction of HLA-DR binding sites. *Bioinformatics*, *17*(12), 1236–1237. doi:10.1093/bioinformatics/17.12.1236

Singh, H., & Raghava, G. P. S. (2003). ProPred1: prediction of promiscuous MHC Class-I binding sites. *Bioinformatics*, *19*(8), 1009–1014. doi:10.1093/bioinformatics/btg108

Smoot, M. E., Ono, K., Ruscheinski, J., Wang, P.-L., & Ideker, T. (2011). Cytoscape 2.8: new features for data integration and network visualization. *Bioinformatics (Oxford, England)*, *27*(3), 431–432. doi:10.1093/bioinformatics/btq675

Sudo, T., Kuramoto, T., Komiya, S., Inoue, A., & Itoh, K. (1997). Expression of MAGE genes in osteosarcoma. *Journal of Orthopaedic Research*, *15*(1), 128–132. doi:10.1002/jor.1100150119

Takahashi, N., Ohkuri, T., Homma, S., Ohtake, J., Wakita, D., Togashi, Y., … Nishimura, T. (2012). First clinical trial of cancer vaccine therapy with artificially synthesized helper/

killer-hybrid epitope long peptide of MAGE-A4 cancer antigen. *Cancer Science*, *103*(1), 150–153. doi:10.1111/j.1349-7006.2011.02106.x

Thun, M. J., DeLancey, J. O., Center, M. M., Jemal, A., & Ward, E. M. (2010). The global burden of cancer: priorities for prevention. *Carcinogenesis*, *31*(1), 100–110. doi:10.1093/carcin/bgp263

Toh, H. C., Wang, W.-W., Chia, W. K., Kvistborg, P., Sun, L., Teo, K., … Claesson, M. H. (2009). Clinical Benefit of Allogeneic Melanoma Cell Lysate–Pulsed Autologous Dendritic Cell Vaccine in MAGE-Positive Colorectal Cancer Patients. *Clinical Cancer Research*, *15*(24), 7726–7736. doi:10.1158/1078-0432.CCR-09-1537

Tomaselli, S., Meli, M., Plescia, J., Zetta, L., Altieri, D. C., Colombo, G., & Ragona, L. (2010). Combined *in silico* and experimental approach for drug design: the binding mode of peptidic and non-peptidic inhibitors to hsp90 N-terminal domain. *Chemical biology & drug design*, *76*(5), 382–391. doi:10.1111/j.1747-0285.2010.01015.x

Van den Eynde, B., Hainaut, P., Hérin, M., Knuth, A., Lemoine, C., Weynants, P., … Boon, T. (1989). Presence on a human melanoma of multiple antigens recognized by autologous CTL. *International journal of cancer. Journal international du cancer*, *44*(4), 634–640.

van der Bruggen, P., Traversari, C., Chomez, P., Lurquin, C., De Plaen, E., Van den Eynde, B., … Boon, T. (1991). A gene encoding an antigen recognized by cytolytic T lymphocytes on a human melanoma. *Science (New York, N.Y.)*, *254*(5038), 1643–1647.

Vita, R., Zarebski, L., Greenbaum, J. A., Emami, H., Hoof, I., Salimi, N., … Peters, B. (2010). The Immune Epitope Database 2.0. *Nucleic Acids Research*, *38*(Database issue), D854–D862. doi:10.1093/nar/gkp1004

von Boehmer, L., Keller, L., Mortezavi, A., Provenzano, M., Sais, G., Hermanns, T., … Wild, P. J. (2011). MAGE-C2/CT10 protein expression is an independent predictor of recurrence in prostate cancer. *PloS one*, *6*(7), e21366. doi:10.1371/journal.pone.0021366

Waterhouse, A. M., Procter, J. B., Martin, D. M. A., Clamp, M., & Barton, G. J. (2009). Jalview Version 2—a multiple sequence alignment editor and analysis workbench. *Bioinformatics*, *25*(9), 1189–1191. doi:10.1093/bioinformatics/btp033

Weinert, B. T., Krishnadath, K. K., Milano, F., Pedersen, A. W., Claesson, M. H., & Zocca, M.-B. (2009). Real-time PCR analysis of genes encoding tumor antigens in esophageal tumors and a cancer vaccine. *Cancer immunity*, *9*, 9.

Xiao, T., Ying, W., Li, L., Hu, Z., Ma, Y., Jiao, L., … Cheng, S. (2005). An Approach to Studying Lung Cancer-related Proteins in Human Blood. *Molecular & Cellular Proteomics*, *4*(10), 1480–1486. doi:10.1074/mcp.M500055-MCP200

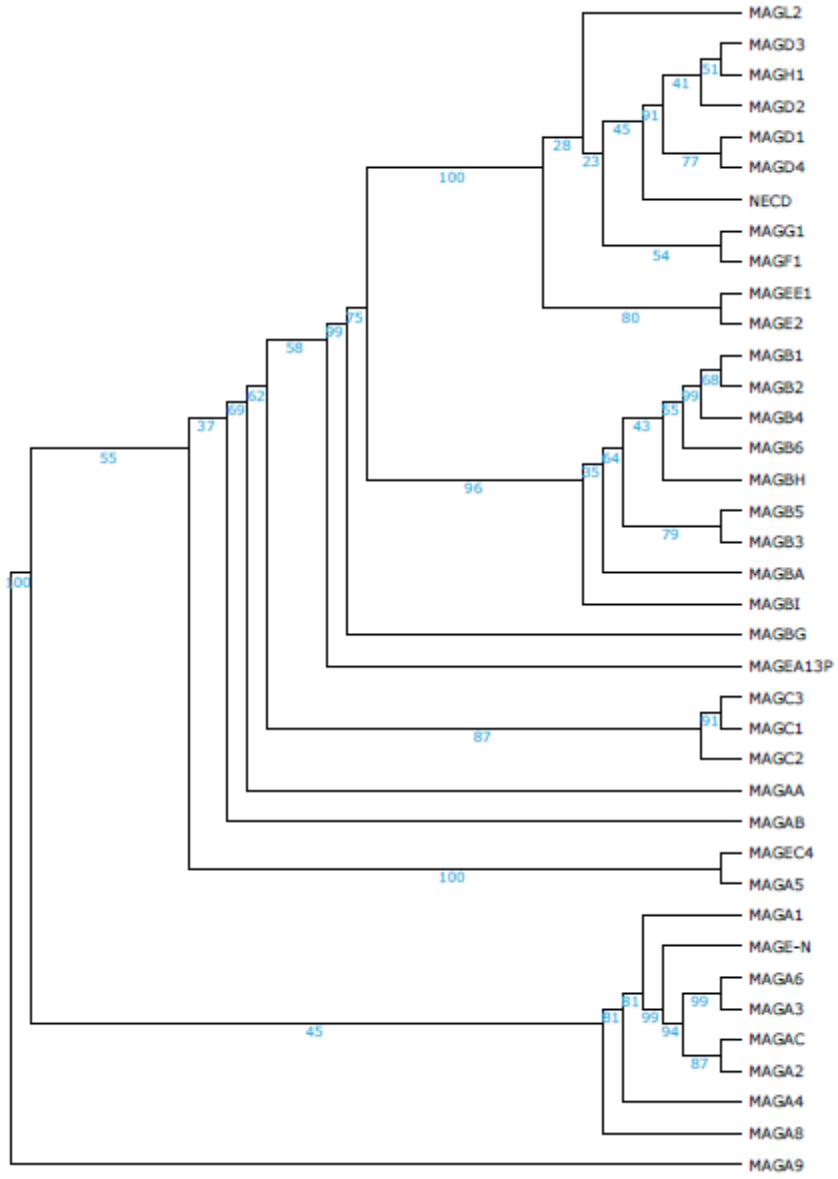
Yang, B., O'Herrin, S., Wu, J., Reagan-Shaw, S., Ma, Y., Nihal, M., & Longley, B. J. (2007). Select cancer testes antigens of the MAGE-A, -B, and -C families are expressed in mast cell lines and promote cell viability *in vitro* and in vivo. *The Journal of investigative dermatology*, *127*(2), 267–275. doi:10.1038/sj.jid.5700548

