Comparison of EBV associated Nasopharyngeal cancers for the identification of therapeutic targets and modification of existing therapies

BY

Maleeha Humayun

Fall-2020-MSCS&E00000330037

Supervised by

Dr. Rehan Zafar Paracha

A THESIS SUBMITTED IN PARTIAL FULFILMENT OF THE REQUIREMENTS FOR THE DEGREE OF MASTER OF SCIENCE

in

Bioinformatics

September, 2022

School of Interdisciplinary Engineering Science(SINES)

National University of Sciences and Technology (NUST)

This thesis is dedicated to my beloved Parents and Dr. Rehan Zafar Paracha for their consistent support and love.

DECLARATION

I Maleeha Humayun, hereby declared that work presented in this thesis is result of my own work except specific reference is made to the work of others wherever due. I also declare that the content presented in this thesis is original and have not been submitted in whole or in part to this university or to any other university for any other degree or qualification.

Maleeha Humayun

September, 2022

Acknowledgement

"So, surely with hardship comes ease" (Quran 94:5) is the one verse that got me through many difficult times. I am immensely thankful to Allah firstly for making it possible for me to able complete my research. I am in debt of my beloved parents who believed in me and sent me to a different city with love and support. I am ever grateful to my siblings as well for their love. I would stand nowhere without them. My dedicated, kind and extremely passionate research instructor Dr. Rehan Zafar deserves all the credit in the world for his efforts, motivational discussions and endless faith in me. I extend my gratitude to you sir. My dearest Dr. Zartasha and Dr. Mehak are one of the most empowered women I met at NUST. I am grateful for their guidance and learned alot from their positive and strong personalities. Last but not least Dr. Zamir, who taught me how to take leadership roles and move forwards with a strategic and positive mindset. I am thankful to you sir.

My beautiful, smart and kind sister Mahrokh and my strong determined brother, Harris. I owe you both my degree for without your help getting a master degree would not have been possible for me. I can only hope I can be as helpful to you both in future as you were for me.

"A friend is someone who knows all about you and still loves you." — Elbert Hubbard. This quote describes the pure relationship based on friendship. Some names worth mentioning here Samana Zahid, Talat Azam and Nausheen Ilyas. It would not be an exaggeration if I say that they are angels send by God for me. Without these people life would not be as colourful as it is now. Samana, I am thankful to you for being kind and supportive. Talat, without you life on campus would have been without colour. Nausheen, I thank you for not letting distance come between us. Despite being in different cities you showed love and support to me. Sameen Ahmed, was one of the first people I become friends with and I am ever glad we decided to stick around each other. Collectively their love and support got me through the two years at NUST.

Contents

Nomenclature

Acronyms / Abbreviations

APCs	Antigen Presenting Cells
BCR	B cell receptor
BRN	Biological Regulatory Network
CAR	Coxsackie-adenovirus receptor
CD4+T cells	Helper T cells
CD8+T cells	Cytotoxic T cells
CNS	Central nervous system
CRS	Cytoreductive surgery
DAMPs	Danger-associated molecular pattern signals
DC	Dendritic Cells
EGF	Epidermal growth factor
FDA	US Food and Drug Administrationa
G-CSF	Granulocyte colony stimulating factor
GM-CSF	Granulocyte monocyte colony stimulating factor
HER2+	Human epidermal growth factor receptor 2+
HN	Hemagglutinin-neuraminidase
IL	Interleukin
INF	Interferon
MAPK	Mitogen-activated protein kinases
MHC	Major Histocompatibility complex
NDV	Newcastle Disease Virus
NK cells	Natural Killer Cells
NSCLC	Non-small cell lung cancer
PAMPs	Pathogen-associated molecular patterns
ROS	Reactive Oxygen specie
SCLC	Small Cell Lung Cancer
SLAM	Signalling lymphocytic activation molecule
STAT3	Signal Transducer and Activator of Transcription 3

SVV	Seneca Valley Virus
T-VEC	Talimogene laherparepvec
TAA	Tumor associated antigens
TAM	Tumor Microenvironment
TCR	T cell receptor
TK	Thymidine kinase
TNF	Tumor Necrosis Factor
VGF	Vaccinia growth factor
VV	Vaccinia Virus
WHO	World Health Organization

List of Tables

List of Figures

ABSTRACT

The Epstein-Barr Virus (EBV), commonly known as human herpesvirus 4, belongs to the herpes virus family. More than 90% of the total human population is affected EBV, making it one of the most widely spread virus. Several attempts have been made to classify EBV categorically: clinical and for epidemiological purposes. The constant mutations in EBV makes it harder to predict the intensity of the disease caused, hence making it difficult to create an all-rounder medication. However, the classification done so far has helped to identify EBV's role in origin and progression of malignant and non-malignant diseases. The classified viral proteins infect epithelial cells and B cells of the human immune system as these viral proteins reside in a dormant state (latency) in memory cells of the immune system for a lifetime after the initial lytic infection. One of the many fatal malignant diseases caused by EBV is Nasopharyngeal carcinoma (NPC). Out of total cases reported, 99% of the cases show elevated levels of antibodies produced against EBV. This project focuses on the pathways involved in the genesis and progression of NPC and identification of genes through High-throughput sequencing approaches and Microarray analysis. Next Generation Sequencing (NGS) high throughput data has been used for holistic analysis to understand the regulation of genes. Moreover, the important genes contributing to onset and progression of NPC are identified and validated through datasets based on different ethnicities. Moreover, the project satisfies the need for development of therapeutics. A contagious virus like EBV requires continued research to prepare the mankind kind for an unforeseeable epidemic.

Chapter 1

INTRODUCTION

1.1 Immune system

The human body depends on numerous systems for its normal functioning. Out of many complex systems, immune system can be described as assortment of cells, operation, enzymes and hormones appointed to protect skin, internal passages, tracts and other internal body environments from parasites, infectious microbes (such as bacteria, fungi, viruses), cancer cells and toxins (1).

In other words, immune system can be viewed as 'line of defence', can be described as chemical and structural barrier that protects the internal body systems. From birth to complete maturity, immune system undergoes several stages innate immunity and adaptive immunity. Upon contact with infection immune system generates a response known as innate immune response. However, after the initial contact with infections immune system generates memory and learns which is known as adaptive immunity(2).

While innate immunity is non-specific defence mechanism that appears initially within hours of encounter with antigen, generates no memory response in case body is exposed to same antigen body will have no recollection of it. On the other hand, adaptive immunity generates antigen dependent and antigen specific response, hence requires more time than innate immunity between exposure (1). Adaptive immunity's distinctive capacity to regenerate memory enables quicker and efficient response to antigens. Innate and adaptive immunities are complementary in their mode of actions in producing an appropriate response (3).

1.2 Cancer immunology

Cancer takes 8.2 million lives annually worldwide, and, with the growing population the toll is only expected to increase (4). Cancers can be broadly divided into two groups: metastatic; that spreads and forms sub subtypes, primary cause of all cancer-related deaths, and nonmetastatic; cancer that does not spread from the primary site of infection (5). Initially metastatic cancer was known to disseminate in the later stage of cancer development; however, with advancement in research it was determined that metastatic spread occurs during stages of tumour formation (6) . Throughout metastasis, cancer cells disseminate from the primary site and travel to distant organs disguised as normal body cells and colonize (7).

Historically, increase in cancer immunology research stems from the immune system acting as a weapon against carcinomas. Cancer immunology goes back as back as to late 1800's where a common term 'magic bullet' was used to describe cancer therapy (8). The tumour microenvironments differ drastically from normal surrounding tissue in both tissues composition and biologic conduct. A basic hallmark of cancer is its genetic instability, tumour specific neo-antigens present at the primary site. Defects in the DNA damage repair system is the most common cause of mutations (9).

