In-Silico Identification of Biomarkers and Vaccine Candidates for Advancement of Lung Cancer Therapeutics

By

Anjali Lathwal

Under the Supervision of Prof. Gajendra P.S. Raghava

Department of Computational Biology

Indraprastha Institute of Information Technology Delhi

New Delhi - 110020

September, 2020

In-Silico Identification of Biomarkers and Vaccine Candidates for Advancement of Lung Cancer Therapeutics

By

Anjali Lathwal

A Thesis

Submitted in Partial Fulfillment of the Requirements for the Degree Of

Doctor of Philosophy

Department of Computational Biology

Indraprastha Institute of Information Technology Delhi

New Delhi - 110020

September, 2020

Certificate

This is to certify that the thesis titled "**In-Silico Identification of Biomarkers and Vaccine Candidates for Advancement of Lung Cancer Therapeutics**" being submitted by **Ms. Anjali Lathwal** to the Indraprastha Institute of Information Technology Delhi, for the award of the degree of Doctor of Philosophy, is an original research work carried out by her under my supervision. In my opinion, the thesis has reached the standards fulfilling the requirements of the regulations relating to the degree.

The results contained in this thesis have not been submitted in part or full to any other university or institute for the award of any degree/diploma.

September, 2020
Month, Year

Prof. Gajendra P.S. Raghava
Supervisor Name

Indraprastha Institute of Information Technology Delhi
New Delhi 110 020

Acknowledgements

"An investment in education, knowledge, research, and learning pays the best interest".

The work documented in this thesis would not have been possible without the encouragement and support of many people. Here, I will take the opportunity to express my sincere thanks to all those people who have made this Ph.D. journey wonderful.

No words can be enough to appreciate a "Guru" who empowers you to chase your dreams. Here, with a feeling of immense respect and gratitude, I bow down to my supervisor **Prof. Gajendra P.S. Raghava** *for his continuous support, motivation, encouragement, and guidance throughout my research. I have no right words that can reflect his contribution to my Ph.D. He motivated me to stay focused, enthusiastic and imaginative about the research. At the same time, he taught me that one must not be afraid of making mistakes and eventually should learn from them. Every discussion with him inspires me to learn more and contribute to the betterment of people around us as he always says that "one can only succeed best and quickest by helping others". Lastly, I want to say that I will be always indebted to him for showing me the right way to learn, progress, and stay motivated for living well.*

I feel a deep sense of gratitude towards **Prof. Pankaj Jalote and Prof. Ranjan Bose** *for their interactive and motivating sessions with the students, which not only inspired them to do the exciting research but shaped their minds to perform great at different platforms in future. I am highly thankful to all the teaching and research staff in IIIT-Delhi, who gave me an exciting experience of learning something new every day. I want to mention special thanks to* **Dr. Subhadip**

Raychaudhuri, Dr. Sriram K, Dr. Ganesh Bagler, Dr. Tavpritesh Sethi, and Dr. Rohit Gupta *for their timely support and motivation that helped me to stay firm during the difficult phases of Ph.D. I want to say thanks to* ***Ms. Priti Patel, Ms. Shipra Jain and Ms. Sheetu*** *along with others in the administrative department for helping me with all the academia related paper works, wherever needed.*

A special thanks to the funding agencies; University Grants Commission and IIIT-Delhi for providing me with financial aid to support my research work. I would like to praise the FMS and mess staff for providing great home-like facilities and food at IIIT-Delhi. I am highly thankful to the IT helpdesk and other system administrator staff for the management of the computer facilities at IIIT-Delhi, which aided in the smooth conduct of the research work.

I want to mention special thanks to my lab mates and colleagues for providing me a great and supportive working environment. I am deeply grateful to my seniors Dr. Pawan Kumar Raghav, Dr. Akshara, Dr. Sherry, Dr. Piyush, and Dr. Salman for their timely help and guidance, wherever required. They have been a constant source of learning in the lab. I want to mention special thanks to Dr. Pawan Kumar Raghav for his encouraging and motivating discussions with me in the lab. I want to thanks my lab mates Dilraj, Rajesh, Harpreet, Chakit, Vinod, Anjali Dhall, Sumeet, and Neelam for their support.

I am feeling profound happiness and pride in acknowledging the role of some great friends that were no less than the pillars of strength during my Ph.D. years. I got many great friends for a lifetime in the form of Mehak, Sarita, Keshav, Rajesh, Dr. Venkatesh (Venky), Harpreet, Dilraj, Divya, Tania, Aakarsh, and Chakit. They have been a great source of happiness, encouragement, strength and support during my Ph.D. journey. Mehak and Sarita were always there to share every high *and low moment during my research time at IIIT-Delhi. I found a great supportive friend in Rajesh, who inspired me to remain calm and focused during one of the most difficult phases of my Ph.D. He is a very hard working person and helped me to bring out the best in me, even when I myself was doubtful about the same. We had lots of fruitful discussions on a variety of topics ranging from our research work to agriculture to Vedas. Venky is the most positive, cheerful and joyful person that I have ever met. He taught me to look beyond his/her sufferings to have a positive outlook on life and people. Harpreet has been a great support both in professional and personal life. She is one genuine friend who can take stand for you even at your back. Dilraj and Chakit have been the constant source of support, whenever needed. My first few outings in Delhi were with Dilraj and Chakit, which I thoroughly enjoyed. I have also met many wonderful people who became good friends over time such as Dr. Monika, Indraprakash, Sphoorti, Aditya, Shiju, Neetesh, Chitrita, Smriti, Krishan, Omkar, Shreya, Raghav, and Dr. Indra. Last but not the least; I want to thank my colleagues at the Department of Computational Biology for being a great source of strength and support.*

No words can ever be strong enough to explain the gratitude to my parents, ***Smt. Anita Devi and Shri. Ravinder Singh Lathwal*** *for their unconditional love, care, and guidance. My brother* ***Aneesh*** *always supported me even when I didn't want him to and was there to just listen whenever I needed it. They have encouraged, motivated, and tolerated me during the most complex and demanding times. I would like to thank my uncle* ***Shri. Parmod Kumar*** *as he is the one who encouraged me to go for higher studies and without his support, I could not have dreamt of going for a Ph.D. degree. I want to extend my gratitude towards the universe and my extended family for their immense blessings.*

Anjali Lathwal

Abstract

According to the World Health Organization report, around 10 million new cases are diagnosed with cancer in the year 2018 alone. Out of these, nearly 45% of the cancer incidences were reported from the Asian countries, 26% from European Union, 15% from the North American continent, 6% from the African countries, and 7% in the Latin American countries. These statistics highlight that cancer is a global problem, and alone is responsible for millions of premature deaths. Among the cancer types, lung cancer ranks first in terms of new incidences and mortality rates in all around the world population. It is highly heterogeneous, and despite the heterogeneity, several other factors such as infectious viruses, smoking, and drinking also correlate with the development of lung cancer. Globally it shares approximately 19% of all cancer-related deaths and 11.6% of all newly diagnosed cases. In terms of mortality rate, it ranks first in men and second in the case of females after breast cancer. The report suggests that the median survival time among the patients at an advanced stage of lung cancer is reduced to just 4.5 months, provided there is no treatment given. However, the addition of bevacizumab, along with other drugs, improved the life expectancy of the patients, but still, it is far from satisfactory. The use of all the targeted chemo-therapies suffers from several limitations - the occurrence of drug resistance, toxic nature of the drug, treatment failure, relapse among the patients, delayed wound healing, and many more. Thus oncologists and researchers all around the world are in continuous search of the alternative molecule that can advance and guides lung cancer therapeutics. After successful application in many cancer types, immunotherapy is considered as an alternate and most advanced strategy for the treatment of cancer. The FDA already approved several immunotherapeutic agents in the form of the oncolytic virus-based drug, interleukin, checkpoint blockade for the treatment of various cancer types. The World Health Organization report suggests that nearly 60% of the mortality rate among the patients can be prevented by improving diagnostic, screening, and therapeutic strategies. Among the therapeutics, vaccinations seem to be an effective measure to prevent new incidences of lung cancer caused by viruses. However, the identification and screening of a new class of immunotherapeutic molecules with experimental studies require a lot of time and resources. Thus the present thesis focuses on the development of computational tools that can find their way in aiding and guiding lung cancer therapeutics. The emphasis is given on the development of a web-resource providing up to date experimental information on oncolytic viruses used in cancer therapy; identification of subunit vaccine candidates against lung cancer-causing oncogenic viruses to be used in providing prophylactic immunity; prognostic biomarkers identification for the major subclasses of non-small-cell lung cancer that can serve the basis of precision therapy; and developing a machine learning-based prediction algorithm for the identification and designing of interleukin-2 inducing peptides. The developed web-resource on the oncolytic virus is "OvirusTdb" which is freely available to the scientific community at https://webs.iiitd.edu.in/raghava/ovirustdb/. It catalogues 5927 records against 25 fields, which were manually curated from the 166 and 27 research articles and patents, respectively. In addition to this, the web-resource holds extensive experimental information on 24 oncolytic virus species, 300 genetically modified oncolytic virus strains, 124 cancer types, 400 cancer cell lines, and 22 model organisms. The web-resource, which holds information about identified proteomic based subunit vaccine candidates against 09 oncogenic virus species, is "VLCvirus," which is freely available to the scientific community at https://webs.iiitd.edu.in/raghava/vlcvirus/. The web-resource provides detailed information on 125 identified best antigenic epitopes having MHC class-I, II binding, B-cell, T-cell, and vaccine adjuvant acting potential. Moreover, the study also identified epitope sequences "VMFVSRVPV," "LRRFMVALI," that shows binding potential to nearly 15 MHC class-I and 49 class II molecules, respectively. In addition to this, the study also identifies 25 promiscuous epitopes that are present in multiple viral strains/species, with the majority of them related to E1 and E6 envelope genes. Further to capture the heterogeneity of lung cancer and its basis in advancing the therapeutics, non-small cell lung carcinoma subtype-specific biomarkers have been identified using the Univariate Cox regression and prognostic index-based models. The Univariate Cox analysis identifies 1334 and 2129 genes of some survival predicting potential in lung squamous cell carcinoma (LUSC) and lung adenocarcinoma (LUAD) datasets, respectively. Using random forest variable hunting technique and iterative search approach, we came up with a minimum number of the gene set that can be used as a subtype-specific prognostic biomarker; the study identifies 05 "(*KIF16B, KLK7, LONRF3, OPLAH, RIPK3*)" and 04 "(*AHSG, DKK1, MGAT5B, NEMP2*)" genes for LUSC and LUAD, respectively. Since literature evidence suggests that the mutant version of interleukin-2 has a high therapeutic index, we developed the machine learning-based prediction algorithm, which is capable of identifying the interleukin-2 inducing potential from a given sequence. The random forest-based hybrid model (dipeptide composition + length) achieved a maximum accuracy of 73.25, with an MCC of 0.46 and AUC of 0.73. The whole prediction algorithm is integrated into the form of a web-server, which is freely available to the scientific community at https://webs.iiitd.edu.in/raghava/il2pred/.

Each of the developed web-resource and algorithms will have the potential to guide the therapeutics of lung cancer. For example, the data stored in the "OvirusTdb" have the potential to be utilized by genetic engineers and biotechnologists for the designing of new oncolytic viruses with the improved anti-cancer response. It can also be serving the basis for the design of experimental protocols for further enhancing the drug efficacy of the existing anti-cancer drugs. The identified 125 best antigenic epitopes can be utilized in clinics for providing immunity against lung cancer-causing viruses. The identified promiscuous epitopes can also be used in offering vaccination to a large human population due to their broad coverage of MHC alleles. Moreover, the identified promiscuous epitopes across the virus strain/species can be utilized in providing heterologous immunity against the concerned pathogenic/viral species. The identified gene biomarkers for non-small cell lung carcinoma have the potential to be investigated for therapeutic and diagnostic possibilities in the form of more subtype-specific interventions. The developed "IL2Pred" server would find its way in clinics for the identification and designing of a mutant version of interleukin-2 inducing peptides more economically as the experimental setup is time-consuming and require a lot of resources. Thus, we conclude that the identified epitopes, biomarker, and interleukin-2 inducing peptides from this study have the potential to be utilized in clinics for aiding and advancing lung cancer therapeutics.

List of Publications

Thesis Related Publications

- ❖ **Lathwal A**, Kumar R, Raghava GPS. Computer-aided designing of oncolytic viruses for overcoming translational challenges of cancer immunotherapy, Drug Discovery Today. (2020)

- ❖ **Lathwal A**, Kumar R, Raghava GPS. OvirusTdb: A database of oncolytic viruses for the advancement of therapeutics in cancer, Virology. (2020)

- ❖ **Lathwal A**, Kumar R, Arora C, Raghava GPS. Identification of prognostic biomarkers for major subtypes of non-small-cell lung cancer using genomic and clinical data, Journal of Cancer Research and Clinical Oncology. (2020)

- ❖ **Lathwal A**, Kumar R, Raghava GPS. Computation of subunit vaccine candidates against lung cancer-associated oncogenic viruses, Computers in Biology and Medicine. (2021)

Other Publications

- ❖ **Lathwal A**, Arora C, Raghava GPS. Prediction of risk scores for colorectal cancer patients from the concentration of proteins involved in mitochondrial apoptotic pathway, PLoS One. (2019)

- ❖ Kumar R, **Lathwal A**, Kumar V, Patiyal S, Raghav PK, Raghava GPS. CancerEnD: A database of cancer associated enhancers, Genomics. (2020)

- ❖ Arora C, Kaur D, **Lathwal A**, Raghava GPS. Risk prediction in cutaneous melanoma patients from their clinico-pathological features: superiority of clinical data over gene expression data, Heliyon. (2020)

- ❖ Kumar R, **Lathwal A**, Raghava GPS. Identification of prognostic potential of alternative splicing in pancreatic adenocarcinoma. [***Under Review***]

Table of Contents

List of Abbreviations

Abbreviations	Explanation
OV	Oncolytic Virus
T-VEC	Talimogene Laherperepvec
FDA	Food and Drug Administration
TAA	Tumour-Associated Antigens
USPTO	United States Patent and Trademark Office
HTML	Hyper Text Markup Language
PHP	Hypertext Pre-Processor
LAMP	Linux Apache MySQL PHP based Server
WHO	World Health Organization
HPV	Human Papillomavirus
HBV	Hepatitis B Virus
HTLV	Human T-lymphotropic Virus
RSV	Raus Sarcoma Virus
IL-2	Interelukin-2
IL-4	Interleukin-4
IFN-Gamma	Interferon Gamma
MV	Measles Virus
EBV	Epstein-Barr Virus
STLV	Simian T-cell Lymphotropic Virus

BLV	Bovine Leukemia Virus
JCV	JC Virus
IEDB	Immune Epitope Database
NSCLC	Non-small-cell Lung Carcinoma
LUAD	Lung Adenocarcinoma
LUSC	Lung Squamous Cell Carcinoma
BAATLE	Biomarker-integrated Approaches of Targeted Therapy for Lung Cancer Elimination
TCGA	The Cancer Genome Atlas
GEO	Gene Expression Omnibus
OS	Overall Survival
GPM	Good Prognostic Markers
BPM	Bad Prognostic Markers
RF-vh	Random Forest Variable Hunting
PI	Prognostic Index
HR	Hazard Ratio
IFN-α	Interferon Alpha
IL-2	Interleukin 2
NK cells	Natural Killer Cells
Treg	Regulatory T-cells
DC	Dendritic Cells

TNF-α	Tumour Necrosis Factor Alpha
AAC	Amino Acid Composition
DPC	Dipeptide Composition
TPC	Tripeptide Composition
CTD	Conjoint Triad Descriptors
CeTD	Composition enhanced Transition and Disritbution
RF	Random Forest
DT	Decision Tress
MLP	Multi-layer Perceptron
KNN	K-nearest Neighbours
SVR	Support Random Vector with Radial Bias
NN	Neural Network
TP	True Positive
TN	True Negative
FP	False Positive
FN	False Negative
MCC	Matthews correlation coefficient
AUC	Area Under Curve
TSL	Two Sample Logo

List of Software

Package	Availability
MERCI	https://dtai.cs.kuleuven.be/software/merci
ProPred	https://webs.iiitd.edu.in/raghava/propred/
ProPred1	https://webs.iiitd.edu.in/raghava/propred1/
LBtope	https://webs.iiitd.edu.in/raghava/lbtope/
CTLPred	https://webs.iiitd.edu.in/raghava/ctlpred/
VaxinPAD	https://webs.iiitd.edu.in/raghava/vaxinpad/
IL4pred	https://webs.iiitd.edu.in/raghava/il4pred/
IFNepitope	https://webs.iiitd.edu.in/raghava/ifnepitope/
IEDB	https://www.iedb.org/
TCGA	https://portal.gdc.cancer.gov/
GEO	https://www.ncbi.nlm.nih.gov/geo/
Programming languages	HTML, PHP, MYSQL, R, PYTHON, CSS and Javascript

List of Figures

List of Tables

1. Introduction

1.1 Overview of Cancer

Worldwide, cancer alone is responsible for millions of premature deaths each year. Reports say that cancer is the 2nd major reason of deaths around the globe after cardiovascular diseases (Nagai & Kim, 2017). It is estimated that around 10 million new patients had been diagnosed with cancer in 2018. Among all diagnosed cases, nearly 53.4% were males and 46.6% were females. Around 45% of all diagnosed cases were from Asia, 26% from European Union, 15% from the North American continent, 6% from the African countries, and 7% in the Latin American countries (R. L. Siegel, Miller, & Jemal, 2020). Cancer is a disease of uncontrolled cell division. Cancer cells arise from the healthy cells of the body due to genomic instability events such as DNA mutation and recombination events (Sharma, Dave, Sanadya, Sharma, & Sharma, 2010). Several viruses, bacteria, UV-rays, chemical compounds contribute to the genomic instability that leads to the development of cancer phenotype. The rapidly proliferating cancer cells even acquire the property of invading the other tissues. If the body's immune system is not capable enough of eliminating these rapidly proliferating cells, the cancer situation can become worse (Gonzalez, Hagerling, & Werb, 2018). The Global Burden of Disease report states that the mortality rate of cancer is high among older people having age above 70 years and in low-income countries (H. Wang et al., 2016). As the world population is in the decline phase of the aging pyramid, and with the advanced age, the body immune system goes weak, we can hypothesize that cancer soon may surpass the cardiovascular disease in terms of mortality rate.

1.2 Prevalence of Lung Cancer

Among all the types of cancers, internationally, lung carcinoma ranks first in terms of prevalence around the world population. It is also the headmost reason of mortality in the world population when both sexes combined (Barta, Powell, & Wisnivesky, 2019). It shares approximately 19% of all cancer-related deaths around the globe. The American Cancer Society report estimated that around 228220 new cases would be diagnosed with lung cancer in 2020 alone in the US, out of which 116300 will be in men, and 112520 will be in women. The report also estimated that around 60% of all diagnosed patients would die because of lung cancer within a year (https://www.cancer.org). Each year lung carcinoma alone is the reason behind more deaths than the other four cancers - breast, colon, prostate, and pancreas combined. Several factors contribute to the high mortality rate among lung cancer patients. The prime factor is that often the patients are diagnosed at a late stage of disease (Barta et al., 2019). There is minimal number of patients which are diagnosed before the age of 45 years. The moderate age at the time of diagnosis for the lung cancer patients is around 70 years. With old age, the body's immune system also becomes weak, which may make cancer cells highly probable to evade the immune system and its response by a variety of mechanisms.

1.3 Risk Factors for Lung Cancer

Knowing the cause of the disorder is an essential step towards finding the cure. The pursuit of finding the decisive risk factor of lung cancer pathogenesis is of utmost importance for both analytical and therapeutics design. In lung cancer pathogenesis, both intrinsic and non-intrinsic risk factors are involved, as reported in several literature studies. Figure 1.1 shows the list of factors that can contribute to lung cancer pathogenesis.

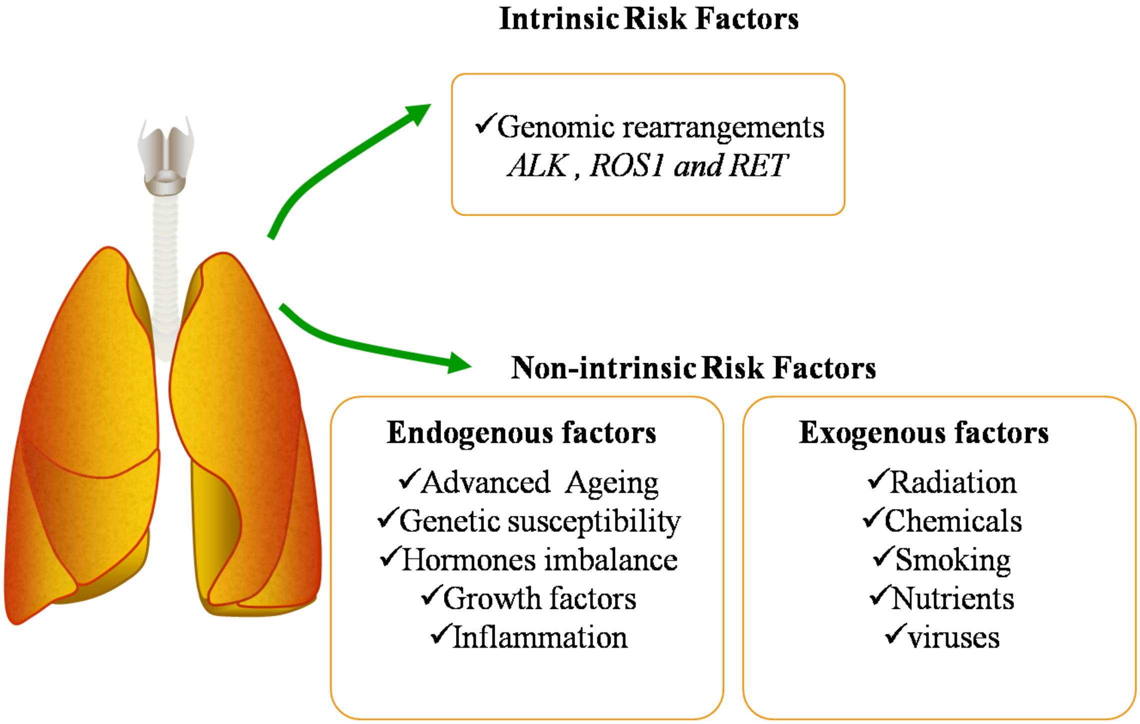


Figure 1.1: The major risk factors involved in lung cancer pathogenesis.

The intrinsic risk factors typically include the random genomic aberrations in the genome of cells, which are not modifiable. Notable examples of genomic aberrations in lung cancer include the *ALK* rearrangements, *EFGR*, and *KRAS* mutations (Devarakonda, Morgensztern, & Govindan, 2015). Ageing, hormone imbalances, radiation, and viruses are some factors that come under non-intrinsic risk factors, which may be altered partially or wholly.

1.4 Therapeutic Strategies for Lung Cancer

Lung cancer is highly metastatic and invasive in nature. Several treatment strategies are available for the lung cancer patients. The basic treatment protocol for the stage I and II lung carcinoma patients includes the surgery. It was observed that the 5-year survival rate among the patient of stage I and II after surgery was 60-80% and 30-50%, respectively (Tanoue & Detterbeck, 2009). Radiotherapy is also one available option for stage I and II lung cancer patients where resectable surgery is not possible. Approximately 70% of lung cancer cases diagnosed at a late stage, such as III and IV. Stage III patients constitute around 20-25% of cases, and the median survival rate among stage III patients is 3-7% (Mountain, 1997). The combination of chemotherapy and radiotherapy is the recommended strategy for stage III lung cancer patients. Stage IV constitutes approximately 40-50% of all newly diagnosed cases. The survival rate of the patients with stage IV is around 4.5 months, provided no treatment was available for the patients. Several treatment regimens are available for patients at stage IV, which includes first-line chemotherapy – such as cisplatin and carboplatin; the combination of chemotherapy with targeted therapy – cisplatin plus bevacizumab; second and third-line treatment options include the use of docetaxel, pemetrexed, and erlotinib, etc. The data from clinical trials shows that all these drugs enhance the survival of the patient.

1.5 Limitation of the Current Therapies

Despite so much advancement in the therapeutic regimens, lung cancers continue to be a challenge for scientists and researchers. The occurrence of drug resistance against the anti cancer drug regimen poses a challenge to scientists and oncologists. The cancer relapse among patients treated with chemo and radiotherapy has also been observed (Uramoto & Tanaka, 2014). The standard chemotherapeutic drugs are often cytotoxic in nature and have severe side effects and thus reduced quality of life. Treatment failure is also observed in several patients. At the advanced stage, surgery is not a viable option for the patient's treatment. All these points highlight that there is an urgent need for some alternative therapies that can overcome the limitation associated with the current conventional approaches and also can activates the immune system against cancer to minimize its relapse after treatment.

1.6 Origin of Proposal

Even though the cure has been achieved in the early stages of lung cancer by using chemotherapy, radiotherapy, and surgery, it still remains a significant challenge for clinicians. According to the World Health Organization report, as the industrialization increases in the world, the cases of lung cancer are an all-time high. All around the globe, a vast amount of resources and money is devoted to cancer research, but still, the success rate of available therapies among cancer patients is not so significant. The clinicians and researchers still are not able to find a definitive cure for cancer, which kills millions of people every year. Hence, providing the proper cancer treatment is a challenging task in the field of biomedical sciences. Oncologists and researchers all around the world are continuously working to ensure better odds of survival for cancer patients. In general, it is estimated that nearly 60% of the mortality rate among the patients can be prevented by improving diagnostic, screening, and therapeutic methods (Colditz & Wei, 2012). Considering the fact that much of the survival of cancer patients is linked with early detection and advanced therapeutics, this opens an opportunity for researchers around the globe to find reliable predictive, prognostic and therapeutic strategies against cancer. With the advancement in the cancer genomics studies, several research groups have identified the predictive and prognostic biomarkers against many cancer types. These biomarkers have been utilized in clinics for guiding the therapeutics. Notable examples of biomarkers include activating *EGFR* mutations predicting response to *TKI* therapy, and activating *KRAS* mutation predicting resistance to *EGFR* directed antibody therapy.

While cancer genomics studies have significantly improved our understanding of cancer biology, but there is still a lot more to understand. An ideal cancer therapeutic molecule is the one that not only serves the basis of targeted therapy but also possesses immune-boosting potential. In this regard, immunotherapy is considered as an alternate and most advanced approach for the cancer treatment. The Food and Drug Administration also passed the Interleukin-2 based therapy for treating melanoma. The literature studies suggest that the mutants of interleukin have a high therapeutic index as compared to wild-type cytokines (Chen et al., 2018). In the advanced stage, cancer cells often become highly heterogeneous, which may account for a varied response towards targeted therapies. The heterogeneity also results in complex behavior of cancer cells, which may further lead to drug resistance (Pucci, Martinelli, & Ciofani, 2019). While the literature studies identify several biomarkers and the causative agents for lung cancer, but still, the mortality rate among patients is high. One possible explanation could be the past studies are still not able to differentiate among the subtypes of lung cancer. The treatment option and targeted therapies should be designed, keeping the tumour heterogeneity in mind. In this regard, a deeper understanding of the cancer tissue type heterogeneity is of fundamental importance, which can further guide precision biomarkers and therapies. The World Health Organization report also suggests that nearly 25% of lung cancer cases are because of virus infection. The role of oncogenic viruses in cancer onset has mostly been unexplored, and fewer efforts are given to design a safe and reliable preventive vaccination strategy. Against these backdrops, the present studies provide some computational approaches that not only advance the cancer immunotherapy but also suggest some prognostic biomarkers and preventive measures that can be utilized in clinical/ research settings.

1.7 Objectives of the Thesis

To fill the lacuna of cancer therapy, the current study focuses on the utilization of available data from various literature resources for the development of computational tools, which can aid in advancing cancer therapeutics. The particular focus is given on the development of a computational resource for oncolytic viruses used in cancer therapy, identifying and prioritizing subunit vaccine candidates against lung cancer-causing oncogenic viruses as a prophylactic measure, prognostic biomarkers identification for the major subclasses of non-small-cell lung cancer that can serve the basis of precision therapy, and developing a prediction method for the identification of interleukin-2 inducing peptides that can further advance the immunotherapeutic against cancer. The following objectives were designed –

- Development of a knowledge base on oncolytic viruses.
- Identification of vaccine candidates against lung cancer-causing viruses.
- Identification of subtype-specific prognostic biomarkers of non-small cell lung cancer
- Prediction and designing of interleukin-2 inducing peptides.

1.8 Organization of Thesis Chapters

For the sake of clarification of the objectives as mentioned above, the whole thesis is split into seven chapters. Overall, the thesis organization is briefly described in Figure 1.2.

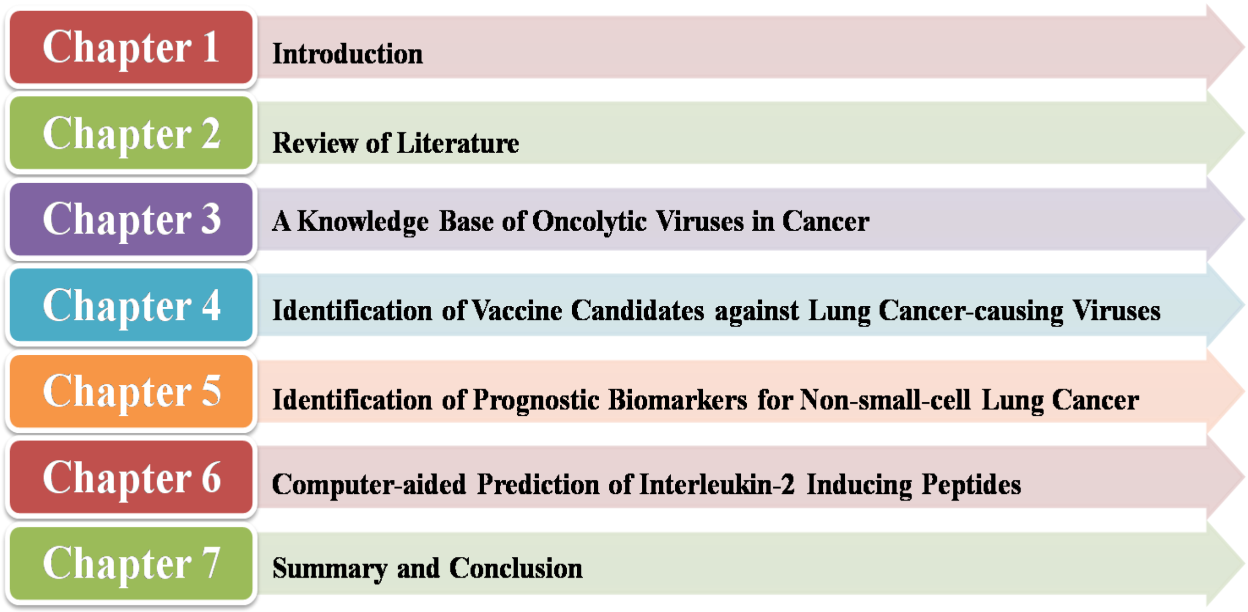

Chapter 1
Introduction
Chapter 2
Review of Literature
Chapter 3
A Knowledge Base of Oncolytic Viruses in Cancer
Chapter 4
Identification of Vaccine Candidates against Lung Cancer-causing Viruses
Chapter 5
Identification of Prognostic Biomarkers for Non-small-cell Lung Cancer
Chapter 6
Computer-aided Prediction of Interleukin-2 Inducing Peptides
Chapter 7
Summary and Conclusion

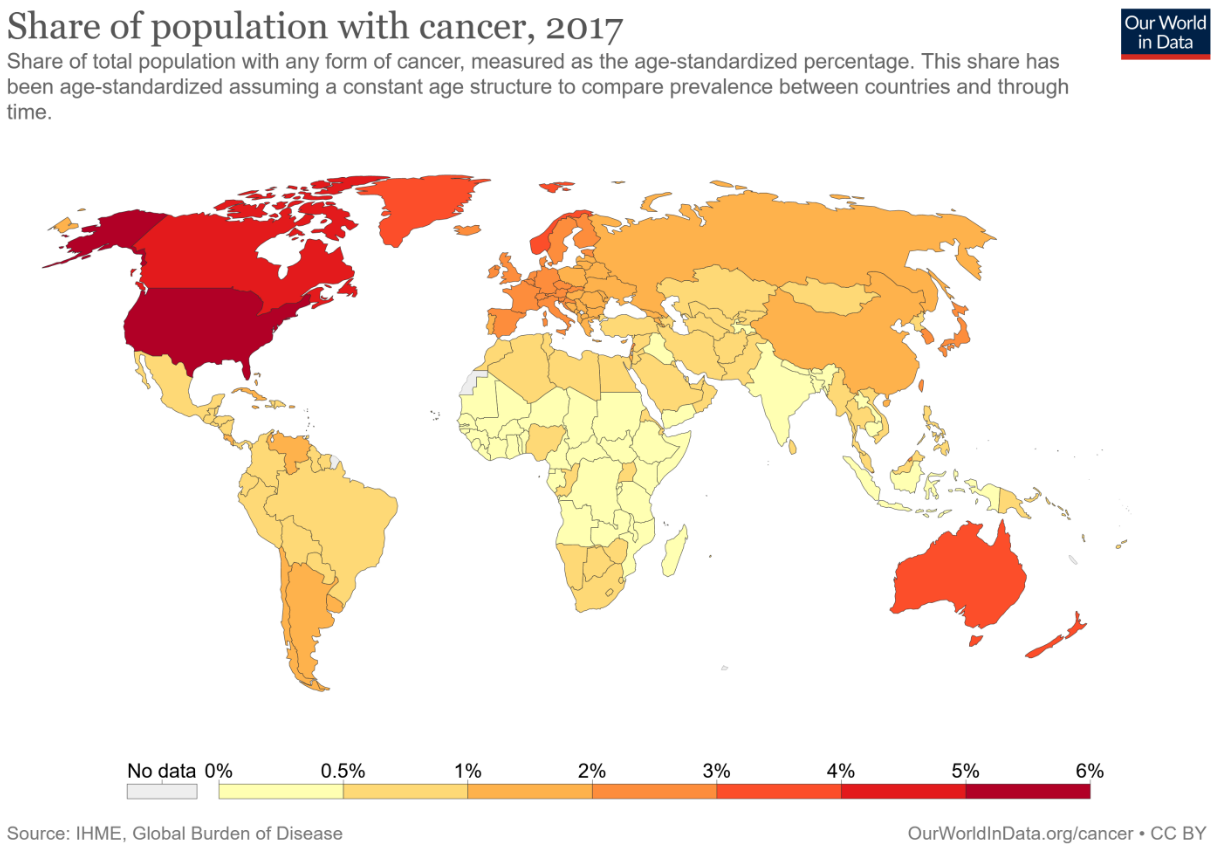
Figure 1.2: Overall organization and flow of the chapters in the thesis.