Yu, J. S., Liu, G., Ying, H., Yong, W. H., Black, K. L., & Wheeler, C. J. (2004). Vaccination with tumor lysate-pulsed dendritic cells elicits antigen-specific, cytotoxic T-cells in patients with malignant glioma. *Cancer research*, *64*(14), 4973–4979. doi:10.1158/0008-5472.CAN-03-3505

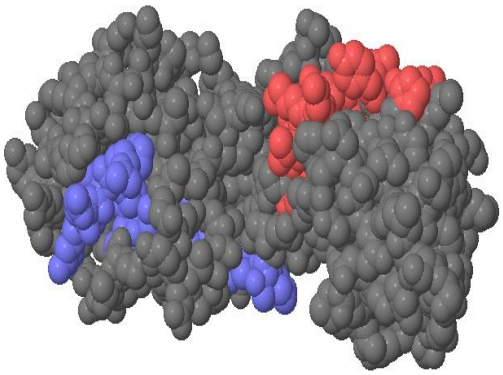
Zhang, S., Zhou, X., Yu, H., & Yu, Y. (2010a). Expression of tumor-specific antigen MAGE, GAGE and BAGE in ovarian cancer tissues and cell lines. *BMC cancer*, *10*, 163. doi:10.1186/1471-2407-10-163

APPENDICES

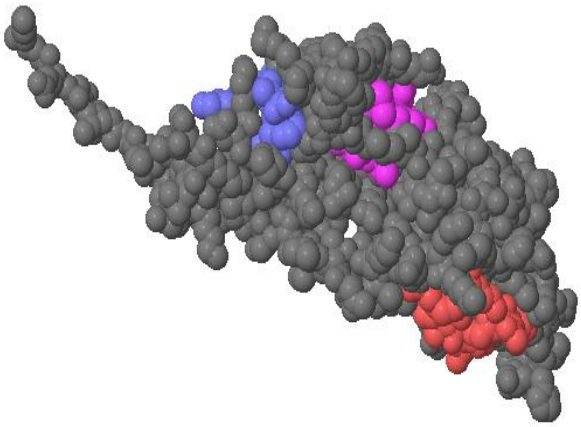
7.1 APPENDIX A: Additional figures



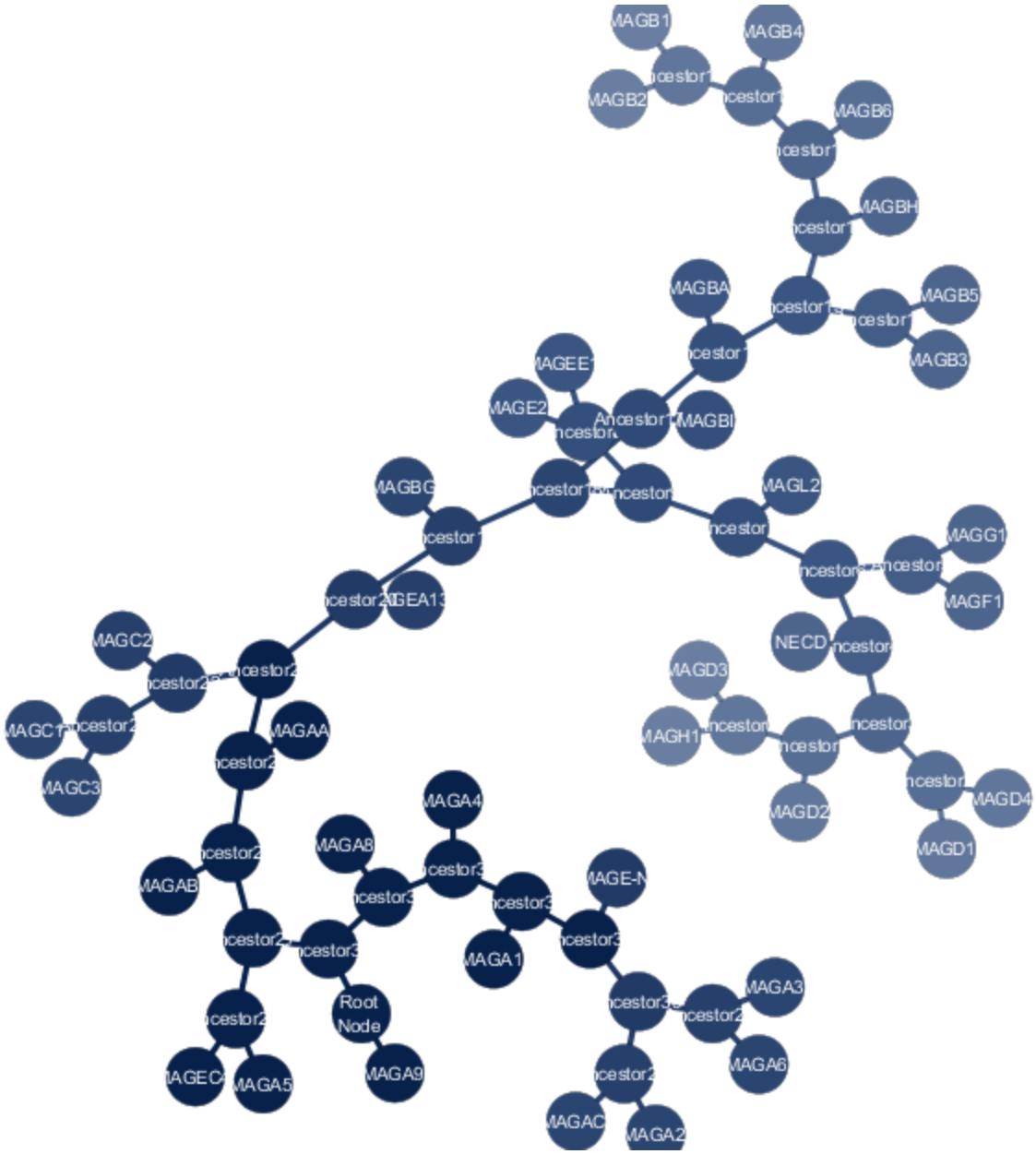
(i) A non-routed phylogenetic tree generated using mobyle@pasteur phylip package (http://mobyle.pasteur.fr) showing the distribution and clusters of MAGE subfamilies.