The real question however, is how do cancer cells avoid cells of immune system? In general principle development of tumours can be controlled by innate an adaptive immunity; however, as the tumour progresses it the microenvironment cells mimic different mechanisms to achieve immune tolerance to protect itself from tumoricidal attack (7).

1.3 Cancer of head and neck

One of the most commonly diagnosed cancer type is head and neck cancer (HNC). Head and neck cancers are the heterogeneous group of upper aerodigestive

tract (10). According to a report 2018 HNC was identified as the 8th most common carcinoma. 3% of all cancer cases reported were of HNC.(11) Worldwide 1.5% death related to cancer were due to HNC.(12) Out of all HNC 90% of the cases are of head and neck squamous cell carcinoma (HNSCC) (11).

HNSCC arises from the upper aerodigestive tract. It can be further subcategorised as

1. Oral squamous cell carcinomas (mouth cancer; arising from lips to retromolar trigone).

2. oropharyngeal squamous cell carcinomas (OPSCC) (rising from tongue to posterior pharyngeal wall.

3. laryngeal squamous cell carcinomas (LSCC), (rising from the supraglottis, glottis, subglottis.

4. nasal squamous cell carcinomas (NSCC), which develop from squamous epithelial cells lining the nasal cavity and paranasal sinuses (13).

A study in 2021 approximated that heavy smokers are 5 to 25 times more at risk for HNC as compared to non-smokers. 85% of all the HNC cases are due to consumption of smoking. In countries with strict policies against smoking and tobacco production like north America and western Europe there is a decrease in HNC cases (14).

The decrease in cases, however, is not observed worldwide, the number of incidents for HNC are significantly higher in Eastern European countries, throughout Asia particularly in India, China, Indonesia, and Pakistan (12). The risk of HNC is not only elevated through tobacco consumption (15) but increased alcohol intake impacts the probability of HNC (11). Moreover, other factors like poor oral hygiene, malnutrition, prolonged irritation, poor fiber diets along with genetic factors are associated with oral cancers (12).

Other past decade cases reported of HNC are seen to be closely associated with cancer causing viruses known as oncogenic viruses like Epstein-Barr virus (EBV)

, Hepatitis C (HCV) and Hepatitis B (HBV) (15)

1.4 Nasopharyngeal cancer

Aero-digestive tract is divided into portions the uppermost part is the nasopharynx. The aero-digestive tract extends from the base of the skull to the soft palate and leads to the posterior aspects of the nasal cavity. On its adjacent walls are the Eustachian tubes also known as pharyngotympanic tubes (16). Each tube, superiorly and posteriorly, is bounded by tubal elevation. Nasopharyngeal recess known as fossa of Rosenmuller is behind the tubal elevation. This site is the most common site for nasopharyngeal cancer (NPC)(17).

1.5 Epidemiology

Nasopharyngeal carcinomas exhibit geographical variation, with the highest incidence rate in Asian counties, Southern China and North Africa a relatively low incidence rate in the European counties and North America (17). A recent study from USA states NPC as the sixth most common type of HNC information from regions like Africa are scanty, but indicates towards low prevalence. With the epidemiological data available supports the fact that HNC is a major public health issue, with a continuous increase in the incident rate, prevalence, and mortality rate. HNC are responsible for 380,000 cancer-related deaths (18).

1.6 Signs and Symptom

In carcinoma, NPC, patients exhibit different signs and symptoms based on the severity of the stages, the spread of tumour. However, signs and symptoms for NPC can generalised as follow: The base nasal level: About 80% of the patients suffer from nasal signs and symptoms. These signs are early indication of the cancer.

Nasal impediment, epistaxis, distortion in nasal voice, decrease in olfactory acuity and several other olfactory disorders can be seen (19).

The otic level: this level exhibits a recurrent ear infection along with ototubal dysfunction, middle ear effusion and tinnitus, ear fullness sensations, otorrhea and otorrhagia (17). .

Other signs and symptoms include the neurological signs that develop as the carcinoma progresses and tumour extension takes place. Disturbance in the nerve structure including vestibule-cochlear nerve, facial and abducens nerve, and with the loss of sensibility (19)

As the NPC evolves into various stages, lymph nodes enlargement is seen. Tumour localizing at the cervical or submandibular level or at further distance sites can be seen (19).

Other general signs and symptoms in NPC patient are: low grade fever, migraines, low RBC count (anaemia) and drastic weight loss (19).

Initial stages of NPC are categorized as low signs and symptoms. However, with NPC progression symptom become more prominent and frequent symptoms include the otic and nasal symptoms. To be noted here that NPC has a propensity to form distant metastasis and higher frequency as compared to other types of HNC. Regarding the frequency metastases are commonly localised in bones, liver and lungs (20).

1.7 Causes

Some of the most common factors include narcotics such as tobacco, alcohol, food containing nitrosamine. Consumption of tobacco and alcohol combined increases the risk of developing NPC. These compounds have synergistic effects on the oral cavity and pharyngeal cancers Other factors include salt preserved foods like fish and meat. Genetic susceptibility host HLA gene on the 6p21, environmental hazards; exposure to chemical substance or carcinogens. Oncogenic viruses like

Epstein-Barr-virus or Human papilloma virus are a leading cause of NPC . NPC grows well in prior set environment that supports the malignant growth created through an infection of oncogenic virus(19).

1.8 Classification

The classification of NPC is based on the histopathological characteristics by the World Health Organization (WHO). Based on the differences the following types are established: keratinizing NPC (25%), non-keratinizing differentiated NPC (up to 15%) and non-keratinizing undifferentiated (up to 65%) . Epstein-Barr virus associated NPC tumours have specific attributes like invasiveness, distant metastases from the primary location, sensitiveness to both radiotherapy and chemotherapy. NPC which is not associated to EBV are less sensitive to chemo and radiotherapy and similar oral cancers (17)

1.9 Diagnosis

Patients come forth for diagnosis in the later stages of cancer. To properly identify the cancer detailed local examination is done through either endoscopy or rhinoscopy followed by the performance of a biopsy obtained from the primary site of the infection. In order to understand physical properties, diameter, precise location and impacts on the surrounding structure (21), of the rhinopharngeal tumour imaging techniques can be used. EBV markers can be used for detection of NPC (17).

Due to the different stages of EBV to find out the presence of virus is through in situ hybridization for EBV- encoding region (EBER). This encoding region is present abundantly in latent-infected EBV cells. Another test used is the PCR test to search EBV in tumour cells. However due to low of number of EBV+ lymphocytes it can generate a false positive. The most recent medical research has emphasized the necessity for plasma EBV deoxyribonucleic acid (DNA) liquid biopsy in population

screening (19).

1.10 Management

1.10.1 *Radiotherapy*

For the management of loco region lesion radiation is used. Radiotherapy proves to be rather effective in all cases expect for targeting distant metastasis ranging from the stages I to IVB. As NPC spread rather quickly in nasopharynx, a small region, spaces musculature and nodes common. Therefore, a dose of nearly 65 Gy for primary tumours with 50 to 55 Gy is also essential for nodal negative necks (22).