Chapter 1 of the thesis provides a brief introduction to the conventional approaches used to treat cancer. This chapter also describes the need for the present work to be done by specifying the lacuna of previous studies. It also elaborated on the overall objectives which are undertaken in the thesis to advance the cancer therapeutics. Chapter 2 of the thesis deals with a review of the literature process. Chapter 3 of the thesis focuses on the first objective of the thesis, which is on the development of a computational resource on oncolytic viruses. The chapter provides the in-depth details of oncolytic viruses used in cancer therapies, which are the newest among the immunotherapy-based cancer treatment strategies. These can also act as a delivery vehicle for targeted anti-cancer therapies, apart from boosting the immune response. The chapter also describes the utility of the developed resource for advancing and designing the oncolytic virus-based cancer immunotherapy. Chapter 4 lies in line with the second objective of the thesis, which is on the identification of subunit vaccine candidates against lung cancer-causing viruses. It describes a detailed methodology of systematical identification of epitope-based vaccine candidates from the proteome of oncogenic viruses. The identified subunit vaccine candidates have the property to stimulate both innate and adaptive arm of the immune system. Thus, the identified epitopes could be utilized for providing epitope-based preventive immunity against oncogenic lung cancer viruses. Moreover, the identified epitopes also have the immune adjuvant potential and can also be used in other combination therapies. Chapter 5 of the thesis corresponds to the third objective of the thesis, which deals with the identification of prognostic biomarkers for each subtype of lung cancer. Lung cancer is highly heterogeneous in nature. It is of utmost importance to identify the subtype specific biomarker for precision medicine. Thus this chapter describes a step by step approach of identifying the prognostic biomarkers using different computational methods. The identified biomarkers have the potential to be utilized in clinics for guiding and advancing the cancer therapeutics. Chapter 6 corresponds to the fourth objective of the thesis, which deals with prediction and designing on interleukin-2 inducing peptides. Finally, a crisp summary of the whole work is in chapter 7 of the thesis, which gives a holistic overview of the overall work.

2. Review of Literature

2.1 Global Burden of Cancer

Cancer is the second topmost killer after cardiovascular diseases around the world (Ferlay et al., 2019). It is estimated that cancer remains the single most non-communicable disease barrier across the globe in the 21st century that hinders the improvement in life expectancy. As the world population is in the decline phase of the age pyramid, the mortality and new cancer incidences are expected to increase throughout the world. It is estimated that occurrences and death due to cancer soon surpasses the coronary heart and cardiovascular diseases. According to the GLOBOCAN statistics containing data on 36 cancer types from 185 countries, there were approximately 18 million new cancer incidences and about 9.6 million deaths due to cancer, all around the world in 2018 alone. When both sexes combined, it is observed that lung cancer is the topmost killer among all cancer types, which solely responsible for around 19% of all cancer-related mortality, followed by colon and rectum cancer (10%), stomach (8.2%) and liver cancer (8.2%). Lung cancer remains at the top among the new incidences of cancer diagnosed. It alone constituting approximately 11.6% of new cases, followed by breast (11%), prostate (7.6%), and colorectal cancer (6%) (Bray et al., 2018). Lung cancer remains the topmost killer among men and ranks seconds in the case of females after breast cancers. From the GLOBOCAN report, it is also observed that patterns and types of cancer cases from the developed countries are shifting rapidly towards developing countries because of increased socio-economic development (Maule & Merletti, 2012). Yet for each country, there exists some common cancer type owing to their geographical distribution, and local risk factors such as infectious agents that also play the biggest role in the onset of incidences of new cancer cases. From all these points, it is very much clear that cancer is a global burden in individual countries, and there is an urgent need to design strategies to manage this disorder to improve the life expectancy of the patients. From the perspective of developing global economies such as India, it is necessary to improve the health care strategies that are well equipped in providing affordable health care to the patients. Figure 2.1 describes the global burden of cancer cases shared by each country in the world map presented by the GLOBOCAN report of 2017.


Share of population with cancer, 2017

Share of total population with any form of cancer, measured as the age-standardized percentage. This share has been age-standardized assuming a constant age structure to compare prevalence between countries and through time.

Source: IHME, Global Burden of Disease

OurWorldInData.org/cancer • CC BY

Figure 2.1: The global burden of cancer incidences around the world population.

2.2 Risk Factors

It is a well-accepted notion that knowing the cause is essential for the devising of curative strategy. By identifying the cause of cancer, preventive strategies can be devised against it. Several factors contribute to the development of the transformed behaviour of a cell. As evident from the literature studies, around 80% of lung cancer deaths are because of heavy smoking. It is also observed that cases of small cell lung carcinoma are rarely developed in person who never smoked. According to the World Health Organization report, infectious agents such as *Hepatitis B Virus*, *Human Papilloma Virus*, etc. alone are responsible for nearly 25% of newly diagnosed cases of cancer (de Martel, Georges, Bray, Ferlay, & Clifford, 2020) and are also involved in the molecular pathogenesis of lung cancer (Robinson et al., 2016). According to the GLOBOCAN report, nearly 2.2 million new cases of cancer are because of infectious agents. Primary infectious agents included in the report were – "*Helicobacter pylori*, *Hepatitis C virus*, *Human* *papilloma virus*, and *Hepatitis B virus*," which are responsible for 0.8, 0.16, 0.69, and 0.3 million new cases, respectively (de Martel et al., 2020). Experimental evidence suggests that these viruses insert their genome in the host cell to drive the tumourigenesis process. The infection of *Helicobacter pylori* is strongly associated with the development of stomach cancer (de Martel et al., 2020). The dietary pattern, obesity, and lack of physical activity also promote carcinogenesis in a complicated way. Heavy alcohol intake was found to be correlated with the high incidence rate of oral, breast, pharynx, larynx, and liver cancer (Meadows & Zhang, 2015). Food contaminants such as aflatoxins, which are produced by fungal species, increased the risk of liver cancer. Harmful radiations such as X-rays, UV are also responsible for the development of melanoma. Table 2.1 enlists the major carcinogen and their related cancer type. From all the above points, it is concluded that all the causative agents of cancer would impart mutations in the genome, which provides a selective advantage to cells, and thus, transformed behaviour happens.

Table 2.1: The primary risk factors linked with the onset of human cancers.

Cancer type	Risk factors
Respiratory system cancers (Lung, Trachea, Bronchi)	Air pollution; Pathogenic viral infections; Tobacco use
Esophagus cancer	Low fruit and vegetable diet; Tobacco use
Breast cancer	Advanced age; Alcohol use; Hormones treatment
Leukemia	Occupational exposure; Radiations; Tobacco use
Pancreatic cancer	Smoking; *Helicobacter pylori*; Age; Obesity and alcohol use
Stomach cancer	*Helicobacter pylori*; Low fruit and vegetable intake
Liver cancer	Aflatoxins; *Hepatitis B virus*; Heavy alcohol intake

2.3 Types of Lung Cancer

As stated above, lung cancer ranks first in terms of mortality rate around the world population. It is highly invasive and metastasizing in nature that alone responsible for more deaths than the four leading cancers (pancreatic, colon, rectal, and breast) combined (Lemjabbar-Alaoui, Hassan, Yang, & Buchanan, 2015). Histologically, it can be categorized into two major groups – Small cell lung carcinoma and Non-small-cell lung carcinoma. Small cell carcinoma contributes to approximately 15% of all diagnosed cases (B. E. Johnson et al., 2006). Its two main subtypes are oat cell cancer and combined small cell carcinoma. Non-small-cell lung carcinoma contributes to nearly 85% of all diagnosed cases and thus is considered as the major lung cancer type. Non-small-cell lung carcinoma further can be classified into three subtypes – squamous cell carcinoma, lung adenocarcinoma, and large cell carcinoma. Among the three subtypes, squamous cell carcinoma contributes to around 25-30% of all diagnosed cases, and lung adenocarcinoma consists of 40% of all new incidences (Zappa & Mousa, 2016). The remaining large cell carcinoma contributes to 5-10% of all newly diagnosed cases. The squamous cell carcinoma occurs in the epithelial cell of the bronchi. At the same time, lung adenocarcinoma arises in type II alveolar and epithelial cells, whereas large cell carcinoma is often observed in the central part of the human lungs (Zappa & Mousa, 2016). The detailed description of the organ localization of each subtype of lung cancer is in Figure 2.2.

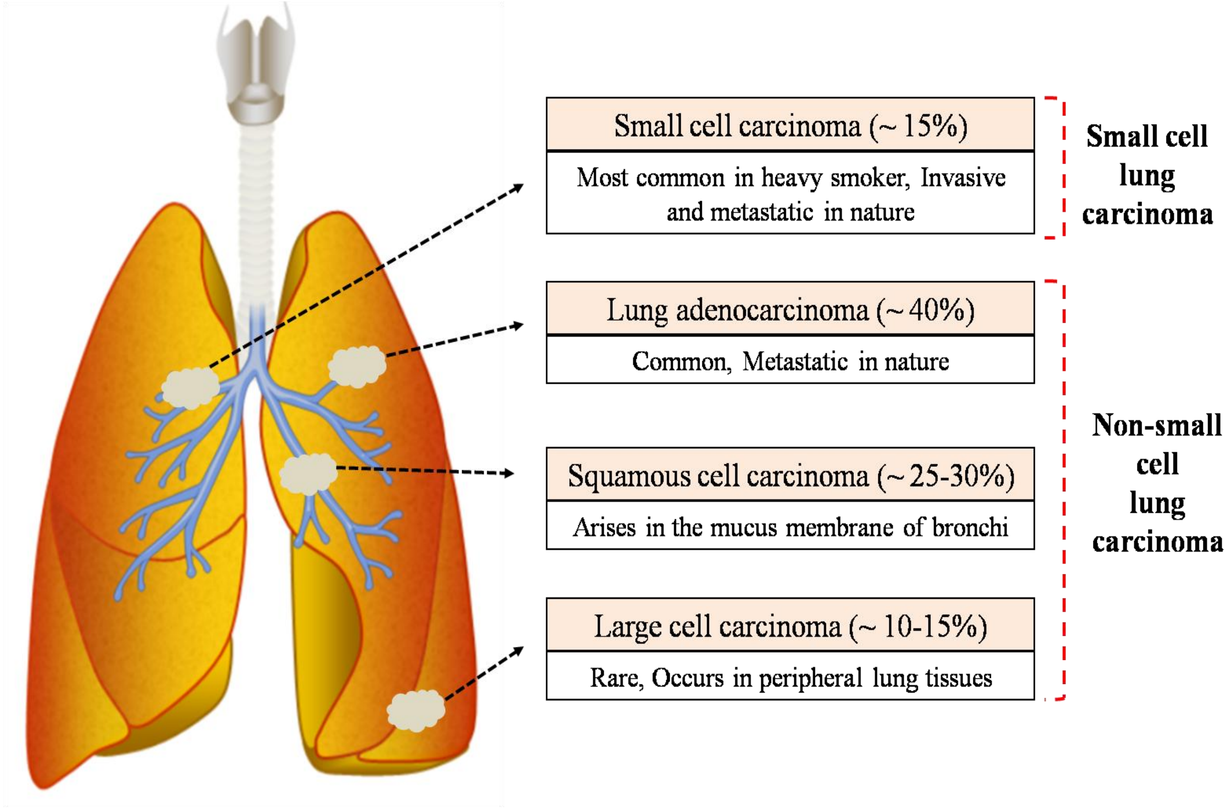

Small cell carcinoma (~ 15%)
Most common in heavy smoker, Invasive and metastatic in nature
Small cell lung carcinoma
Lung adenocarcinoma (~ 40%)
Common, Metastatic in nature
Squamous cell carcinoma (~ 25-30%)
Arises in the mucus membrane of bronchi
Non-small cell lung carcinoma
Large cell carcinoma (~ 10-15%)
Rare, Occurs in peripheral lung tissues

Figure 2.2: Schematic representation of histopathological subtypes of lung cancer.

2.4 Mutational Landscape of Lung Cancer Types

The mutational landscape varies greatly across the subtypes of lung cancer. Small cell lung cancer carries certain genetic mutations that are found in nearly all patients. The typical example includes *p53, Rb,* and *cMYC* gene mutations that are predominately found in almost all patients. (Karachaliou et al., 2016). The genomic profiling of the non-small-cell lung carcinoma also leads to the identification of several genomic alterations specific to each subtype. The specific example includes gene mutation of *ALK*, *MET*, *NTRK2*, *PRKCB*, *RET*, *ROS1* in case of lung adenocarcinoma, and *FGFR1*, *FGFR2*, *FGFR3*, *FGR*, *PKN1*, *PRKCA,* and *PRKCB* gene mutation in case of lung squamous cell carcinoma (Lemjabbar-Alaoui et al., 2015). Genomic studies also identify several genomic alterations that can be used in targeted therapies. Notable examples include the activating mutations of proto-oncogenes such as *BRAF*, *KRAS*, *EGFR*, *HER2*, *PI3K*, etc. Table 2.2 provides the details of targetable genes that can be utilized for developing targeted lung cancer therapies.

Table 2.2: Genomic targets and respective pathways targeted in different lung cancers.

Gene name	Lung cancer type	Pathway
PTEN	Adenocarcinoma; squamous cell carcinoma	PI3K signalling pathway
ARIDIA	Adenocarcinoma	Epigenetic pathway
CEBBP	Small cell carcinoma	Epigenetic pathway
ALK	Adenocarcinoma	RTK signalling pathway
BRAF	Adenocarcinoma; Squamous cell carcinoma	RAF signalling pathway
EGFR	Adenocarcinoma; Squamous cell carcinoma	RTK signalling pathway
MLL	Small cell lung cancer	Epigenetic pathway

2.5 Role of Viruses in Cancer

As explained above, many DNA and RNA genome-based virus species are responsible for millions of cases of cancer around the globe. Oncogenic virus species are responsible for the onset or progression of different cancer types - bladder carcinoma, vulvar cancer, cervical, head and neck carcinoma, hepatocellular carcinoma, Kaposi sarcoma, lung cancer, and several others. Figure 2.3 explains the mechanism of the carcinogenesis adopted by the retroviral infection.

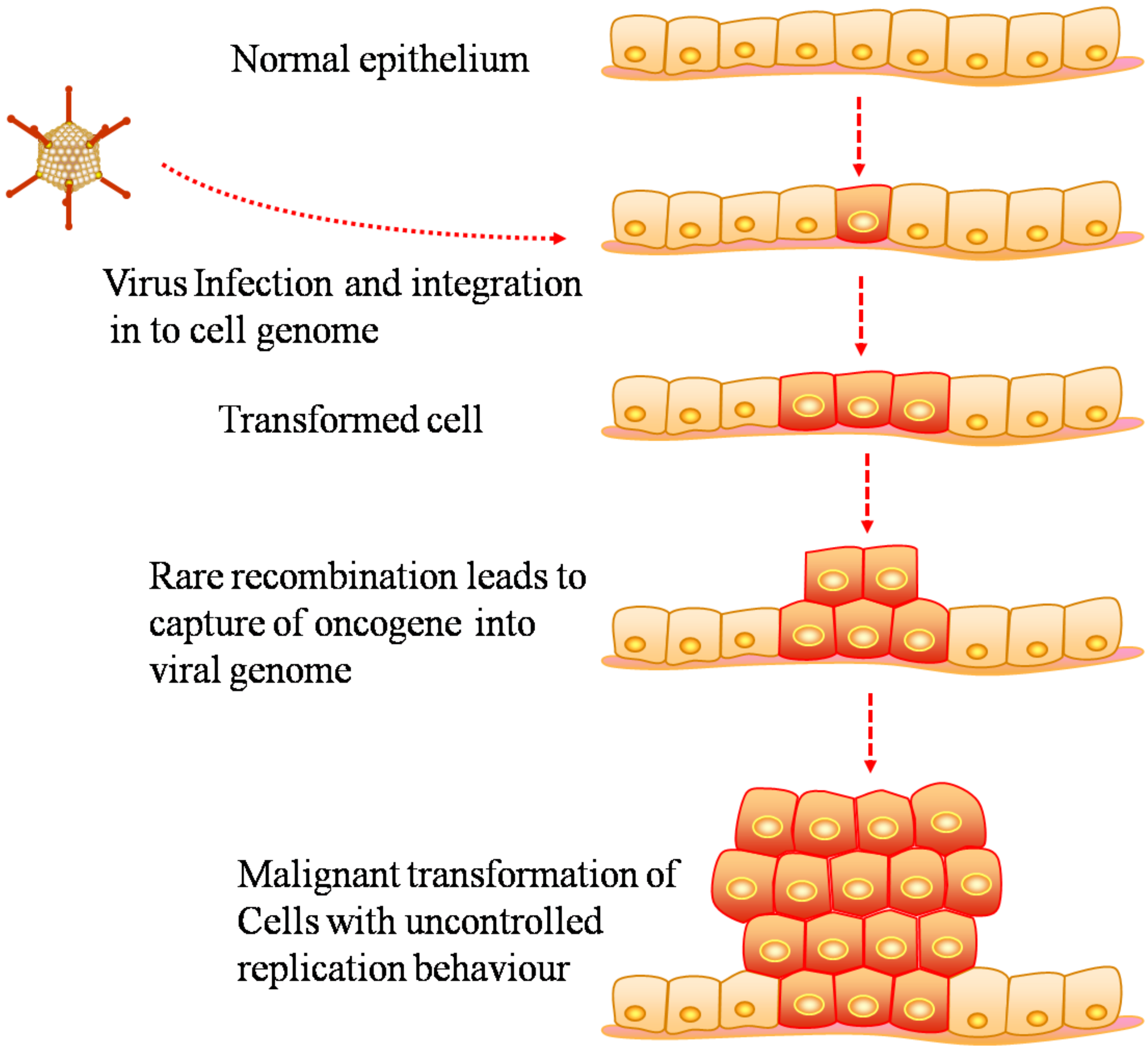

Normal epithelium
Virus Infection and integration in to cell genome
Transformed cell
Rare recombination leads to capture of oncogene into viral genome
Malignant transformation of Cells with uncontrolled replication behaviour

Figure 2.3: Mechanism of retroviral infection induced carcinogenesis.

There is a complex mode of the mechanism employed by each virus for carcinogenesis. It involves the infection, cooperative interaction among the immune cells with viruses so that viruses become tolerable in the human body, and many more. In general, DNA genome-based oncogenic viruses inhibit the expression of tumour suppressor genes such as *p53*, *pRB* to bring about the transformation in the cells, which ultimately leads to cancer development. RNA viruses, after infection into the cells, transcribe their genomic content with the help of reverse transcriptase enzyme. The newly developed double-strand DNA of the virus genome is then inserted into the host cell's genome with the help of enzyme integrase. There are abundant sites present in the human genome for viral DNA insertion and integration, and these sites are known as fragile sites. The integrated copy of the virus genome is known as provirus – a chimeric host and virus DNA molecule. During the subsequent round of replication, the virus may acquire oncogene. The captured oncogene by the virus may undergo some mutations and can escape from the normal cellular regulatory machinery and continue to start expressing itself, and thus carcinogenesis occurs. Sometimes, the provirus also infects healthy cells that do not have the control mechanism to regulate the expression of the proto-oncogene. Thus proto-oncogene is expressed in high amount without any control and leads towards the transformed behaviour of the cell. In past studies, multiple drugs were used to control the infection that also targets the cancer hallmarks induced by virus infection. Table 2.3 provides a summary of these drugs, along with their cancer hallmark target.

Table 2.3: A list of drugs used to target cancer hallmark induced by virus infection.

Virus type	**Cancer type**	**Drug name**	**Cancer hallmark**
Hepatitis B virus	Liver cancer	Sorafenib	Angiogenesis
Epstein-Barr virus	Non-hodgkin lymphoma	Butvrate	Cell proliferation
Human papilloma virus	Oropharyngeal carcinoma	Bevacizumab	Proliferation
Human T-cell lymphotropic virus	Acute T-cell leukemia	Zidovudine	Proliferation

2.6 Oncolytic Viruses

Researchers and clinicians are continuously working to provide better odds of survival to cancer patients. From the experimental studies, there arises a striking observation that viruses have the natural tendency to infect and proliferate inside the cells. Their natural tendency of infection can be used as a weapon against fast proliferating cancer cells. Based on several observations, scientists have investigated the cancer-killing potential of viruses. They have termed these viruses as oncolytic viruses. In this regard, oncolytic viruses emerge as a relatively new class of anti-cancer molecules for providing the therapeutics in cancer patients. In literature, it was observed that some oncolytic viruses have a natural greater tendency to infect the fast proliferating cells such as cancer cells (Howells, Marelli, Lemoine, & Wang, 2017). At the same time, some other oncolytic viruses utilize the cancer-specific phenomenon for infecting the cells. These include I) receptor targeting mechanism where receptors necessary for virus entry are selectively expressed on tumour cell, e.g., CD46 receptors are utilized by *Measles virus*; II) abnormal signalling pathways, e.g., *AKT* signalling exploited by *Myxoma virus*; III) hypoxic tumour environment used by *Vesicular stomatitis virus*; IV) defective type I interferon signalling pathway utilized by *Vaccinia virus* (Howells et al., 2017). Despite the natural tendency of many viruses towards cancer cells, some oncolytic virus species were also genetically engineered to target the cancer cells to achieve a better therapeutic outcome. The genetic modification of the virus genome even allows carrying the foreign gene of a high therapeutic index. An example includes – integration of interleukin-2/24 genes in oncolytic adenovirus has a better therapeutic outcome as compared to wild-type oncolytic virus (Ashshi, El-Shemi, Dmitriev, Kashentseva, & Curiel, 2016). Adenovirus expressing the interleukin-18 gene has a better therapeutic result as compared to wild-type when tested in in vitro condition (J. N. Zheng et al., 2010). The other advantage of using oncolytic viruses for cancer therapeutics is that they also boost the antitumour immune response and thus can overcome the limitation posed by the immunosuppressive tumour microenvironment. Since oncolytic viruses can accept foreign immune genes, they can overcome the barriers of immune suppression poised by cancer cells during treatment. Upon selective infection in cancer cells, oncolytic viruses replicate inside the cell. The selective replication inside the cells leads to the cell lysis and release of tumour-specific antigens, which later on are engulfed by the nearby macrophages and thus enhances the antitumour immune response

(Lathwal, Kumar, & Raghava, 2020). Owing to their greater benefit to risk ratio, the use of oncolytic viruses has gained much more momentum in cancer therapeutics. Figure 2.4 explains the mechanism of action of oncolytic viruses as well as their advantages and challenges in cancer therapy.

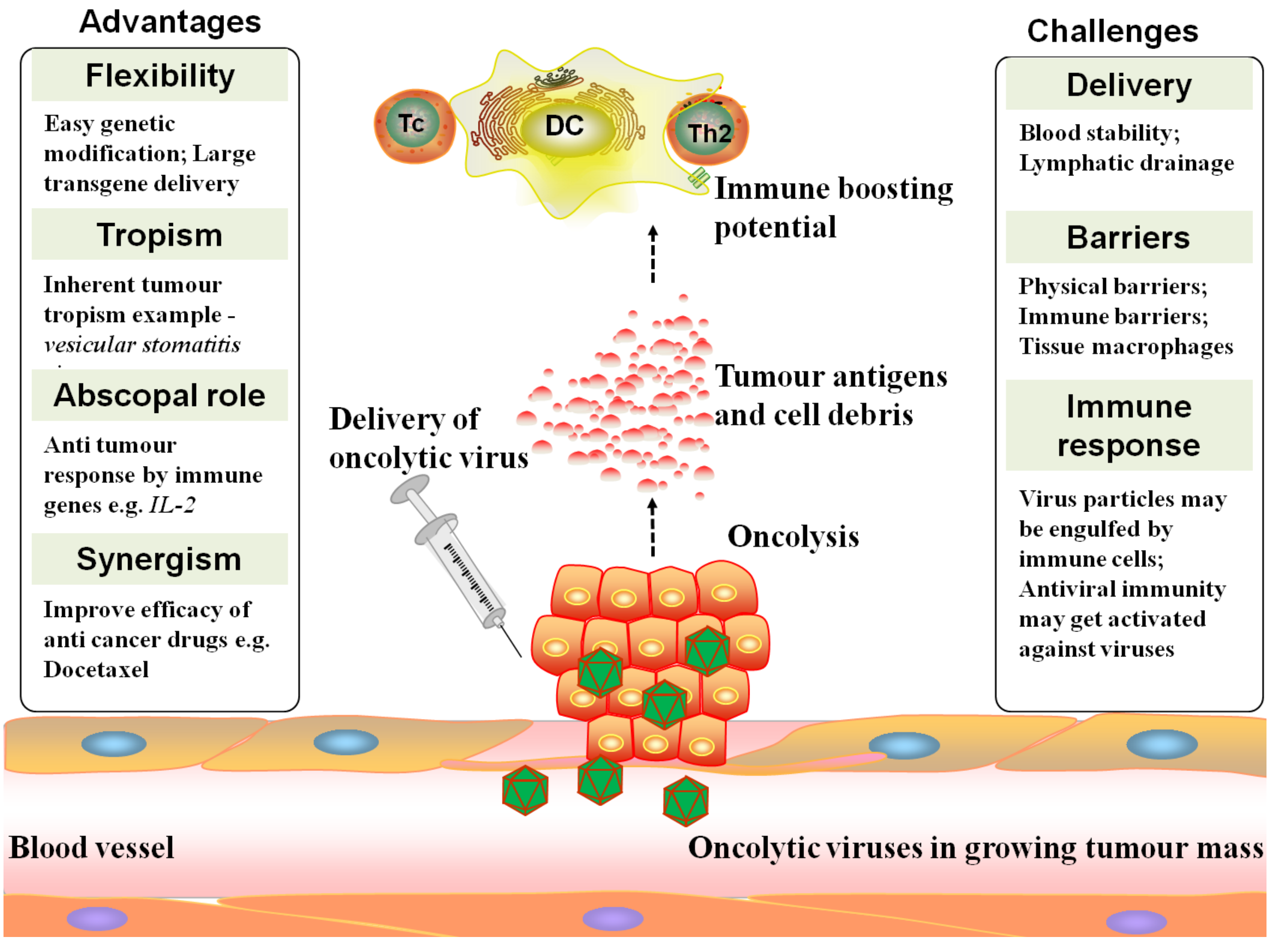

Advantages
Flexibility
Easy genetic modification; Large transgene delivery
Tropism
Inherent tumour tropism example - *vesicular stomatitis*
Abscopal role
Anti tumour response by immune genes e.g. *IL-2*
Synergism
Improve efficacy of anti cancer drugs e.g. Docetaxel
Tc
DC
Th2
Immune boosting potential
Tumour antigens and cell debris
Delivery of oncolytic virus
Oncolysis
Challenges
Delivery
Blood stability; Lymphatic drainage
Barriers
Physical barriers; Immune barriers; Tissue macrophages
Immune response
Virus particles may be engulfed by immune cells; Antiviral immunity may get activated against viruses
Blood vessel
Oncolytic viruses in growing tumour mass

Figure 2.4 Mechanism of action of oncolytic virotherapy, its advantages, and challenges.

2.7 Conventional Therapeutic Interventions

Lung cancer is highly heterogeneous, and often patients are diagnosed when they are already at an advanced stage of cancer. Certain standard conventional treatment options exist for lung cancer in the literature. These include surgery, radiotherapy, and combinatorial treatment approaches. The conventional therapeutic strategy depends upon the severity and the stage at which disease is diagnosed. The surgical resection of lung cancer can be further divided into several types based on the severity of the diseases. The lobectomy includes where one lobe of the infected lung is removed following the surgical procedure. The pneumonectomy is a surgical procedure where one of the lungs is removed from the body of the patient. The wedge resection is another type of surgical procedure where some portion of both the lobes of a lung is removed. Segmentectomy is a class of surgical operation for lung cancer where the area of the infected lung along with veins is removed. The surgical removal of the lung tissue is possible only at stage I and II of lung cancer. However, since it was observed that a greater number of patients with lung cancer are recognized at stage III and IV, which is considered as a late stage of the disease, so surgical removal is not possible at that time (Zappa & Mousa, 2016). Nearly 70% of lung cancer patients are diagnosed at stage III. The standard conventional therapeutic options available for stage III lung cancer is the combination of chemotherapeutics drugs with resectable surgery, if possible. The combination of drugs with radiotherapy is recommended for patients of stage III, where surgery is not possible. Approximately 40% of patients, when diagnosed with lung cancer, were already at stage IV. Stage IV of lung cancer is considered as advanced stage, and the survival rate at this stage remains to only 4.5 months when left untreated. At this advanced stage combinatorial approach of cytotoxic chemotherapeutic drugs is the preferred choice. The first-line chemotherapy drugs include cisplatin, docetaxel, and paclitaxel. However, with stage IV disease severity, the second line of drugs are also given to the patients. These include docetaxel, erlotinib, gefitinib, etc. (Lemjabbar-Alaoui et al., 2015).

2.8 Genomic Therapies for Lung Cancer

The genomic studies have identified several gene mutations that are specific to each lung cancer subtype. Tremendous work has been done in the literature to translate this acquired genomic information related to genetic alteration for the improvement of patient's healthcare. The information from these genes serves the basis of screening, early diagnosis, prognosis, and the development of targeted therapies. Since it is a well-observed phenomenon in the literature that patient survival time gets increased if cancer diagnosis and prognosis happened at an early stage. Keeping this fact is mind; several prognostic and predictive markers have been identified for lung cancer. Typical examples include the *EGFR* and *ALK* gene mutations that have shown good prognostic power and found their way in clinical settings. The gene fusion product of *ALK*-

EML4 in non-small-cell lung carcinoma is a bad prognostic indicator of poor response towards *EGFR* and *TKI* therapy (Lemjabbar-Alaoui et al., 2015). The *EGFR* and *ALK* gene mutation targeting therapy by the erlotinib drug is approved for lung cancer treatment. The overexpressed *RRM1* gene of non-small-cell lung cancer is the main target of the drug gemcitabine (Chan & Hughes, 2015). As mentioned in the mutational landscape section above, there is an enormous level of heterogeneity present at the genomic level among the subtypes of non-small-cell lung cancer. These differences suggest that the subtypes will be going to have separate outcomes against similar treatment regimes. These differences also emphasize the requirement of a deep understanding of the primary mechanism to form the basis of the subtype-specific therapeutic strategies that can improve the overall survival of the patients. Distinct histology's and nature of origins demand a therapeutic regimen that accounts for all these factors to enhance the specificity and curative potential of a given therapy with minimal side-effects. There are nearly 100 clinical trials that are still ongoing, which consider the gene mutation as a target for providing healthcare to lung cancer patients. A brief list of targeted therapies related to lung cancer is in Table 2.4.

Table 2.4: A list of targeted therapeutic drugs currently used in lung cancer treatment.