(ii) Pepitope image of surface exposed epitope clusters in MAGE B2

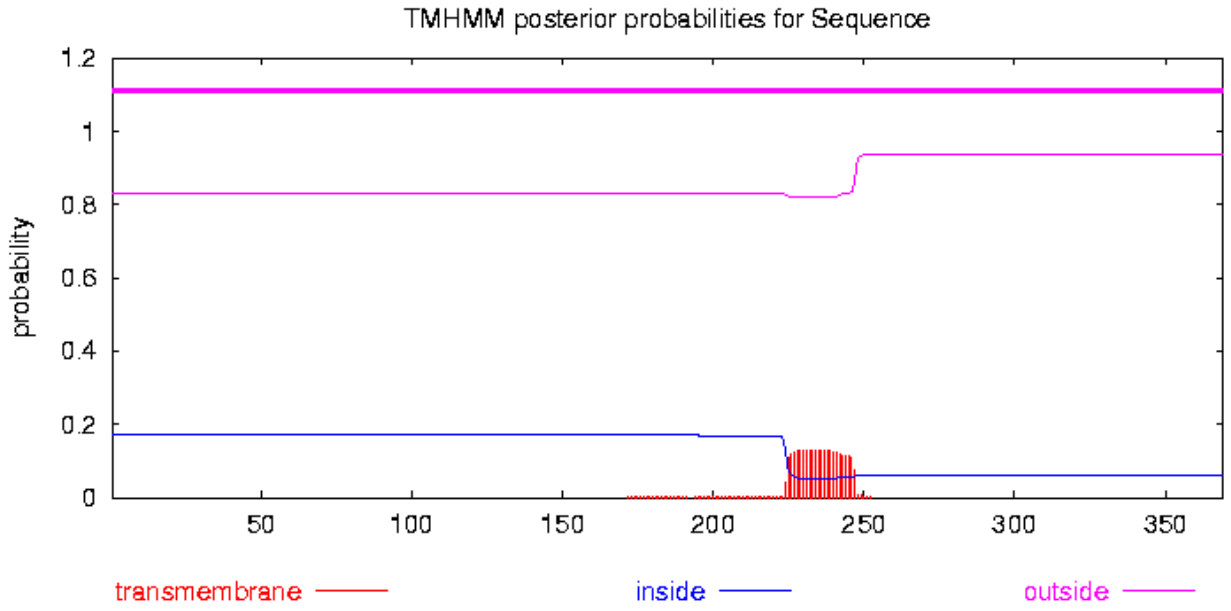


(iii) Pepitope image of epitope clusters in MAGE C1



(iv) Cytoscape network showing monopolic clusters of MAGE family

```
# Sequence Length: 369
# Sequence Number of predicted TMHs:  0
# Sequence Exp number of AAs in TMHs: 2.91997
# Sequence Exp number, first 60 AAs:  0
# Sequence Total prob of N-in:        0.17019
Sequence        TMHMM2.0        outside      1    369
```



(v) A graph plot of TMHMM server prediction of transmembrane helices (red line) and inner membrane helix (blue line) in MAGE H1.

APPENDIX B: Additional tables

Table 7.1: Cluster scores from Pepitope server for individual candidate epitopes

Epitope	MAGE A	Scores	MAGE B	Scores	MAGE C	Scores
8-mer	CLGLSYDG	26.487 (2)	SKMKVLEF	56.207 (1)	VIFIKGNC	68.991
8-mer	HLA-B*27:05-RYEFLWGP	7.793 (3)	ETSKMKVL	56.207		
9-mer	CLGLSYDGL/LGLSYDGLL	31.502 (1)	HLA B-2705-TSKMKVLEF	56.207	VIFIKGNCA	68.991
9-mer (2^{nd})	LLLKYRARE	31.502	MPRGQKSKL	22.366	SVIFIKGNC	68.991
	HLA B-2705-LSYDGLLGD	31.502	TSKMKVLEF	56.207	SVIFIKGNC	
10-mer	TCLGLSYDGL	26.487	HLA-B*27:05-RAHAETSKMK	56.207	SVIFIKGNCA	68.991
11-mer	TCLGLSYDGLL	26.487	AETSKMKVLEF	56.207	FIKGNCASEEV	68.991

Table 7.2: Template proteins used for homology modeling

MAGE family	SCOP Code/ Protein PDB ID	Fold/PDB header	PDB molecule
A1,A2,A3,A4,A6,A8,A9,A11,A12 C1,C2,C4,N ,D3,D1	c2wa0A	Immune system	Melanoma-associated antigen 4
A10,A13p,B1,B2,B3,B4,B5,B6,B10,B16, B17,B18,H1,F1,G1,Necdin,E2,L2,D2,D4	c2wa0A	Immune system	Melanoma-associated antigen 4
	c3nw0B	Metal binding protein	Melanoma-associated antigen g1
A5	d1y02a1	LEM/SAP HeH motif	SAP domain/ super family
L2	c3iynN	Cell cycle	Inner centromere protein
D2,D4	d1np3a1	Phosphogluconate dehydrogenase C-terminal domain-like	6-phosphogluconate dehydrogenase C-terminal domain-like/super family
D4	c2wsiA	Transferase	Fad synthetase
	c2qfaC	Cell cycle	Inner centromere protein
	c1xezA	Toxin	Hemolysin

Table 7.3: Studies on MAGE vaccines

Vaccine	Condition	Reference
DC-based vaccine targeting six HLA-A*0201 epitopes from CEA, **MAGE**, and HER2/neu	Metastatic colorectal cancer	(Toh *et al*., 2009)
DC-based vaccine targeting **MAGE A1**, gp100, and HER-2	Malignant glioma	(Yu *et al*., 2004)
Multi-peptide vaccine containing **MAGE A1**, **MAGE A10** and gp100	Melanoma	(Chianese-Bullock *et al*., 2005)
HLA-Cw*0702- **MAGE A12** epitope vaccine	Metastatic melanoma	(Bettinotti *et al*., 2003)
DC containing **MAGE A3**	Gastrointestinal carcinomas	(Toh *et al*., 2009)
MAGE B DNA vaccine	Breast cancer	(S. H. Kim *et al*., 2008a)
DC vaccine targeting six **MAGE A** antigens	Advanced colorectal cancer	(Kavanagh *et al*., 2007)
Synthetic helper/killer-hybrid epitope long peptide (H/K-HELP) from **MAGE A4**	Colon cancer	(Takahashi *et al*., 2012)

Table 7.4: Reported T-cell epitopes from TANTIGEN and Cancer Immunity Epitope databases