1.10.2 *Chemotherapy*

NPC is very susceptible to radiation and chemotherapy. In locally advance stages of NPC a simultaneous dose of both radiation and chemotherapy is administrated. The tumour shrinking responds to it better. Cisplatin is common chemotherapeutic for initial line of control. The standard dost is 100mg every third week. As radiation therapy doesn't cater to the distant tumours however chemotherapy solves that issue through palliative chemotherapy. The doctor usually has a choice between cisplatin and 5-fluorouracil. Despite the available therapy the survival rate is not more than a year (23)

1.10.3 *Surgical intervention*

In NPC surgical intervention is only appointed as last resort. Its considered as a salvage option. Assession to nasopharynx is difficult as it's a small camber making surgery a complicated challenge. However, in cases of recurring disease, patients are given the option of surgery. For the management of oligo-metastasis surgery is done concurrent with radiotherapy and chemotherapy (24).

1.11 Epstein-Barr virus

Epstein-Barr virus, was first identified in 1964, is a member of herpes virus family. Around the world nearly 90% of the human population in infected by the virus (25). However, not everyone shows disease symptoms. Its transmission is mainly through saliva in case of NPC (26).

EBV is known for its capabilities to exhibit dual tropism. Which basically means that it effects, both B cells and epithelial cells (27). The virus has a life cycle of two condition lytic and latent. Lytic life cycles the virus actively causes disease and aggressively multiply on the other hand, during latent cycle the virus stays dormant and doesn't cause disease. Under the latency conditions the virus exists quietly in the infected memory B cells (27). Although both B cells and epithelial cells can get infected, B cells seem to be more susceptible. By switching its emvelop proteins the virus cell entry mechanism. The virus is known to bind to the complement receptor type 2 proteins using its protein gp350 which are membrane protein present on the B cells. While it uses gp40 envelop protein to bind to surface integrins on epithelial cells. These alternating methods of cell entry are impertinent to EBV's survival in humans (4)

1.12 Genome and structure

Viral genomes comprise of a double helix deoxyribonucleic acid (DNA) and is about 122-180 nm in size. The genome contains 85 genes and 172,000 base pairs. The DNA has coverings around it. Protein nucleocapsid firstly surrounded by tegument made of protein which is then covered by an envelop which contains both lipids and glycoprotein surface projections which are essential for causing infections in the host cell (11). In 2020 researchers completed the first atomic model of the nucleocapsid of the virus (28)

1.13 Life cycle

EBV has two phases in life cycles lytic and latency. B-cells or the epithelial cells are infected during the lytic stages of EBV whereas the B-lymphocytes predominantly get infected in latency stages. Lytic cycles are the active infection spreading stages they comprise of viral amplification, around 80 EBV proteins express and production of virions. BZLF1 is a transcriptional activator that promotes the expression of virus in the early genes (29).

Where lytic infection cycle is actively disease-causing latency stages are complicated and is preparation phase for EBV. Latent cycle induces cell proliferation and immortalization as shown in figure 1. In the infected cells the viral genome replicates and exists as circular episome at a constant copy number. There are three latency stages I, II, III all categorised based on the protein expression pattern as seen in EVB induced tumours (30).

Latency I: the only viral protein expressed is the Epstein-Barr nuclear antigens EBNA1 in NPC

Latency II: along with EBNA1 three other latent membrane proteins are also expressed; LMP1, 2A and 2B.

Latency III: infection in primary rest B cells takes place. All six EBVNA; EBVNA1, EBVNA2 EBVNA3a EBVNA3b EBVNA3c and EBVNA-LP along with previously expressed LPM. post-transplant lymphoproliferative disease (PTLD) and AIDS-associated diffuse large B-cell lymphoma is associated with EBV (29).

1.14 Oncogenesis progression

The first human virus to be identified as cancer causing is EBV. It is classified as group 1 oncovirus (25). Usually, EBV immortalise in B cells in vitro. Commonly the virus resides in the cells without causing any symptoms known as latency state of the virus. However, at any given moment the virus can revert to the active disease-

causing cycle known a lytic. During the lytic cycle EBV can cause tumours and turn oncogenic. In all EBV related cancer at molecular level viral gene products can be seen. Oncogenesis is triggered and promoted by the viral gene products through blocking of apoptosis, creating genomic instabilities, production of uncontrolled cell proliferation and migration. This results in tumour initiation and leads to tumour maintenance (31).

Common feature of all cancers is the typical display of mechanisms through which it escapes the immune recognition and promote tumour progression. EBV is known to only display a few proteins in initial lytic stage to avoid altering the immune system. In B-cells the virus wields immunomodulatory effects for silencing of anti-EBV effect causes by interferon-gamma (INF-γ)(18).

In additions, to these mechanism it also alters antiviral cytokines TNF-$\alpha, IL-1\beta, and IL-6$(27)

The viral cytokines are able to mimic the properties of IL-10 that allows the virus to escape the hosts antiviral response. An immunocompromised host system due to other medical conditions or inflammatory microenvironment promote the progression of the pathogenesis of the virus (11).

1.15 Treatment

EBV associated NPC is targeted by radiation and/or chemotherapy. 15-30% pf the NPC patients show weak prognosis and develop at various sites. 5-15% exhibit local failure. Available treatment has side effects. Therefore, development of novel therapies and targets with less side effects is the interest globally (13).

1.16 Bio-markers

There exist certain bio-molecules like RNA, DNA, peptides, proteins or any biochemical molecules that serve as a measure of normal or abnormal biological states

in a living body these are known as bio-markers (32).

The definition for bio-markers has been evolving over the time however, WHO defines bio-markers as "A bio-marker is any element, composition or activity that can be measured in the body or its effects and affect or foresee the prevalence of outcome or disease." (29)

In clinical terms a cancer bio-marker can be used to measure the probability of cancer development in a specific tissue, be used to prediction of progression of cancer and for prediction of response to the appointed therapy. Moreover, cancer bio-markers are being used to justify the use of therapeutic strategies by linking them to molecular pathway deregulation or cancer pathogenesis (33).

According to their functions bio-markers can be categories into two groups: discovery of mechanism of disease action and drug targets, for prediction, early diagnosis, and prognosis. Several articles cite both types of bio-markers for NPC (78,79,80,81). Recent advancement in the technology have led to development of novel bio-markers like circulating EBV DNA and EBNA levels (7).

1.17 NGA

The next generation sequencing (NGS) technology refers to non-Sanger based DNA sequencing methods which have replaced conventional sequencing methods. They have been vividly used for analyses of complete genome (whole genome sequencing), the coding exons within already reported genes (whole exome sequencing), and only coding regions of selected genes (targeted panel). Conventional sequencing methods were replaced by the non-sanger-based DNA sequencing method; the technology called next generation sequencing (NGS). This technology has been used for whole genome sequencing; a complete analysis of genome, the coding exons that area already reported and selective coding regions of the genes (20).

NGS technology is working to revolutionize the prospective to sort out genetics related to cancer, epigenetic and transcriptomic level studies. In a matter of days

information like mutations, copy number and somatic aberration at base pair level can be obtained. High-throughput chromatid assays and whole genome methylation assays can be performed (34).

1.18 RNA sequencing technology

The advances in science have led to the formation of RNA sequencing technology. It has proved be an indispensable tool for the Differential gene expression analysis and splicing of messenger RNAs. With the development of NGS, RNA-seq to has progressed. RNA-seq tools are available for studying many different characteristics including but not limited to single cell gene expression, the translatome and structurome of RNA. With time more advanced RNA-seq technologies are being developed to attain a fuller understanding of RNA and solving more complicated questions regarding biology (35).

1.19 Microarray

Microarray is used for data analysis for biological and other fields. To process massive data set produced from functional genomic experiments. Each microarray experiment can measure information from hundreds to thousands of gene simultaneously to generate unprecedented amount of biological data. However, general limitation microarray has the small percentage of the gene present on any given array are identified as differentially regulated (36).