Drug	**Drug target**	**Type of drug**
Axitinib	*PDFGR*	TKI inhibitor
Bevacizumab	*VEGF*	Monoclonal antibody
CI-994	HDAC	Small molecule
Docetaxel	Tubulin	Small molecule
Erlotinib	*EGFR*	TKI inhibitor
Etoposide	Topoisomerase II	Small molecule
Gefitinib	*EGFR*	TKI inhibitor
Iniparib	*PARP-1*	Small molecule
Ipilimumab	*CTLA-4*	Monoclonal antibody
Paclitaxel	Tubulin	Small molecule
Pemetrexed	Nucleotide inhibitor	Small molecule
Nintedanib	*PDGFR*; *VEGFR*	TKI inhibitor
Vinblastine	Tubulin	Small molecule

The use of targeted therapies has significantly improved the patient's survival, but several problems also exist with these therapies. One such issue is the development of drug resistance among the cancer cells after long term exposure to targeted therapies. Another major challenge that is faced by the targeted approach is the relapse among the patients. Another challenge that is faced by the targeted approach is that the target gene sometimes gets mutated in cancer cells, and hence drugs can no longer be effective. Despite this, several other side effects are also observed with targeted approaches, which include – skin problems, hair depigmentations, delayed wound healing, high blood pressure, and gastrointestinal perforations. All these limitations suggest the requirement of some alternative means of therapeutics, which not only overcome the problem as mentioned earlier but also provide long-lasting immunity effects.

2.9 Immunotherapy for Lung Cancer

The immunotherapy becomes an appealing and attractive strategy to fight against cancer. The advantage of using immunotherapy is that it helps to educate the immune system to fight against cancer as well as it provides long-lasting antitumour immunity also (Liu & Guo, 2018). Moreover, the immunotherapeutic agents does not show any kind of toxicity, as seen in the case of conventional chemo and radiotherapies. Cancer immunotherapy has shown tremendous results in clinical trials, thus selected as the breakthrough in the year 2013 by science magazine (Sukari, Nagasaka, Al-Hadidi, & Lum, 2016). Recently, several immunotherapeutic agents for the treatment of wide varieties of cancer have been passed by the Food and Drug Administration of USA. Notable examples include the checkpoint inhibitors, CAR T-cell therapy, and the use of interleukins. The Nobel Prize of the year 2018 in physiology and medicine was awarded to scientists for their discovery of cancer therapy by negative inhibition of immune regulation. Several studies have identified the role of CAR T-cell therapy in improving the overall life expectancy of the patients, even in the refractory lung cancer type (Brahmer et al., 2010; Thomas & Hassan, 2012). Several clinical trials are studying the effect of CTLA-4 in combination with various drugs for lung cancer treatment (Thomas & Hassan, 2012). The immunotherapy regimen using the interleukin-2 has also been passed for treating the advanced stage of melanoma. However, literature evidence supports the notion that the mutant versions of interleukin-2 have a high therapeutic index as compared to the wild-type. The favourable biological impact of immunotherapy motivated the researchers around the world to identify several units that can be used in cancer clinics. One such group is the oncolytic viruses. Oncolytic viruses are genetically engineered to express the foreign gene and are also used as a targeted delivery system to impart the desired immune effect in the cold tumour microenvironment. Recently an oncolytic virus-based drug has been approved for the treatment of melanoma. This leads to the attention of researchers all around the world to design and use oncolytic viruses as a delivery vehicle for providing better immunotherapeutic results in cancer patients. Numerous clinical trials are investigating the use of oncolytic viruses for the treatment of subtype-specific lung cancer. Table 2.5 enlist the currently ongoing clinical trials for treating different types of lung cancer using oncolytic viruses.

Table 2.5: Ongoing clinical trials using oncolytic viruses for the treatment of lung cancer

Oncolytic virus	Cancer type	Intervention	Clinical trial
Adenovirus	Non-small cell lung cancer	Valacyclovir; Pembrolizumab	NCT03004183
Adenovirus–MAGEA3	Non-small cell lung cancer	Pembrolizumab	NCT02879760
Reovirus	Non-small cell lung cancer	Carboplatin; Paclitaxel	NCT00625456
Vaccinia virus	Squamous cell carcinoma	GM-CSF expressing	NCT03029871
Adenovirus	Mesothelioma	-	NCT01503177

In spite of oncolytic virus-based therapy, the vaccination approach is an exciting class of cancer immunotherapy. The vaccination strategy can save millions of lives. As for the advanced stage of lung cancer, the median survival rate is around 3.5%, so developing vaccination to improve the life expectancy of pathogen-related cancer origin can have dramatic effects. The non-small-cell lung carcinoma is a non-immunogenic tumour type, thus developing vaccine candidates against infectious agents responsible for such cancers can have favourable outcomes. Table 2.6 summarizes some of the ongoing vaccine development strategies for treating non-small-cell lung cancer, which utilizes tumour-specific antigen.

Table 2.6: Vaccines in developmental stages against non-small-cell lung cancer targeting tumour-specific antigens.

Vaccine name	Vaccine type
ClimaVax	Allogenic
CRS-207	EGF vaccine
MAGE-A3	Antigen vaccine
TG4010	Antigen vaccine
Emepeimut-S	Synthetic peptide
PRAME	Adjuvant vaccine

2.10 Conclusion

Despite the improvements in lung cancer treatment by using targeted therapies, the 5-year survival rate among patients diagnosed at an advanced stage remains low. Several factors, such as age, sex, body's immune system, pathological stage, infectious agents, may affect the survival of lung cancer patients. Thus, it is necessary to identify and design strategies that can help in improving the patient's health. The use of immunotherapy is considered as an optimal choice for the same. The literature revealed that oncolytic viruses and interleukin-2 based therapies are already approved for treating several types of cancer. There are abundant clinical and pre-clinical studies that utilize oncolytic viruses as an immunotherapeutic agent. But, there is a lack of a unified platform for scientists, where all the available information regarding the therapeutic application of oncolytic viruses is stored. Moreover, literature studies also report that mutants of interleukin-2 have a high therapeutic index as compared to the wild-type protein. Despite the approval of interleukin-2 based immunotherapy in 1992, there is still a lack of computational tools that can help scientists and researchers in designing and identifying interleukin-2 inducing potential of peptides for further advancing the immunotherapy. Immunotherapy employing vaccination is also a fascinating research area that can help in improving patient's lives. Several tumour-antigens associated vaccine candidates are in clinical trials. Still, no study in the literature identified the subunit epitope vaccine candidates from the proteome of the oncogenic viruses against lung cancer. As explained in the above sections, the success of targeted therapy in treating lung cancer depends largely on the timely diagnosis/prognosis of the patients as well as on the level of heterogeneity present at the subtype level. Thus, it is of utmost importance to identify the lung cancer subtype-specific biomarkers to further advance the therapeutic process. In the literature, no study identified the non-small-cell lung cancer subtype-specific biomarkers, which can be used in the prognostification of patients. Thus, the present thesis work focuses on developing computational strategies that can aid in advancing the lung cancer therapeutics by identifying the subtype-specific biomarkers, predicting and designing interleukin-2 inducing peptides, identifying vaccine candidates against lung cancer-causing oncogenic viruses and building a web-resource on oncolytic viruses used in cancer theraputics.

3. A Knowledge Base of Oncolytic Viruses in Cancer

3.1 Background

Over the past decade, immunotherapy has become the standard choice of treatment for cancer patients. Literature data show the continuous surge of increase in clinical trials related to immunotherapy for cancer treatment. According to the US National Library of Medicine, nearly 1473 ongoing clinical trials consider immunotherapy at the forefront for the management of cancer. Out of these 1473, nearly 400 studies have completed the clinical trials up to stage II (https://clinicaltrials.gov/). Further, data analysis reveals that almost 308 clinical trials are alone in literature for lung cancer neoplasm. However, data from published clinical trials show that patients with an immunologically cold tumour microenvironment do not get benefits from conventional immunotherapeutic approaches (Russell, Peng, Russell, & Diaz, 2019). This highlights the need for some alternative immunotherapy approach, which can also be rationally designed and combined with conventional treatment methods.

Oncolytic viruses (OV) has recently emerged and taken the forefront of modern cancer immunotherapeutic strategies. In the literature, OV based cancer therapy is also considered a major breakthrough after the checkpoint inhibitors (Fukuhara, Ino, & Todo, 2016). OV can be defined as natural or genetically modified viruses by recombinant DNA technologies that have the capacity to selectively infect and replicate within the cancer tissues without harming the healthy tissues (Chiocca & Rabkin, 2014). The use of viruses to control cancer progression is a decade back observation. Early studies suggest the regression of tumour with the natural infection of viruses (Kelly & Russell, 2007). Data from recent studies showed that several other virus species such as *Herpes simplex virus, Measles virus* can also have the therapeutic potential for cancer immunotherapy. Recently, the “Food and Drug Administration (FDA)” has approved a drug Talimogene Laherperepvec (T-VEC) for the management of melanoma patients (Bommareddy, Patel, Hossain, & Kaufman, 2017). T-VEC is a type I *Herpes simplex virus* modified by recombinant DNA technologies for the expression of granulocyte-macrophage stimulating factors and enhanced antigen loading capacity for MHC class-I (Conry, Westbrook, McKee, & Norwood, 2018). The data from the clinical trial of T-VEC showed that it not only suppresses the tumour growth but also improves the overall patient health in terms of survival. T-VEC is currently being investigated for combination with adjuvant and chemotherapies (D. B. Johnson, Puzanov, & Kelley, 2015). Despite this, several other OV species such as JX-594 for liver cancer, CG0070 for bladder cancer, and a wild-type *Reolysin* virus for head and neck squamous cell carcinoma are closing clinical approval in the North American continent (Fukuhara et al., 2016).

The clinical safety of OV therapy has now been well established, with hundreds and thousands of cancer patients being treated with the use of OV. The use of OV in cancer therapeutics has several advantages over other conventional therapies. Firstly, that there is no chance of developing resistance against OV based treatment regimen as OV affects cancer cells in multiple ways. Secondly, the minimal toxicity as OV selectively infects and replicates within cancer cells. Thirdly, it was observed that virus dose in cancer tissue increases over time due to in situ replication of the virus. Fourthly, the safety features can also build up in viruses such as drug safety and immune sensitivity. Fifth and the most important one is that OVs, such as *Adenovirus*, *Herpes simplex virus*, offers large genome size that can be modified to achieve higher benefit ratio (Chiocca & Rabkin, 2014). The general mechanism of action of OV is shown in Figure 3.1.

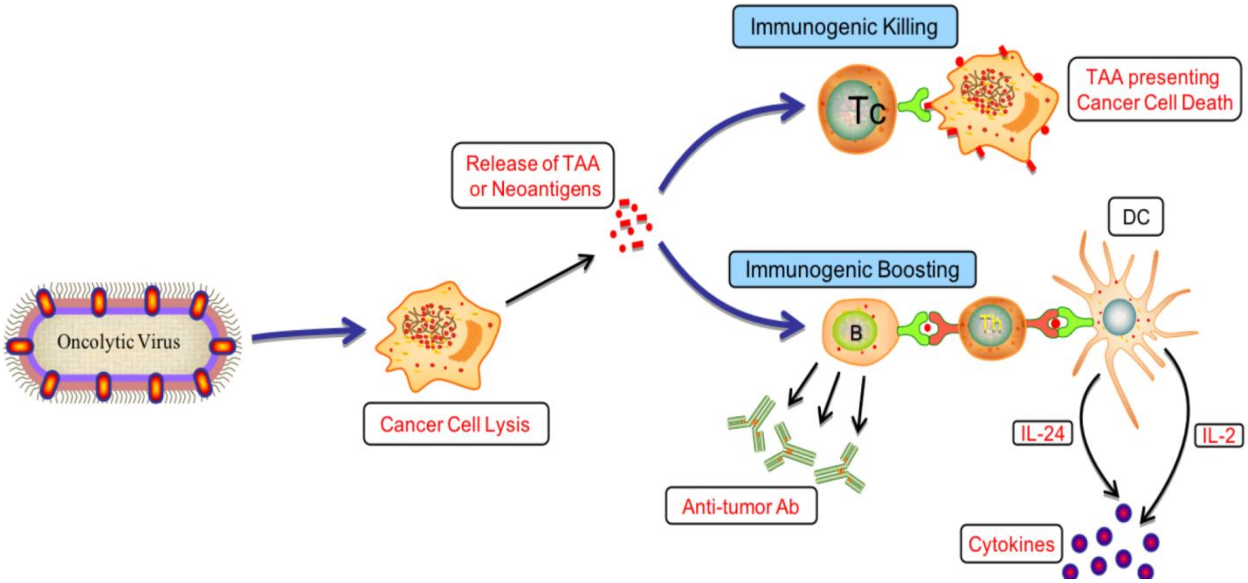


Figure 3.1: A schematic representation of mechanism of action of oncolytic viruses [*Source - Lathwal et.al. Virology -2020*].

The selective replication of OV in cancer tissues leads to the bursting of it, which releases Tumour-Associated Antigens (TAA). These TAAs can act as a danger signal and subsequently get captured by macrophages and dendritic cells, which serve the basis for the generation of the prolonged antitumour immune response (Bai, Hui, Du, & Su, 2019). Considering the enhanced safety, improved antigenic response, minimal side effects, and higher benefit ratio helps in promoting the use of OV in cancer management and therapy. Data from several literature studies, pre-clinical and clinical studies show impressive results that further advocate the use of OV as a potential anticancer treatment regimen. However, all the results from the above-mentioned studies are widely distributed in literature. Their scattered distribution poised a problem to researchers and clinicians in identifying and design new OV based cancer treatment regimens. To fill this lacuna in the OV based cancer treatment strategy, we have developed a resource that aims to provide all available information on OV in terms of their experimental details. All the available data curated from the literature studies are presented to the scientific community in the form of a web-resource, i.e., "OvirusTdb." The web-resource is freely available to the scientific community at (https://webs.iiitd.edu.in/raghava/ovirustdb/).

3.2 Materials and Methods

3.2.1 Data Collection

PubMed was searched using a string of keywords, namely "oncolytic viruses,” "oncolytic virotherapy." The total number of abstracts that come after this search criterion was 4514. Moreover, data regarding OV's clinical potential was also present in the patents. The same combination of strings was also explored in the United States Patent and Trademark Office (USPTO), which resulted in 1100 patents. We have only downloaded the PubMed abstract and patents, which were available in the English language. The articles and patents which are not available in the English language were rejected from the study.

3.2.2 Data Reviewing and Inclusion

Firstly, each of the downloaded abstracts was screened for the research articles. We have excluded all review articles, perspectives, and letters to editors from our study. The remaining PubMed abstracts were examined manually for the presence of relevant experimental details regarding OV. This stringent initial filtering criteria lead towards the remaining 1604 PubMed abstracts. Patents that did not hold any relevant experimental details were also excluded from the study. All this leads to the remaining of only 644 patents. OVs hold great promise in cancer immunotherapy and are also used in combination with chemotherapeutics. Literature data showed that OVs could be engineered to express immune-stimulatory genes such as interleukins, colony-stimulating factors. Thus, the remained research articles and patents were manually examined for the presence of desired experimental details. Thus, the final data that constitutes the OvirusTdb web-resource has come from 166 PubMed articles and 27 patents.

3.2.3 Data Curation

All the relevant experimental details from the PubMed article and patents were catalogued in seven major sections. The overall architecture of the OvirusTdb is in Figure 3.2.

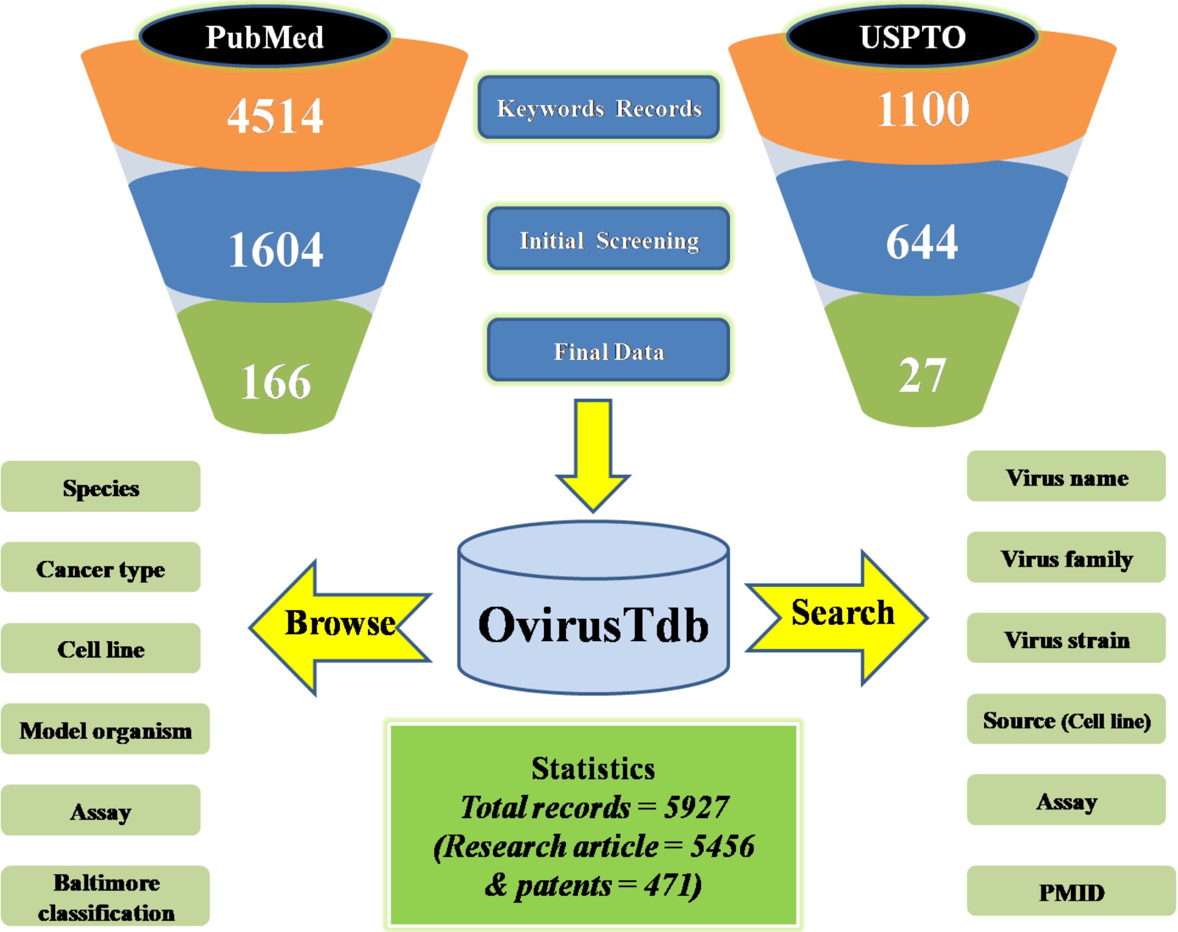


Figure 3.2: Visual representation of database architecture and its facilities.

Broadly, OvirusTdb seven major sections are - i) virus details – includes virus name (Adenovirus), strain (AR339), genome type (DNA or RNA), family (Adenoviridae); ii) virus genomic modifications expression and deletion of any gene (*IL-2* insertion and *ICP34*, *E1A* promoter); iii) virus used as monotherapy and in combination with other chemotherapeutics (gemcitabine); iv) in vitro assay details – includes the source of the cell line (ATCC), cell line use (PANC-1), concentration tested, and toxicity; v) in vivo assay details – includes model organisms used (BALB/c), route of administration (intratumoural, intravenous), toxicity, virus concentration; vi) immunogenic effect which includes major cell death pathway activation and immune-boosting potential and vii) PubMed/patent details.

3.3 Results

3.3.1 Data Statistics

All the manually curated experimental details regarding OV were catalogued in 25 fields. OvirusTdb web-resource holds experimental details of OV manually curated from literature studies. Web-resource holds 5927 records against 25 fields, which were manually curated from the 166 and 27 research articles and patents, respectively. Out of 5927, 5456 records extracted from research articles, and 471 from patents. Moreover, OvirusTdb holds promising experimental details on 24 OV species, where the majority of virus species were found to be genetically modified for the treatment of cancer. In contrast, few species, such as the *Vaccinia virus*, were used in their wild-type form for the treatment of cancer cells. The database holds extensive information on 300 genetically modified OV strains. The genetic modifications have been introduced into the genome of OVs to enhance their efficacy towards cancer cells. Moreover, OvirusTdb provides in-depth experimental details on 124 cancer types and 400 cancer cell lines on which the oncolytic effect of viruses have been tested in the clinical studies. Also, our web-resource provides a piece of extensive information on model organisms that were used to measure the efficacy of the OV treatment regimens, either alone or in combination as immune adjuvant and chemotherapeutic agents. The vital statistics of the data stored in the web-resource are explained in Table 3.1.

3.3.2 Data Analysis

The major advantage of using OV in cancer immunotherapy is their mode of action as an immune stimulant. The immune stimulant nature of OV is also evident from the data analysis of total records present in web-resource. The data analysis reveals that around 1506 (around 25%) of total records provide information on the immune-stimulation potential of OVs.

Table 3.1: The data statistics of the OvirusTdb web-resource.

Property	Records
No. of entries	5927
Virus species	24
DNA genome virus species	09
RNA genome virus species	15
Wild-type strains	15
Genetically modified strains	300
Cancer types	124
Cancer cell line	427
Biological assays	30
Model organisms	22

Further data analysis reveals that around 3243 (~ 55%) of total records show apoptosis induction as a mechanism of action of oncolysis by OV. Data analysis further reveals that BALB/c mice were the preferred choice of a model organism for testing the oncolytic capability of viruses. Similarly, for measuring the in vitro oncolytic efficacy of OVs, Methyl Tetrazolium (MTT) assay was the preferred choice. Literature studies signify that OVs, in combination with chemotherapeutics, not only enhanced their efficacy but also improves the overall survival of patients along with boosting the immune response (Phan, Watson, Alain, & Diallo, 2018). In a combinatorial setting, the drug given in combination provides enough time to the virus for replication within the tumour cells. This way, after oncolysis boosting the immune system may happen as measured in clinical studies (M. Zheng, Huang, Tong, & Yang, 2019). Data analysis from the OvirusTdb also reveals that the majority of OV species are used in combination with anticancer drugs. The combinatorial settings were found to improve the efficacy of drugs, as evident from the increased survival and decrease tumour load in model organisms. The results of the data analysis of the OvirusTdb web-resource are shown in Table 3.2.

Figure 3.2: Oncolytic virus and combination settings used in literature studies.

Virus species	Drug in combination
Adenovirus	Cisplatin; 5-FU; Paclitaxel; Doxorubicin; Luteolin; Gemcitabine; Gancliovar; Etoposide; Cyclosporin; Campothecin; Decarbazine
Herpes simplex virus	5-FU; Rapamycin; Mitomycin; ATN224; SN-38; Flutamide; CI994
Vaccinia virus	5-FC; Paclitaxel; Gemcitabine; Cisplatin; 5-FU
Reovirus	Paclitaxel; Docetaxel; Vinblastine
Parovirus	Reservatol; Norfloxacin
Measles virus	Campothecin; Alisertib
Vesicular stomatitis virus	Ruxolinitib; Cisplatin
Poxvirus	I-131
Sendai virus	PAI-1
Newcastle disease dirus	Campothecin
Alphavirus	Campothecin; H-89 protein kinase inhibitor
Herpesvirus	Fluorouracil; Tubacin; Rapamycin
Maraba virus	Paclitaxel

Analysis of the data stored in the OvirusTdb, highlights that the preferred mode of delivery of OV in the model organism was intratumoural followed by intravenous, intraperitoneal, and intramuscular. The OV offers large genome size to be genetically modified, thus offers several great benefits for the insertion of large foreign genes and drugs of high therapeutic value. Several literature studies support the notion that deletion on the *E1A* promoter and non-virulent gene enhances the therapeutic index of OV (Rojas et al., 2010; Ulasov et al., 2008). We have enlisted all the genetic modification done in OV for increasing their effectiveness in cancer therapy in Table 3.3.

Table 3.3: All different deletion and insertion mutants of oncolytic viruses.

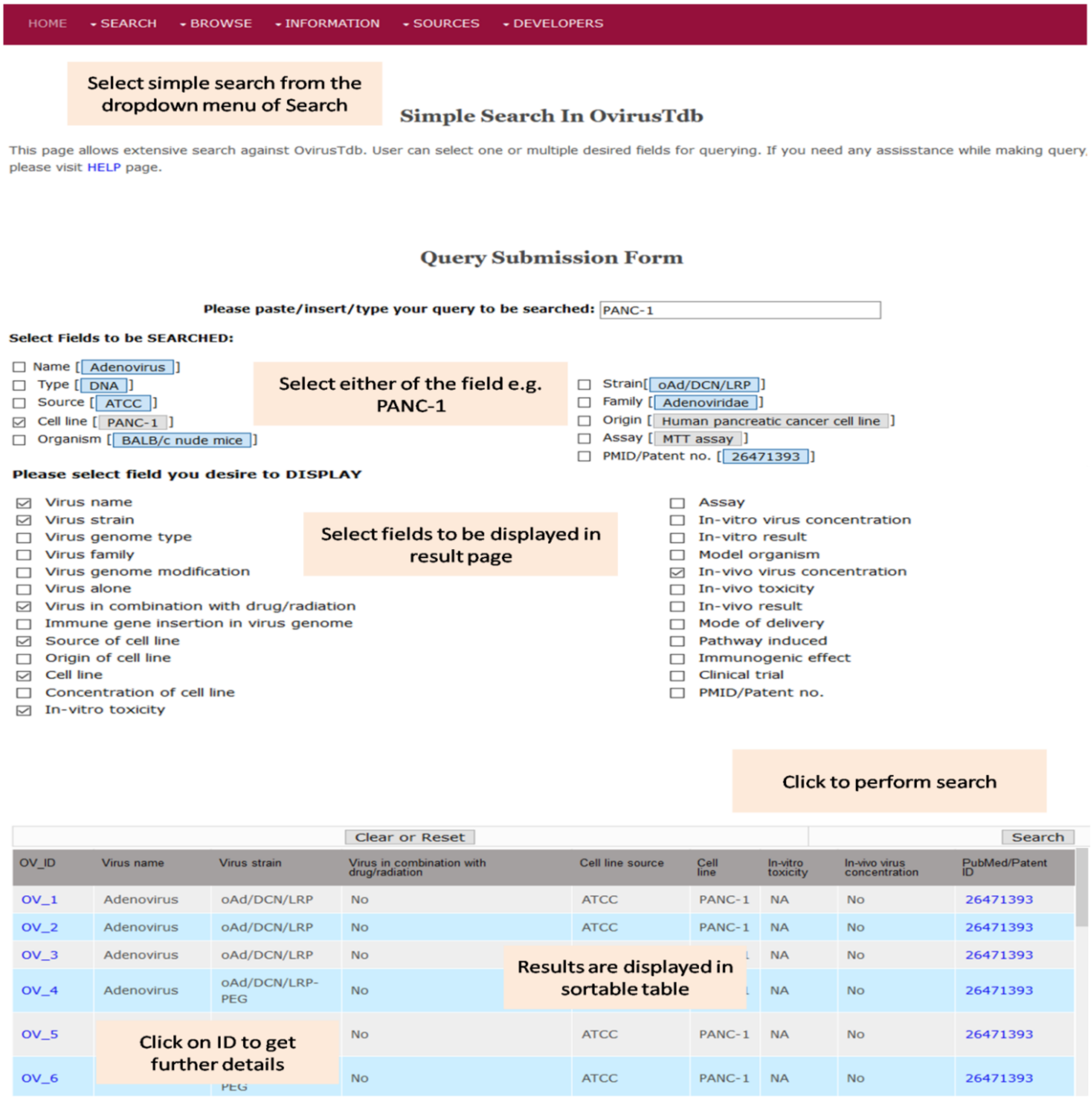
Virus species	Deletion mutant	Insertion mutant
Adenovirus	*E1A*; *E1B*; *MDR1* promoter; *hTERT* promoter; *E1AR2*	*DCN*; *LRP*; *ZD55*; *EGFP*; CPP; *MYCN*; *GOLPH2*; *dNK*; *CCL20*; *TRAIL*; *DCN*; IL-18; IL-2
Herpes simplex virus	*ICP34.5*; *ICPC6*; *ICP4*; *ICP47*	*EGFP*; *GALV.fus*; *SNORD44*; *GAS5*
Vaccinia virus	NA	*TK*; *glaf-2*; *Ruc-GFP*; *FCU1*; *DAI*
Parovirus	*ICP34.5*; *ICP4*; *ICP47*	*hNIS*; *VEGF*
Measles virus	NA	*Etag*; *P*; *P*, *N* and *L*
Vesicular stomatitis virus	G protein; *M51* gene	*P*; GFP; *SVF-5*; *Sox10*; *cytC*; *TYRP-1*
Sendai virus	*M* and *F* gene	NA
Newcastle disease virus	NA	IL-2; *TRAIL*; *hNIS*; *VEGF*

3.4 Implementation of web-resource

All the manually curated data related to the OV is stored in one single MySQL table and presented to the scientific community at (https://webs.iiitd.edu.in/raghava/ovirustdb/). The front end of the web-resource is developed using a responsive HyperText Markup Language (HTML). The advantage of using a responsive template is that it adjusts the screen ratio by sensing the user's device and thus provides interactive searching and visualization. All the back end queries related to the data retrieval were managed with the help of Hypertext Pre-processor (PHP). The developed web-interface is powered by Linux Based Apache Server, popularly known as LAMP. Moreover, the developed web-resources are equipped with various searching and browsing facilities. The description of all the implemented modules in OvirusTdb is below.

3.4.1 Data Retrieval Services

Data retrieval services of OvirusTdb include both data searching and data browsing. Users can search in and against all the fields of the OvirusTdb web-resource. From the single search page of the OvirusTdb (https://webs.iiitd.edu.in/raghava/ovirustdb/simple_search.php) user can search against any of the fields shown on the web interface. Figure 3.3 describes the whole methodology of data searching in OvirusTdb web-resource.


HOME
SEARCH
BROWSE
INFORMATION
SOURCES
DEVELOPERS
Select simple search from the dropdown menu of Search
Select either of the field e.g. PANC-1
Select fields to be displayed in result page
Click to perform search
Clear or Reset
Search
Results are displayed in sortable table
Click on ID to get further details

Simple Search In OvirusTdb

This page allows extensive search against OvirusTdb. User can select one or multiple desired fields for querying. If you need any assisstance while making query, please visit HELP page.