MAGE FAMILY	HLA ALLELE	EPITOPE	POSITION
MAGE-A1	A1	EVDPIGHLY	168-176
	A2	FLWGPRALV	271-279
	A2	QLVFGIELMEV	159-169
	A2	IMPKAGLLIIV	188-198
	A2	KVAELVHFL	112-120
	A24	TFPDLESEF	97-105
	A24	IMPKAGLLI	195-203
	A24	VAELVHFLL	113-121
	A52	WQYFFPVIF	143-151
	B18	MEVDPIGHLY	167-176
	B35	EVDPIGHLY	168-176
	B37	REPVTKAEML	127-136
	B40	AELVHFLLL	114-122
	B44	MEVDPIGHLY	167-176
	B52	WQYFFPVIF	143-151
	Cw7	EGDCAPEEK	212-220
	DP4	KKLLTQHFVQENYLEY	243-258
	DQ6	KKLLTQHFVQENYLEY	243-258
	DR1	ACYEFLWGPRALVETS	267-282
	DR4	RKVAELVHFLLLKYR	111-125
	DR4	VIFSKASSSLQL	149-160
	DR7	VIFSKASSSLQL	149-160
	DR7	VFGIELMEVDPIGHL	161-175
	DR11	GDNQIMPKAGLLIIV	191-205
	DR11	TSYVKVLHHMVKISG	281-295
	DR13	RKVAELVHFLLLKYRA	111-126
	DR13	AELVHFLLLKYRAR	114-127
	DR13	FLLLKYRAREPVTKAE	119-134
MAGE A2	A*0201	FLWGPRALV	271-279
	A*0201	IMPKAGLLIIV	195-205
	A*0201	KIWEELSVL	220-228
	A*0201	KVAELVHFL	112-120
	A*0201	LVFGIELMEV	160-169
	A*0201	QLVFGIELMEV	159-169
	A*0202	FLWGPRALV	271-279
	A*0202	IMPKAGLLIIV	195-205
	A*0202	KIWEELSVL	220-228
	A*0202	KVAELVHFL	112-120
	A*0202	LVFGIELMEV	160-169
	A*0202	QLVFGIELMEV	159-169
	A*0203	FLWGPRALV	271-279
	A*0203	IMPKAGLLIIV	195-205
	A*0203	KVAELVHFL	112-120
	A*0203	LVFGIELMEV	160-169
MAGE A3	A*0203	QLVFGIELMEV	159-169
	A*0206	FLWGPRALV	271-279
	A*0206	IMPKAGLLIIV	195-205
	A*0206	KIWEELSVL	220-228
	A*0206	KVAELVHFL	112-120
	A*0206	LVFGIELMEV	160-169
	A*0206	QLVFGIELMEV	159-169
	A*0207	FLWGPRALV	271-279
	A*0209	FLWGPRALV	271-279
	A24	IMPKAGLLI	195-203
	A24	VAELVHFLL	113-121
	A*2402	TFPDLESEF	97-105
	A*6802	FLWGPRALV	271-279
MAGE A5	A*6802	KVAELVHFL	112-120
	A*6802	LVFGIELMEV	160-169
	A*6802	QLVFGIELMEV	159-169
	DPB1*0401	TQHFVQENYLEY	247-258
	DPB1*0401	LLKYRAREPVTKAEMLGSVVGNWQ	121-144
	DPB1*0401	YRAREPVTKAEMLG	124-137
MAGE A4	A1	EVDPASNTY	169-177
	A4	GVYDGREHTV	230-239
	A24	NYKRCFPVI	143-151
	B37	SESLKMIF	156-163

MAGE A6	A2	QLVFGIELMEV	159-169
	A*2402	NYKRCFPVI	143-151
	A*0201	LVHFLLLKY	116-124
	Cw7	EGDCAPEEK	212-220
	A*0201	KIWEELSVL	220-228
	A*0201	LVFGIELMEV	160-169
	A*0201	QLVFGIELMEV	159-169
	A*0202	KIWEELSVL	220-228
	A*0202	LVFGIELMEV	160-169
	A*0202	QLVFGIELMEV	159-169
	A*0203	LVFGIELMEV	160-169
	A*0203	QLVFGIELMEV	159-169
	A*0206	KIWEELSVL	220-228
	A*0206	LVFGIELMEV	160-169
	A*0206	QLVFGIELMEV	159-169
	DRB1*0401	LLKYRAREPVTKAEMLGSVVGNWQ	121-144
	DRB1*0401	YRAREPVTKAEMLG	124-137
	DRB1*0401	ESEFQAALSRKVAKL	102-116
MAGE-A8	DRB1*0401	FFPVIFSKASDSLQLVFGI	146-164
	DRB1*0401	IFSKASDSLQLVFGIE	150-165
	DRB1*0401	LTQYFVQENYLEYRQVPG	246-263
	DRB1*0401	VGNWQYFFPVIFSKASDSLQLVFGIELMEVD	140-170
	A2	QLVFGIELMEV	159-169
	A24	TFPDLESEF	97-105
	A52	WQYFFPVIF	143-151
	A34	MVKISGGPR	290-298
	B35	EVDPIGHVY	168-176
	B37	REPVTKAEML	127-136
	Cw7	EGDCAPEEK	212-220
	Cw16	ISGGPRISY	293-301
	DR13	LLKYRAREPVTKAE	121-134
MAGE-A9	A2	ALSVMGVYV	223-231
	A*0201, A*0202 A*0203, A*0206	KVAELVHFL	111-119
	A24	VAELVHFLL	112-120
	A*6802	KVAELVHFL	111-119
MAGE-A10	A2	GLYDGMEHL	254-262
	B53	DPARYEFLW	290-298
MAGE-A12	A2[g]	FLWGPRALV	271-279
	A2	YLQLVFGIEV	157-166
	A24	EYLQLVFGI	156-194
	Cw7	VRIGHLYIL	170-178
	Cw7	EGDCAPEEK	212-220
	DP4	REPFTKAEMLGSVIR	127-141
	A*0201	KASEYLQLV	153-161
	A*0201	LVFGIEVVEV	160-169
	A*0201	LVHFLLLKY	116-124
	A*0201	LVQENYLEY	250-258
	A*0201	YLQLVFGIEV	157-166
	A*0203	KASEYLQLV	153-161
	A*0203	LVFGIEVVEV	160-169
	A*0203	YLQLVFGIEV	157-166
	A*0206	KASEYLQLV	153-161
	A*0206	LVFGIEVVEV	160-169
	A*0206	YLQLVFGIEV	157-166
	A*2402	EYLQLVFGI	156-164
	A*6802	LVFGIEVVEV	160-169
	B40	AELVHFLLL	114-122
	DR1	ACYEFLWGPRALVETS	267-282
	DRB1*1301	AELVHFLLLKYRAR	114-127
	DR13	AELVHFLLLKYRAR	114-127
MAGE B1	A2	FLWGPRAYA	271-279
MAGE B2	A2	FLWGPRAYA	271-279
MAGE-C1	DQ6	SSALLSIFQSSPE	137-149
	DQ6	SFSYTLLSL	450-458
	DR15	VSSFFSYTL	779-787
MAGE-C2	A2	LLFGLALIEV	191-200
	A2	ALKDVEERV	336-344
	A*0201	FLAKLNNTV	317-325
	B*4403	SESIKKKVL	307-315
	B44	SESIKKKVL	307-315