1.20 Docking

Molecular docking or MD is formation of 3D structural complexes produced by the interaction between two or more interacting molecules. The main purpose of producing 3D structures is to observe the interaction between two or more molecules. MD is mostly used in drug formation or improvement. Molecules and uncomplicated

entrance to structural databases has befallen vital system. Molecular docking has greatly contributed to structural molecular biology and structure-based drug discovery. Docking has been substantially facilitated by tremendous growth in computers and ever-growing access to public molecular databases. The central objective of molecular docking to predict molecular recognition, discovering likely binding modes and predicting binding affinity energetically. Docking is dispatched between a protein and a ligand. Molecular docking has tremendous application in drug discovery structure activity analysis, lead identification and optimization, binding hypothesis to clear the way for mutagenesis, chemical mechanism studies (37)

Chapter 2

LITERATURE REVIEW

2.1 Nasopharyngeal Carcinoma

Nasopharyngeal carcinoma (NPC) is a malignancy of the nasopharynx epithelium. NPC remains infrequent worldwide, however patterns of endemic have been reported in North Africa, East and southeast Asia. Depicting that the NPC follows an epidemiological pattern and genetic susceptibility in certain regions. Irrespective of the geographical distribution NPC is significantly different from other cancer due rigorous infiltration of the immune system. Another clinically important characteristic is its low degree of differentiation (38).

Head and neck cancer (HNCs) like NPC are ranked as the sixth most common type of carcinomas globally (39). The sensitivity due to the location of these cancers makes the deadly if not caught at an early stage. The prognosis and survival rates are poor worldwide making HNC very difficult to deal with. Nevertheless, the most recent improvements in the chemotherapy and radiotherapy have improved the quality of the given treatment (40)

2.2 Epstein Bar Virus

EBV infection is substantially associated with NPC, shapes the tumor microenvironment through chronic activation of immune system directly impacting the prognosis and therapeutic outcomes. Furthermore, it is clinically observed that an alarming amount of supporting stromal cells are often admixed with the malignant cells in the epithelium. (18).

The Epstein Barr virus (EBV) belongs to the Herpesviridae family; subfamily Gammaherpesvirinae commonly known as Human herpesvirus 4 (HHHV4). The

EBV's unique mode of action is residing in the B cells of the immune system. The virus has latency stages which allows it to be dormant while it resides in the immune cells without alerting the normal immune reaction in a carrier but non infected person. Any disruption in the cycle which can be caused by environmental or genetic factors leads to EBV associated tumors in B-cell (41).

The Epstein-Barr virus is one of the most commonly present viruses throughout the world. EBV is diffused between different ethnic regions causing several non-malignant infectious diseases; infectious mononucleosis (IM) and post-transplant lymphoproliferative disease (PTLD) and various malignant diseases; Burkitt's lymphoma (BL), nasopharyngeal carcinoma (NPC) Gastric cancer (GC), Hodgkin's lymphoma (HL) and Lung carcinoma (LC) (42)

2.3 EBV Association With NPC

Nasopharyngeal carcinoma (NPC) is the pathogenesis of epithelial tumors mainly caused by EBV however, the role EBV plays in the progression of the disease remains unclear. Previous studies have established a link between presence of latent viral infection in cells with existing premalignant tendencies due to genetic changes. Even though the pinpointed participation of EBV isn't known, however, the presence of the virus in all tumors cells creates hopes to produce novel therapeutic and specific diagnostic tactics for treatment (43).

Literature suggests that Epstein-Barr virus (EBV) infection takes place before the malignant cell cloning indicating that tumour cells already carry monoclonal viral genomes. EBV is repeatedly observed in high grade invasive lesions in NPC (26). NPC is related to multiple genetic changes with recurrent chromosomal deletions and hypermethylation of promotors of specific genes residing on the chromosomes 3p, 9p, and 11q (44).

Repeated evidence collected through literature review have indicated a strong association of EBV with development of NPC (31). EBV infection in early life is

typically related to high incidence area, is usually critical. (32, 33). Although 99% of NPC are linked to EBV however, EBV alone is not the only contributing factor. Globally all adults and children are infected with the virus or are carriers yet relatively a small population of individuals develop NPC forcing a conclusion that other factors supporting NPC like genetic mutations and environmental factors (45).

2.4 Detection of Malignancy of HNC

Cancer detected at advanced stage often tend to reappear and metastasis making the 5-year survival rate relatively low as compared other types of cancers. Liquid biopsies have recently been used to detect the biomarkers in HNCs. Some of the common biomarkers in NPC are miRNA, EBV DNA and certain active proteins. Liquid biopsies are done in hopes of early diagnosis to improve patient's disease-free survival chances (46).

Although the process of liquid biopsies is least invasive and safer for the patient, there are certain downsides to each of these biomarkers. For the Circulating tumor DNA (ctDNA) to be circulating in the bloodstream the chances are that the cancer has spread and reached a level of chronic illness. Other limitations include high rates of false positives and low true positives. This process is relatively new leading to equipment limitations and lack of experienced operators. Therefore, there is still a need for development of effective early diagnostic techniques using the biomarkers present in serum of HNCs (47)

2.5 Genetics

Long noncoding RNAs (LncRNAs) have observed to contribute to the devolvement of HNCs. For clinical diagnostic and prognostic application, a promising approach is detection of LncRNAs in serum and, or other protein present in bodily fluids. Biomarkers like; serum MALAT1, AFAP1-AS1, and AL359062 have been

reported to have been used for detection of NPC. LOC284454 LncRNAs is associated with molecular functions and regulation of genes that are significantly associated with NPC, oral cancers, and thyroid cancers. The upregulation of the LOC284454 is directly linked to the intensity of each of these cancers. These upregulations are used for clinical diagnostics and can be evaluated (48)

2.6 Prognosis

Despites the improvements made over the decade to control disease and to increase the survival rate there still remain high mortality rate in NPC patients. Additionally, the unfavourable effects of radiotherapy, high risk of drug toxicity and probable resistance to chemotherapy are constant risks for the NPC patients in general deal with. Considering these adverse effects there exists a scope for personalised and specific sensitive therapy for NPC patients (49).

Precision therapy and prognosis that focuses on special circumstance for each patient. Such treatments rely on validated biomarkers with improved capability to screen, diagnose and monitor tumours to provide patient specific treatment. Using techniques like bioinformatic analysis, gene chips and RNA sequencing have shown promising results in comprehensive screening of biomarkers along with development of a technique exposing details of underlying role of biomarkers in pathology of cancers (50).

Another study identified hub DEGs, DNALI1, RSPH4A, RSPH9, DNAI2, using analysis and PPI revealed that these genes are closely related and are involved in the key pathway linked with NPC. The literature emphasises on the need to study to host and viral factors responsible for causing NPC in respect to different ethnicities. Moreover, a better understanding is required to develop better therapeutic targets that respond well to therapeutics high in efficiency (47)

2.7 RNA Sequencing Analysis

During the revolution of diagnosis and molecular targeted therapy of several diseases including multiple types of cancer, RNA sequencing analysis has played a crucial in identification of genes. Previous studies have reported many novel genes identified through RNA-seq analysis that elucidated how such genes can be used in diagnosis and treating for nasopharyngeal cancer. Baoyu He and his team in China predicted the prognosis for NPC in clinic basing the prediction on RNA seq analysis outputs (51).

RNA-seq is helpful in the early treatment decisions for its specific ability to detect not only early mutations, but also mutations that pose a high molecular risk. The detection of mutations aids in the discovery of novel cancer biomarkers and possible therapeutic targets. It also helps in monitoring of progression of disease and in early treatment decisions through targeted therapy. RNA-seq can be used to explore the application of Tumor mutation burden (TMB), which is considered a good biomarker for immune checkpoint therapy and prognosis (52).