Query Submission Form

Please paste/insert/type your query to be searched: PANC-1

Select Fields to be SEARCHED:

- ☐ Name [Adenovirus]
- ☐ Type [DNA]
- ☐ Source [ATCC]
- ☑ Cell line [PANC-1]
- ☐ Organism [BALB/c nude mice]
- ☐ Strain[oAd/DCN/LRP]
- ☐ Family [Adenoviridae]
- ☐ Origin [Human pancreatic cancer cell line]
- ☐ Assay [MTT assay]
- ☐ PMID/Patent no. [26471393]

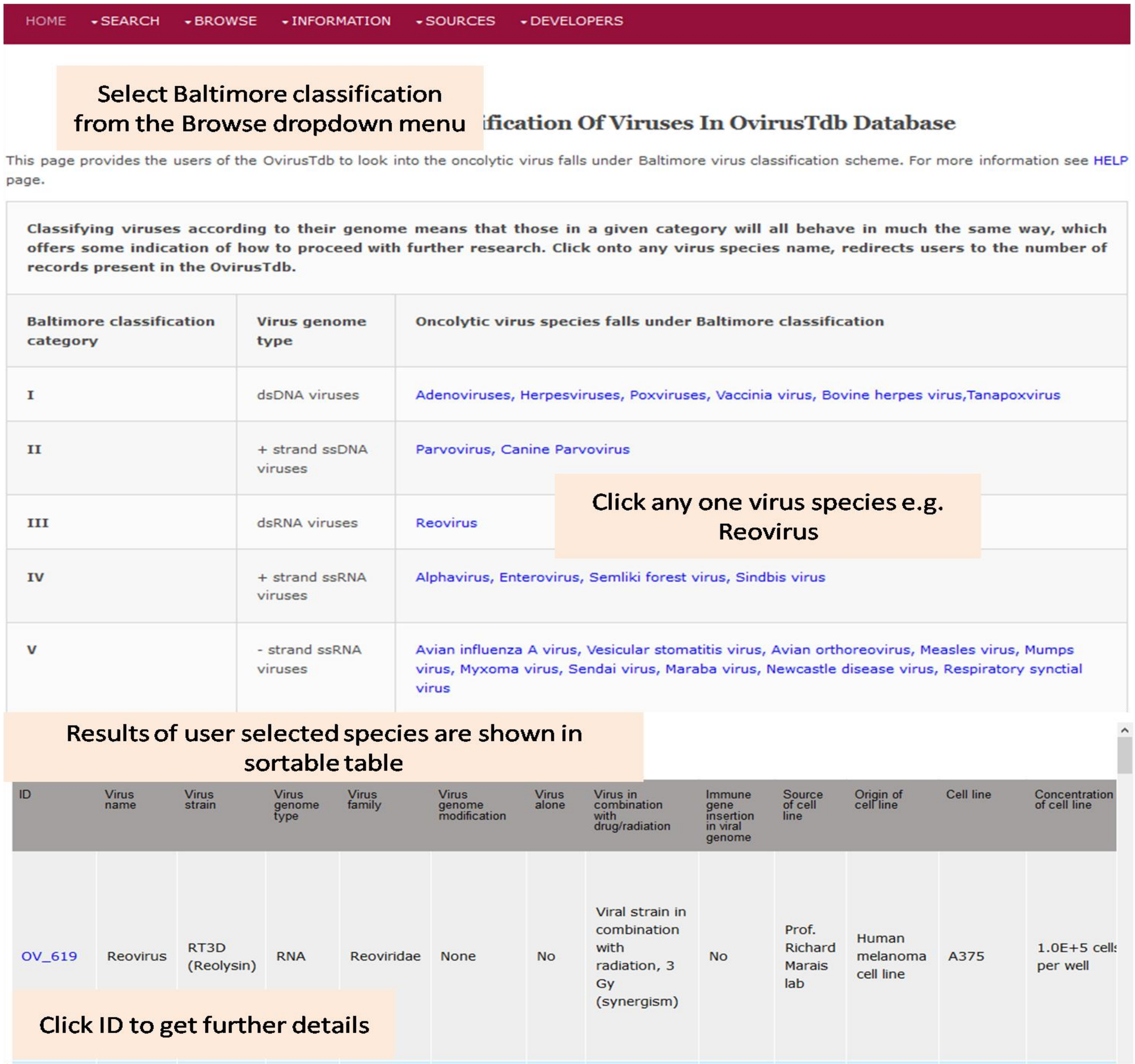
Please select field you desire to DISPLAY

- ☑ Virus name
- ☑ Virus strain
- ☐ Virus genome type
- ☐ Virus family
- ☐ Virus genome modification
- ☐ Virus alone
- ☑ Virus in combination with drug/radiation
- ☐ Immune gene insertion in virus genome
- ☑ Source of cell line
- ☐ Origin of cell line
- ☑ Cell line
- ☐ Concentration of cell line
- ☑ In-vitro toxicity
- ☐ Assay
- ☐ In-vitro virus concentration
- ☐ In-vitro result
- ☐ Model organism
- ☑ In-vivo virus concentration
- ☐ In-vivo toxicity
- ☐ In-vivo result
- ☐ Mode of delivery
- ☐ Pathway induced
- ☐ Immunogenic effect
- ☐ Clinical trial
- ☐ PMID/Patent no.

OV_ID	Virus name	Virus strain	Virus in combination with drug/radiation	Cell line source	Cell line	In-vitro toxicity	In-vivo virus concentration	PubMed/Patent ID
OV_1	Adenovirus	oAd/DCN/LRP	No	ATCC	PANC-1	NA	No	26471393
OV_2	Adenovirus	oAd/DCN/LRP	No	ATCC	PANC-1	NA	No	26471393
OV_3	Adenovirus	oAd/DCN/LRP	No			NA	No	26471393
OV_4	Adenovirus	oAd/DCN/LRP-PEG	No			NA	No	26471393
OV_5			No	ATCC	PANC-1	NA	No	26471393
OV_6		PEG	No	ATCC	PANC-1	NA	No	26471393

Figure 3.3: Data searching facility of the OvirusTdb web-resource. URL (https://webs.iiitd.edu.in/raghava/ovirustdb/simple_search.php)

If the user is interested, for example in finding all the experimental details regarding the PANC-1 cell line. The user is advised to just click on the PANC-1 from the selected fields and can also customize other options for display. By clicking onto the submit button, the user will be redirected to the result page containing all the entries of the user-defined query. Users can further get the details of each query by clicking on the ID. Moreover, an advanced search page is also there in the web-resource, which provides the output of the user-defined query by using boolean expressions.



Select Baltimore classification from the Browse dropdown menu

ification Of Viruses In OvirusTdb Database

This page provides the users of the OvirusTdb to look into the oncolytic virus falls under Baltimore virus classification scheme. For more information see HELP page.

Classifying viruses according to their genome means that those in a given category will all behave in much the same way, which offers some indication of how to proceed with further research. Click onto any virus species name, redirects users to the number of records present in the OvirusTdb.

Baltimore classification category	Virus genome type	Oncolytic virus species falls under Baltimore classification
I	dsDNA viruses	Adenoviruses, Herpesviruses, Poxviruses, Vaccinia virus, Bovine herpes virus,Tanapoxvirus
II	+ strand ssDNA viruses	Parvovirus, Canine Parvovirus
III	dsRNA viruses	Reovirus
IV	+ strand ssRNA viruses	Alphavirus, Enterovirus, Semliki forest virus, Sindbis virus
V	- strand ssRNA viruses	Avian influenza A virus, Vesicular stomatitis virus, Avian orthoreovirus, Measles virus, Mumps virus, Myxoma virus, Sendai virus, Maraba virus, Newcastle disease virus, Respiratory synctial virus

Click any one virus species e.g. Reovirus

Results of user selected species are shown in sortable table

ID	Virus name	Virus strain	Virus genome type	Virus family	Virus genome modification	Virus alone	Virus in combination with drug/radiation	Immune gene insertion in viral genome	Source of cell line	Origin of cell line	Cell line	Concentration of cell line
OV_619	Reovirus	RT3D (Reolysin)	RNA	Reoviridae	None	No	Viral strain in combination with radiation, 3 Gy (synergism)	No	Prof. Richard Marais lab	Human melanoma cell line	A375	1.0E+5 cells per well

Click ID to get further details

Figure 3.4: Browsing by Baltimore classification of OvirusTdb web-resource. URL (https://webs.iiitd.edu.in/raghava/ovirustdb/baltimore.php)

Despite this, various data browsing facilities are also implemented in the OvirusTdb web-resource. These include – Browse by species, cell line, cancer type, model organism, assay, and Baltimore classification. The Baltimore classification scheme classifies the virus species according to their genomic nature. The complete schema of browsing by the Baltimore classification is in Figure 3.4. One key application of the Baltimore classification of viruses is that every virus which falls within each same classification scheme would behave in a nearly similar manner and thus may help clinicians /researchers in guiding the experimental setup.

The Entity-Relationship diagram of the database is a high-level data representation model that defines data elements and their relationships.

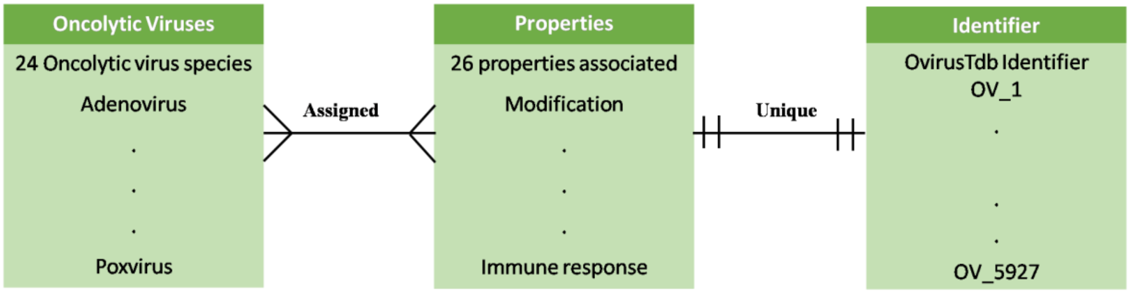


Figure 3.5: ER relationship diagram of the OvirusTdb illustrating entities and associated relationships.

The ER diagram entity represents the definable things or the main things in the database, to which various other parameters/attributes are associated. Thus, the ER diagram illustrates the relationships between the entities and their associated attributes. The relationships could be one-to-one, one-to-many, many-to-one, and many-to-many. Unique ID is in one-to-one mapping with other properties, whereas, rest of the variables exhibits many-to-many relationship pattern. The abstract ER diagram of the Ovirustdb is given above.

3.5 Conclusion and Summary

The clinical manifestation of immunotherapeutic approaches has been broadly employed in the handling of cancer. Several therapies, such as interleukin based, CAR T-cell based therapies, are already showing promising results in clinical trials. At the same time, interleukin-2 (Jiang, Zhou, & Ren, 2016) and cell cycle checkpoint inhibitors (Mills, Kolb, & Sampson, 2018) were approved for cancer treatment. Still, some challenges remained there. One major challenge faced by all cancer therapeutics is that patients with the compromised immune system do not respond well. Moreover, cancer cells also develop resistance to targeted cancer therapies (Manstein et al., 2014). Thus, a cancer treatment strategy is urgent in need, which can overcome the translational barriers of all conventional therapies such as side effects, toxicity, reduced immune stimulation, etc. The use of OV as a new class of immunotherapeutic agents can overcome all the limitations associated with conventional therapies and also have a greater benefit to risk ratio. Several experimental studies in the literature have documented promising results related to the use of OV as an immunotherapeutic agent. But scattered information in the literature is challenging to analyze for further advancement in OV based cancer treatment strategies. The high therapeutic index and lack of a unified platform regarding OV motivated us to make a single largest repository of OVs used in cancer treatment. The OvirusTdb holds manually curated information regarding experimental details on 24 OV species. The developed resources have the manifold application in advancing cancer treatment strategies. Firstly, data from the developed web-resource could be utilized for designing new OVs as OvirusTdb holds genetic modification information on 300 virus species. Secondly, it could help the genetic engineers and clinicians to design their protocols as developed web-resource provides experimental information on experimental details of model organisms, 127 cancer types, and 427 cancer cell lines. Moreover, it could also help the genetic engineers in choosing the use of a combinatorial setting to improve the therapeutic index. Thus, we conclude that the developed web-resource may help the clinicians and researchers to advance the cancer immunotherapy.

4. Identification of Vaccine Candidates against Lung Cancer-causing Viruses

4.1 Background

Lung cancer ranks first in terms of mortality rate among all the cancer types, as suggested by the World Health Organization (WHO). Lung cancer alone is the reason for a large number of deaths than the four other cancer types pancreatic, colon, rectal, and breast combined**.** WHO report also suggests that lung cancer stands first in terms of mortality among both the sexes, i.e., 28% and 26% of all cancer-related casualties among men and women, respectively. There is a continuous surge in patients of lung cancer among developing countries. It is estimated that the pervasiveness of lung cancer has increased by up to 44% in men and 76% in women since 1985 (Dela Cruz, Tanoue, & Matthay, 2011).

Lung cancer is highly metastasizing and invasive. It is estimated that nearly fifty percent of the patients with lung cancer die within 12 months, and has a 5-year overall survival rate of approximately 17.8% (Zappa & Mousa, 2016). One such prime factor that contributes towards high mortality among the patients is that they are mainly diagnosed at a later stage (Jones & Baldwin, 2018). However, it is suggested that if diagnosis happened at an early stage, the mortality rate could be brought down to very low. Numerous factors lead to the pathogenesis of lung cancer. Notable examples include occupational hazards and chemical pollutants. It is estimated that approximately 25% of all new cases of lung cancers are because of virus infections (Y. Kim, Pierce, & Robinson, 2018). Molecular experimental evidence suggests that chief viruses that drive lung cancer pathogenesis include – "*Human papilloma virus* (HPV), *Hepatitis B virus* (HBV), *Human T-lymphotropic virus* (HTLV), and *Rous sarcoma virus* (RSV)" (Robinson et al., 2016)**.** It is a general notion that nearly 30-35% of all cancer cases can be reduced by avoiding risk factors and designing prophylactic measures against causative agents. The vaccination is the preferred choice of preventive strategy against viruses that causes cancer phenotypes (Apostolopoulos, 2019).

Thus designing the vaccine candidates against viruses has remained a forefront research area for scientists. In general, developing the delivery methods that can effectively stimulate the immune system has remained a major focused area of research. Identification of TAA, soluble factors and immune-stimulant cytokines such as interleukin-2 (IL-2), interleukin-4 (IL-4), Interferon Gamma (IFN-Gamma), and TRAIL proteins has also been utilized by scientists and clinicians to design the effective cancer therapeutics. One successful example is the *MUC1* gene, which was found to be overexpressed in adenocarcinoma. Several clinical/pre-clinical studies are in the literature that considered this gene as an effective immunotherapeutic agent (Apostolopoulos, Stojanovska, & Gargosky, 2015; Tang & Apostolopoulos, 2008). Figure 4.1 summarizes the current ongoing immunotherapeutic approaches to design vaccine candidates against cancer types.

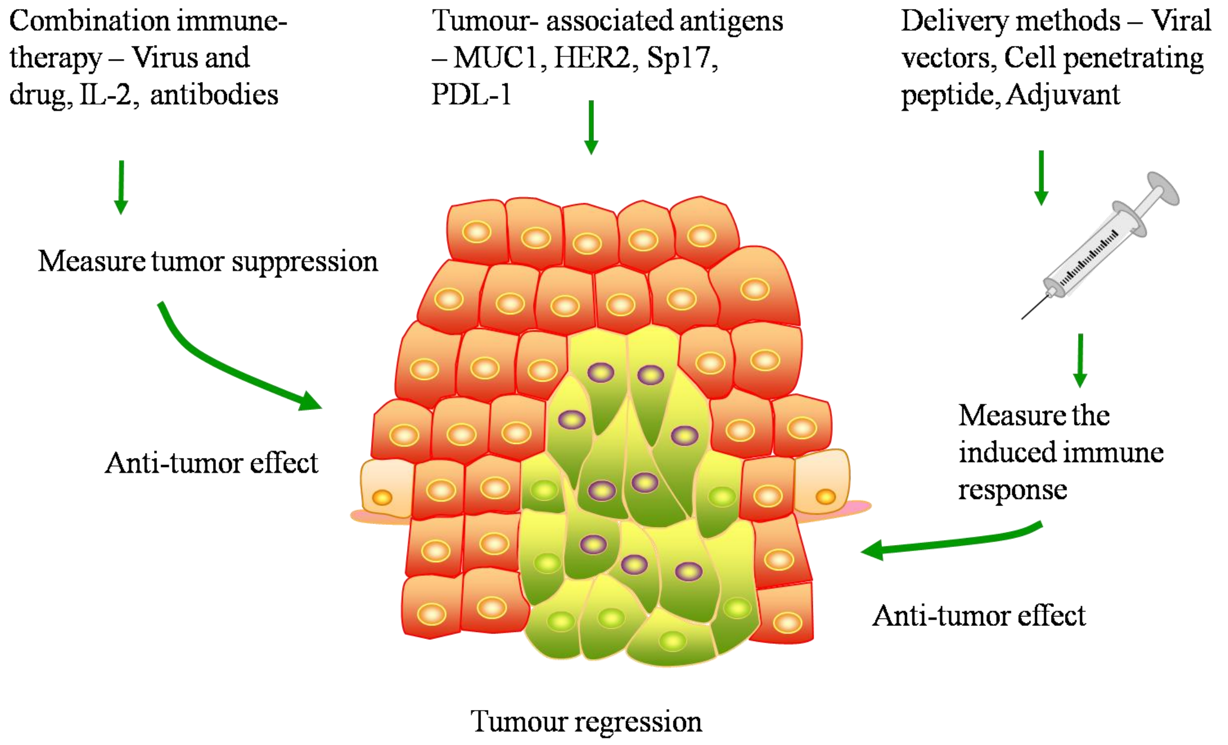

Combination immune-therapy – Virus and drug, IL-2, antibodies
Tumour- associated antigens – MUC1, HER2, Sp17, PDL-1
Delivery methods – Viral vectors, Cell penetrating peptide, Adjuvant
Measure tumor suppression
Anti-tumor effect
Measure the induced immune response
Anti-tumor effect
Tumour regression

Figure 4.1: Schematic representation of various strategies used for cancer vaccine development.

Among the vaccine category, proteomic based vaccine design seems to be useful as it possesses a greater benefit to risk ratio. The development of proteome based epitope-focused vaccine generation is more attractive for the cure of diseases where conventional approaches face several obstacles. One particular feature of epitope-based vaccines is their ability to elicit antibodies and stimulation of immune response toward sequences that otherwise is less immunogenic, such as vaccination based on attenuated/inactivated/killed pathogens. Over the past decade, the use of bioinformatics methods to identify the epitopes from the proteome of the pathogenic organism has gained the attention of clinicians and researchers. Since the pathogenic organism's proteins are foreign to the body, the immune system can be triggered against them.

The present study focuses on the identification of subunit vaccine candidate epitopes from the proteome of oncogenic lung cancer-causing viruses. The virus species included in this study are as follows – "HPV, HBV, HTLV, RSV, *Measles virus* (MV), *Epstein- Barr virus* (EBV), *Simian T-cell lymphotropic virus* (STLV), *Bovine leukemia virus* (BLV), and *JC virus* (JCV)". With the best of our knowledge, no study in the literature identifies and prioritizes the proteome based epitope vaccine candidates against the lung cancer-causing viruses. All the identified vaccine candidates were stored and presented to the scientific community in the form of a web-resource, which is freely available for use at (http://webs.iiitd.edu.in/raghava/vlcvirus).

4.2 Materials and Methods

4.2.1 Proteomic Data Extraction

We have searched the literature evidence against the viruses which are linked with lung cancer pathogenesis. We found 09 virus species that have molecular evidence of viral integration and were involved in lung cancer pathogenesis. Details of oncogenic virus species are in Table 4.1.

Table 4.1: Oncogenic viruses considered in VLCvirus study.

Virus species	Proteins	Cancer involved
Human papilloma virus (HPV)	621	Squamous cell carcinoma, Adenocarcinoma
Hepatitis B virus (HBV)	05	Squamous cell carcinoma, Adenocarcinoma
Rous sarcoma virus (RSV)	04	Adenocarcinoma, Squamous cell carcinoma
Simian T-cell lymphotropic virus (STLV)	06	Squamous cell carcinoma, Adenocarcinoma
Bovine leukemia virus (BLV)	06	Squamous cell carcinoma
Human T-cell lymphotropic virus (HTLV)	06	Squamous cell carcinoma
Epstein-Barr virus (EBV)	171	Lung adenocarcinoma
Measles virus	22	Lung adenocarcinoma
JC virus	97	Squamous cell carcinoma, Adenocarcinoma

The standard reference proteome data on oncogenic virus species have been downloaded from the UniProt database using the virus name as the keyword. We have downloaded the 100 reference proteome against the 09 oncogenic virus species containing a total of 945 proteins.

4.2.2 Selection of Candidates Proteins

Since virus proteins are an integral part of their infection machinery, targeting such proteins is an exciting research area in proteome based virus vaccine research (N. Wang, Shang, Jiang, & Du, 2020). Table 4.2 summarizes the major components of the viral proteomes utilized in our study.

Table 4.2: Oncogenic viruses proteins used in the VLCvirus study.

Virus species	**Protein components**
Human papilloma virus (HPV)	NS1, VP1, Replication protein E1, E7, E6, E4, E2, E8, Minor capsid protein L2, L1, E5A & E5B
Hepatitis B virus (HBV)	Large envelope protein, Capsid protein, Antigenic protein, Protein P & Protein X
Rous sarcoma virus (RSV)	Gag-Pol polyprotein, Gag polyprotein, Envelope glycoprotein & Src tyrosine kinase
Simian T-cell lymphotropic virus (STLV)	Tax protein, Gag polyprotein, Pro protein, Rex protein, Pol protein & Envelope glycoprotein
Bovine leukemia virus (BLV)	RT-IN, Pr66, Gp60 SU, Gag polyprotein, p18 & p34
Human T-cell lymphotropic virus (HTLV)	Gag-pro-pol, Protein Rex, Gag polyprotein, Envelope gp63 glycoprotein & Protein tax-2
Epstein-Barr virus (EBV)	BRRF2, BKRF4, BTRF1, BRRF1, BNLF2b, BZLF1, BLRF2, Capsid vertex 2, N protein, EBNA4, M protein, EBNA6, GP350, & BPLF1 protein
Measles virus	Phosphoprotein, L, N, M, H, Protein C, Fusion glycoprotein F0 & Non-structural protein V
JC virus	DnaB-like helicase, Portal protein, Minor tail protein, Beta-lactamase protein, Capsid maturation, Integrase protein, DNA primase & Uncharacterized proteins

Several ongoing clinical trials against various viral species suggest the importance of protein component-based subunit vaccine candidates. Majorly, these studies target the essential virus protein components such as non-structural proteins, glycoproteins, and receptor binding proteins to design epitope-based subunit vaccine candidates, as these proteins are integral components for virus pathogenesis (Galassie & Link, 2015). But other non-structural proteins can also have the potential to acts as an immunogen (da Silva, da Silva, Mendes, & Pena, 2020; Klemm, Boergeling, Ludwig, & Ehrhardt, 2018). In our study, we have included essential protein components as well as all other protein components of the viruses for the identification of vaccine candidates.

4.2.3 Epitopes Prediction Pipeline

The proteomic data of the oncogenic viruses were downloaded from the UniProt, and the immune-stimulatory potential of the generated nine-mers was assessed using a stringent pipeline. The complete picture of the developed pipeline is explained in Figure 4.2.

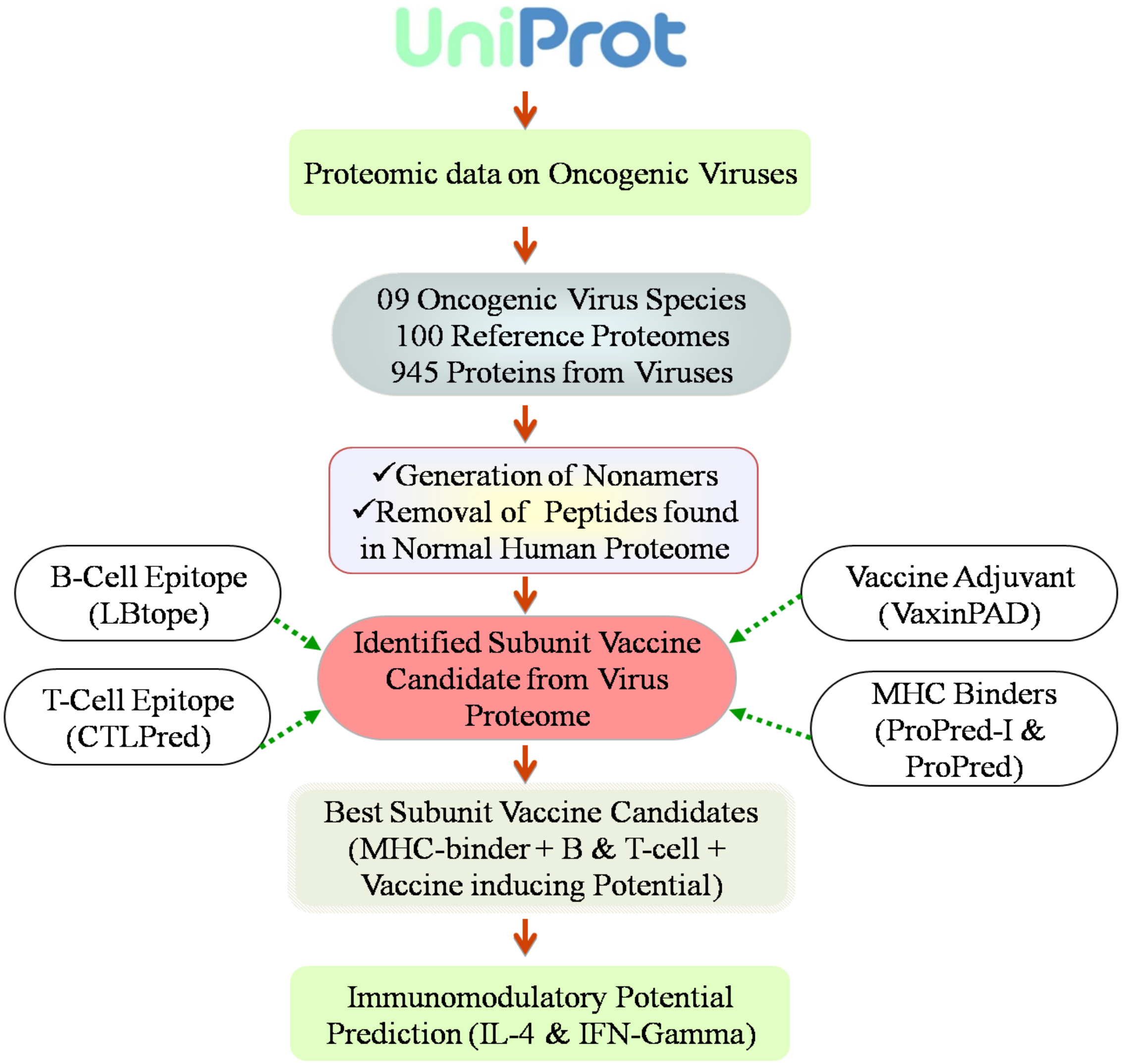

UniProt
Proteomic data on Oncogenic Viruses
09 Oncogenic Virus Species
100 Reference Proteomes
945 Proteins from Viruses
✓Generation of Nonamers
✓Removal of Peptides found in Normal Human Proteome
B-Cell Epitope (LBtope)
T-Cell Epitope (CTLPred)
Identified Subunit Vaccine Candidate from Virus Proteome
Vaccine Adjuvant (VaxinPAD)
MHC Binders (ProPred-I & ProPred)
Best Subunit Vaccine Candidates (MHC-binder + B & T-cell + Vaccine inducing Potential)
Immunomodulatory Potential Prediction (IL-4 & IFN-Gamma)

Figure 4.2: Workflow of the entire pipeline used to build VLCvirus web-resource.

We have generated the nona-peptides from the virus proteome using in house developed PYTHON script. The nona-peptide is a continuous stretch of overlapping nine residue long proteomic subparts from the whole length protein. It was observed that viruses might evade the immune system by producing ligands and molecules that mimic the cellular and biological pathways of the host (Christiaansen, Varga, & Spencer, 2015). Thus, self-tolerance should be considered while designing the proteome based subunit vaccine candidates. To avoid the self-tolerance mechanism, we first generated all overlapping nine-mers from the normal 1000 human reference proteome. Secondly, the generated nona-peptides from the virus proteomes were mapped onto the nine-mers obtained from the human proteomes. The nona-peptides, which are 100% identical with the human nine-mers, were removed from the study. The remaining nona-peptides of the virus proteomes are further checked for redundancy, and if found, they were removed. Our next aim was to identify the nona-peptides that could act as an immune stimulant. We have developed an integrative pipeline that predicts the immune-stimulant potential of all the nona-peptides. The pipeline includes – integration of ProPred1 (Harpeet Singh & Raghava, 2003), ProPred (Harpreet Singh & Raghava, 2002), CTLPred (Bhasin & Raghava, 2004), LBTope (Harinder Singh, Ansari, & Raghava, 2013), and VaxinPAD (Nagpal, Chaudhary, Agrawal, & Raghava, 2018), which predicts the MHC class-I, II binder, T-cell epitopes, B-cell epitopes, and vaccine adjuvant potential of all the nona-peptides. Literature evidence supports that IL-4, IFN-Gamma, can act as immune-activator cytokines. Further, to identify the anticancer property of the best subunit vaccine epitopes, we have utilized the IL-4 and IFN-Gamma in our prediction pipeline.

4.3 Results

4.3.1 Identification of Immune Stimulation Properties

All the generated nona-peptides after exclusion from the normal 1000 human proteome were fed into the developed prediction pipeline for the identification of different immune stimulation properties. The integrated tools in the pipeline help us to find out the immunogenic epitopes from the generated nona-peptides of all virus species under consideration. We have also predicted the combinational properties of the nona-peptides such as – a nona-peptide having B-cell inducing potential along with T-cell inducing potential and can act as a vaccine adjuvant as well. In our analysis, we termed these epitopes as multifunctional epitopes. The result obtained from this analysis was provided in Table 4.3.

Table 4.3: The complete statistics of the unifunctional and multifunctional epitopes.

Properties	BLV	HBV	HPV	HTLV	RSV	STLV	EBV	JCV	MV
Unifunctional Epitopes Identified From Oncogenic Viruses Proteome									
T-Cell Epitope	717	427	48347	665	729	500	17559	3429	3396
B-cell Epitope	228	251	17985	445	297	229	7208	1102	1397
Vaccine Epitope	324	256	25723	293	401	234	7312	1466	1833
MHC_I	10943	6332	665437	11384	9769	8293	270008	48869	49913
MHC_II	4241	3043	272264	4835	3736	3297	117045	16236	24343
Multifunctional Epitopes Identified From Oncogenic Viruses Proteome									
MHC_I_II	292	228	21472	289	282	298	5265	1615	776
B_MHC_I	113	137	9771	197	141	145	2203	767	369
B_MHC_II	19	19	1815	41	21	35	335	103	86
B_MHC_I_II	18	16	1443	34	18	30	272	82	66
T_MHC_I	509	357	38753	468	486	490	9057	3243	1311
T_MHC_II	97	71	6628	82	83	88	1554	466	236
T_MHC_I_II	97	69	6480	82	81	87	1535	455	228
BV_MHC_I	12	27	1130	15	13	8	191	70	52
BV_MHC_II	2	2	329	4	4	2	59	11	14
BV_MHC_I_II	2	2	272	4	4	2	50	8	11
TV_MHC_I	86	55	6743	60	75	51	1246	437	221
TV_MHC_II	17	9	1578	8	17	9	311	96	58
TV_MHC_I_II	17	9	1539	8	17	8	306	94	54
TBV	5	11	494	4	5	2	75	34	26

However, the identified nona-peptides having immune stimulation potential were large enough in numbers to be utilized in the clinical setup. Thus, to prevent such limitations and enhance the scalability of our epitopes in the clinical setup, we have applied stringent criteria. Those epitopes were retained, which shows T-cell, B-cell, vaccine adjuvant, and MHC-I, II binding potential. We termed these epitopes in our study as the best antigenic epitopes. These epitopes were final recommendations for therapeutics and can be taken further for experimental validations. The results obtained from this analysis were provided in Table 4.4. Another immune stimulation and anticancer properties such as IL-4 and IFN-Gamma inducing potential of best antigenic epitopes were also predicted using IL4pred and IFNepitope, respectively.

Table 4.4. The data statistics of best antigenic epitopes obtained after stringent criteria.

Properties	BLV	HBV	HPV	HTLV	RSV	STLV	EBV	JCV	MV
TBV_MHC_I	5	10	473	4	5	2	72	34	26
TBV_MHC_II	1	0	99	1	0	1	18	3	5
TBV_MHC_I_II	1	0	97	1	0	1	17	3	5

4.3.2 Identification of Promiscuous Best Antigenic Epitopes

Earlier studies suggested that epitope-based immunization was capable of eliciting a protective immune response. Yet, experimental evidence supports the notion that further to increase the efficacy of epitope-based vaccine design, promiscuous epitopes should be considered (Ebrahimi, Mohabatkar, & Behbahani, 2019). Promiscuous epitopes show the binding affinity with a large number of HLA molecules, thus posing great therapeutic benefits. In our study, we have also identified the promiscuous epitopes, which can bind to large numbers of HLA molecules. These epitopes are necessary for boosting the T-cell mediated immune response. The most frequently expressed alleles in human populations are HLA-A1, HLA-A2, DRB1, and DRB4 (Oprea & Antohe, 2013). Out of 125 best-identified epitopes, some can bind to nearly 15 MHC-type I ("*VMFVSRVPV*") and 49 ("*LRRFMVALI* ") MHC-type 2 alleles.

4.3.3 Promiscuous Epitopes and Immunodominant Genes across Strains

We have also identified the presence of promiscuous epitopes across the virus species along with their combination of various stimulating properties. The result of this analysis was provided in the Figure 4.3. An immune response generated against one virus species may provide immunity against another related virus species is termed as heterologous immunity (Ngono & Shresta, 2019). We have also accessed the cross-reactive immunity generation potential of identified best subunit vaccine candidates. We have analyzed the best antigenic epitopes for their presence across the virus species as they can serve the basis for the generation of heterologous immunity.