APPENDIX C: PERL SCRIPTS

 # Module containing subroutines

```
use strict;
use warnings;
package mage1;
require exporter;
our @ISA=qw(Exporter);
our @EXPORT_OK=qw(getseqs splitseqs compare);
our @EXPORT=qw(getseqs splitseqs compare);
```

 # Subroutine to retrieve sequences and store in a file

```
sub getseqs ($)
{ my ($filename)=@_; my @data=();
open (FILE,"<$filename") or die "Can't open $filename $!";
@data=<FILE>;
close FILE;
return @data; }
```

 # Subroutine to retrieve sequences from fasta format

```
sub splitseqs (@)
{ my (@data)=@_;
my @headings; my @protein; my $i=0;
foreach (@data)
{ if ($_=~/^>/) {$headings[$i]=$_; next;}
else {$protein[$i]=$_;} $i++; }
return (\@headings, \@protein); }
```

 # Subroutine to do a simple alignment and compare sequences

```
sub compare ($$)
{ my ($headingsRef, $proteinRef)=@_;
my @headings=@$headingsRef; my @protein=@$proteinRef;
my $last= scalar @protein;
my $i=0; chomp ($headings[0]);
print "$headings[0]   $protein[0]";
do
{ my $first=$protein[$i];
my $second=$protein[$i+1];
my $headings=$headings[$i+1];
chomp ($headings);
my $count= (length $headings)+1;
my $length= length $second;
printf '%31s';
```

```
for my $j (0..$length-1)
{
my $base1=substr($first, $j, 1);
my $base2=substr($second, $j, 1);
if ($base1=~/^\W/ or $base2=~/^\W/) { print " "; next;} #skip gaps
else {
if ($base1 ne $base2)
{ print "|";}
else { print " ";}
}}
print "\n$headings   $second";
$i++;
}while($i<$last);
}
```

 # Subroutine to remove dashes in the sequences

```
sub remove_dashes($)
{
my $filename=@_; my $file2="clean.txt"; my $count=0;
open (FILE,"<$filename") or die "Can't $!";
open (FILE2,">$file2") or die "Can't $!";
while (<FILE>)
{
if ($_=~/^>/) {next;}else
 if ($_=~/-/) {$_=~s/-//g;} # to remove dashes
print FILE2 ($_);
}
close FILE; close FILE2;
return $file2;
}
```

 # Subroutine to change simple files into a simple fasta format

```
sub fasta ($)
{
my $filename=@_; my $filename2="fasta.txt";
open (FILE, "<$filename") or die "Can't $!";
open (FILE2, ">$filename2") or die "Can't $!";
while (<FILE>)
{ print FILE2 ("> \n $_"); }
close FILE; close FILE2; }
```

APPENDIX D: MAGE SEQUENCES

>sp|P43355|MAGA1_HUMAN Melanoma-associated antigen 1 OS=Homo sapiens GN=MAGEA1 PE=1 SV=1
MSLEQRSLHCKPEEALEAQQEALGLVCVQAATSSSSPLVLGTLEEVPTAGSTDPPQSPQGASAFPTTINFTRR
QPSEGSSSREEEGPSTSCILESLFRAVITKKVADLVGFLLLKYRAREPVTKAEMLESVIKNYKHCFPEIFGKAS
ESLQLVFGIDVKEADPTGHSYVLVTCLGLSYDGLLGDNQIMPKTGFLIIVLVMIAMEGGHAPEEEIWEELSV
MEVYDGREHSAYGEPRKLLTQDLVQEKYLEYRQVPDSDPARYEFLWGPRALAETSYVKVLEYVIKVSARR
FFFPSLREAALREEEEGV
>sp|P43356|MAGA2_HUMAN Melanoma-associated antigen 2OS=Homo sapiens GN=MAGEA2 PE=1 SV=1
MPLEQRSQHCKPEEGLEARGEALGLVGAQAPATEEQQTASSSSTLVEVTLGEVPAADSPSPPHSPQGASSFS
TTINYTLWRQSDEGSSNQEEEGPRMFPDLESEFQAAISRKMVELVHFLLLKYRAREPVTKAEMLESVLRNC
QDFFPVIFSKASEYLQLVFGIEVVEVVPISHLYILVTCLGLSYDGLLGDNQVMPKTGLLIIVLAIIAIEGDCAPE