2.8 Microarray Analysis

Utilizing the Human Genome Project, microarrays provide a versatile platform for monitoring the RNA expression to benefit human health. A microarray experiment is based on the arrangement of known and unknown sample of DNA (53).

RNA expression levels for hundreds of thousands of genes can be monitored by microarray technology not only for primary tumor but for cell line data as well. Data obtained from the studies based on tumor and cell lines has revolutionized cancer diagnosis. The basic working principle of microarray is expression or lack of expression of identified genes that elucidates the functions of the genes involved in the tumor progression (54)

A demonstration by Golub et al who used a series of acute myeloid leukemias'

(AML) to determine RNA expression profiles using the approach of expression differences using a discovery program that automatically identified the distinction between different leukemia groups. Using this very program Alizadeh et al identified a new class of leukemia case (53)

Chapter 3

MATERIALS AND METHODS

3.1 Materials and Methods

For the fulfillment of objectives illustrated in chapter 1, a workflow was plotted. The targets of this research are listed in figure 3.1. The pattern of the workflow involves identification of therapeutic targets of Nasopharyngeal carcinoma, docking of existing ligands against the selected protein and development of a machine learning model for the designation of the prime ligand from the catalogue of ligands against any protein based upon the binding affinity.

3.2 Identification of Therapeutic Targets

Identification of therapeutic targets is to distinguish differentially expressed genes. The upregulated or down regulated genes were termed as differentially expressed genes. In each disease there were a few genes whose expression level is divergent in contrary to a normal condition of a human body. Such genes are earmarked as therapeutic targets. For this motive microarray and Next Generation Sequencing are the two profound approaches. The approaches for the identification of therapeutic targets of nasopharyngeal cancer are listed in figure 3.1. Agilent the state-of-the-art microarray, mRNA seq and microRNA seq were implemented.

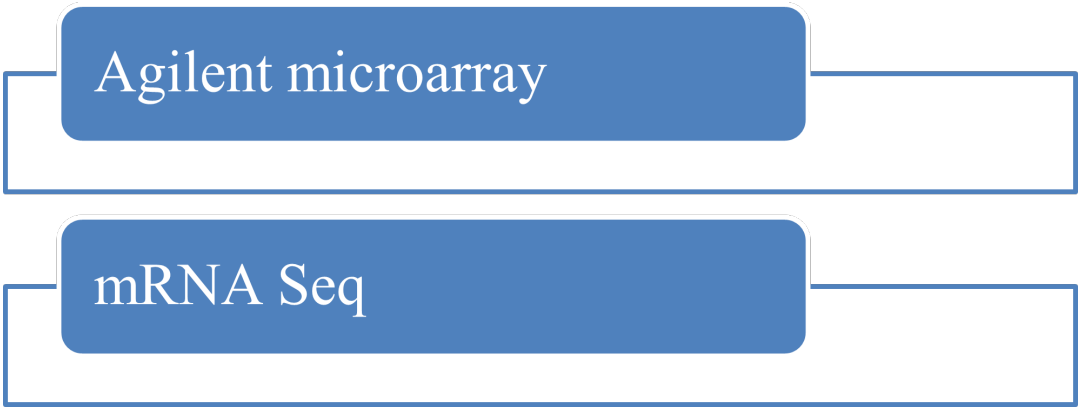

Figure 3.1. Approaches for identification of therapeutic targets of Nasopharyngeal cancer

3.3 Agilent microarray

Agilent microarray involves two color sample hybridization of oligonucleotides. Two different fluorescent samples were used for hybridization and measurement of differential expression against the relative abundance of the hybridized oligonucleotide. The workflow for microarray is described in figure 3.2

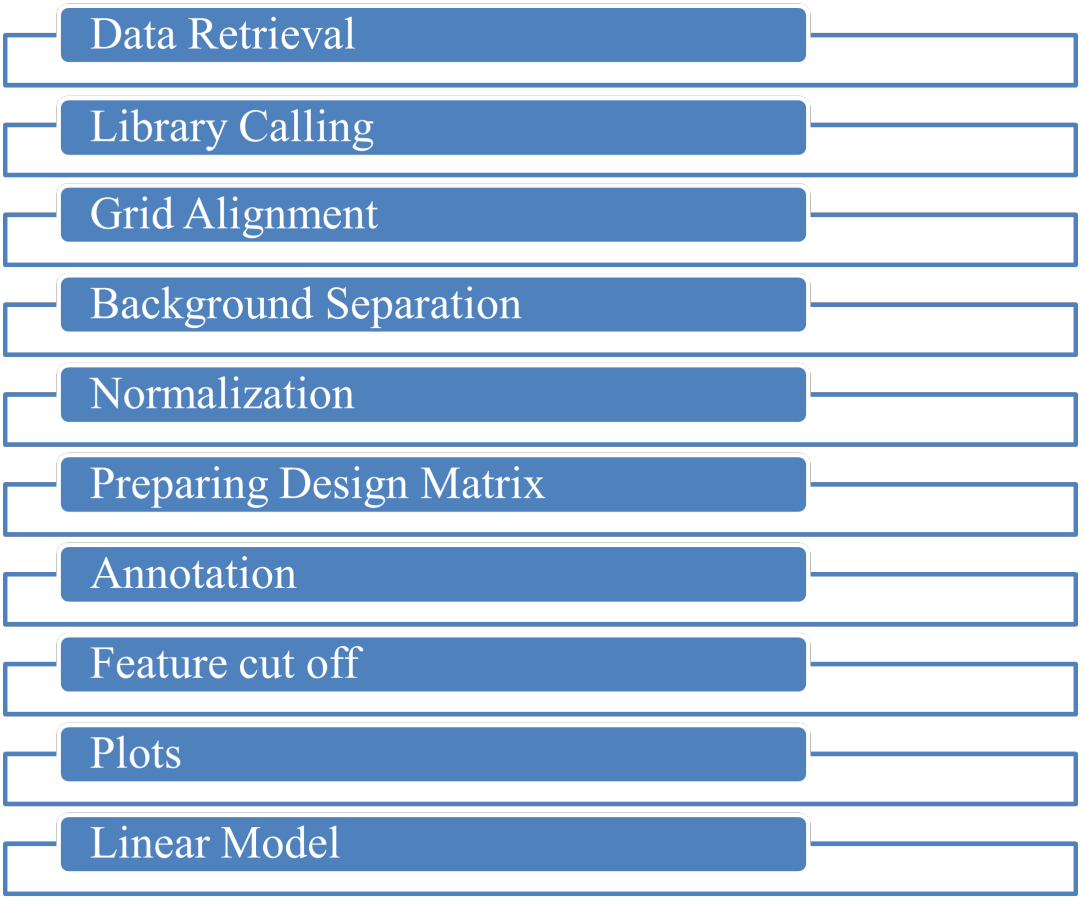

Figure 3.2. Workflow of Microarray

3.3.1 Data Retrieval

The first step for the measurement of differential expression is retrieval of dataset. There are two ways to retrieve the data. In wet lab by direct sequencing of organisms that are to be analyzed or by using the databases that contain the required sequences.

In this case GEO database was used to retrieve the data.