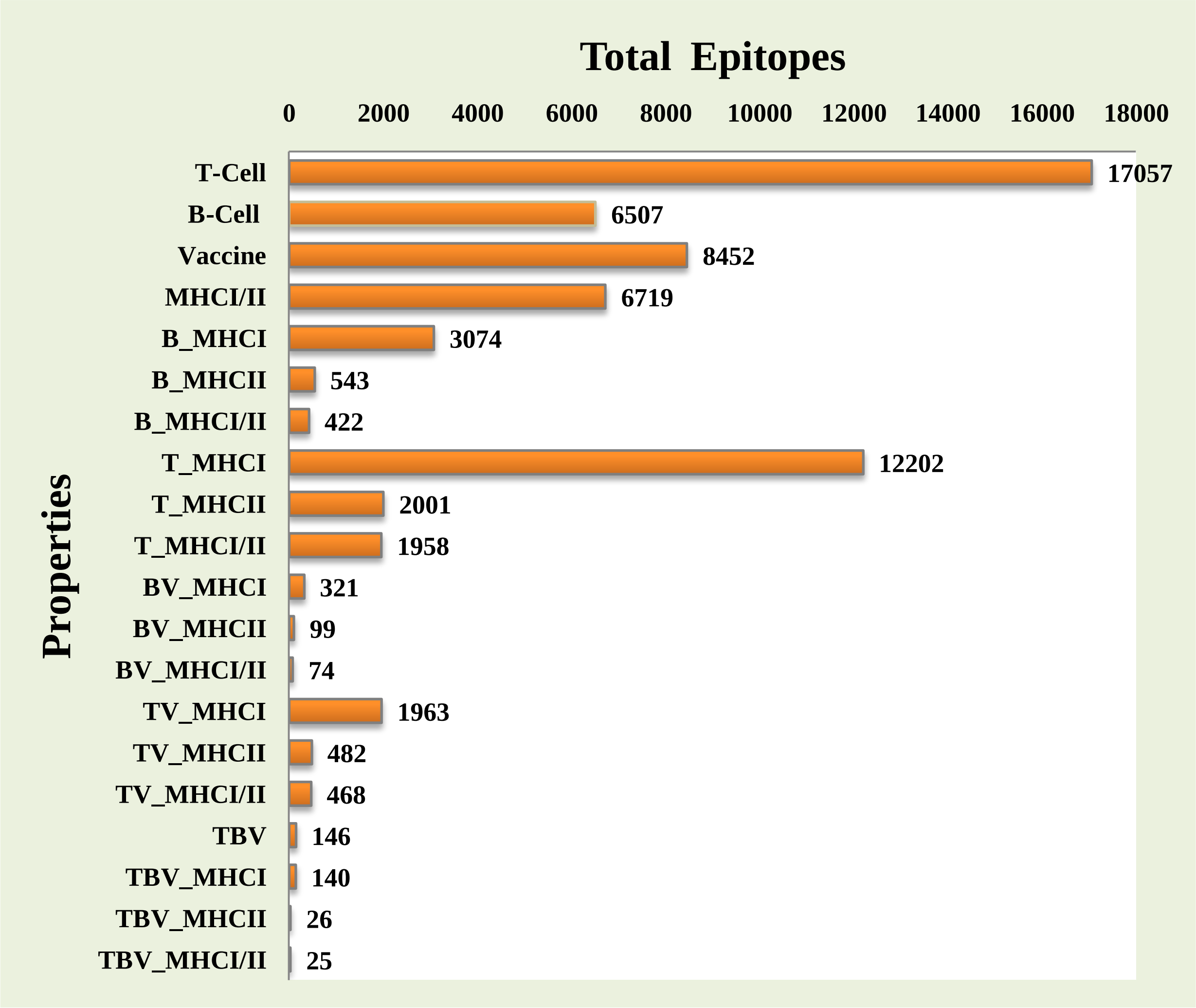


Figure 4.3: Distribution of promiscuous epitopes across the different lung cancer-causing viral strains/species

Results suggest that the majority of the epitopes belonged to *E1* and *E6* genes, with several literature studies, backed the importance of these envelope proteins in designing epitope-based vaccines against the respective pathogenic/virus strains (Lei et al., 2019; Sabah et al., 2018). Thus, the identified genes are not only used for developing cross-species immunity but also act as Immunodominant genes. The visual representation of the distribution of the epitopes across the various genes is in Figure 4.4.

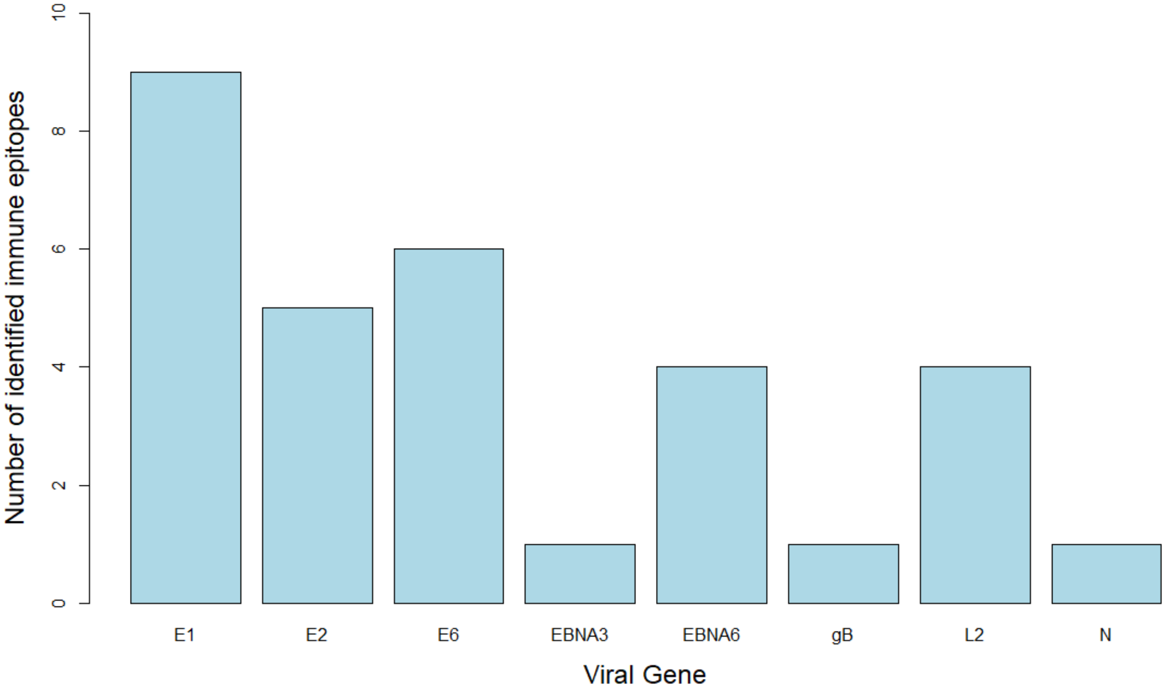


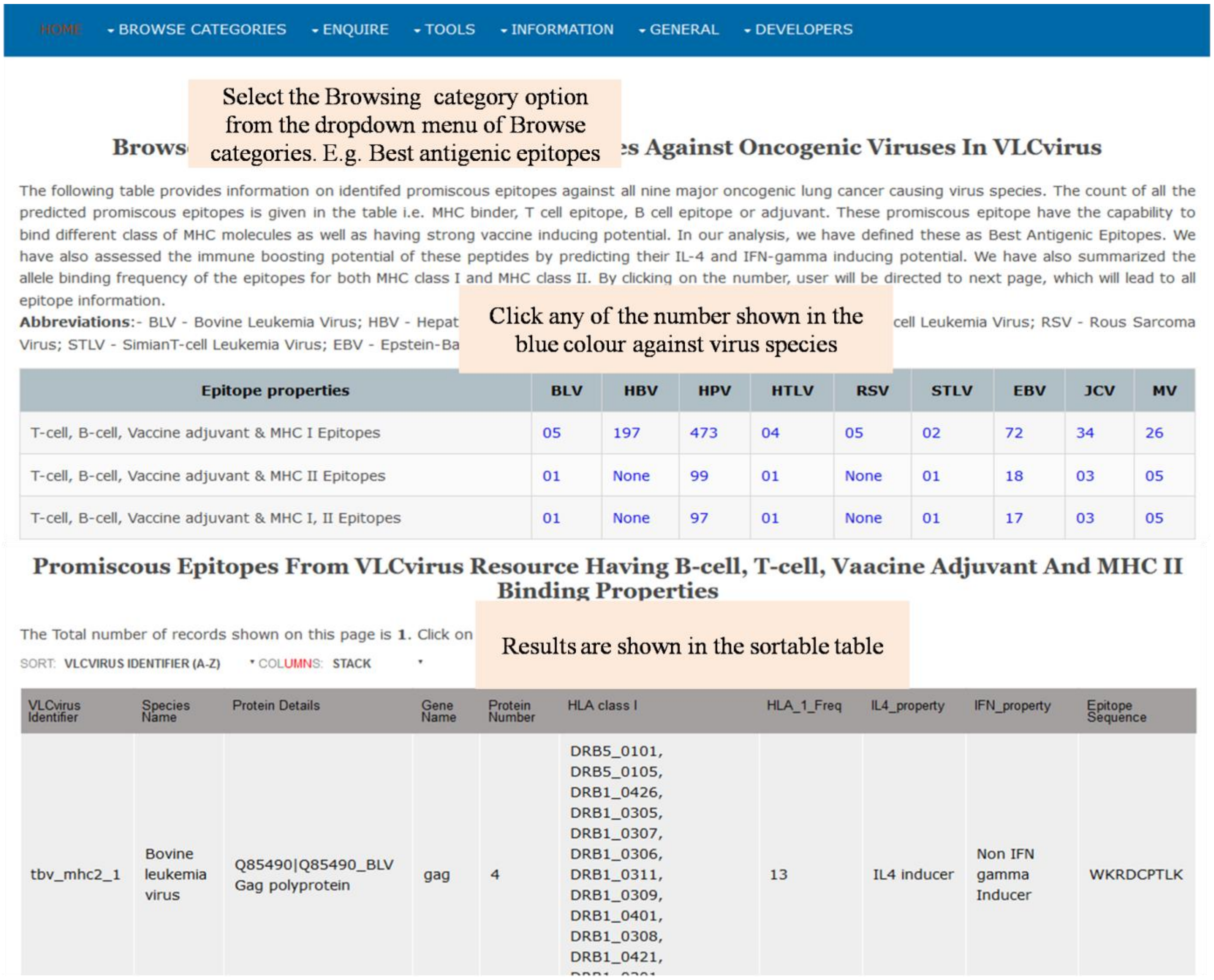
Figure 4.4: Immunodominant genes across the virus species.

4.4 Implementation of the Web-resource

All the results obtained from the analysis were stored in multiple MySQL tables and presented to the scientific community in the form of a free to use web-resource VLCvirus, which is provided at http://webs.iiitd.edu.in/raghava/vlcvirus. The VLCvirus web-resource is built on a LAMP server of 2.4.7 version. The web-resource front-end was created using a bootstrap responsive HTML5.0 template, which adjusts the screen ratio by sensing the user's device. The back-end data related queries were managed by MySQL server of version 5.5.55-0 ubuntu 0.14.04.1 - (Ubuntu).

4.4.1 Data Retrieval Services

The VLCvirus web-resource is equipped with various data searching and data browsing facilities. Users can browse the data stored in the web-resource in an interactive manner with seamless speed. The browsing option provided in the resource includes – Browse by Multifunctional epitopes, Promiscuous epitopes, and best antigenic epitopes. A step by step procedure of the browse functionality by best antigenic epitopes is shown in Figure 4.5. Despite data browsing, various enquire modules are also provided in the VLCvirus resource.



Select the Browsing category option from the dropdown menu of Browse categories. E.g. Best antigenic epitopes

Brows… …es Against Oncogenic Viruses In VLCvirus

The following table provides information on identifed promiscous epitopes against all nine major oncogenic lung cancer causing virus species. The count of all the predicted promiscous epitopes is given in the table i.e. MHC binder, T cell epitope, B cell epitope or adjuvant. These promiscous epitope have the capability to bind different class of MHC molecules as well as having strong vaccine inducing potential. In our analysis, we have defined these as Best Antigenic Epitopes. We have also assessed the immune boosting potential of these peptides by predicting their IL-4 and IFN-gamma inducing potential. We have also summarized the allele binding frequency of the epitopes for both MHC class I and MHC class II. By clicking on the number, user will be directed to next page, which will lead to all epitope information.

Abbreviations:- BLV - Bovine Leukemia Virus; HBV - Hepat… …cell Leukemia Virus; RSV - Rous Sarcoma Virus; STLV - SimianT-cell Leukemia Virus; EBV - Epstein-Ba…

Click any of the number shown in the blue colour against virus species

Epitope properties	BLV	HBV	HPV	HTLV	RSV	STLV	EBV	JCV	MV
T-cell, B-cell, Vaccine adjuvant & MHC I Epitopes	05	197	473	04	05	02	72	34	26
T-cell, B-cell, Vaccine adjuvant & MHC II Epitopes	01	None	99	01	None	01	18	03	05
T-cell, B-cell, Vaccine adjuvant & MHC I, II Epitopes	01	None	97	01	None	01	17	03	05

Promiscous Epitopes From VLCvirus Resource Having B-cell, T-cell, Vaacine Adjuvant And MHC II Binding Properties

The Total number of records shown on this page is **1**. Click on

SORT: VLCVIRUS IDENTIFIER (A-Z) COLUMNS: STACK

Results are shown in the sortable table

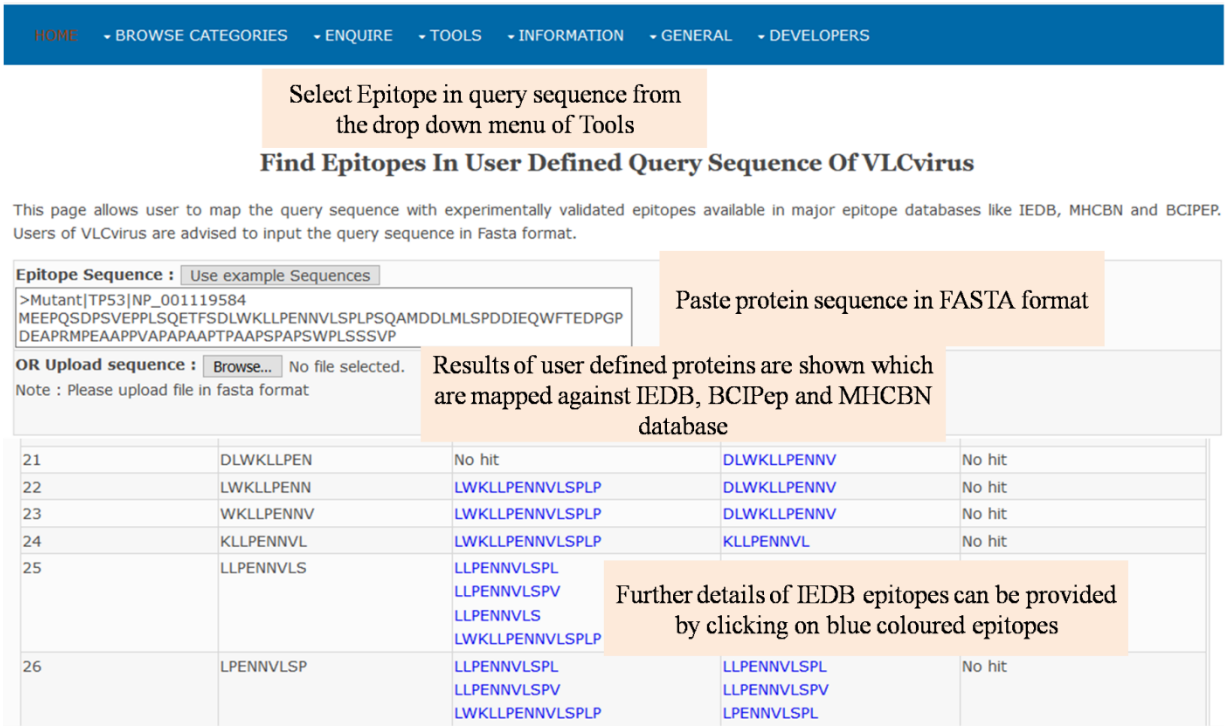
VLCvirus Identifier	Species Name	Protein Details	Gene Name	Protein Number	HLA class I	HLA_1_Freq	IL4_property	IFN_property	Epitope Sequence
tbv_mhc2_1	Bovine leukemia virus	Q85490|Q85490_BLV Gag polyprotein	gag	4	DRB5_0101, DRB5_0105, DRB1_0426, DRB1_0305, DRB1_0307, DRB1_0306, DRB1_0311, DRB1_0309, DRB1_0401, DRB1_0308, DRB1_0421,	13	IL4 inducer	Non IFN gamma Inducer	WKRDCPTLK

Figure 4.4: Screenshots of the browsing facility implemented in VLCvirus.

The data enquire module includes – Multiquery search, Immunomodulatory epitopes, and Immune Epitope Database (IEDB) mapped epitopes. The multi-query search has its advantages as it provides the users with the flexibility in searching the resource by giving various parameters for selection.

4.4.2 Integration of Tools

Moreover, VLCvirus web-resource also equipped with several data analysis tools such as BLAST. Another interesting tool that was integrated into the VLCvirus web-resource is the identification of the epitopes in user-defined query protein. The identified epitopes in the user given protein sequences were mapped on to the IEDB, MHCBN, and BCIPEP databases. Thus this will helps the user in identifying and prioritizing the identified epitopes. Figure 4.6 shows the step by step procedure to use this facility implemented in VLCvirus.



Select Epitope in query sequence from the drop down menu of Tools

Find Epitopes In User Defined Query Sequence Of VLCvirus

This page allows user to map the query sequence with experimentally validated epitopes available in major epitope databases like IEDB, MHCBN and BCIPEP. Users of VLCvirus are advised to input the query sequence in Fasta format.

Epitope Sequence :

```
>Mutant|TP53|NP_001119584
MEEPQSDPSVEPPLSQETFSDLWKLLPENNVLSPLPSQAMDDLMLSPDDIEQWFTEDPGP
DEAPRMPEAAPPVAPAPAAPTPAAPSPAPSWPLSSSVP
```

Paste protein sequence in FASTA format

OR Upload sequence : No file selected.

Note : Please upload file in fasta format

Results of user defined proteins are shown which are mapped against IEDB, BCIPep and MHCBN database

21	DLWKLLPEN	No hit	DLWKLLPENNV	No hit
22	LWKLLPENN	LWKLLPENNVLSPLP	DLWKLLPENNV	No hit
23	WKLLPENNV	LWKLLPENNVLSPLP	DLWKLLPENNV	No hit
24	KLLPENNVL	LWKLLPENNVLSPLP	KLLPENNVL	No hit
25	LLPENNVLS	LLPENNVLSPL LLPENNVLSPV LLPENNVLS LWKLLPENNVLSPLP		
26	LPENNVLSP	LLPENNVLSPL LLPENNVLSPV LWKLLPENNVLSPLP	LLPENNVLSPL LLPENNVLSPV LPENNVLSPL	No hit

Further details of IEDB epitopes can be provided by clicking on blue coloured epitopes

Figure 4.6: Screenshots of the epitope in the query sequence facility of VLCvirus.

4.5 Conclusion and Summary

Despite the improvement in cancer therapy, in terms of incidences and mortality rates, lung cancer stands first. Several treatment options, such as drugs like paclitaxel, carboplatin, and cisplatin, have improved the lives of patients. But, even today, there is no definitive cure for lung cancer. Additionally, chemotherapy has several side effects like toxicity, immune suppression, along with many others. Contrary to chemotherapy, instigating the host's immune response against the pathogen offers several advantages. The advantage of vaccination can be seen in terms of successful usage of DPT, BCG, and polio vaccine for the management of human diseases. Owing to the greater benefit ratio makes the peptide-based subunit vaccine drug development as an exciting area of research. Several literature studies highlight the use of bioinformatics based techniques for the identification of subunit vaccine candidates against pathogenic organisms. A notable successful example is the validation of predicted epitopes of SARS-Cov-2 viruses to design vaccine candidates and which additionally also serves the basis of heterologous immunity generation (Grifoni et al., 2020). Literature studies also advocate the use of proteomic based approaches for identifying the vaccine epitopes as a prophylactic measure. By taking insight from the literature evidence, the present research focuses on the identification of subunit vaccine candidates against the lung cancer-causing virus species. The present study identifies 125 best antigenic epitopes that can be utilized in clinics for providing immunity against lung cancer viruses. We have also accessed the immune booster and anticancer cytokine induction analysis of these best antigenic epitopes. Out of 125 epitopes, 103 and 38 epitopes were found to have IL-4 and IFN-gamma inducing potential, whereas 32 were identified to be the inducers of both. We have also assessed the presence of the best antigenic epitopes in the IEDB database. We found 38 matched records against 125 best antigenic epitopes and 29,193 records for the epitopes (193437) having any of the computed properties, i.e., T-cell epitopes; B-cell epitopes, vaccine candidates, and MHCI/II binders in IEDB database. Full details of IEDB matched epitopes are provided on the server (https://webs.iiitd.edu.in/raghava/vlcvirus/iedb_mapped.php). This analysis supports the reliability and credibility of our identified epitopes to a greater extent. Thus we conclude that the identified vaccine candidates may serve the basis for the generation of preventive therapy against lung cancer. Since each prediction algorithm has its limitations, we suggest that identified epitopes should be verified experimentally in the clinical setup before to be used as a therapeutic agent. Despite such limitations, we hope that data stored in the VLCvirus will be helpful for researchers to fasten the vaccine designing process against lung cancer.

5. Identification of Prognostic Biomarkers for Non-small-cell Lung Cancer

5.1 Introduction

Lung cancer in both men and women worldwide is the most common cause of death (Barta et al., 2019). A report from the American Cancer Society estimated the death of nearly 60% of the lung-cancer patients within one year of their diagnosis (https://www.cancer.org). It is well-observed that the mortality rate among the patient of lung cancer is high as compared to the other four prevalent cancers (breast, prostate, colon, and pancreas) combined. Lung cancer was primarily divided into two types – small and non-small-cell lung carcinoma on the basis of histological features and immunological markers. The classification is also impacted by the availability of targeted therapies and the molecular characteristics of cancer. Non-small-cell lung cancer (NSCLC) is the most common type of lung cancer that accounts for nearly 84% of diagnosed lung cancer patients. NSCLC can be further categorized into three major classes – "Lung Adenocarcinoma (LUAD), Lung Squamous Cell Carcinoma (LUSC), and Large Cell Carcinoma (LCC)," where LUAD and LUSC share the most significant load with 70% of all the NSCLC cases. NSCLC subtypes have different origins, lung locations, and even growth patterns, which suggest that both LUAD and LUSC are different diseases that progress through distinct molecular mechanisms (Pikor, Ramnarine, Lam, & Lam, 2013). Until very recently, owing to the distinct histopathologies, genomic patterns, and supporting clinical evidence suggesting the impact of these differences on response rates and toxicity of the particular treatment, histology has been recognized as a vital feature in suggesting possible therapeutic interventions for the NSCLC patients. Considering all the underlying differential genomic and histological mechanism in the LUSC and LUAD, it will be an unrealistic assumption that both the subtypes will respond similarly when subjected to a given therapeutic regimen. These differences highlight the dire need for information on the underlying mechanism to guide the informed therapeutic decisions in a subtype-specific manner that can improve the overall survival and response of the patient.

It is a well-known fact that early detection of the NSCLC resulted in the increased chances of successful treatment. Once the diagnosis is made, the timely prognosis can help in deciding and anticipating the response of the patient towards a given treatment. Various advancements in molecular diagnostic, prognostic, and genome-directed therapies have provided us with hope for the effective management of the advanced NSCLC (Mehta et al., 2019). The immense level of heterogeneity present among the NSCLC subtypes and lack of subtype-specific therapeutic interventions contribute to the poor survival rate among the patients. Also, nearly two-thirds of the NSCLC patients are diagnosed in their advanced stage, which leads to poor prognosis, limited therapeutic options, and, ultimately, the low five-year survival rates in the patients.

Thus, there is a strong need for new contra-positive biomarkers for both subtypes that may aid in early diagnosis, improved prognosis, and facilitate more personalized treatment for the patients. Biomarkers can be a characteristic used to indicate the biological state of the tumour and patient. Prognostic biomarkers indicate the outcome and the progression rate of the diseases, irrespective of the treatment given to the patient. In contrast, predictive biomarkers provide the likelihood of the benefit for a patient when subjected to a particular treatment. “The Biomarker-integrated Approaches of Targeted Therapy for Lung Cancer Elimination (BATTLE)” trial got completed in 2011, which integrated the molecular data into the clinical trials to locate the patients that may benefit from personalized therapies (E. S. Kim et al., 2011). This project made the real-time biomarker analysis quite feasible and, at the same time, also identified the potential new markers that may aid in the progression of the personalized lung cancer treatment. Results from these studies, along with the latest next-generation sequencing technologies, resulted in a better biological understanding of various molecular mechanisms of the diseases. Also, with the technological advancement, computational analysis, and interpretation of sequencing results can be made in a time-efficient and cost-effective manner. Our ongoing efforts in further investigating the heterogeneity in the genomic landscape of LUSC and LUAD will keep on improving our understanding of the NSCLC biology, thus improving the overall outcome in the patients.

In this regard, the current study aims to assess the prognostic potential of the genes involved in the major subtypes of the NSCLC, i.e., LUSC and LUAD. We included different cancer-driving pathways and the clinical information in the study to identify the best possible prognostic signature that can successfully assess risk in the patients. The overall significance of the study is that it not only prioritize the biological pathways significant to the overall survival but also identifies the potential prognostic genes that can discriminate among NSCLC’s subtypes to devise better personalized therapeutic strategies.

5.2 Materials and Methods

5.2.1 Data Extraction and Pre-Processing

We extracted the RNAseq and clinical-pathological data of “LUSC (488 samples) and LUAD (497 samples)” from The Cancer Genome Atlas (TCGA) via TCGA assembler (Zhu, Qiu, & Ji, 2014). We analyzed 20530 genes and shortlisted only those that were present in more than 50% of the samples; thus, only 17982 and 17756 genes were selected for LUSC and LUAD sets, respectively. Gene Expression Omnibus (GEO) dataset with accession number GSE42127 was used to validate the final results of the study. The validation dataset consisted of a total of 176 samples, including 43 of LUSC and 133 of LUAD cancer types. Data had 16295 intersecting genes among LUSC test and validation sets, whereas 16104 genes among LUAD set. The ProbeID was mapped onto the gene symbol using the schema provided in the GEO database, and then quantile normalization of the test and validation datasets was done.

5.2.2 Prognostic Markers and Model Framework

We screened the genes linked to the overall survival (OS) of the patients in TCGA datasets using Univariate Cox regression via the “Survival” package (V.2.42-6) in R. Genes that were significantly linked to the OS of the patients (p-value < 0.05) were selected for further analysis. The median of the gene expression was used as a threshold criterion for the identification of survival-related genes. Prognostic genes were classified among good and bad prognostic markers, i.e., GPM, BPM, and combined (BPM+GPM) set. BPM genes are the ones that were found to be directly correlated with the low survival of the patients, whereas GPM genes were correlated with the better outcome in the patients. We implemented random-forest variable hunting (RF-vh) with 100 iterations using the “randomForestSRC” package of R to refine the gene pool further. We iteratively searched for the minimum gene set to build a robust survival model to enhance the scalability of the model in the clinical/research setup. We also incorporated information on the pathways which are likely to be cancer driver and therapeutic targets, as suggested by various literature studies. The pathways are, namely Apoptotic pathways, PI3K-AKT, MYC, Cell Cycle, WNT, HIPPO, NRF2, NOTCH, P53, RAS, and TGF-Beta. We utilized different clinical features available in the downloaded file from the TCGA to access their contribution in predicting the OS of the NSCLC patients.

We have developed the prognostic index (PI) based models which were formulated as follows:

$$PI = \beta_1 Z_1 + \beta_2 Z_2 + \ldots + \beta_n Z_n$$

Where β value represents the regression coefficient for any gene Z, calculated via univariate cox regression

For each set, PI was used to classify cancer patients in high and low-risk sets. Patients with PI ≤ median (PI) were categorized as low-risk groups, whereas patients having PI > median (PI) were labeled as high-risk patients. We validated our final results on the GEO dataset as well.

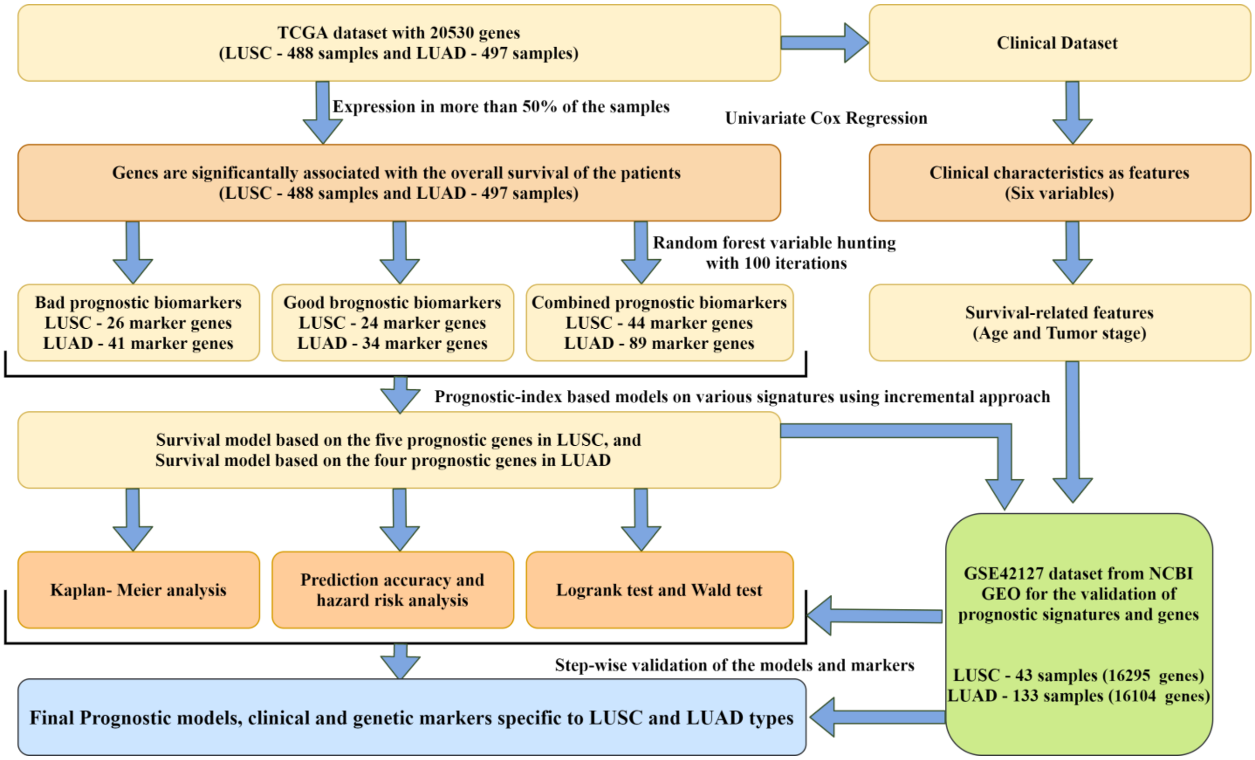

TCGA dataset with 20530 genes
(LUSC - 488 samples and LUAD - 497 samples)
Clinical Dataset
Expression in more than 50% of the samples
Univariate Cox Regression
Genes are significantly associated with the overall survival of the patients
(LUSC - 488 samples and LUAD - 497 samples)
Clinical characteristics as features
(Six variables)
Random forest variable hunting
with 100 iterations
Bad prognostic biomarkers
LUSC - 26 marker genes
LUAD - 41 marker genes
Good brognostic biomarkers
LUSC - 24 marker genes
LUAD - 34 marker genes
Combined prognostic biomarkers
LUSC - 44 marker genes
LUAD - 89 marker genes
Survival-related features
(Age and Tumor stage)
Prognostic-index based models on various signatures using incremental approach
Survival model based on the five prognostic genes in LUSC, and
Survival model based on the four prognostic genes in LUAD
Kaplan- Meier analysis
Prediction accuracy and
hazard risk analysis
Logrank test and Wald test
GSE42127 dataset from NCBI
GEO for the validation of
prognostic signatures and genes
LUSC - 43 samples (16295 genes)
LUAD - 133 samples (16104 genes)
Step-wise validation of the models and markers
Final Prognostic models, clinical and genetic markers specific to LUSC and LUAD types

Figure 5.1: A schematic depiction of the computational pipeline of the study.

5.2.3 Evaluation Metrics

The performance of the developed PI model was accessed based on several statistical parameters such as hazard ratio (HR), p-values, log-rank, concordance, and Wald test. HR was used to assess the relative risk among high and low-risk groups. The p-value is preferred choice in medical articles and gives information on the correctness of the hypothesis. The log-rank test explained the statistical relevance of the survival curves for the risk sets. Wald-test was used to describe the importance of variables employed in calculating HR. Concordance tells us about the performance level of the stratification of the model among high and low-risk sets. Lower log-rank and high concordance depict a good prognostic model. We have validated our results on the GEO datasets, and the overall workflow of the study is in Figure 5.1.