EKIWEELSMLEVFEGREDSVFAHPRKLLMQDLVQENYLEYRQVPGSDPACYEFLWGPRALIETSYVKVLH
HTLKIGGEPHISYPPLHERALREGEE
>sp|P43357|MAGA3_HUMAN Melanoma-associated antigen 3 OS=Homo sapiens GN=MAGEA3 PE=1 SV=1
MPLEQRSQHCKPEEGLEARGEALGLVGAQAPATEEQEAASSSSTLVEVTLGEVPAAESPDPPQSPQGASSLP
TTMNYPLWSQSYEDSSNQEEEGPSTFPDLESEFQAALSRKVAELVHFLLLKYRAREPVTKAEMLGSVVGN
WQYFFPVIFSKASSSLQLVFGIELMEVDPIGHLYIFATCLGLSYDGLLGDNQIMPKAGLLIIVLAIIAREGDCA
PEEKIWEELSVLEVFEGREDSILGDPKKLLTQHFVQENYLEYRQVPGSDPACYEFLWGPRALVETSYVKVL
HHMVKISGGPHISYPPLHEWVLREGEE
>sp|P43358|MAGA4_HUMAN Melanoma-associated antigen 4 OS=Homo sapiens GN=MAGEA4 PE=1 SV=2
MSSEQKSQHCKPEEGVEAQEEALGLVGAQAPTTEEQEAAVSSSSPLVPGTLEEVPAAESAGPPQSPQGASA
LPTTISFTCWRQPNEGSSSQEEEGPSTSPDAESLFREALSNKVDELAHFLLRKYRAKELVTKAEMLERVIKN
YKRCFPVIFGKASESLKMIFGIDVKEVDPASNTYTLVTCLGLSYDGLLGNNQIFPKTGLLIIVLGTIAMEGDS
ASEEEIWEELGVMGVYDGREHTVYGEPRKLLTQDWVQENYLEYRQVPGSNPARYEFLWGPRALAETSYV
KVLEHVVRVNARVRIAYPSLREAALLEEEEGV
>sp|P43359|MAGA5_HUMAN Melanoma-associated antigen 5 OS=Homo sapiens GN=MAGEA5 PE=2 SV=1
MSLEQKSQHCKPEEGLDTQEEALGLVGVQAATTEEQEAVSSSSPLVPGTLGEVPAAGSPGPLKSPQGASAIP
TAIDFTLWRQSIKGSSNQEEEGPSTSPDPESVFRAALSKKVADLIHFLLLKY
>sp|P43360|MAGA6_HUMAN Melanoma-associated antigen 6 OS=Homo sapiens GN=MAGEA6 PE=1 SV=1
MPLEQRSQHCKPEEGLEARGEALGLVGAQAPATEEQEAASSSSTLVEVTLGEVPAAESPDPPQSPQGASSLP
TTMNYPLWSQSYEDSSNQEEEGPSTFPDLESEFQAALSRKVAKLVHFLLLKYRAREPVTKAEMLGSVVGN
WQYFFPVIFSKASDSLQLVFGIELMEVDPIGHVYIFATCLGLSYDGLLGDNQIMPKTGFLIIILAIIAKEGDCA
PEEKIWEELSVLEVFEGREDSIFGDPKKLLTQYFVQENYLEYRQVPGSDPACYEFLWGPRALIETSYVKVLH
HMVKISGGPRISYPLLHEWALREGEE
>sp|P43361|MAGA8_HUMAN Melanoma-associated antigen 8 OS=Homo sapiens GN=MAGEA8 PE=2 SV=2
MLLGQKSQRYKAEEGLQAQGEAPGLMDVQIPTAEEQKAASSSSTLIMGTLEEVTDSGSPSPPQSPEGASSSL
TVTDSTLWSQSDEGSSSNEEEGPSTSPDPAHLESLFREALDEKVAELVRFLLRKYQIKEPVTKAEMLESVIK
NYKNHFPDIFSKASECMQVIFGIDVKEVDPAGHSYILVTCLGLSYDGLLGDDQSTPKTGLLIIVLGMILMEGS
RAPEEAIWEALSVMGLYDGREHSVYWKLRKLLTQEWVQENYLEYRQAPGSDPVRYEFLWGPRALAETSY
VKVLEHVVRVNARVRISYPSLHEEALGEEKGV
>sp|P43362|MAGA9_HUMAN Melanoma-associated antigen 9 OS=Homo sapiens GN=MAGEA9 PE=2 SV=1
MSLEQRSPHCKPDEDLEAQGEDLGLMGAQEPTGEEEETTSSSDSKEEEVSAAGSSSPPQSPQGGASSSISVY
YTLWSQFDEGSSSQEEEEPSSSVDPAQLEFMFQEALKLKVAELVHFLLHKYRVKEPVTKAEMLESVIKNYK
RYFPVIFGKASEFMQVIFGTDVKEVDPAGHSYILVTALGLSCDSMLGDGHSMPKAALLIIVLGVILTKDNCA
PEEVIWEALSVMGVYVGKEHMFYGEPRKLLTQDWVQENYLEYRQVPGSDPAHYEFLWGSKAHAETSYEK
VINYLVMLNAREPICYPSLYEEVLGEEQEGV
>sp|P43363|MAGAA_HUMAN Melanoma-associated antigen 10 OS=Homo sapiens GN=MAGEA10 PE=2
MPRAPKRQRCMPEEDLQSQSETQGLEGAQAPLAVEEDASSSTSTSSSFPSSFPSSSSSSSSSCYPLIPSTPEEVS
ADDETPNPPQSAQIACSSPSVVASLPLDQSDEGSSSQKEESPSTLQVLPDSESLPRSEIDEKVTDLVQFLLFKY
QMKEPITKAEILESVIRNYEDHFPLLFSEASECMLLVFGIDVKEVDPTGHSFVLVTSLGLTYDGMLSDVQSM
PKTGILILILSIVFIEGYCTPEEVIWEALNMMGLYDGMEHLIYGEPRKLLTQDWVQENYLEYRQVPGSDPAR
YEFLWGPRAHAEIRKMSLLKFLAKVNGSDPRSFPLWYEEALKDEEERAQDRIATTDDTTAMASASSSATGS
FSYPE
>sp|P43364|MAGAB_HUMAN Melanoma-associated antigen 11 OS=Homo sapiens GN=MAGEA11 PE=1