GEO Database

GEO abbreviated as Gene Expression Omnibus is an online, international communal repository. It freely distributes microarray and next generation sequencing and various forms of high throughput genomics data put forward by the research community. The main objectives of GEO are

- Data Organization
- Submission Guide
- Query Analysis

It has a platform record that is composed of a description of the sequencer or array, a table describing the array template. Every platform is assigned a unique GEO accession ID (GPLxxx). A sample record defines the conditions in which every individual sample was handled and manipulated and measurement of abundance it had gone through. Each sample is designated with a unique GEO accession number (GSMxxx). Series record combines related samples and furnishes the focal point and interpretation of whole analysis. It is composed of tables that describes the summary conclusions. A dataset represents an organized collection of statistically and biologically comparable samples based on GEO suite of data representation and analysis tools. A Profile provides a measurement of gene expression of every single gene across samples provides two platforms GEO datasets and GEO profiles. GEO datasets

provides a study-based platform for the researchers to perceive the study of their interest. However, GEO profiles offers gene level database that provides the expression measurement of genes.

The link to GEO database is: https://www.ncbi.nlm.nih.gov/geo/

3.3.2 Library Calling

Microarray analysis for differential expression is mainly executed employing various packages of R language. It is based on BiocManager project built in R language that provides numerous tools for microarray data analysis. The required libraries for the analysis are installed in R language using the command "BiocManager::install("library name")" and called "library(library name)" respectively. Multiple libraries are required for this purpose table 3.1 provides a list of libraries along with their corresponding function. The libraries that will be used for microarray analysis are discussed below

Preprocessing of microarray data is regulating raw intensities of microarray that its biology significance starts to make sense

- Grid alignment represents microarray data as a 2D array of spots by registering uneven intensities with the 2D image content
- Background correction adjusts data for ambient intensities revolving around each feature of microarray data
- Normalization is fine-tuning microarray data that is as a result of technical biases as opposed to biological biases

The microarray data is preprocessed to exclude all the technical biases. The biological meaning of the data can be interpreted through statistical processes.

Preparing Design Matrix

Microarray data after preprocessing is all set for analysis that involves the construction of design matrix. Design matrix specified RNA samples were pertained

to each channel on every array. Design matrix leads to the construction of contrast matrix that allows the comparison between RNA samples.

Annotation

Microarray data should be annotated. Annotation is the description of microarray samples. Microarray data is annotated that depicts which biological entities are portrayed on microarray chip. This step is essential for inferring the biological significance of the data.

Feature Cut off

Differentially expressed genes are identified by choosing an appropriate cut off. The crucial goal of microarray is the distinguish of differentially expressed genes. The identification is based on plumping of a threshold value and ruling out all the genes below it because their expression is normal. The genes crossing the cut off value corresponds to abnormal expressions and they are considered as differentially expressed gene.

Plots

The microarray data is plotted in the form of heat map, histogram and boxplot. To visualize all microarray intensities of all the gene R provides suitable package which plots the gene expression and they can be analyzed in the forms of various plots.

Linear Model

After all the above mentioned steps are performed linear model is built based upon design and contrast matrix. It helps in analysis for comparing the intensities.

3.4 mRNA Seq

mRNA Seq is a 6 step process that was performed for the identification of DEGs. The general workflow for the identification of DEGs is shown in Figure 3.3

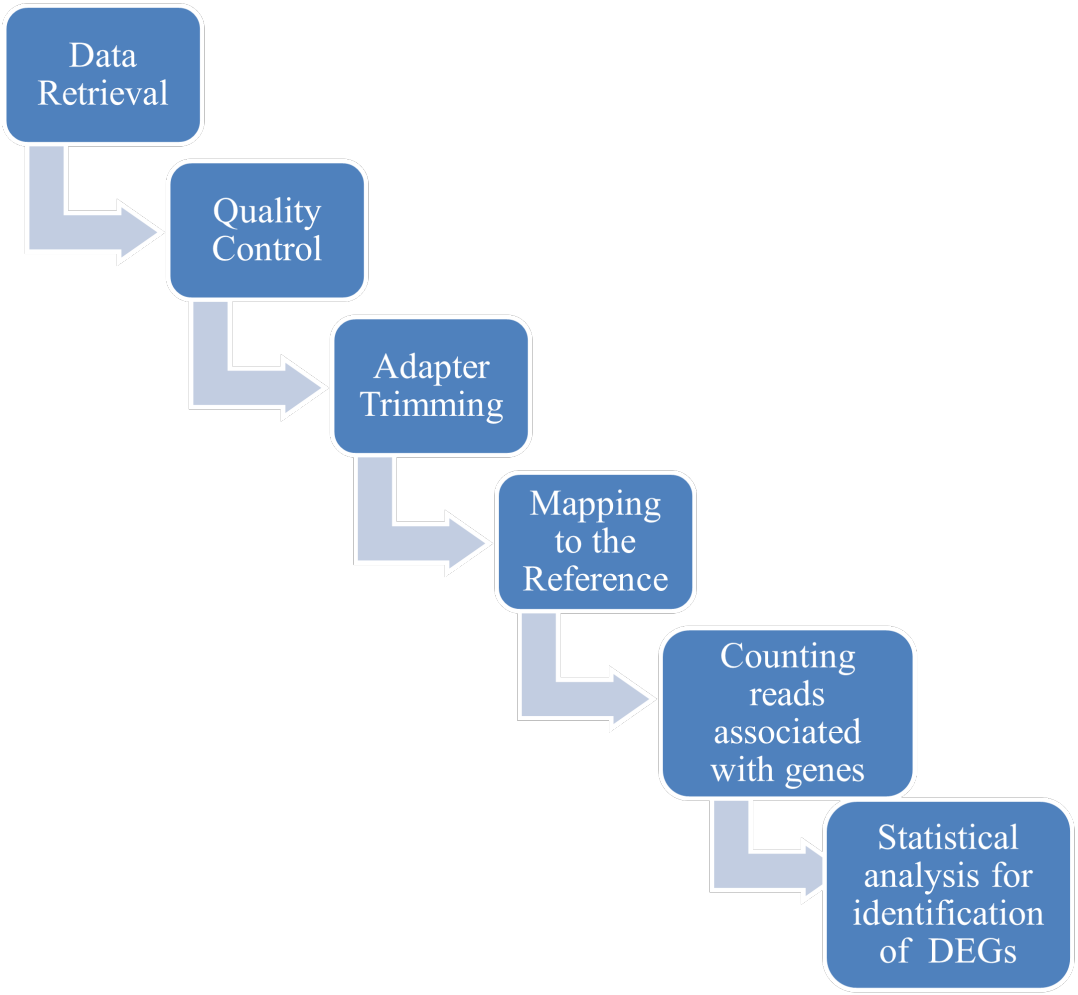

Figure 3.3. RNA Seq Workflow

The workflow that was adopted for the identification of DEGs along with the tools used is shown in figure 3.4