5.3 Results

5.3.1 Identification of Prognostic Biomarkers

To identify the prognostic potential of the genes, the survival analysis was done on each dataset (TCGA- LUSC & LUAD). Using the median cut-off of gene expressions in cox regression analysis, we obtained 1334 and 2129 prognostic genes in LUSC and LUAD, accordingly. These genes were catalogues into BPM and GPM classes using the approach mentioned in the materials and method section above. Using RF-vh, we got 26 genes (out of 637 BPM genes), 24 genes (out of 697 GPM genes), and 44 genes (out of the combined set of 1334 genes) in LUSC, whereas 41 genes (out of 1153 BPM genes), 34 genes (out of 976 GPM genes), and 89 genes (out of the combined set of 2129 genes) were obtained in LUAD. We build various PI-based models on these sets of genes, and their results are listed in Table 5.1.

Table 5.1 Statistics detail of the prognostic index-based model for each TCGA dataset.

LUSC	**p-value**	**HR**	**PA (%)**
BPM (n=26)	1.83e-06	2.21	61
GPM (n=24)	2.23 e-06	2.20	59
Combined (n=44)	1.50 e-05	2.04	60
LUAD	**p-value**	**HR**	**PA (%)**
BPM (n=41)	1.08 e-03	1.85	60
GPM (n=34)	3.78 e-03	1.70	58
Combined (n=89)	9.58 e-05	2.10	63

n = number of genes

In order to arrive at a minimal number of gene sets that can segregate the patients in high and low-risk sections in their respective cohorts, the best performing LUSC (BPM) and LUAD

(Combined) sets were searched reiteratively. Initially, we have started the iterative search process using two genes and stopped the incremental iteration process, where the model performed comparably to the previously developed existing model. In this way, we have obtained the best survival models based on 5 genes (“*KIF16B, KLK7, LONRF3, OPLAH, and RIPK3”*) for LUSC [HR=2.10 & p-value = 1.86x10-5] and 4 genes (“*AHSG, DKK1, MGAT5B, and NEMP2”*) for LUAD [HR=2.70 & p-value = 3.31x10-7]. We assessed the functional role of these genes in different biological pathways using the Enrichr (Kuleshov et al., 2016). Five genes in the LUSC-prognostic model played key roles in “*Glutathione metabolism, Cytosolic DNA-damage sensing pathway, tumour necrosis related signaling pathway*” while four genes in the LUAD-survival model have functional roles in “*Mannose type O-glycan biosynthesis, N-Glycan biosynthesis, Wnt signaling pathway*.” The Kaplan-Meier plot of the minimal number of best performing model was also developed and provided in Figure 5.2.

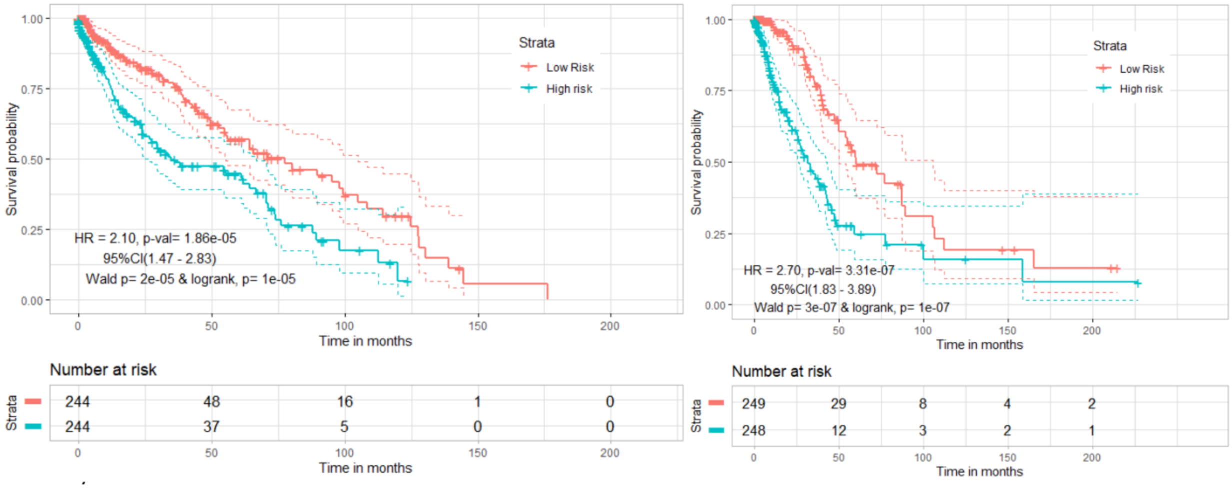


Figure 5.2: Kaplan-Meier plots against best models for NSCLC patients’ risk stratification for TCGA-dataset (a) For LUSC, patients with “PI > median (PI)” are at greater risk to the patients having “PI ≤ median (PI)” with HR = 2.10 (p-value = 1.86x10^{-5}; PA = 61%). (b) For LUAD, patients with “PI > median (PI)” are at greater risk to the patients having “PI ≤ median (PI)” with HR = 2.70 (p-value = 3.31x10-7; PA = 68%). [*Source - Lathwal et.al. JOCR -2020*]

To further evaluate the subtype-specific discriminating potential of the model, the performance of the developed model was also evaluated on the counter cohorts. We observed that the LUSC-specific model did not work well in LUAD and contrariwise. Thus, it can be hypothesized that the developed model will dedicatedly work well on the respective cohorts.

5.3.2 Poor Discrimination of NSCLC Model among Subtypes

We have also examined if a universal PI model can work for the NSCLC dataset. Among the subtypes of NSCLC, we have not found any common gene after applying RF-vh. We retrieved 1381 survival-related genes from the NSCLC full cohort by implementing the univariate cox regression. This set includes 682 BPM and 699 GPM genes, which were used to construct a survival model that can be applied uniformly in all NSCLC patients. We identified 68 overlapping survival-related genes among subtypes and the NSCLC full cohort, which eventually laid the foundation of the universal models.

Table 5.2 Statistics obtained for each dataset for universal NSCLC survival model.

Cohorts	p-value	HR	PA (%)
LUSC	7.46e-04	1.73	58
LUAD	1.28 e-03	1.82	58
NSCLC Cohort	1.87 e-06	1.79	58

All the results of various statistical measures for the universal model are in Table 5.2. From this table, it is clearly evident that a universal model was not able to discriminate well among the subtypes of NSCLC when compared in the form of HR and PA to the subtype-specific model.

5.3.3 Profiling of Survival-related Cancer-specific Pathways Genes

The prognostic potential of 11 cancer-specific pathway genes was evaluated to explore the histological differences in NSCLC subtypes by using the methodology as explained above. We observed that none of the pathway-based models outperformed the previously developed subtype-specific models in HR and PA metrics. Only the model developed using the genes of the apoptotic pathway performed well in both LUSC and LUAD in terms of HR > 2.0 in both the LUSC and LUAD. Thus the obtained result from this analysis further suggests that the NSCLC diagnosed patients should be considered in a subtype-specific manner to provide better therapeutic options.

5.3.4 Clinical Factors in Risk Stratification

The correlation between different clinical features and the survival of NSCLC patients was examined using the Cox Univariate regression model. The input data was transformed to binary using the criteria mentioned in the strata column of Table 5.3.

Table 5.3 Statistics of clinical features-based risk stratification model for NSCLC subtypes.

LUSC (Samples)	**Strata**	**p-value**	**HR**	**PA (%)**
Age (*488*)	<65 vs >=65	1.70e-01	1.28	52
Gender (*488*)	Female vs Male	7.70 e-01	1.06	51
N stage (*482*)	N0 vs N3, N2, N1	1.00 e-01	1.32	51
T stage (*488*)	T2, T1 vs T4, T3	5.04 e-02	1.48	53
Tumour Stage (*484*)	II, I vs IV, III	3.50 e-02	1.48	53
Organ Subdivision (*459*)	Left vs Right	6.14 e-01	1.51	50
LUSC (Samples)	**Strata**	**p-value**	**HR**	**PA (%)**
Age (*497*)	<65 vs >=65	1.38 e-01	1.32	54
Gender (*497*)	Female vs Male	7.29 e-01	0.94	49
N stage (*485*)	N0 vs N3, N2, N1	1.97 e-07	2.62	63
T stage (*494*)	T2, T1 vs T4, T3	2.69 e-04	2.39	57
Tumour Stage (*489*)	II, I vs IV, III	4.84 e-08	2.80	64
Organ Subdivision (*483*)	Left vs Right	6.79 e-01	1.08	51

We analyzed that not even a single clinical feature is of much significance in the LUSC patients except tumour stage. In LUAD cohort, N stage and tumour gives the fairly equal performance as the best survival model**.**

5.3.5 Validation of the Survival Models

In order to assess the robustness of the best concluding prognostic signatures and clinical makers-based survival models, we extracted the GEO dataset having the accession number

GSE42127. The dataset has been converted and normalized based on the target matrix of the TCGA dataset in both LUSC and LUAD. We implemented the Cox Univariate regression algorithm to realize the power of prediction of our gene signatures. All the relevant results of this analysis are in Table 5.4. The ultimate best models in the context of LUSC and LUAD also performed well in our validation GEO cohorts.

Table 5.4 Statistics results of the PI-based models in both TCGA and GEO cohorts.

LUSC (n=5)	**p-value**	**HR**	**PA (%)**
TCGA	1.86e-05	2.10	61
GEO	4.00 e-02	2.53	60
LUAD (n=4)			
TCGA	3.31 e-07	2.70	68
GEO	4.00 e-03	2.50	63

n = number of genes

We have implemented the Kaplan-Meier curves for the pictorial representation of the comparative analysis in the high and low-risk sets among the LUSC and LUAD datasets of the validation samples in Figure 5.3**.** We observed that the identified gene signature acted reasonably well in the validation data also with a good predictive score. We also validated the reliability of our clinical markers-based survival models in the case of NSCLC subtypes. GEO dataset had only three clinical evaluator markers, i.e., gender, age, and tumour Stage. In the validation set, the significant association of not even a single clinical factor was seen with the OS of the patients while implementing the Univariate Cox regression. However, in both TCGA and GEO survival analysis, the two features that were having some risk stratification potential were age and tumour stage only.

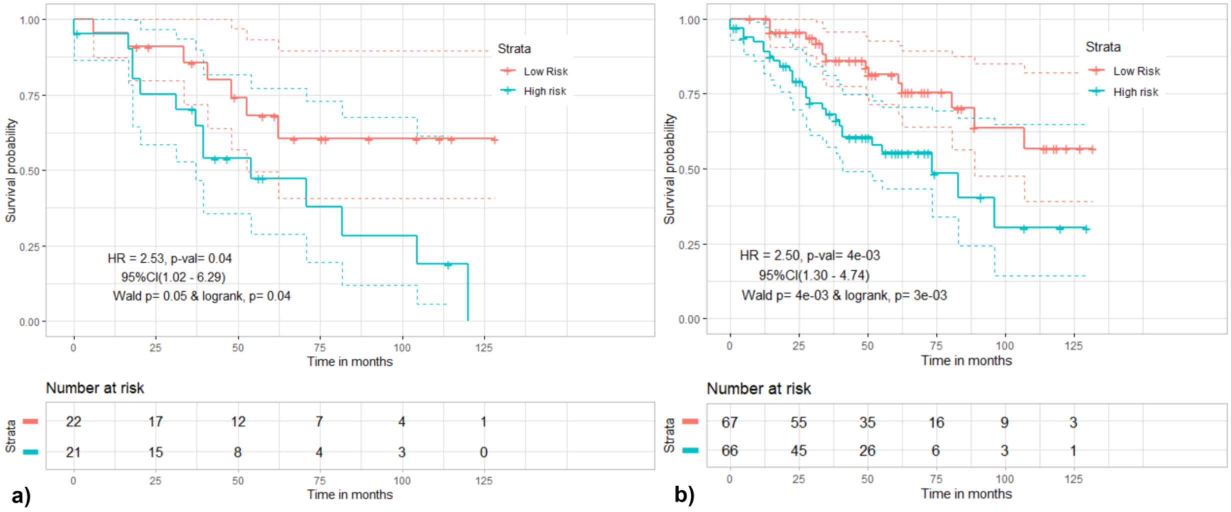


Figure 5.3: Kaplan-Meier curves of the best models employed in prognostication of NSCLC cohorts within GEO dataset (a) In LUSC, patients having "PI > median (PI)" are at higher risk in comparison to the patients having "PI ≤ median (PI)" with HR = 2.53 (p-value = 4.00x10-2; PA = 60%). (b) In LUAD, patients having "PI > median (PI)" are at higher risk to the patients with "PI ≤ median (PI)" with HR = 2.50 (p-value = 4.00x10-3; PA = 63%). [*Source - Lathwal et.al. JOCR -2020]*

5.3.6 Age and Tumour Stage based Survival Model for both Subtypes

We attempted to build a survival model on the combination of various clinical factors to find a simplistic and robust signature for the risk prediction among NSCLC cohorts. Entries according to each clinical element under consideration is categorized as -1, 1, and 0 for low risk, high risk, and not available, respectively.

Table 5.5 Results of the combinatorial models based on clinical factors for NSCLC types.

LUSC (clinical factors- Tumour Stage, Age)	**P-value**	**HR**	**PA (%)**
TCGA	1.06e-02	1.74	53
GEO	6.29 e-03	3.90	60
LUSC (clinical factors- Tumour Stage, Age)			
TCGA	2.85 e-06	2.72	61
GEO	1.53 e-02	2.40	56

The missing data were handled using the classification schema mentioned above, which also ensured the fixed-length vectors for the evaluation purpose. Numerous linear combinations of many factors were evaluated to retrieve the best results. “SUM” was termed coined to represent the linear combinations, which eventually formed the foundation of risk-group classification in the respective cohorts. We classified the samples having “SUM ≤ median (SUM)” as high-risk samples and low-risk samples otherwise. We have also measured the predicting potential of tumour stage and age in the stratification of patient samples in training and testing risk cohorts, the results of which are in Table 5.5. The analysis depicts that the tumour stage and age are the two survival-related factors that are best in predicting the outcome for both LUSC and LUAD patients, irrespective of their different histologies and oncogenic origins.

5.4 Therapeutic Potential of Subtype-specific Prognostic Markers

All the analysis, until now, projected the importance of various prognostic models in predicting the outcome of lung cancer patients. We retrieved the 32 and 271 unique subtype-specific survival-related markers in LUSC and LUAD cohorts. These markers were also validated using an external GEO dataset. A large number of published studies proved the importance of around 9-10% of these identified markers in the risk-stratification of NSCLC patients. One such example is of the “*AHSG* gene,” which is used in the prognostic model, has been altered explicitly in a clinical trial (CA2847188A1) for the treatment of LUAD patients.

This study stratifies the patients among two risk groups based on the expression levels of the various genes in the groups such as “*APH1A* gene” was found to have high expression levels in samples with greater risk as compared to the lower-risk ones. Thus, potent inhibitors that can alter the levels of such genes in high-risk sets with greater power can be prescribed to achieve better outcomes in these patients. Accordingly, in LUSC and LUAD, we identified 32 and 271 genes that were having a greater level of up-regulation in poor outcome patients when compared to ones with good survival. “DGIdb resource (http://www.dgidb.org/search_interactions),” which houses the latest details of various experimentally validated drug-gene interactions, was explored to assess the therapeutic potential of the identified subtype-specific marker genes. We concluded that 6 (LUSC) and 58 (LUAD) survival-related marker genes already had inhibitory drugs/molecules approved against them. Table 5.6 and 5.7 enlists some prognostic genes which have a larger number of the inhibitory drug approved against them for LUSC and LUAD, respectively.

Table 5.6 LUSC-specific prognostic genes (five) targeted by a large number of inhibitory drugs.

Genes	Active drug inhibitors
APH1A	Tarenflurbil; Nirogacestat; Chembl247471; Pinitol; Avagacestat
ELANE	Freselestat; Depelestat; Chembl27885; Moxalactam; Filgrastim; Nicotine; Pegfilgrastim; Nifedipine; Sivelestat; Tiprelestat
HRH1	Mirtazapine; Amoxapine; Zuclopenthixol; Chlophedianol; Flunarizine; Bepotastine; Tesmilifene; Iloperidone; Cariprazine; Mianserin; Esmirtazapine; Antazoline; Chloropyramine; Dimethindene; Isothipendyl; Chlorcyclizine; Butriptyline; Acrivastine; Dexchlorpheniramine Maleate; Bilastine; Tiapride; Ketotifen; Mk-0249; Molindone; Perphenazine; Pimozide; Pipamperone; Tiprolisant; Promethazine; Quetiapine; Sertindole; Thioridazine; Thiothixene; Trifluoperazine; Triprolidine; Ziprasidone; Zotepine; A-349,821; Chembl351231; Cetirizine; Chembl1334217; Galnon; Levocetirizine Dihydrochloride; Ly2624803; Esmirtazapine Maleate; Trimeprazine Tartrate; Epinastine Hydrochloride; Cyclizine Lactate; Fexofenadine Hydrochloride; Cyproheptadine Hydrochloride; Azatadine Maleate; Dexbrompheniramine Maleate; Promethazine Hydrochloride; Methdilazine Hydrochloride; Chlorpheniramine Maleate; Levocabastine Hydrochloride; Carbinoxamine Maleate; Triprolidine Hydrochloride; Tripelennamine Citrate; Hydroxyzine Hydrochloride; Diphenhydramine Hydrochloride; Doxylamine Succinate; Antazoline Phosphate; Bepotastine Besylate
CACNA1C	Cinnarizine; Nitrendipine; Nilvadipine; Drotaverine; Mibefradil; Ibutilide; Celecoxib; Dronedarone; Amlodipine
STAG2	Olaparib; Camptothecin; Vemurafenib; Trametinib; Veliparib

Our analyses highlights that the subtype-specific treatment can be advanced by employing these identified drugs against the established marker genes within each subtype as most of these drugs are either already passed or are in clinical trials for one or the other diseases. It projects another critical aspect of repurposing the already passed drugs for the identified marker genes in treating NSCLC patients.

Table 5.7 LUAD-specific prognostic genes (five) targeted by a large number of inhibitory drugs.

Genes	Active drug inhibitors
AURKB	Tozasertib; Sns-314; Hesperidin; At-9283; Reversine; Cyc-116; Sunitinib Malate; Danusertib; Gsk-1070916; Pf-03814735; Amg-900; Cenisertib; Tak-901; Tandutinib; Sunitinib; 4sc-203; Barasertib; Chembl2062155; Azd-1152-Hqpa; Mk-5108; Midostaurin; Orantinib; Anilinoquinazoline1; Chembl1765740; Hesperadin; Kw-2449; Mln-8054; Chembl605003; Chembl202721; Ilorasertib; Chiauranib; Ttp-607; Enmd-2076; Bi-811283; Mk-6592
AURKA	Chembl369507; At-9283; Cyc-116; Alisertib; Chembl223147; Chembl472193; Sns-314; Anilinopyrimidine1; Chembl521105; Chembl495758; Chembl522891; Mk-5108; Enmd-2076; Danusertib; Pf-03814735; Tozasertib; Amg-900; Orantinib; Cenisertib; Xl-228; Alisertib Sodium; Mln-8054; Tak-901; Fluorouracil; Tamoxifen; Paclitaxel; Phenylthiourea; Anilinoquinazoline1; Barasertib; Gsk-1070916; Chembl1765740; Chembl359482; Kw-2449; Pf-562271; Chembl202721; Enmd-981693; Ilorasertib; Tas-119; Ttp-607; Mk-6592; Rg-1530
TYMS	Raltitrexed; Floxuridine; Trifluridine; Trimethoprim; Gemcitabine; Fluorouracil; Capecitabine; Thymidine Monophosphate; Chembl389051; Deoxyuridine Monophosphate; Folitixorin Calcium; Chembl169896; Pralatrexate; Chembl22148; Phentolamine; Reserpine; Tamoxifen; Topotecan; Sodium Beta-Nicotinamide Adenine Dinucleotide Phosphate; Verapamil; Pemetrexed (Chembl1201258); Pemetrexed Disodium; Cytarabine; Prednisone; Etoposide; Leucovorin; Folic Acid; Tegafur; Oxaliplatin; Asparaginase; Sulfasalazine; Vincristine; Anthracene-9-Carboxylic Acid; Daunorubicin; Dexamethasone; Irinotecan; Methotrexate
TOP2A	Teniposide; Amsacrine; Valrubicin; Epirubicin; Enoxacin; Pefloxacin; Trovafloxacin; Lomefloxacin; Norfloxacin; Levofloxacin; Ofloxacin; Podofilox; Mitoxantrone; Sparfloxacin; Doxorubicin; Etoposide; Genistein; Fleroxacin; Lucanthone; Banoxantrone; Amonafide; Elsamitrucin; Berubicin Hydrochloride; Amrubicin; Finafloxacin; Daunorubicin; Dexrazoxane; Hydroquinone; Vincristine; Vosaroxin; Amrubicin Hydrochloride; Aldoxorubicin; Becatecarin; Mm-302; Etoposide Phosphate; Daunorubicin Citrate; Daunorubicin Hydrochloride; Doxorubicin Hydrochloride; Mitoxantrone Dihydrochloride; Idarubicin; Paclitaxel; C-1311 (Chembl3545337); Dactinomycin; Camptothecin; Idarubicin Hydrochloride; Idronoxil
SLC2A1	Estradiol; Diazepam; Genistein; Gentamicin; Quazepam; Phenytoin; Pioglitazone; Progesterone; Tretinoin; Rosiglitazone; Thymidine; Triamcinolone; Glufosfamide

One such example is the *AURKB* gene-targeting drug - “Paclitaxel drug,” which is a well-accepted therapeutic for treating lung cancer. Another example is of “Alisertib drug” (also targeting the *AURKB* gene), which, when given in combination with “Oimertinib” can treat EGFR mutant-based lung cancer and is already in a clinical trial (NCT04085315).

5.5 Conclusion and Summary

In this work, we have attempted to systematically profile the risk-predicting potential of various genes, pathways, and clinical factors specific to LUSC and LUAD. In the case of LUSC, we have shown that p53 and apoptotic pathway genes were having greater performance in risk estimation of the patients, whereas, for LUAD, the PI3K-AKT, WNT, and apoptotic pathway-related genes performed well. One analysis showed that age and tumour stage works best for both subtypes; even there are differences in their origins and histologies. Different gene sets and pathways that were coming out are important survival-predictors in LUSC, and LUAD highlights the importance of more subtype-specific therapeutic regimes for NCLC subtypes. We had evaluated the comparative performance of our established survival models with the existing ones and found that our models out-performed them when looked in terms of HR and PA. We have only considered the expression data and the clinical features here, which is the one significant limitation of the work. We hope that shortly, we will implement a more integrated, comprehensive, and robust model by incorporating the information of the copy number change, methylation, mutation data, and miRNA data in our study.

The major highlight of our work is that it can not only explain the underlying basics of heterogeneity in the NSCLC subtypes but also helps in devising more subtype-specific therapeutics. The established prognostic marker genes may possess tremendous potential in designing therapeutic approaches based on inhibitors, agonists, and antagonists. Also, established marker genes have been investigated poorly in the lab settings and thus demand the attention of researchers and clinicians to suggest better reliable predictive, prognostic as well as therapeutic regimens for treating NSCLC patients in a more subtype-specific manner.

6. Computer-aided Prediction of Interleukin-2 Inducing Peptides

6.1 Background

Cancer cells are often heterogeneous in nature. Often cancer cells show complex behaviour in terms of interaction with normal cells and immune cells. The conventionally used cancer therapeutic approaches such as chemotherapy or radiation therapy solely kill fast proliferating cells inside the body and thus have several limitations. The major limitation faced by conventional therapies is that they also kill normal fast proliferating cells such as blood cells, sperm cells, etc. (Pucci et al., 2019), thus affects the overall health of the patients during treatment. Due to such limitation of conventional therapies, protecting the normal cells during treatment procedure becomes an important area of research for providing better healthcare.

In this regard, tumour immunotherapy has great potential for providing better healthcare to the patients. Cancer immunotherapy aims to augment the body's immune system and its response towards the cancer cells. There are several options available for cancer immunotherapy, which consists of the use of checkpoint inhibitors, monoclonal antibodies, adoptive T-cell therapy, vaccination, OV based therapies, etc. (Lathwal et al., 2020). One another strategy that is also approved by the FDA is the use of interleukins for the treatment of cancer. Currently, two interleukin based drugs are approved by the FDA for the treatment of human malignancies. These include the interferon-alpha (IFN-alpha) based drug “Roferon-A” and interleukin-2 (IL-2) based drug “Aldesleukin” for the treatment of advanced melanoma.

The IL-2 is an immune activator cytokine that belongs to the Gamma chain family of immune mediators. The IL-2 is the major cytokine that regulates the proliferation, differentiation, and survival of T-cell and Natural killer (NK) cells (Malek, 2008). The IL-2 receptors such as alpha are present on the cell surface of regulatory T-cell (Treg), CD4+ and CD8+ T-cells, and dendritic cells (DCs) (Malek, 2008), (Rudensky, 2011). The beta receptors of IL-2 are mainly expressed on cell surfaces of monocytes, lymphocytes, NK cells (J. P. Siegel, Sharon, Smith, & Leonard, 1987). The gamma receptors of IL-2 are mainly expressed by the hematopoietic cells. The binding of IL-2 on its cell surface receptors activates the downstream pathways such as JAK-STAT, PI3K-AKT, and MAP kinase, which further results in the activation of the immune cells. The complete flow diagram of IL-2 based immune activation is depicted in Figure 6.1.

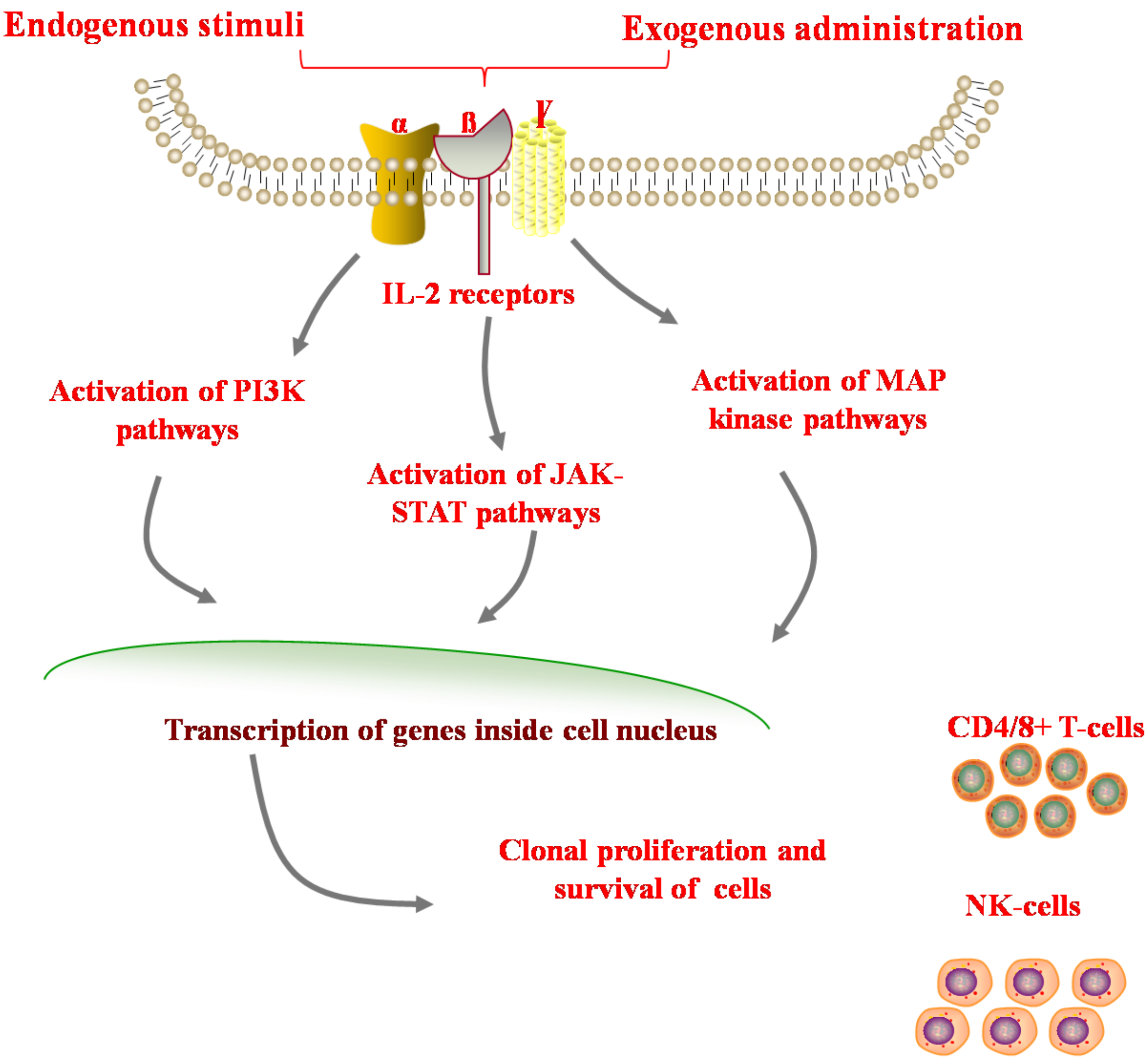

Endogenous stimuli
Exogenous administration
α
β
γ
IL-2 receptors
Activation of PI3K pathways
Activation of JAK-STAT pathways
Activation of MAP kinase pathways
Transcription of genes inside cell nucleus
CD4/8+ T-cells
Clonal proliferation and survival of cells
NK-cells

Figure 6.1: The mechanism of immune activation by interleukin-2.

IL-2 in clinical settings has been used as an immune-stimulating agent that can counteract the immune suppression mediated by cancer cells. The well-elaborated mechanism of action of IL-2 and its potent anti-tumour immune properties led to the development of several IL-2 based treatment regimens that are under clinical trials for advancing the cancer therapeutics. The IL-2 has been utilized in clinical settings as monotherapy agents as well as in combination with drugs and antibodies to enhance the immune activation and life expectancy of the patients. Table 6.1 enlists major IL-2 based cytokines therapies that are currently being evaluated for their cancer therapeutic potentials.

Table 6.1: Interleukin-2 based cancer immunotherapy under evaluation in clinical trials.

Cancer type	Cytokine based treatment regimen
Renal cancer	IL-2 as a monoadjuvant
Metastatic renal cell carcinoma	IL-2 in combination with IFN-alpha 2b
Melanoma	IL-2 in combination with IFN-alpha2b and chemotherapy
Melanoma stage IV	IL-2 in conjugation with Ipilimumab
Melanoma stage III	IL-2 in combination with gp100 vaccine
Colorectal carcinoma; Melanoma stage IIIB/C/IVM1a; Renal cancer	IL-2 in combination with Daromum and Darleukin

Despite the high therapeutic index value of IL-2 in cancer treatment, it also suffers from the limitation of short half-life in blood serum. Exogenous administration of IL-2 is rapidly cleared from the circulation by the kidneys, thus pose a limit in its success as a therapeutic agent. Often the problem of short half-life is counteracted by the external administration of a high amount of IL-2. But, this high dose of IL-2 suffers from toxicities related problems. The high dose of IL-2 also results in the appearance of vesicular leakage syndrome, edema in lung airways and hypotension, etc. (Jiang et al., 2016). In order to overcome the constraint of IL-2 based therapy, several approaches are used in literature. In one study, researchers use the oncolytic adenovirus as a vector to express the IL-2 gene endogenously inside the cancer cells (Chaurasiya et al., 2016). In another study, oncolytic adenovirus is used to express the IL-2 and Tumour necrosis factor-alpha (*TNF-α*) gene; the results of the study showed 100% treatment of experimental mice models with no side effect (Tähtinen et al., 2016). Another drawback associated with IL-2 based therapy is that it can also activate the expression of Treg cells, which brings about the immune suppression instead of immune activation. To tackle this limitation, IL-2 mutants have been generated in literature studies. These IL-2 mutants, popularly known as “superkines” bind differently to the IL-2 receptors to efficiently stimulate the production of NK cells, CD8+, and CD4+ T-cells but not Treg cells (Jiang et al., 2016). It was also observed that IL-2 mutants even free from the toxicities related to the high dose administration.

Overall, the aforementioned points suggest that IL-2 mutants possess a high therapeutic index as compared to wild-type. Thus identification of IL-2 inducing mutant peptides is necessary for overcoming the translational challenges associated with the current conventional therapy. The computational methods can help the scientist and clinicians in this regard as they can reduce the cost and time for generating and identifying the properties of peptides having interleukin inducing potential. With the best of the author's knowledge, there are no such computational methods available in the literature that can predict the IL-2 inducing potential of peptides. Against this backdrop and to advances the IL-2 based cancer treatment strategy, the present chapter focuses on the development of a prediction algorithm for the identification of peptides having IL-2 inducing potential. The best predictive model, along with various analysis tools, was implemented in the form of a web-server that is freely available to the scientific community at https://webs.iiitd.edu.in/raghava/il2pred/.