METQFRRGGLGCSPASIKRKKKREDSGDFGLQVSTMFSEDDFQSTERAPYGPQLQWSQDLPRVQVFREQA
NLEDRSPRRTQRITGGEQVLWGPITQIFPTVRPADLTRVIMPLEQRSQHCKPEEGLQAQEEDLGLVGAQALQ
AEEQEAAFFSSTLNVGTLEELPAAESPSPPQSPQEESFSPTAMDAIFGSLSDEGSGSQEKEGPSTSPDLIDPESF
SQDILHDKIIDLVHLLLRKYRVKGLITKAEMLGSVIKNYEDYFPEIFREASVCMQLLFGIDVKEVDPTSHSYV
LVTSLNLSYDGIQCNEQSMPKSGLLIIVLGVIFMEGNCIPEEVMWEVLSIMGVYAGREHFLFGEPKRLLT
QNWVQEKYLVYRQVPGTDPACYEFLWGPRAHAETSKMKVLEYIANANGRDPTSYPSLYEDALREEGEGV
>sp|P43365|MAGAC_HUMAN Melanoma-associated antigen 12 OS=Homo sapiens GN=MAGEA12 PE=2
MPLEQRSQHCKPEEGLEAQGEALGLVGAQAPATEEQETASSSSTLVEVTLREVPAAESPSPPHSPQGASTLP
TTINYTLWSQSDEGSSNEEQEGPSTFPDLETSFQVALSRKMAELVHFLLLKYRAREPFTKAEMLGSVIRNFQ
DFFPVIFSKASEYLQLVFGIEVVEVVRIGHLYILVTCLGLSYDGLLGDNQIVPKTGLLIIVLAIIAKEGDCAPEE
KIWEELSVLEASDGREDSVFAHPRKLLTQDLVQENYLEYRQVPGSDPACYEFLWGPRALVETSYVKVLHH
LLKISGGPHISYPPLHEWAFREGEE
>sp|P43366|MAGB1_HUMAN Melanoma-associated antigen B1 OS=Homo sapiens GN=MAGEB1 PE=1
MPRGQKSKLRAREKRRKAREETQGLKVAHATAAEKEECPSSSPVLGDTPTSSPAAGIPQKPQGAPPTTTAA
AAVSCTESDEGAKCQGEENASFSQATTSTESSVKDPVAWEAGMLMHFILRKYKMREPIMKADMLKVVDE
KYKDHFTEILNGASRRLELVFGLDLKEDNPSGHTYTLVSKLNLTNDGNLSNDWDFPRNGLLMPLLGVIFLK
GNSATEEEIWKFMNVLGAYDGEEHLIYGEPRKFITQDLVQEKYLKYEQVPNSDPPRYQFLWGPRAYAETT
KMKVLEFLAKMNGATPRDFPSHYEEALRDEEERAQVRSSVRARRRTTATTFRARSRAPFSRSSHPM
>sp|O15479|MAGB2_HUMAN Melanoma-associated antigen B2 OS=Homo sapiens GN=MAGEB2 PE=1
MPRGQKSKLRAREKRRKARDETRGLNVPQVTEAEEEEAPCCSSSVSGGAASSSPAAGIPQEPQRAPTTAAA
AAAGVSSTKSKKGAKSHQGEKNASSSQASTSTKSPSEDPLTRKSGSLVQFLLYKYKIKKSVTKGEMLKIVG
KRFREHFPEILKKASEGLSVVFGLELNKVNPNGHTYTFIDKVDLTDEESLLSSWDFPRRKLLMPLLGVIFLN
GNSATEEEIWEFLNMLGVYDGEEHSVFGEPWKLITKDLVQEKYLEYKQVPSSDPPRFQFLWGPRAYAETSK
MKVLEFLAKVNGTTPCAFPTHYEEALKDEEKAGV
>sp|O15480|MAGB3_HUMAN Melanoma-associated antigen B3 OS=Homo sapiens GN=MAGEB3 PE=2
MPRGQKSTLHAREKRQQTRGQTQDHQGAQITATNKKKVSFSSPLILGATIQKKSAGRSRSALKKPQRALST
TTSVDVSYKKSYKGANSKIEKKQSFSQGLSSTVQSRTDPLIMKTNMLVQFLMEMYKMKKPIMKADMLKIV
QKSHKNCFPEILKKASFNMEVVFGVDLKKVDSTKDSYVLVSKMDLPNNGTVTRGRGFPKTGLLLNLLGVI
FMKGNCATEEKIWEFLNKMRIYDGKKHFIFGEPRKLITQDLVKLKYLEYRQVPNSNPARYEFLWGPRAHA
ETSKMKVLEFWAKVNKTVPSAFQFWYEEALRDEEERVQAAAMLNDGSSAMGRKCSKAKASSSSHA
>sp|O15481|MAGB4_HUMAN Melanoma-associated antigen B4 OS=Homo sapiens GN=MAGEB4 PE=1
MPRGQKSKLRAREKRQRTRGQTQDLKVGQPTAAEKEESPSSSSSVLRDTASSSLAFGIPQEPQREPPTTSAA
AAMSCTGSDKGDESQDEENASSSQASTSTERSLKDSLTRKTKMLVQFLLYKYKMKEPTTKAEMLKIISKKY
KEHFPEIFRKVSQRTELVFGLALKEVNPTTHSYILVSMLGPNDGNQSSAWTLPRNGLLMPLLSVIFLNGNCA
REEEIWEFLNMLGIYDGKRHLIFGEPRKLITQDLVQEKYLEYQQVPNSDPPRYQFLWGPRAHAETSKMKVL
EFLAKVNDTTPNNFPLLYEEALRDEEERAGARPRVAARRGTTAMTSAYSRATSSSSSQPM
>sp|Q9BZ81|MAGB5_HUMAN Melanoma-associated antigen B5 OS=Homo sapiens GN=MAGEB5 PE=2
MTSAGVFNAGSDERANSRDEEYPCSSEVSPSTESSCSNFINIKVGLLEQFLLYKFKMKQRILKEDMLKIVNP
RYQNQFAEIHRRASEHIEVVFAVDLKEVNPTCHLYDLVSKLKLPNNGRIHVGKVLPKTGLLMTFLVVIFLK
GNCANKEDTWKFLDMMQIYDGKKYYIYGEPRKLITQDFVRLTYLEYHQVPCSYPAHYQFLWGPRAYTET
SKMKVLEYLAKVNDIAPGAFSSQYEEALQDEEESPSQRCSRNWHYCSGQDCLRAKFSSFSQPY
>sp|Q8N7X4|MAGB6_HUMAN Melanoma-associated antigen B6 OS=Homo sapiens GN=MAGEB6 PE=2
MPRGHKSKLRTCEKRQETNGQPQGLTGPQATAEKQEESHSSSSSSRACLGDCRRSSDASIPQESQGVSPTGS
PDAVVSYSKSDVAANGQDEKSPSTSRDASVPQESQGASPTGSPDAGVSGSKYDVAANGQDEKSPSTSHDV
SVPQESQGASPTGSPDAGVSGSKYDVAAEGEDEESVSASQKAIIFKRLSKDAVKKKACTLAQFLQKKFEKK
ESILKADMLKCVRREYKPYFPQILNRTSQHLVVAFGVELKEMDSSGESYTLVSKLGLPSEGILSGDNALPKS
GLLMSLLVVIFMNGNCATEEEVWEFLGLLGIYDGILHSIYGDARKIITEDLVQDKYVVYRQVCNSDPPCYEF
LWGPRAYAETTKMRVLRVLADSSNTSPGLYPHLYEDALIDEVERALRLRA
>sp|Q96LZ2|MAGBA_HUMAN Melanoma-associated antigen B10 OS=Homo sapiens GN=MAGEB10 PE=2
MPRGQKSKLRAREKRRQARGGLEDLIDALDILEEEEESPPSASACLKDVFQSSLDGASNNPHGLREAQSTST
SATAASHTRHPEGVNDQMEERPICTQDLEATDSFPRGPVDEKVIILVHYLLYKYQMKEPITKADMLRNVTQ
MSKSQFPVILSRASEHLELIFGLDLKEVEPNKHIYVLVNKLDLGCDAKLSDETGVPKTGLLMTVLGIIFTNG
NCVAEEEVWKVFNTMGLYDGIEHFMFGEPRKLLTKDLVKENYLEYQQVPNSDPPRYQFLWGPRAHAETS
KMKVLEFLAKVNDTAPSEFSNWYTEALQDEEERARARVAAKARVSATAGARSKVKSSKSSQLQ