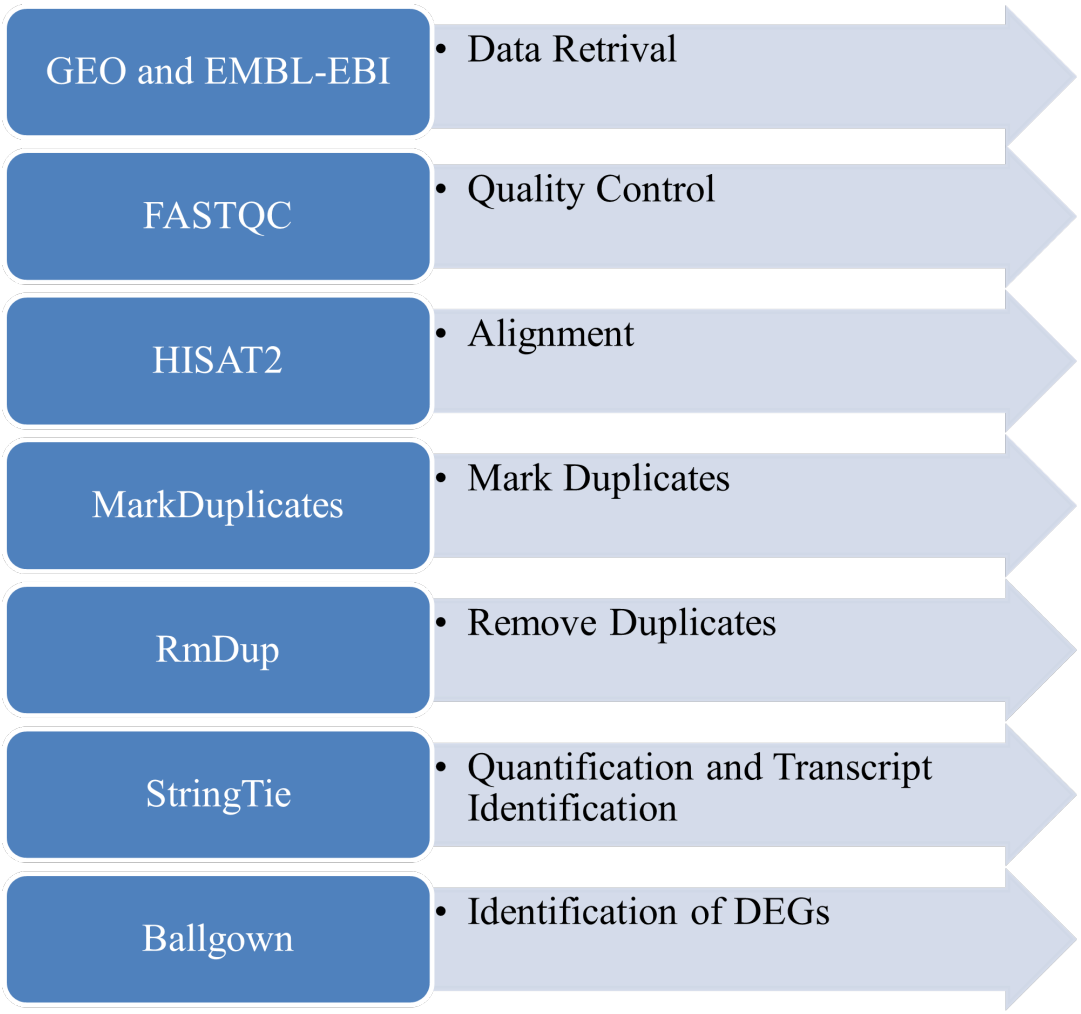

Figure 3.4. RNA Seq Workflow and Tools

3.4.1 Data Retrieval

The very first step of mRNA Seq is retrieval of data. There are numerous ways to get genomic sequences in wet lab or dry lab. In wet lab direct sequencing is possible on the other hand from dry lab a variety of databases are available that contain genomic sequences. For this research two data sets will be retrieved from two databases "GEO" and "EMBL-EBI".

EMBL-EBI

EMBL is European Bioinformatics Institute is an international, interdisciplinary and innovational of open data for life sciences. EMBL abbreviated as European Molecular Biology Laboratory. EMBL-EBI provides free bioinformatics services and data, research programs in personal genomics and computational translational bioinformatics, hands on bioinformatics training, industrial partnership and ELIXER hub.

The link to EMBL-EBI database is: https://www.ebi.ac.uk/

The information about GEO database is provided obove

3.4.2 Quality Control

Posterior to data retrieval, the data is progressed through quality control operation. Quality control is performed to scrutinize the quality of data and for the removal of technical biases. In NGS data there are abundant factors that influence the analysis. The flaws can be in the sequencing platform, the sequencer, the nucleotides. The major obstacle is the assurance of reliability of the incorporated nucleotide in the corresponding read.

To perform this task FastQC will employed that investigates the quality of the reads.

FastQC

FastQC examines the quality of raw and sequenced reads processed from the sequencing pipelines. It gives a set of analysis which provides an insight into the

quality of the sequencing process. The main task of FastQC is

- Imports data in variety of file formats like SAM, BAM and FastQ
- Provides an insight into problematic region of the data
- Visual representation to assess the reads in the form of graphs and plots
- Exports the quality results into an HTML report
- Automatic report generation is performed offline without processing interactive application

3.4.3 Alignment

After deriving data from preprocessing the succeeding task is to align the reads to the reference genome. This is performed to observe the resemblance between the data that we are analyzing and the reference genome. This done to reflect light on genetic diversity among the population for the identification of mutations, isoforms, SNPs

HISAT2 will be employed to perform the alignment between the reads under study and the reference genome which is Hg39.

HISAT2

HISAT2 is graph based genome alignment algorithm. It is a splice aware aligner that employs graph-based strategy, alignment algorithm and indexing algorithm. It implements hierarchical indexing which incorporates global index and local index.

- Global index covers the whole genome
- Local indexes encompass 56kb regions covering whole genome

Initially HISAT2 uses global index to map the read to the genome and further uses local index to map the read that refines the location of the read mapping. Local

index comes into play where the splicing strategy comes in. As local index covers short genomic regions so it can easily detect where the read is spliced. This twostep indexing algorithm helps makes HISAT2 as a fast and memory efficient algorithm.

3.4.4 Mark Duplicates and Duplicates Removal

RNA seq depends on PCR for the amplification of oligonucleotides for the sequencing process to initiate. PCR results in the expansion of reads but not every single read is amplified equally. This results in some reads to be overrepresented giving an insight that expression level of the resulting gene is higher in quantity. When it is not the case. Such sequences are termed as overrepresented sequences. Such sequences should be excluded from the analysis. Computational removal of these duplicates based on their alignment coordinates is important. These duplicates will be marked and then removed by markduplicates and Rmdup respectively.

MarkDuplicates and RmDup

MarkDuplicates tags duplicate reads in SAM/BAM files. The tools work by comparing sequences at the 5 prime positions of both reads in SAM/BAM files. After the duplicates reads are accumulated, they are distinguished between primary and duplicate reads by ranking the sum of their base quality score. MarkDuplicates locate the duplicated reads and rank them based upon their base quality, Rmdup then removes the low base quality overrepresented reads .

3.4.5 Quantification and Transcript Identification

In human body introns are removed and selective exons are retained from pre-mRNA resulting in RNA splicing. This splicing leads to different versions of RNA referred as transcripts and isoforms. Majority of the human genome is alternatively spliced which leads to the expression of a single gene to different types of proteins. These proteins have unique functions and involve in different pathways and implicated in numerous diseases involving cancer. Identification and quantification

of transcripts is of paramount importance because they are translated to different proteins. Isoforms presence and quantity varies in different samples that leads to novel biomarkers. Different biomarkers lead to level biological process that is not detectable at gene level.

Transcript Identification and quantification will be done through String Tie.

StringTie

StringTie identifies gene isoforms in the data and computes their abundance. StringTie assembles transcript fragments using splice map reads and infers isoforms for each sample. StingTie algorithm implements network flow algorithm and denovo approach. SringTie works mapping the reads to the reference genome and constructing a graph for all the possible isoforms of the gene. The splice graph possesses nodes representing exons and the path between the nodes represents splicing sites. The reads are classified into clusters, constructing a graph for each cluster leading to transcript identification. It creates a individual flow to estimate expression using maximum flow.

3.4.6 Identification of Differentially Expressed Genes

The quantification of the expression level of gene is followed by statistical procedure to test the difference between the quantification values of the samples. To consider a gene differentially expressed there should be a statistically significant difference between the read counts or quantification values between two experimental conditions. Statistical methods are employed for gene expression. Differential gene expression patterns are approximated by statistical methods. The genes are short-listed based upon the expression change cutoff and score threshold.