6.2 Materials and Methods

6.2.1 Data Extraction and Pre-processing

We extracted a dataset containing experimentally validated IL-2 inducing and non-IL-2 inducing epitopes from the IEDB database (Vita et al., 2015), which is the most authentic and huge repository of different immune epitopes. We initially specifically extracted a total of 5427 peptides having MHC class-II binding ability to trigger IL-2 secretion. These epitopes were termed as IL-2 inducing peptides and grouped under a positive dataset. Similarly, we extracted a total of 3568 experimentally validated epitopes as a negative dataset that did not trigger IL-2 secretion and named them as non-IL-2 inducers. We then screened the positive and negative dataset for the peptides of length 8-25; these are the most suitable length peptides to be used by MHC molecule for antigen presentation as suggested in various literature studies. This way, we landed up 5205 and 3501 IL-2 inducing and non-inducing peptides, respectively. We then removed the redundant peptides to get the final datasets containing 2528 IL-2 inducing epitopes and 2104 non-IL-2 inducing peptides.

6.2.2 Length and Positional Conservation Analysis of Peptides

We analyzed the distribution of length among the positive and negative dataset peptides using in-house R script. We have also computed the average composition of different amino acids present in the positive and negative datasets. To gain insights about the position-specific preference of each residue, we implemented a two-sample logo using the R package "ggseqlogo" (Wagih, 2017). This package processes only fixed-length peptide sequences as input, and since the smallest length peptide is of eight residue long, we considered the combination of 8-residue length sequences from both N and C terminals for plotting the two-sample logo.

6.2.3 Motif Analysis

Proper identification of the motifs within peptides that are having some functionality is very crucial in the annotating functions of these peptides/proteins. In our study, we have used a tool named MERCI (Vens, Rosso, & Danchin, 2011) for locating the motifs that were exclusive to the IL-2 inducing and non-IL-2 inducing peptides. MERCI software uses both negative and positive datasets at the same time, but results are provided for the positive set only. So, we implemented the MERCI software in two steps; firstly, we extracted the motifs for a positive dataset by providing the IL-2 inducing peptides as a positive set and non-IL-2 inducing peptides as negative set. In the second iteration, we retrieved the motifs for the negative dataset by inputting non-IL-2 inducing peptides as a positive set and IL-2 inducing peptides as a negative set. Using the above-mentioned approach, we calculated the motifs in both positive and negative datasets. We have used two different classification techniques for motif searching using MERCI: a) None, and b) Koolman-Rohm. All these methods gave us different motifs in the positive and negative datasets. We have screened peptides containing unique motifs from both the sets so as to have an idea of overall coverage of the various motifs in the complete data.

6.2.4 Feature Estimation and Selection

We have calculated the different properties from the IL-2 inducing and non-IL-2 inducing peptides. These properties formed the basis of our machine learning-based classification models. The features were computed via sophisticated software Pfeature, which can estimate thousands of descriptors – "i) Composition based descriptors - Amino Acid Composition (AAC), Dipeptide

Composition (DPC), Tripeptide Composition (TPC), Autocorrelation, Conjoint Triad Descriptors (CTDs), and Composition enhanced Transition and Distribution (CeTD); ii) Binary profiles – Amino acids, Atom and Bonds, and Residue properties; iii) Structural descriptors – SMILES, Surface Accessibility and Secondary; iv) Pattern-based descriptors – Binary profiles, PSSM profiles, and Universal properties from the peptides." We computed nearly 10000 features using Pfeature, which were further, screened using random forest feature selection function available in the R caret package. All the features mentioned above were tested independently as well as in combination with other features after feature selection step, for their classification capabilities among IL-2 inducing and non-inducing peptides.

6.2.5 Machine Learning based Classification Models

We have implemented different machine learning algorithms using "caret – Classification and Regression Training" package of R. Classification algorithms used to build different models are "decision trees (DT), random forest (RF), multi-layer perceptron (MLP), eXtreme gradient boosting (XGBoost), K-nearest neighbours (KNN), support random vector wit[h racial basis (SVR), neural network (NN), Ridge, Lasso, and Elastic Net". Different parameters were optimized using "expand Grid" functionality of the "caret" package of R. Eighty-percent of the overall data has been used in training, whereas as remaining twenty percent is used in the testing of the models. The different classifiers were trained and evaluated using a ten-fold cross-validation method, which is a well-accepted technique for the optimization of parameters and performance of the models. In ten-fold cross-validation, the whole training dataset was divided into ten equal parts, where nine parts were used in training and one in the validation, which optimizes the parameters of the models. This process was repeated ten times to ensure that every set has been used in the training and validation. All these classifiers were implemented using an in-house R script. We have explained the overall workflow of the study in Figure 6.2.

6.2.6 Evaluation Metrics

We have assessed the performance of our models using various metrics such as true positives (TP), true negatives (TN), false positives (FP), true negatives (FN), specificity, sensitivity, accuracy, Matthews correlation coefficient (MCC), and area under the curve (AUC). TPs are the number of peptides that were originally IL-2 inducing and are correctly classified as IL-2 inducing peptides by the classifiers.

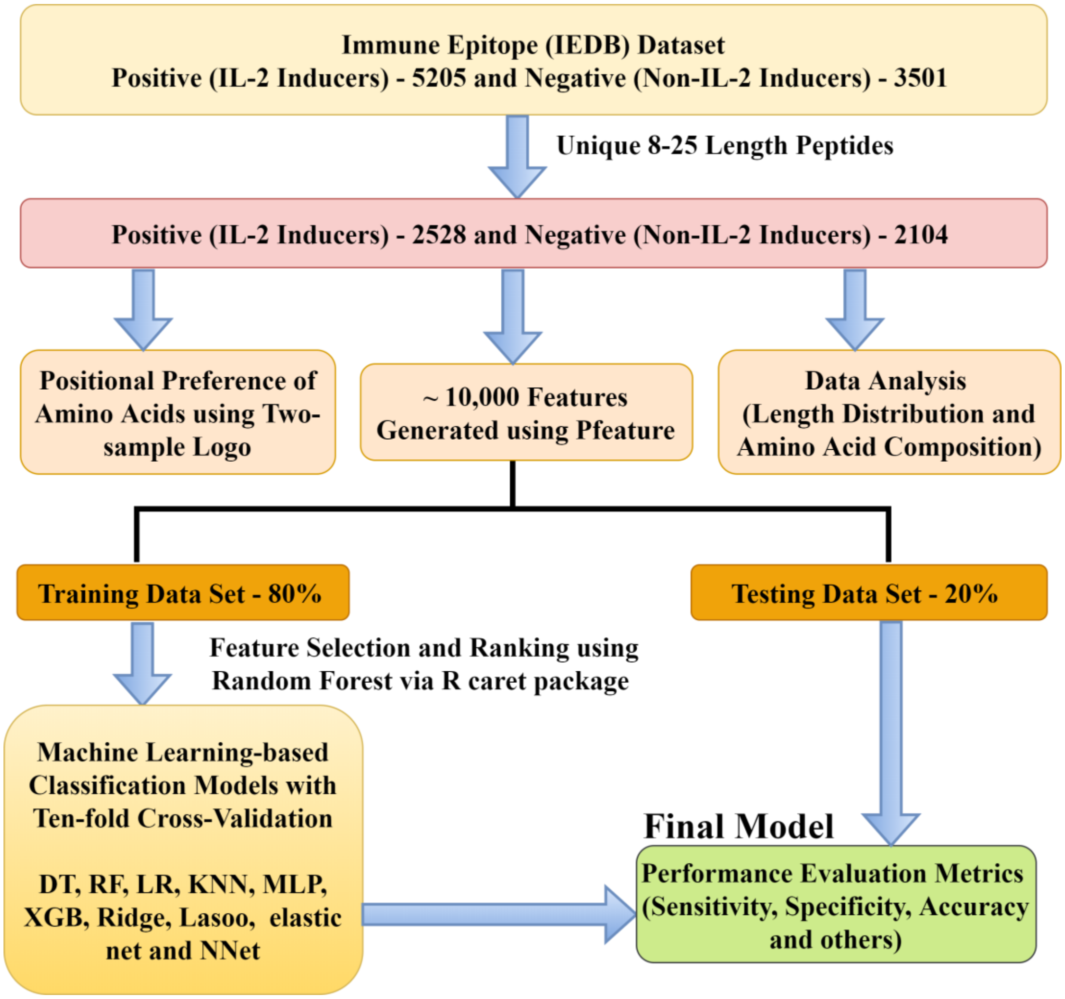

Immune Epitope (IEDB) Dataset
Positive (IL-2 Inducers) - 5205 and Negative (Non-IL-2 Inducers) - 3501
Unique 8-25 Length Peptides
Positive (IL-2 Inducers) - 2528 and Negative (Non-IL-2 Inducers) - 2104
Positional Preference of Amino Acids using Two-sample Logo
~ 10,000 Features Generated using Pfeature
Data Analysis (Length Distribution and Amino Acid Composition)
Training Data Set - 80%
Testing Data Set - 20%
Feature Selection and Ranking using Random Forest via R caret package
Machine Learning-based Classification Models with Ten-fold Cross-Validation
DT, RF, LR, KNN, MLP, XGB, Ridge, Lasoo, elastic net and NNet
Final Model
Performance Evaluation Metrics (Sensitivity, Specificity, Accuracy and others)

Figure 6.2: Overall pipeline employed in building IL-2 inducers prediction model.

TNs are those IL-2 inducing peptides that were incorrectly grouped under non-IL-2 inducers by the models. FPs is the non-IL-2 inducing peptides that were rightly identified as non-IL-2 inducers by the classification models. FNs are those non-IL-2 inducers that were incorrectly categorized as IL-2 inducers by the classifiers. We have explained other parameters below:

$$\text{Sensitivity} = \frac{TP}{(TP+FN)}$$

$$\text{Specificity} = \frac{TN}{(TN+FP)}$$

$$\text{Accuracy} = \frac{TP+TN}{(TP+TN+FP+FN)}$$

$$\text{MCC} = \frac{TP\times TN-FP\times FN}{\sqrt{(TN+FP)(TP+FN)(TN+FP)(TN+FN)}}$$

6.3 Results

In our study, we have utilized 2528 IL-2 inducing epitopes as a positive dataset and 2104 non-IL-2 inducing peptides as a negative set while making classification models. We have performed all the analysis and model building on these final datasets. On the basis of the overall analysis, different prediction models were built, trained, and tested for their performance in predicting IL-2 inducing peptides.

6.3.1 Data Analysis

We have analyzed the final IL-2 inducing and non-IL-2 inducing peptides to interpret their key features. It was well documented in the literature that peptide length plays a key role in discriminating the peptides. Data analysis (Figure 6.3) revealed that the majority of negative peptides comprises of 15, 17, and 20 length amino-acid sequences, whereas positive peptides are more widely distributed in length ranging from 12-20 (15 lengths being the most abundant).

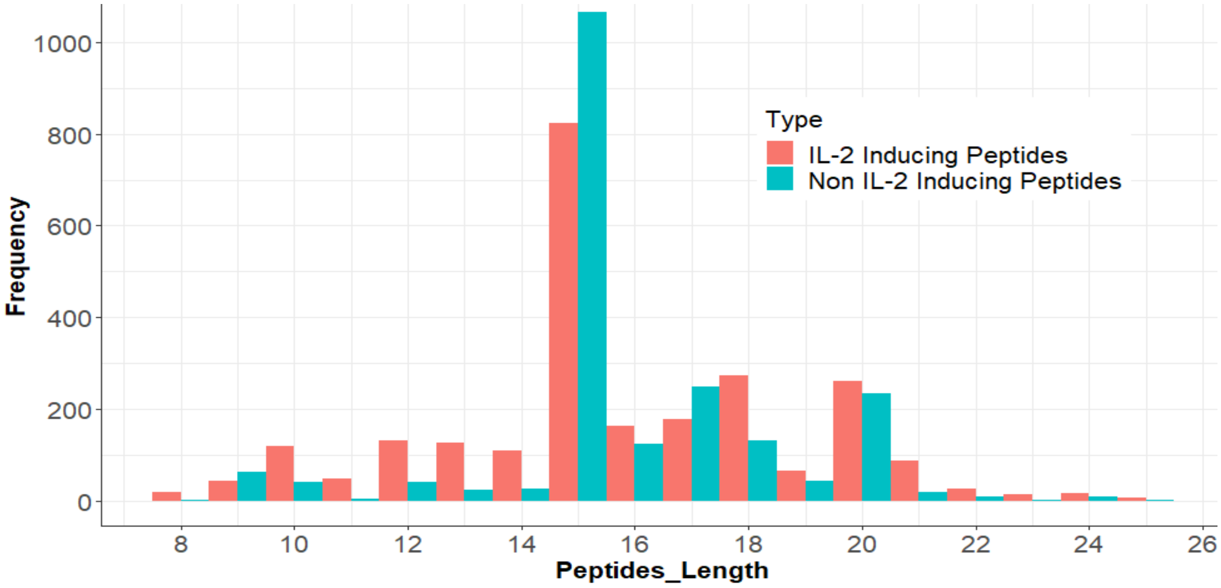
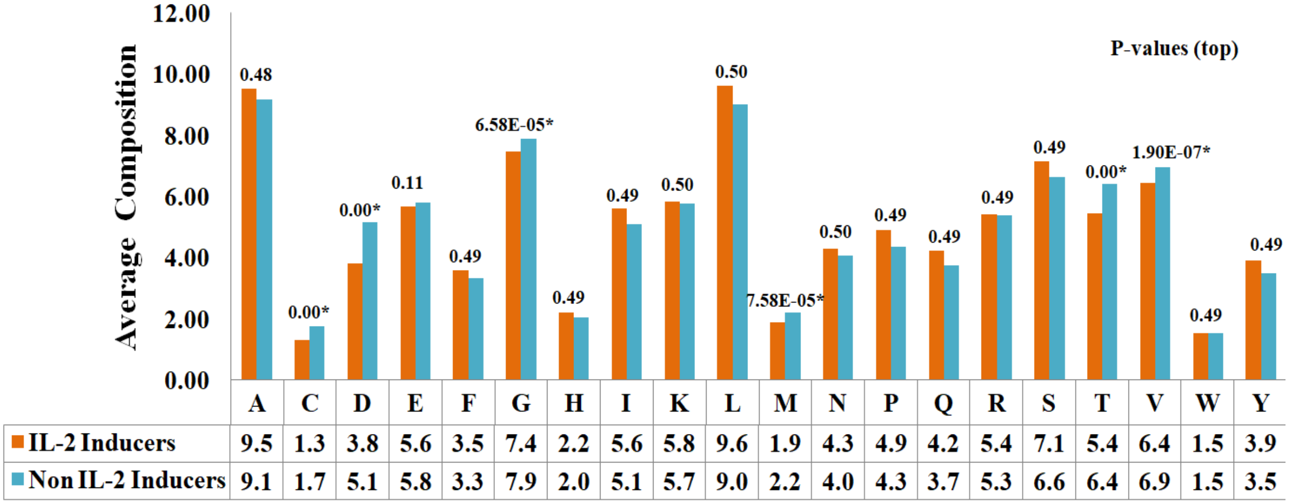


Figure 6.3: Graphical representation of the length of peptides and their frequency in the dataset.

The x-axis of the figure represents the different length of the peptides that are present in the complete dataset. Y-axis depicts the frequency of occurrence of a particular length peptide in positive and negative datasets. IL-2 inducing and non-IL-2 inducing peptides were analyzed for their average single amino acid composition, as shown in Figure 6.4. The X-axis of the figure stores the information of the different single amino acid residues in each positive and negative dataset, whereas bars adjacent to the Y-axis are the graphical representation of the values. In IL-2 inducers, amino acids such as A, L, S, and Y are high as compared to non-IL-2 inducers, whereas, residues like D and M have a higher average composition in the non-IL-2 inducing peptides than the IL-2 inducers.



	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y
🟧 IL-2 Inducers	9.5	1.3	3.8	5.6	3.5	7.4	2.2	5.6	5.8	9.6	1.9	4.3	4.9	4.2	5.4	7.1	5.4	6.4	1.5	3.9
🟦 Non IL-2 Inducers	9.1	1.7	5.1	5.8	3.3	7.9	2.0	5.1	5.7	9.0	2.2	4.0	4.3	3.7	5.3	6.6	6.4	6.9	1.5	3.5

Figure 6.4: Visual representation of the average single amino acid composition of the peptides in the positive and negative datasets.

6.3.2 Positional Preference Analysis

In our study, we have generated the two-sample logo (TSL) for the positive and negative epitopes to investigate the positional preference of specific residues. We have used an in-house script using “ggseqlogo” package of R to generate the TSL for the positive and negative datasets. In TSL, the overall height of any amino acid is a symbolic representation of its relative abundance in the dataset. Since TSL requires fixed-length peptides as input, we have used minimum length epitope, i.e., eight-length from C and N-terminal of the peptides and joined them to make a total of 16 length sequence to be given as input to the program. It can be seen in Figure 6.5, G and S are highly preferential at 16th and 11th position in the IL-2 inducing peptides, whereas no such trend is observed in the non-IL-2 inducers. Similarly, A and L are most prominent in the non-IL-2 inducers. Overall, polar residues are more prominent in the positive datasets, and hydrophobic is the abundant ones in the negative dataset.

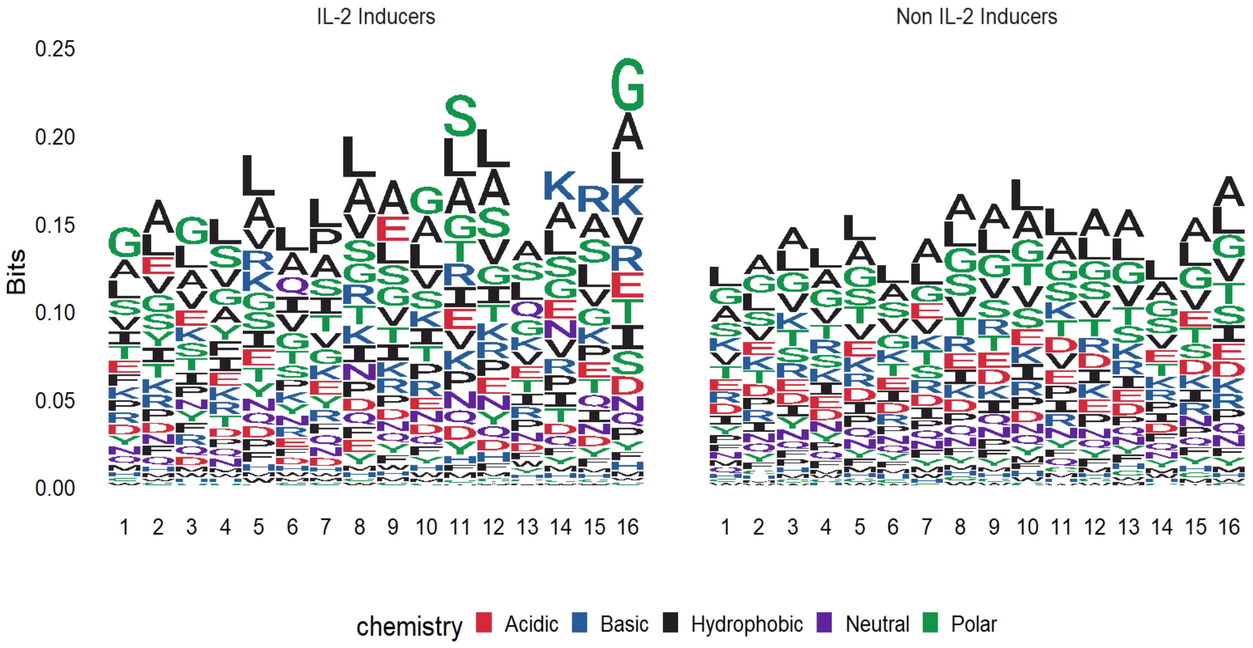


Figure 6.5: Two-sample logo of the IL-2 and non-IL-2 inducing peptides reflecting the relative abundance of single residues at different positions.

6.3.2 Motif Analysis

In our study, we have used MERCI software to search the motifs present exclusively in the IL-2 inducing peptides but not in non-IL-2 inducing peptides. In a similar manner, we have computed the motifs that are exclusive to non-IL-2 inducing peptides. It is observed for the motif analysis that S and G are more prominent residues in the positive dataset, which was also found from the two-sample logo analysis.

Table 6.2: Motifs exclusive to interleukin-2 inducing and non-inducing peptides along with the frequency of their occurrence in the peptides searched using MERCI software.

MERCI MOTIF NONE			
IL-2 Inducers - Motifs	**Freq**	**Non-IL-2 Inducers - Motifs**	**Freq**
L-E-G-S	99	D-C-L	19
A-L-E-G	98	K-D-C	19
A-L-E-G-S	97	C-L-F	18
E-G-S-L-Q	96	E-A-Y-F	18
L-A-L-E-G	96	F-K-D-C	18
S-L-Q-K	96	L-E-A-Y	18
-		Y-F-K-D	18

6.3.3 Machine Learning based Classification Models

In our study, we have developed different classifiers using "caret" package of R. The most widely used classifiers such as DT, RF, KNN, MLP, Ridge, Lasso, Elastic net have been implemented using the in-house R script. As explained in the above sections, we have first generated nearly 10000 features using the Pfeature software. Feature selection and ranking have been implemented on these features using SVC-LI functionality in Pfeature and also through random forest functionality available in "caret" package of R. Apart from these selected features, classification models were built and tested separately on independent features that were widely followed in the several past studies (Dhanda, Vir, & Raghava, 2013; Nagpal et al., 2017). After the feature selection process, we came up with the top 10 and 100 features based on their importance in identifying the subgroups. All the statistical details of the selected features and the corresponding best classifiers have been provided in Table 6.3. As evident from Table 6.3, it can be seen that overall, DPC performed well in classifying the IL-2 inducing peptides versus non-IL-2 inducing peptides with overall accuracy, MCC and AUC of 71.29%, 0.42, and 0.71, respectively. AAC based classification model performed the second-best with an overall accuracy of 69.00 %, MCC of 0.39, and AUC of 0.69. These results reflect the importance of single and dipeptide based composition in classifying the IL-2 and non-IL-2 inducers.

Table 6.3: Performance measures of best classifiers using the different features of the peptides.

Features [Classifier]	Sensitivity (%)	Sensitivity (%)	Accuracy (%)	MCC	AUC
Training					
AAC [RF]	74.70	64.13	69.90	0.39	0.69
DPC [RF]	**80.07**	**60.72**	**71.29**	**0.42**	**0.71**
PCP [RF]	70.05	57.42	64.32	0.27	0.64
Adv PCP [RF]	69.96	44.66	58.46	0.15	0.58
Struct PCP [XGB]	66.11	56.23	61.62	0.22	0.61
RRI [RF]	76.43	51.60	65.16	0.29	0.65
CeTD [XGB]	67.34	56.41	62.38	0.24	0.62
DOOR [RF]	70.15	54.81	63.18	0.25	0.63
Pfeature_top10 [RF]	67.63	54.04	61.46	0.22	0.61
Pfeature_top100 [RF]	77.56	57.48	68.45	0.35	0.68
Testing					
AAC [RF]	72.72	63.33	68.47	0.36	0.68
DPC [RF]	**78.81**	**60.95**	**70.70**	**0.40**	**0.70**
PCP [RF]	70.94	55.23	63.82	0.26	0.63
Adv PCP [RF]	72.52	40.95	58.21	0.14	0.57
Struct PCP [XGB]	59.28	56.19	57.88	0.15	0.57
RRI [RF]	74.30	51.90	64.14	0.26	0.63
CeTD [XGB]	70.94	52.85	62.74	0.24	0.62
DOOR [RF]	68.37	57.14	63.28	0.25	0.63
Pfeature_top10 [RF]	69.16	53.09	61.87	0.22	0.61
Pfeature_top100 [RF]	79.44	52.62	67.28	0.33	0.67

**Bold faced are the best models* — *Bold faced are the best models*

Hybrid Model – It has been observed from the results of the previous sections and was also evident from the feature selection process that length plays a crucial role in discriminating the IL-2 and non-IL-2 inducing peptides. Thus, we have developed various hybrid models that are based on the dipeptide composition (best model, refer Table 6.3) and length of the peptides in the dataset. The RF-based hybrid model achieved a maximum accuracy of 73.25, with an MCC of 0.46 and AUC of 0.73. The statistical details of the different models using various classifiers have been presented in Table 6.4, which clearly reflects the importance of length in combination with dipeptide composition while discriminating IL-2 and non-IL-2 inducers.

Table 6.4: Performance measures of the hybrid model (DPC+Length) on different classifiers.

Classifiers	**Sensitivity (%)**	**Sensitivity (%)**	**Accuracy (%)**	**MCC**	**AUC**
			Training		
DT	55.28	66.80	60.52	0.22	0.61
RF	**74.60**	**71.61**	**73.25**	**0.46**	**0.73**
KNN	72.43	64.96	69.04	0.37	0.69
MLP	64.77	60.27	62.79	0.25	0.62
SVR	72.97	59.02	66.63	0.32	0.66
XGB	73.12	64.20	69.07	0.37	0.68
LASSO	67.69	53.92	61.43	0.22	0.61
			Testing		
DT	58.30	68.09	62.74	0.26	0.63
RF	**74.11**	**71.42**	**72.89**	**0.45**	**0.72**
KNN	73.91	64.76	69.76	0.39	0.69
MLP	62.84	65.47	64.03	0.28	0.64
SVR	73.91	58.57	66.95	0.33	0.67
XGB	73.12	66.67	70.19	0.39	0.69
LASSO	66.99	52.61	60.47	0.20	0.60

**Bold faced are the best models*

Also, we have assessed the robustness of our models using ten folds cross-validation and testing the performance of the models on a dataset that have not been used in the training model. All the results related to the training and testing of the models have been listed in Table 6.3 and 6.4.

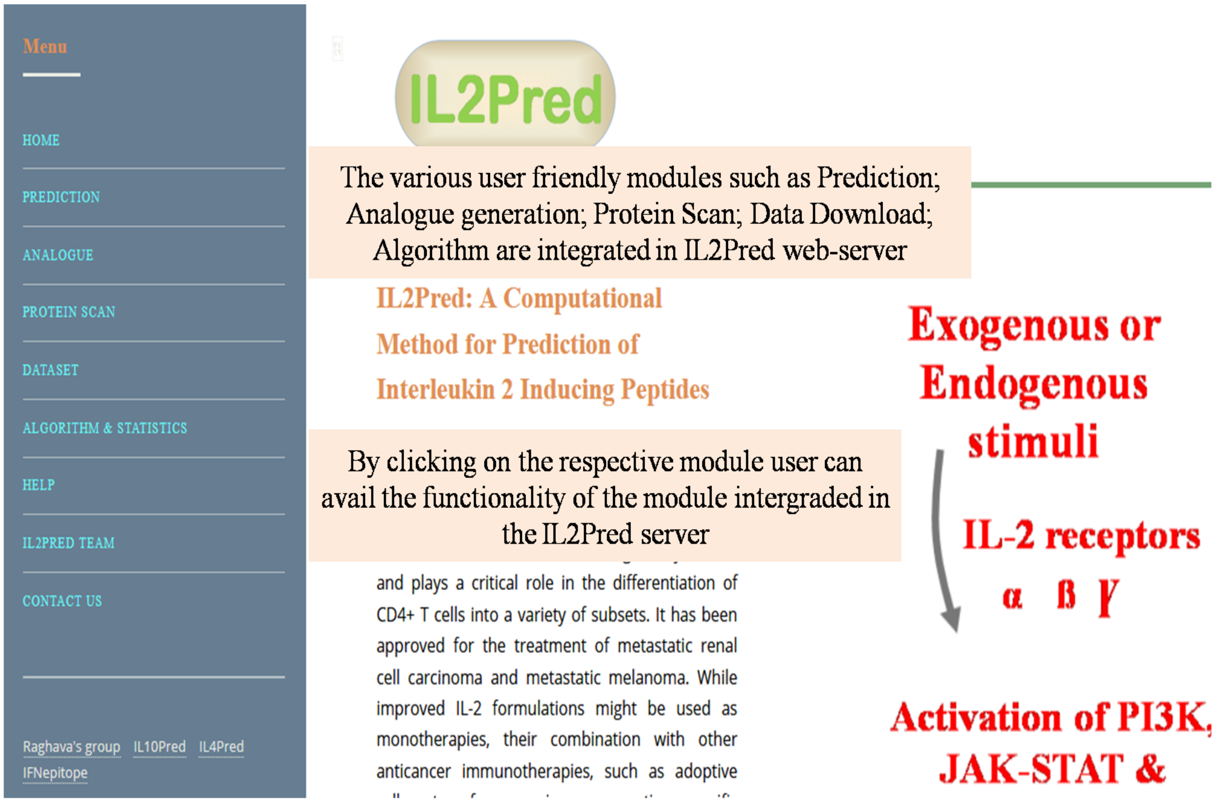
6.4 Implementation of Web-resource

The best performing machine learning-based model, along with various analysis tools, has been designed in the shape of a web-resource. The front end of the web-resource is developed using a responsive template, which is based on HTML version 5.0. The advantage of using a responsive template is that it provides access to data and its integrated facilities in a user-friendly way by sensing the user's device and thus adjusting the screen ratio accordingly. The back end data related computation facilities have been implemented with the help of UNIX commands, R, Java, Perl, and R/Python codes.

6.4.1 Web Server Modules

For the sake of easy data analysis, the IL2Pred web-server is implemented with various tools. The primary integrated tools include the I) Prediction module, II) Analogue generation module, III) Protein scan module. Using Prediction module facilities, users of the IL2Pred server can predict the IL-2 inducing potential of the user-defined query protein. Keeping in mind the importance of mutants of IL-2 in cancer treatment regimens, the analog generation module is integrated into the IL2Pred web-server. Using this module, the user can generate all possible mutants of query protein at each residue. In addition to this, the user can also get the information regarding the IL-2 inducing potential of generated mutants. Thus, the analogue generation module is beneficial and desirable for the experimental scientists who wish to design and identify mutants with IL-2 inducing potential.

Moreover, the integrated protein scan module also be very applicable for clinicians and experimental scientist who wishes to identify the regions of a protein that are enriched in IL-2 inducing peptide. In this way, it helps in identifying the immunodominant region in a query protein. In addition to the modules mentioned above, the IL2Pred server is also equipped with a data download module, algorithm, and statistics module. Figure 6.6 provides the key details of the integrated modules of the IL2Pred web-server.



The various user friendly modules such as Prediction; Analogue generation; Protein Scan; Data Download; Algorithm are integrated in IL2Pred web-server

IL2Pred: A Computational Method for Prediction of Interleukin 2 Inducing Peptides

By clicking on the respective module user can avail the functionality of the module intergraded in the IL2Pred server

and plays a critical role in the differentiation of CD4+ T cells into a variety of subsets. It has been approved for the treatment of metastatic renal cell carcinoma and metastatic melanoma. While improved IL-2 formulations might be used as monotherapies, their combination with other anticancer immunotherapies, such as adoptive

Figure 6.6: Description of user-friendly modules of IL2Pred server. (URL: https://webs.iiitd.edu.in/raghava/il2pred/index.html)

6.4.2 Prediction Functionality

The identification and designing of potent IL-2 inducing peptide is at the forefront of interleukin based cancer treatment strategy. However, identifying the peptide with IL-2 inducing potential via experimental setup is time-consuming and cumbersome. Thus, to advance the identification interleukin based cancer treatment approach and to cater to the need of experimental scientists in identifying IL-2 inducing potential of a given sequence, we have specifically integrated functionality in the IL2Pred server that we named as "Prediction module." The prediction is powered by the machine learning-based model, which is integrated at the backend of the web-server. The user has to submit the query protein sequence as defined in the web-server section, and just by clicking, the user will get the information of IL-2 inducing potential of the provided sequence within a matter of seconds. Users are also provided with the facility to download the results from the web-server for further analysis.

6.5 Discussion and Conclusion

Systemic response via immunosuppression is more desirable in some cases like allergic and asthmatic conditions, whereas, in other cases, like cancer, it may cause serious problems. Immunotherapies based on peptides are gaining much importance in past decades. It has been proven that the immune system can be activated against neoplastic cells. The first cytokine based therapy that was used for treating cancer patients was based on the IL-2 (Choudhry et al., 2018). Patients treated with IL-2 have shown better outcomes for diseases such as metastatic melanoma, renal cancer (Rosenberg, 2014). It is a well-known fact that IL-2 can induce T-lymphocytes to grow in large amounts and activate anti-tumour response (Trinchieri, 1989). IL-2 alone or when combined with other chemotherapeutic-drugs, immune cytokines, and/or checkpoint blockades exhibits synergistic and enhanced anti-tumour activity. These points highlight the importance of employing IL-2 or IL-2 inducing peptides in therapeutic regimes to improve the overall patient's outcome even in the advanced stages of the cancers.