>sp|A2A368|MAGBG_HUMAN Melanoma-associated antigen B16 OS=Homo sapiens GN=MAGEB16 PE=4
MSQDQESPRCTHDQHLQTFSETQSLEVAQVSKALEKTLLSSSHPLVPGKLKEAPAAKAESPLEVPQSFCSSSI
AVTTTSSSESDEASSNQEEEDSPSSSEDTSDPRNVPADALDQKVAFLVNFMLHKCQMKKPITKADMLKIIIK
DDESHFSEILLRASEHLEMIFGLDVVEVDPTTHCYGLFIKLGLTYDGMLSGEKGVPKTGLLIIVLGVIFMKGN
RATEEEVWEVLNLTGVYSGKKHFIFGEPRMLITKDFVKEKYLEYQQVANSDPARYEFLWGPRAKAETSKM
KVLEFVAKVHGSYPHSFPSQYAEALKEEEERARARI
>sp|A8MXT2|MAGBH_HUMAN Melanoma-associated antigen B17 OS=Homo sapiens GN=MAGEB17
MKHAGECHGQGAILCKVCGRKTVKEIAGIWHFPGSLELMRQEPRCEAPCRRTEDTVIPEYWRSNRDQPVA
ALRLRGPQAGLRLSISTIILQGSQEHERFQEDKLRCPPPAPSCLLPLSTVIMPRGQASKRRAREKRRQARGED
QCLGGAQATAAEKEKLPSSSSPACQSPPQSFPNAGIPQESQRASYPSSPASAVSLTSSDEGAKGQKGESPNSF
HGPSSSESTGRDLLNTKTGELVQFLLNKYIRKEPITREAMLKVINRKYKQHFPEILRRSTENVEVVFGLYLKE
MDPSRQSYVLVGKLDFPNQGSLSDGGGFPLSGLLMVLLSTIFMHGNRATEEEMWECLNALGMYKGRKHFI
YGEPQELVTKDLVREGYLEYQQVPSSDPPRYEFLWGPRARAETSKMKVLEFVAKLNDTVASTYKSRYEEA
LREEEEQARARAVARDSARARASRSFQP
>sp|Q96M61|MAGBI_HUMAN Melanoma-associated antigen B18 OS=Homo sapiens GN=MAGEB18 PE=1
MPRGQKSKLRAREKRHQARCENQDLGATQATVAEGESPSPAYLLFGDRPQNLPAAETPSIPEALQGAPSTT
NAIAPVSCSSNEGASSQDEKSLGSSREAEGWKEDPLNKKVVSLVHFLLQKYETKEPITKGDMIKFVIRKDKC
HFNEILKRASEHMELALGVDLKEVDPIRHYYAFFSKLDLTYDETTSDEEKIPKTGLLMIALGVIFLNGNRAP
EEAVWEIMNMMGVYADRKHFLYGDPRKVMTKDLVQLKYLEYQQVPNSDPPRYEFLWGPRAHAETSKM
KVLEFVAKIHDTVPSAFPSCYEEALRDEEQRTQARAAARAHTAAMANARSRTTSSSFSHAK
>sp|O60732|MAGC1_HUMAN Melanoma-associated antigen C1 OS=Homo sapiens GN=MAGEC1 PE=1
MGDKDMPTAGMPSLLQSSSESPQSCPEGEDSQSPLQIPQSSPESDDTLYPLQSPQSRSEGEDSSDPLQRPPEG
KDSQSPLQIPQSSPEGDDTQSPLQNSQSSPEGKDSLSPLEISQSPPEGEDVQSPLQNPASSFFSSALLSIFQSSPE
STQSPFEGFPQSVLQIPVSAASSSTLVSIFQSSPESTQSPFEGFPQSPLQIPVSRSFSSTLLSIFQSSPERTQSTFEG
FAQSPLQIPVSPSSSSTLLSLFQSFSERTQSTFEGFAQSSLQIPVSPSFSSTLVSLFQSSPERTQSTFEGFPQSPLQI
PVSSSSSSTLLSLFQSSPERTHSTFEGFPQSLLQIPMTSSFSSTLLSIFQSSPESAQSTFEGFPQSPLQIPGSPSFSS
TLLSLFQSSPERTHSTFEGFPQSPLQIPMTSSFSSTLLSILQSSPESAQSAFEGFPQSPLQIPVSSSFSYTLLSLFQS
SPERTHSTFEGFPQSPLQIPVSSSSSSSTLLSLFQSSPECTQSTFEGFPQSPLQIPQSPPEGENTHSPLQIVPSLPE
WEDSLSPHYFPQSPPQGEDSLSPHYFPQSPPQGEDSLSPHYFPQSPQGEDSLSPHYFPQSPPQGEDSMSPLYFP
QSPLQGEEFQSSLQSPVSICSSSTPSSLPQSFPESSQSPPEGPVQSPLHSPQSPPEGMHSQSPLQSPESAPEGEDS
LSPLQIPQSPLEGEDSLSSLHFPQSPPEWEDSLSPLHFPQFPPQGEDFQSSLQSPVSICSSSTSLSLPQSFPESPQS
PPEGPAQSPLQRPVSSFFSYTLASLLQSSHESPQSPPEGPAQSPLQSPVSSFPSSTSSSLSQSSPVSSFPSSTSSSL
SKSSPESPLQSPVISFSSSTSLSPFSEESSSPVDEYTSSSDTLLESDSLTDSESLIESEPLFTYTLDEKVDELARFL
LLKYQVKQPITKAEMLTNVISRYTGYFPVIFRKAREFIEILFGISLREVDPDDSYVFVNTLDLTSEGCLSDEQG
MSQNRLLILILSIIFIKGTYASEEVIWDVLSGIGVRAGREHFAFGEPRELLTKVWVQEHYLEYREVPNSSPPR
YEFLWGPRAHSEVIKRKVVEFLAMLKNTVPITFPSSYKDALKDVEERAQAIIDTTDDSTATESASSSVMSPSF
SSE
>sp|Q9UBF1|MAGC2_HUMAN Melanoma-associated antigen C2 OS=Homo sapiens GN=MAGEC2 PE=1
MPPVPGVPFRNVDNDSPTSVELEDWVDAQHPTDEEEEEASSASSTLYLVFSPSSFSTSSSLILGGPEEEEVPS
GVIPNLTESIPSSPPQGPPQGPSQSPLSSCCSSFSWSSFSEESSSQKGEDTGTCQGLPDSESSFTYTLDEKVAEL
VEFLLLKYEAEEPVTEAEMLMIVIKYKDYFPVILKRAREFMELLFGLALIEVGPDHFCVFANTVGLTDEGSD
DEGMPENSLLIIILSVIFIKGNCASEEVIWEVLNAVGVYAGREHFVYGEPRELLTKVWVQGHYLEYREVPHS
SPPYYEFLWGPRAHSESIKKKVLEFLAKLNNTVPSSFPSWYKDALKDVEERVQATIDTADDATVMASESLS
VMSSNVSFSE
>sp|Q8TD91|MAGC3_HUMAN Melanoma-associated antigen C3 OS=Homo sapiens GN=MAGEC3 PE=1
MLLPCHWVLDATFSDGSLGQWVKNTCATYALSPVVLPPQPQPRKKATDKDYSAFHLGHLREVRLFLRGG
TSDQRMDSLVLCPTYFKLWRTLSGSPGLQLSDLHFGSQPEGKFSLRRAVSVKQREEPQDWPLNEKRTLWK
DSDLPTWRRGTGYTLSLPAVSPGKRLWGEKAGSLPESEPLFTYTLDEKVDKLVQFLLLKYQAKEPLTRAE
MQMNVINTYTGYFPMIFRKAREFIEILFGISLTEVDPDHFYVFVNTLDLTCEGSLSDEQGMPQNRLLILILSVI
FIKGNCASEEVIWEVLNAIGPWSALAGFADVLSRLALWESEGPEAFCEESGLRSAEGSVLDLANPQGLAGH
RQEDGRRGLTEASPQQKKGGEDEDMPAAGMPPLPQSPPEIPPQGPPKISPQGPPQSPPQSPLDSCSSPLLWTR
LDEESSSEEEDTATWHALPESESLPRYALDEKVAELVQFLLLKYQTKEPVTKAEMLTTVIKKYKDYFPMIF
GKAHEFIELIFGIALTDMDPDNHSYFFEDTLDLTYEGSLIDDQGMPKNCLLILILSMIFIKGSCVPEEVIWEVL
SAIGPIQRPAREVLEFLSKLSSIIPSAFPSWYMDALKDMEDRAQAIIDTTDDATAMASASPSVMSTNFCPE
LAMLKNTVPITFPSSYKDA