In this regard, the present study focuses on the rigorous attempt to formulate a pipeline and prediction model that can assess the IL-2 inducing potential of any given peptide. As per the best of the author's knowledge, it's the first computer-aided study on predicting the IL-2 inducing peptides. In our study, we have extracted the data for both IL-2 inducers and non-IL-2 inducers from the largest experimentally validated immune epitope database IEDB. We have analyzed nearly 10000 features for their capability in discriminating IL-2 inducers from IL-2 inducers. We have analyzed the positional preference of these amino acid residues in the peptides in positive and negative datasets using the two-sample logo. Length and amino acid composition analysis was carried out to investigate the abundance of certain residues in one class over the other. We have employed a variety of features and classifiers to build the prediction models that can reliably identify IL-2 inducing and non-IL-2 inducing peptides. As stated in the results section, random forest-based classification models using dipeptide composition and length s combined features performed the best in discriminating positive and negative sets, followed by the RF-DPC based classifier. Although there are several methods that were built for the prediction of numerous immune cytokines in the past (Dhanda, Gupta, Vir, & Raghava, 2013; Dhanda, Vir, et al., 2013), yet there is no computational model that can predict the IL-2 inducing and non-IL-2 inducing peptides. We have provided the pipeline in the form of an interactive web server, i.e., IL2pred, to aid the scientific community, where users can estimate the IL-2 inducing potential of the concerned peptides. We hope that our study and web server will help the researchers in designing better IL-2 inducing epitopes, which further will improve the survival of cancer patients.

7. Summary and Conclusion

Each year, cancer alone is responsible for millions of casualties around the globe. Despite the improvements in cancer therapeutics, the survival rate in the patients is not at par. Lung cancer is the top-most prevailing cancer across the world and has a low 5-year survival rate of 10% when diagnosed at stage IV. During progression, cancer cells become highly heterogeneous, which may further contribute towards low therapy response and subsequently high mortality. In this regard, clinicians and researchers are putting a lot of effort into finding the new diagnostic and therapeutic strategies that can reduce the side effects of conventional therapies along with improving the overall survival of the patients.

Immunotherapy is considered as the fourth and the most advanced pillar of cancer therapy. Over the past decade, the enthusiasm in the research of cancer immunotherapy has increased as it showed remarkable improvement in patient's survival even in refractory cancer types. One of the advanced immune-therapeutics approaches widely documented in the literature studies is the use of oncolytic viruses. Oncolytic viruses have the potential to infects and replicate within the cancer cells selectively. Upon inducing cancer cell lysis, oncolytic viruses release tumour-specific antigens, which lay the foundation for the generation of long-lasting anti-tumour immunity. A strong preferential anti-tumour response can also be induced when GM-CSF, chemotherapeutic drugs, and other cytokines (IL-2, IL-12, and IL-21) are delivered via oncolytic viruses either in the form of recombinant viral vectors or as vaccine adjuvant. Another popular immune-therapeutics approach is the use of IL-2 as a monotherapy or in combination with other therapies. Various experimental studies showed that the mutant of IL-2, known as "superkine" has a high therapeutic index with minimal side-effects.

Moreover, World Health Organization reports suggest that viruses such as *Human papilloma virus* contribute towards 25% of the new cancer cases around the world. Literature evidence suggests that the mortality rate among cancer patients can be reduced if proper prognosis and therapies are provided to patients at an early stage. Thus, finding a reliable biomarker and therapeutic regimens are of utmost importance, especially when it accounts for tumour heterogeneity. In this regard, the present work focuses on the development of computational strategies which aid in advancing the cancer immunotherapy process such as finding subunit epitopes against cancer-causing viruses as prophylactic measures; developing of knowledgebase on cancer-killing viruses which find application in designing of oncolytic virus-based treatment regimes; predicting the IL-2 inducing mutant peptides for immunotherapies, and in finding reliable prognostic biomarkers for heterogeneous cancer types. We hope that the results from our studies will be utilized by clinicians and researchers in designing advanced cancer treatment strategies.

As discussed above, chapter 3 of the thesis focuses on one of the emerging technologies that help in fighting cancer is immunotherapy based on oncolytic viruses, which can lyse tumour cells directly. Lately, an oncolytic virus called T-VEC has been cleared by the FDA for treating melanoma patients. To aid the scientific community in fighting cancer, we build a knowledge base on oncolytic viruses named OvirusTdb (https://webs.iiitd.edu.in/raghava/ovirustdb/). It is a manually curated database that stores the information retrieved from the PubMed research papers and US patents. The current version of the database houses comprehensive information on therapeutically critical oncolytic viruses. It has 5927 records, and each record has 25 fields, such as the species of virus, cell line of cancers, synergism with anti-cancer drugs, and several other. OvirusTdb houses data on 09 DNA and 15 RNA type viruses; 300 recombinant strains and 09 wild-type viral strains that were tested against 427 cancer cell lines and 124 cancer types. Nearly 1047 entries reflect enhanced anti-cancer effect by adopting the combinatorial strategy where chemotherapeutic agents were given along with the viruses. Approximately 3243 and 1506 entries showed the death of cancer cells via activation of apoptosis and the immune system, accordingly. Overall, a user-friendly web-resource on oncolytic viruses where one can retrieve and analysis data was developed to aid the researchers in suggesting and devising new oncolytic viruses for advancing treatment strategies against cancer.

Chapter 4 of the thesis focused on several species of viruses that are known to cause lung cancer and how this information can be exploited to suggest better therapeutic options to the patients. The most robust, documented causal relationship was found for HPV, which can cause 99% of cervical cancers. Nearly 20-25% of the newly diagnosed cancers are caused by infection from viruses. The present study aims to identify the subunit vaccine candidates against 09 oncogenic virus species causing lung cancer. We have utilized the reverse vaccinology approach to identify the subunit vaccine candidates from the proteomes of the virus species. We identified 125 best antigenic epitopes that can stimulate both arms of the immune system, acting as both immune booster (IL-4 and IFN-Gamma inducer) and vaccine adjuvant. Some of the identified epitopes may bind up to 15 MHC-type I and 49 MHC-type 2 alleles. This broad population coverage of the MHC molecule makes the identified epitopes an attractive candidate for vaccinating the large human population. Moreover, we have also identified 25 promiscuous epitopes available across multiple virus species. These epitopes can be used to provide heterologous immunity among different strains/species of oncogenic viruses. We have also validated the predicted epitopes on the IEDB dataset and found nearly 38 matched records against the experimentally validated IEDB datasets. The mapping of predicted epitopes on the IEDB dataset further strengthens the reliability of our results. Taken together, we conclude that identified best antigenic epitopes may aid in the advancement of cancer immunotherapy.

Chapter 5 of the thesis investigated the possible prognostic markers that are specific to the major subtypes of non-small-cell lung cancers. Considering the fact that much of the survival of cancer patients is linked with early detection and advanced therapeutics, this opens an opportunity for researchers around the globe to find reliable predictive, prognosis, and therapeutic strategies against cancer. With the advancement in the cancer genomics studies, several research groups have identified the predictive and prognostic biomarkers against many cancer types. These biomarkers have been utilized in clinics for guiding the therapeutics. Notable examples of biomarkers include activating EGFR mutations predicting response to TKI therapy and activating KRAS mutation predicting resistance to EGFR directed antibody therapy. Several drugs have also been approved by the FDA, which targets the specific biomarker – such as Cobimetinib, Erlotinib for melanoma, and non-small-cell lung carcinoma, which targets activating BRAF and EGFR mutations. But all these targeted therapies suffer from several limitations such as toxicity and cancer drug resistance. In the advanced stage, cancer cells become highly heterogeneous, which may account for a varied response towards targeted therapies. The heterogeneity also results in complex behavior of cancer cells, which may further lead to drug resistance. While the literature studies identify several biomarkers and causative agents for lung cancer, but still, the mortality rate among patients is high. One possible explanation could be the past studies are still not able to differentiate among the subtypes of lung cancer. In this regard, a deeper understanding of the cancer tissue type heterogeneity is of fundamental importance, which can further guide precision biomarkers and therapies.

Chapter 6 of the thesis deals with the development of a web-resource for the identification and designing of interleukin-2 inducing potential of peptides. The interleukin-2 holds great promise in cancer treatment. Experimental studies suggest that there is a need for a better version of interleukin-2, to enhance its capability in treating and managing the cancers. Thus, to aid the clinicians and scientists, we have developed a machine learning-based prediction algorithm for assessing the interleukin-2 inducing potential of a user-defined query sequence. The predictive algorithm is trained using experimentally verified interleukin-2 inducing sequences extracted from the IEDB database. The 2528 interleukin-2 inducing peptides serve the basis of the positive dataset, and 2104 non-IL-2 inducing peptides serve the basis of the negative dataset for the development of a machine learning-based model. Approximately 10000 peptide features have been computed from the positive and negative datasets. After applying feature selection on the computed features, various machine learning algorithms have been applied for the development of the correct prediction model. We have achieved maximum sensitivity of 74.60 and 74.11 on training and external datasets, respectively, using dipeptide composition and length as a hybrid feature on the random forest-based machine learning algorithm. The best predictive algorithm, along with the mutant generation and protein scan module, has been integrated into the form of a web-server that is fully available for the scientific community at https://webs.iiitd.edu.in/raghava/il2pred/index.html. We hope that the developed web-resource will help the researchers in designing better interleukin-2 inducing epitopes/peptides, which finds their way in clinics for advancing the cancer therapeutics.

Further, we conclude that since most of the study included in this thesis focuses on developing the various computational strategies for and identification for the identification and advancement of cancer therapeutic strategies. Computationally identified peptide-based vaccine candidates, subtype-specific biomarkers, and interleukin-2 inducing peptides need to be validated in the experimental setup before using in clinics.

References

Apostolopoulos, V. (2019, August 1). Cancer vaccines: Research and applications. *Cancers*, Vol. 11. https://doi.org/10.3390/cancers11081041

Apostolopoulos, V., Stojanovska, L., & Gargosky, S. E. (2015, August 21). MUC1 (CD227): A multi-tasked molecule. *Cellular and Molecular Life Sciences*, Vol. 72, pp. 4475–4500. https://doi.org/10.1007/s00018-015-2014-z

Ashshi, A. M., El-Shemi, A. G., Dmitriev, I. P., Kashentseva, E. A., & Curiel, D. T. (2016). Combinatorial strategies based on CRAd-IL24 and CRAd-ING4 virotherapy with anti-angiogenesis treatment for ovarian cancer. *Journal of Ovarian Research*, *9*(1). https://doi.org/10.1186/s13048-016-0248-5

Bai, Y., Hui, P., Du, X., & Su, X. (2019, May 1). Updates to the antitumour mechanism of oncolytic virus. *Thoracic Cancer*, Vol. 10, pp. 1031–1035. https://doi.org/10.1111/1759-7714.13043

Barta, J. A., Powell, C. A., & Wisnivesky, J. P. (2019). Global Epidemiology of Lung Cancer. *Annals of Global Health*, *85*(1). https://doi.org/10.5334/aogh.2419

Bhasin, M., & Raghava, G. P. S. (2004). Prediction of CTL epitopes using QM, SVM and ANN techniques. *Vaccine*, *22*(23–24), 3195–3204. https://doi.org/10.1016/j.vaccine.2004.02.005

Bommareddy, P. K., Patel, A., Hossain, S., & Kaufman, H. L. (2017). Talimogene Laherparepvec (T-VEC) and Other Oncolytic Viruses for the Treatment of Melanoma. *American Journal of Clinical Dermatology*, *18*(1). https://doi.org/10.1007/s40257-016-0238-9

Brahmer, J. R., Drake, C. G., Wollner, I., Powderly, J. D., Picus, J., Sharfman, W. H., … Topalian, S. L. (2010). Phase I study of single-agent anti-programmed death-1 (MDX-1106) in refractory solid tumours: Safety, clinical activity, pharmacodynamics, and immunologic correlates. *Journal of Clinical Oncology*, *28*(19), 3167–3175. https://doi.org/10.1200/JCO.2009.26.7609

Bray, F., Ferlay, J., Soerjomataram, I., Siegel, R. L., Torre, L. A., & Jemal, A. (2018). Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: A Cancer Journal for Clinicians*, *68*(6), 394–424. https://doi.org/10.3322/caac.21492

Chan, B. A., & Hughes, B. G. M. (2015). Targeted therapy for non-small cell lung cancer: Current standards and the promise of the future. *Translational Lung Cancer Research*, Vol. 4, pp. 36–54. https://doi.org/10.3978/j.issn.2218-6751.2014.05.01

Chaurasiya, S., Hew, P., Crosley, P., Sharon, D., Potts, K., Agopsowicz, K., … Hitt, M. M. (2016). Breast cancer gene therapy using an adenovirus encoding human IL-2 under control of mammaglobin promoter/enhancer sequences. *Cancer Gene Therapy*, *23*(6), 178–187. https://doi.org/10.1038/cgt.2016.18

Chen, X., Ai, X., Wu, C., Wang, H., Zeng, G., Yang, P., & Liu, G. (2018). A novel human IL-2 mutein with minimal systemic toxicity exerts greater antitumour efficacy than wild-type IL-2. *Cell Death and Disease*, *9*(10). https://doi.org/10.1038/s41419-018-1047-2

Chiocca, E. A., & Rabkin, S. D. (2014). Oncolytic viruses and their application to cancer immunotherapy. *Cancer Immunology Research*, *2*(4), 295–300. https://doi.org/10.1158/2326-6066.CIR-14-0015

Choudhry, H., Helmi, N., Abdulaal, W. H., Zeyadi, M., Zamzami, M. A., Wu, W., … Jamal, M. S. (2018). Prospects of IL-2 in Cancer Immunotherapy. *BioMed*

Research International, Vol. 2018. https://doi.org/10.1155/2018/9056173

Christiaansen, A., Varga, S. M., & Spencer, J. V. (2015, October 1). Viral manipulation of the host immune response. *Current Opinion in Immunology*, Vol. 36, pp. 54–60. https://doi.org/10.1016/j.coi.2015.06.012

Colditz, G. A., & Wei, E. K. (2012, April 21). Preventability of cancer: The relative contributions of biologic and social and physical environmental determinants of cancer mortality. *Annual Review of Public Health*, Vol. 33, pp. 137–156. https://doi.org/10.1146/annurev-publhealth-031811-124627

Conry, R. M., Westbrook, B., McKee, S., & Norwood, T. G. (2018, April 3). Talimogene laherparepvec: First in class oncolytic virotherapy. *Human Vaccines and Immunotherapeutics*, Vol. 14, pp. 839–846. https://doi.org/10.1080/21645515.2017.1412896

de Martel, C., Georges, D., Bray, F., Ferlay, J., & Clifford, G. M. (2020). Global burden of cancer attributable to infections in 2018: a worldwide incidence analysis. *The Lancet Global Health*, *8*(2), e180–e190. https://doi.org/10.1016/S2214-109X(19)30488-7

Dela Cruz, C. S., Tanoue, L. T., & Matthay, R. A. (2011, December). Lung Cancer: Epidemiology, Etiology, and Prevention. *Clinics in Chest Medicine*, Vol. 32, pp. 605–644. https://doi.org/10.1016/j.ccm.2011.09.001

Devarakonda, S., Morgensztern, D., & Govindan, R. (2015, July 1). Genomic alterations in lung adenocarcinoma. *The Lancet Oncology*, Vol. 16, pp. e342–e351. https://doi.org/10.1016/S1470-2045(15)00077-7

Dhanda, S. K., Gupta, S., Vir, P., & Raghava, G. P. (2013). Prediction of IL4 inducing peptides. *Clinical & Developmental Immunology, 2013*, 263952.

https://doi.org/10.1155/2013/263952

Dhanda, S. K., Vir, P., & Raghava, G. P. S. (2013). Designing of interferon-gamma inducing MHC class-II binders. *Biology Direct*, *8*(1). https://doi.org/10.1186/1745-6150-8-30

Ebrahimi, S., Mohabatkar, H., & Behbahani, M. (2019). Predicting Promiscuous T Cell Epitopes for Designing a Vaccine Against Streptococcus pyogenes. *Applied Biochemistry and Biotechnology*, *187*(1), 90–100. https://doi.org/10.1007/s12010-018-2804-5

Ferlay, J., Colombet, M., Soerjomataram, I., Mathers, C., Parkin, D. M., Piñeros, M., … Bray, F. (2019, April 15). Estimating the global cancer incidence and mortality in 2018: GLOBOCAN sources and methods. *International Journal of Cancer*, Vol. 144, pp. 1941–1953. https://doi.org/10.1002/ijc.31937

Fukuhara, H., Ino, Y., & Todo, T. (2016, October 1). Oncolytic virus therapy: A new era of cancer treatment at dawn. *Cancer Science*, Vol. 107, pp. 1373–1379. https://doi.org/10.1111/cas.13027

Gonzalez, H., Hagerling, C., & Werb, Z. (2018). Roles of the immune system in cancer: From tumour initiation to metastatic progression. *Genes and Development*, Vol. 32, pp. 1267–1284. https://doi.org/10.1101/GAD.314617.118

Grifoni, A., Weiskopf, D., Ramirez, S. I., Mateus, J., Dan, J. M., Moderbacher, C. R., … Sette, A. (2020). Targets of T Cell Responses to SARS-CoV-2 Coronavirus in Humans with COVID-19 Disease and Unexposed Individuals. *Cell*, *181*(7), 1489-1501.e15. https://doi.org/10.1016/j.cell.2020.05.015

Howells, A., Marelli, G., Lemoine, N. R., & Wang, Y. (2017, September 8).

Oncolytic viruses-interaction of virus and tumour cells in the battle to eliminate cancer. *Frontiers in Oncology*, Vol. 7. https://doi.org/10.3389/fonc.2017.00195

Jiang, T., Zhou, C., & Ren, S. (2016, June 2). Role of IL-2 in cancer immunotherapy. *OncoImmunology*, Vol. 5. https://doi.org/10.1080/2162402X.2016.1163462

Johnson, B. E., Crawford, J., Downey, R. J., Ettinger, D. S., Fossella, F., Grecula, J. C., … Williams, C. C. (2006). Small cell lung cancer: Clinical Practice Guidelines in Oncology^{TM}. *JNCCN Journal of the National Comprehensive Cancer Network*, Vol. 4, pp. 602–622. https://doi.org/10.6004/jnccn.2006.0050

Johnson, D. B., Puzanov, I., & Kelley, M. C. (2015). Talimogene laherparepvec (T-VEC) for the treatment of advanced melanoma. *Immunotherapy*, *7*(6), 611–619. https://doi.org/10.2217/imt.15.35

Jones, G. S., & Baldwin, D. R. (2018, April 1). Recent advances in the management of lung cancer. *Clinical Medicine, Journal of the Royal College of Physicians of London*, Vol. 18, pp. s41–s46. https://doi.org/10.7861/clinmedicine.18-2-s41

Karachaliou, N., Pilotto, S., Lazzari, C., Bria, E., de Marinis, F., & Rosell, R. (2016). Cellular and molecular biology of small cell lung cancer: An overview. *Translational Lung Cancer Research*, Vol. 5, pp. 2–15. https://doi.org/10.3978/j.issn.2218-6751.2016.01.02

Kelly, E., & Russell, S. J. (2007, April). History of oncolytic viruses: Genesis to genetic engineering. *Molecular Therapy*, Vol. 15, pp. 651–659. https://doi.org/10.1038/sj.mt.6300108

Kim, E. S., Herbst, R. S., Wistuba, I. I., Jack Lee, J., Blumenschein, G. R., Tsao, A., … Hong, W. K. (2011). The BATTLE trial: Personalizing Therapy for Lung Cancer. *Cancer Discovery*, *1*(1), 44–53. https://doi.org/10.1158/2159-8274.CD-10-0010

Kim, Y., Pierce, C. M., & Robinson, L. A. (2018). Impact of viral presence in tumour on gene expression in non-small cell lung cancer. *BMC Cancer*, *18*(1). https://doi.org/10.1186/s12885-018-4748-0

Kuleshov, M. V., Jones, M. R., Rouillard, A. D., Fernandez, N. F., Duan, Q., Wang, Z., … Ma'ayan, A. (2016). Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. *Nucleic Acids Research*, *44*(W1), W90–W97. https://doi.org/10.1093/nar/gkw377

Lathwal, A., Kumar, R., & Raghava, G. P. S. (2020, July 1). Computer-aided designing of oncolytic viruses for overcoming translational challenges of cancer immunotherapy. *Drug Discovery Today*, Vol. 25, pp. 1198–1205. https://doi.org/10.1016/j.drudis.2020.04.008

Lei, Y., Zhao, F., Shao, J., Li, Y., Li, S., Chang, H., & Zhang, Y. (2019). Application of built-in adjuvants for epitope-based vaccines. *PeerJ*, *2019*(1). https://doi.org/10.7717/peerj.6185

Lemjabbar-Alaoui, H., Hassan, O. U. I., Yang, Y. W., & Buchanan, P. (2015, December 1). Lung cancer: Biology and treatment options. *Biochimica et Biophysica Acta - Reviews on Cancer*, Vol. 1856, pp. 189–210. https://doi.org/10.1016/j.bbcan.2015.08.002

Liu, M., & Guo, F. (2018). Recent updates on cancer immunotherapy. *Precision Clinical Medicine*, *1*(2), 65–74. https://doi.org/10.1093/pcmedi/pby011

Malek, T. R. (2008). The biology of interleukin-2. *Annual Review of Immunology*, Vol. 26, pp. 453–479. https://doi.org/10.1146/annurev.immunol.26.021607.090357

Manstein, V., Yang, C., Richter, D., Delis, N., Vafaizadeh, V., & Groner, B. (2014). Resistance of Cancer Cells to Targeted Therapies Through the Activation of Compensating Signaling Loops. *Current Signal Transduction Therapy*, *8*(3), 193–202. https://doi.org/10.2174/1574362409666140206221931

Maule, M., & Merletti, F. (2012, August). Cancer transition and priorities for cancer control. *The Lancet Oncology*, Vol. 13, pp. 745–746. https://doi.org/10.1016/S1470-2045(12)70268-1

Meadows, G. G., & Zhang, H. (2015). Effects of alcohol on tumour growth, metastasis, immune response, and host survival. *Alcohol Research: Current Reviews*, *37*(2). Retrieved from https://pubmed.ncbi.nlm.nih.gov/26695753/

Mehta, A., Sriramanakoppa, N., Agarwal, P., Viswakarma, G., Vasudevan, S., Panigrahi, M., … Suryavanshi, M. (2019). Predictive biomarkers in nonsmall cell carcinoma and their clinico-pathological association. *South Asian Journal of Cancer*, *8*(4), 250. https://doi.org/10.4103/sajc.sajc_373_18

Mills, C. C., Kolb, E. A., & Sampson, V. B. (2018, January 15). Development of chemotherapy with cell-cycle inhibitors for adult and pediatric cancer therapy. *Cancer Research*, Vol. 78, pp. 320–325. https://doi.org/10.1158/0008-5472.CAN-17-2782

Mountain, C. F. (1997). Revisions in the international system for staging lung cancer. *Chest*, *111*(6), 1710–1717. https://doi.org/10.1378/chest.111.6.1710

Nagai, H., & Kim, Y. H. (2017, March 1). Cancer prevention from the perspective of global cancer burden patterns. *Journal of Thoracic Disease*, Vol. 9, pp. 448–451. https://doi.org/10.21037/jtd.2017.02.75

Nagpal, G., Chaudhary, K., Agrawal, P., & Raghava, G. P. S. (2018). Computer-aided prediction of antigen presenting cell modulators for designing peptide-based vaccine adjuvants. *Journal of Translational Medicine*, *16*(1). https://doi.org/10.1186/s12967-018-1560-1

Nagpal, G., Usmani, S. S., Dhanda, S. K., Kaur, H., Singh, S., Sharma, M., & Raghava, G. P. S. (2017). Computer-aided designing of immunosuppressive peptides based on IL-10 inducing potential. *Scientific Reports, 7*. https://doi.org/10.1038/srep42851

Ngono, A. E., & Shresta, S. (2019). Cross-reactive T Cell immunity to dengue and zika viruses: New insights into vaccine development. *Frontiers in Immunology*, Vol. 10. https://doi.org/10.3389/fimmu.2019.01316

Oprea, M., & Antohe, F. (2013). Reverse-vaccinology strategy for designing T-cell epitope candidates forStaphylococcus aureus endocarditis vaccine. *Biologicals*, *41*(3), 148–153. https://doi.org/10.1016/j.biologicals.2013.03.001

Phan, M., Watson, M. F., Alain, T., & Diallo, J. S. (2018). Oncolytic Viruses on Drugs: Achieving Higher Therapeutic Efficacy. *ACS Infectious Diseases, 4*(10), 1448–1467. https://doi.org/10.1021/acsinfecdis.8b00144

Pikor, L. A., Ramnarine, V. R., Lam, S., & Lam, W. L. (2013, November). Genetic alterations defining NSCLC subtypes and their therapeutic implications. *Lung Cancer*, Vol. 82, pp. 179–189. https://doi.org/10.1016/j.lungcan.2013.07.025

Pucci, C., Martinelli, C., & Ciofani, G. (2019, September 10). Innovative approaches for cancer treatment: Current perspectives and new challenges. *Ecancermedicalscience*, Vol. 13. https://doi.org/10.3332/ecancer.2019.961

Robinson, L. A., Jaing, C. J., Pierce Campbell, C., Magliocco, A., Xiong, Y., Magliocco, G., … Antonia, S. (2016). Molecular evidence of viral DNA in non-small cell lung cancer and non-neoplastic lung. *British Journal of Cancer*, *115*(4), 497–504. https://doi.org/10.1038/bjc.2016.213

Rojas, J. J., Guedan, S., Searle, P. F., Martinez-Quintanilla, J., Gil-Hoyos, R., Alcayaga-Miranda, F., … Alemany, R. (2010). Minimal RB-responsive E1A promoter modification to attain potency, selectivity, and transgene-arming capacity in oncolytic adenoviruses. *Molecular Therapy*, *18*(11), 1960–1971. https://doi.org/10.1038/mt.2010.173

Rosenberg, S. A. (2014). IL-2: The First Effective Immunotherapy for Human Cancer. *The Journal of Immunology*, *192*(12), 5451–5458. https://doi.org/10.4049/jimmunol.1490019

Rudensky, A. Y. (2011). Regulatory T cells and Foxp3. *Immunological Reviews*, *241*(1), 260–268. https://doi.org/10.1111/j.1600-065X.2011.01018.x

Russell, L., Peng, K. W., Russell, S. J., & Diaz, R. M. (2019, October 1). Oncolytic Viruses: Priming Time for Cancer Immunotherapy. *BioDrugs*, Vol. 33, pp. 485–501. https://doi.org/10.1007/s40259-019-00367-0

Sabah, S. N., Gazi, M. A., Sthity, R. A., Husain, A. B., Quyyum, S. A., Rahman, M., & Islam, M. R. (2018). Designing of Epitope-Focused Vaccine by Targeting E6 and E7 Conserved Protein Sequences: An Immuno-Informatics Approach in Human Papillomavirus 58 Isolates. *Interdisciplinary Sciences: Computational Life Sciences*, *10*(2), 251–260. https://doi.org/10.1007/s12539-

016-0184-5

Sharma, G. N., Dave, R., Sanadya, J., Sharma, P., & Sharma, K. K. (2010, April). Various types and management of breast cancer: An overview. *Journal of Advanced Pharmaceutical Technology and Research*, Vol. 1, pp. 109–126. Retrieved from https://pubmed.ncbi.nlm.nih.gov/22247839/

Siegel, J. P., Sharon, M., Smith, P. L., & Leonard, W. J. (1987). The IL-2 receptor β chain (p70): Role in mediating signals for LAK, NK, and proliferative activities. *Science*, *238*(4823), 75–78. https://doi.org/10.1126/science.3116668

Siegel, R. L., Miller, K. D., & Jemal, A. (2020). Cancer statistics, 2020. *CA: A Cancer Journal for Clinicians*, *70*(1), 7–30. https://doi.org/10.3322/caac.21590

Singh, Harinder, Ansari, H. R., & Raghava, G. P. S. (2013). Improved Method for Linear B-Cell Epitope Prediction Using Antigen's Primary Sequence. *PLoS ONE*, *8*(5). https://doi.org/10.1371/journal.pone.0062216

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

Singh, Harpreet, & Raghava, G. P. S. (2002). ProPred: Prediction of HLA-DR binding sites. *Bioinformatics*, *17*(12), 1236–1237. https://doi.org/10.1093/bioinformatics/17.12.1236

Sukari, A., Nagasaka, M., Al-Hadidi, A., & Lum, L. G. (2016, November 1). Cancer immunology and immunotherapy. *Anticancer Research*, Vol. 36, pp. 5593–5606. https://doi.org/10.21873/anticanres.11144

Tähtinen, S., Blattner, C., Vähä-Koskela, M., Saha, D., Siurala, M., Parviainen, S., … Hemminki, A. (2016). T-Cell Therapy Enabling Adenoviruses Coding for IL2 and TNFα Induce Systemic Immunomodulation in Mice with Spontaneous Melanoma. *Journal of Immunotherapy*, *39*(9), 343–354. https://doi.org/10.1097/CJI.0000000000000144

Tang, C. K., & Apostolopoulos, V. (2008, September). Strategies used for MUC1 immunotherapy: Preclinical studies. *Expert Review of Vaccines*, Vol. 7, pp. 951–962. https://doi.org/10.1586/14760584.7.7.951

Tanoue, L. T., & Detterbeck, F. C. (2009). New TNM classification for non-small-cell lung cancer. *Expert Review of Anticancer Therapy*, Vol. 9, pp. 413–423. https://doi.org/10.1586/ERA.09.11

Thomas, A., & Hassan, R. (2012, July). Immunotherapies for non-small-cell lung cancer and mesothelioma. *The Lancet Oncology*, Vol. 13. https://doi.org/10.1016/S1470-2045(12)70126-2

Trinchieri, G. (1989). Biology of Natural Killer Cells. *Advances in Immunology*, *47*(C), 187–376. https://doi.org/10.1016/S0065-2776(08)60664-1

Ulasov, I. V., Tyler, M. A., Rivera, A. A., Nettlebeck, D. M., Douglas, J. T., & Lesniak, M. S. (2008). Evaluation of E1A double mutant oncolytic adenovectors in anti-glioma gene therapy. *Journal of Medical Virology*, *80*(9), 1595–1603. https://doi.org/10.1002/jmv.21264

Uramoto, H., & Tanaka, F. (2014). Recurrence after surgery in patients with NSCLC. *Translational Lung Cancer Research*, Vol. 3, pp. 242–249. https://doi.org/10.3978/j.issn.2218-6751.2013.12.05

Vens, C., Rosso, M. N., & Danchin, E. G. J. (2011). Identifying discriminative classification-based motifs in biological sequences. *Bioinformatics*, *27*(9), 1231–1238. https://doi.org/10.1093/bioinformatics/btr110

Vita, R., Overton, J. A., Greenbaum, J. A., Ponomarenko, J., Clark, J. D., Cantrell, J. R., … Peters, B. (2015). The immune epitope database (IEDB) 3.0. *Nucleic Acids Research*, *43*(D1), D405–D412. https://doi.org/10.1093/nar/gku938

Wagih, O. (2017). Ggseqlogo: A versatile R package for drawing sequence logos. *Bioinformatics*, *33*(22), 3645–3647. https://doi.org/10.1093/bioinformatics/btx469

Wang, H., Naghavi, M., Allen, C., Barber, R. M., Carter, A., Casey, D. C., … Zuhlke, L. J. (2016). Global, regional, and national life expectancy, all-cause mortality, and cause-specific mortality for 249 causes of death, 1980–2015: a systematic analysis for the Global Burden of Disease Study 2015. *The Lancet*, *388*(10053), 1459–1544. https://doi.org/10.1016/S0140-6736(16)31012-1

Wang, N., Shang, J., Jiang, S., & Du, L. (2020, February 28). Subunit Vaccines Against Emerging Pathogenic Human Coronaviruses. *Frontiers in Microbiology*, Vol. 11. https://doi.org/10.3389/fmicb.2020.00298

Zappa, C., & Mousa, S. A. (2016). Non-small cell lung cancer: Current treatment and future advances. *Translational Lung Cancer Research*, *5*(3), 288–300. https://doi.org/10.21037/tlcr.2016.06.07

Zheng, J. N., Pei, D. S., Mao, L. J., Liu, X. Y., Sun, F. H., Zhang, B. F., … Han, D. (2010). Oncolytic adenovirus expressing interleukin-18 induces significant antitumour effects against melanoma in mice through inhibition of angiogenesis. *Cancer Gene Therapy*, *17*(1), 28–36. https://doi.org/10.1038/cgt.2009.38

Zheng, M., Huang, J., Tong, A., & Yang, H. (2019, December 20). Oncolytic Viruses for Cancer Therapy: Barriers and Recent Advances. *Molecular Therapy - Oncolytics*, Vol. 15, pp. 234–247. https://doi.org/10.1016/j.omto.2019.10.007

Zhu, Y., Qiu, P., & Ji, Y. (2014). TCGA-assembler: Open-source software for retrieving and processing TCGA data. *Nature Methods*, Vol. 11, pp. 599–600. https://doi.org/10.1038/nmeth.2956