Ballgown will be used to identify differentially expressed genes.

Ballgown

Ballgown is a R package designed to perform statistical comparisons between transcripts. Its underlying model in general linear model approach. It contains built in visualization routines for displaying transcript abundance and structure.

3.5 Identification of Common Genes

Microarray, mRNA seq and miRNA seq results in DEGs. The ultimate goal is to detect common genes among all the dataset of microarray, Mrna seq and miRNA seq. The utmost purpose of extracting common genes is to further is pathway analysis and selection of protein.

The tool that will be employed to extract common genes among all the datasets was Draw Venn Diagram

Draw Venn Diagram

Draw Venn Diagram is an online tool that calculates list of intersecting elements. A textual output is generated that indicates intersecting and unique elements. When the lists are lesser than 7 a graphical output is also generated in the shape of venn diagram. The user can choose between non-symmetric and symmetric diagram. Intersection between a maximum of 30 lists can be calculated. The graphical output can be downloaded in PNG or SVG format.

The link to visit Draw Venn Diagram is: https://bioinformatics.psb.ugent.be/webtools/Venn/

3.6 Pathway Analysis

A schematic representation of organized, well characterized segment of physiological machinery at molecular level. In general, a pathway is describing as model where an extracellular signaling molecule results in activation of a specific receptor triggering a chain of molecular interactions. The ultimate goal of pathway analysis is the identification of genes within a known pathway corresponding to a specific pathological condition.

The set of common genes identified among all the datasets is all set for pathway analysis for the identification of genes/ proteins that ca be used for further analysis. Reactome will used for pathway analysis.

Reactome

Reactome is freely available open access, peer reviewed pathway database. The major unit of reactome is the reaction. The fundamental unit of Reactome is reaction Biological entities participating in a in biochemical reaction configure a network of biological interactions and are classified in reaction. The pathways that are possessed by reactome include intermediary, metabolism, classical, transcriptional, apoptosis, regulation and disease. The link to visit reactome is : https://reactome.org/ When it comes to pathway analysis Reactome a list of DEGs is provided as an input and it gives the output in the form of pathways list with the entities involved and p-value.

3.7 Selection of Therapeutic Target

Statistically significant pathway is shortlisted and the entities participating in the pathway are analyzed. The analysis of the genes participating in the pathway results is the selection of 1 gene that has

- Experimentally determined structure with a good resolution
- Domains
- Reference from literature about the involvement of the gene is he corresponding disease

The gene possessing the above mentioned credentials is brought forward in analysis

3.8 Docking

Docking is a type of bioinformatics modelling that allows the interaction of two or more than two molecules to form a stable adduct. Molecular docking involves a target and a ligand. Docking predicts the 3D structure of the complex. Variety of adducts are formed and they are queued using the scoring function based upon the total energy of the system. The general workflow of docking is represented in Fig 3.5

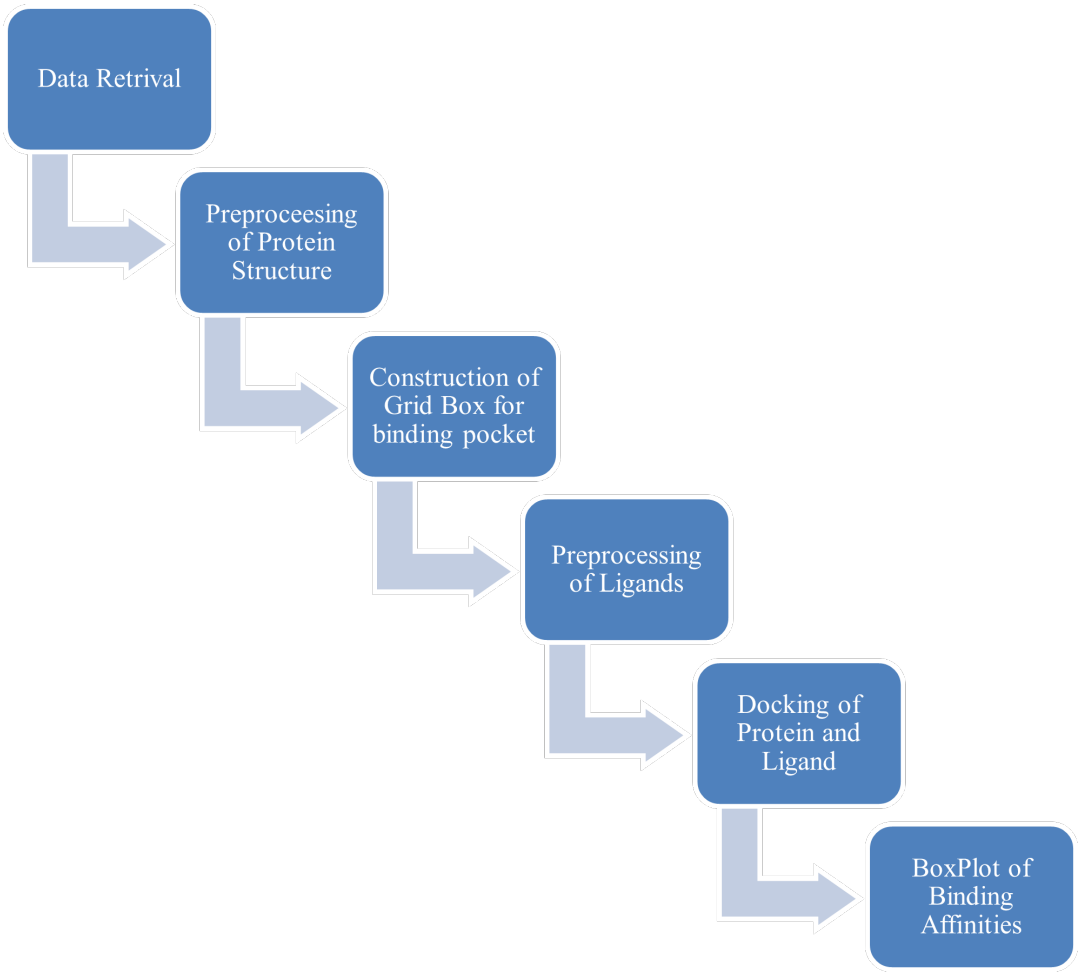

Figure 3.5. Docking Workflow

3.8.1 Data Retrieval

Molecular docking requires 3D structure of target (protein) and ligands in the form of PDBQT. The 3D structure of protein will be extracted from UniProt and ligands from databases like DrugBank and BindingDB.

UniProt

UniProt also known as Universal Protein Resource is database for protein sequence and annotation. The UniProt databases are a collection of UniProt KnowledgeBase, UniProt Reference Clusters and UniProt archive. The 3D structure of molecular docking will be retrieved form UniProt

The link to UniProt is: https://www.uniprot.org/

Binding DB

Binding Database (Binding DB) publics experimental data on the non-covalent interactions of the molecules in solution with the help of world wide web. The core unit of bimolecular systems, information of host-guest supramolecular systems is added overtime. The aim of Binding DB is to ease drug discovery, self-assembling system assembly and predictive compute models development. The ligands smilies will be retrieved from Binding DB

The link to visit Binding DB is: http://bdb2.ucsd.edu/bind/index.jsp

DrugBank

DrugBank is an online, free, comprehensive database on drugs and drug targets where bioinformatics meets cheminformatics. It possesses chemical, pharmaceuticals and pharmacological data.

The ligand similes will be extracted from DrugBank: https://go.drugbank.com/

3.8.2 Preprocessing of Protein Structure

The preprocessing of molecular structure of target is illustrated in fig 3.6.

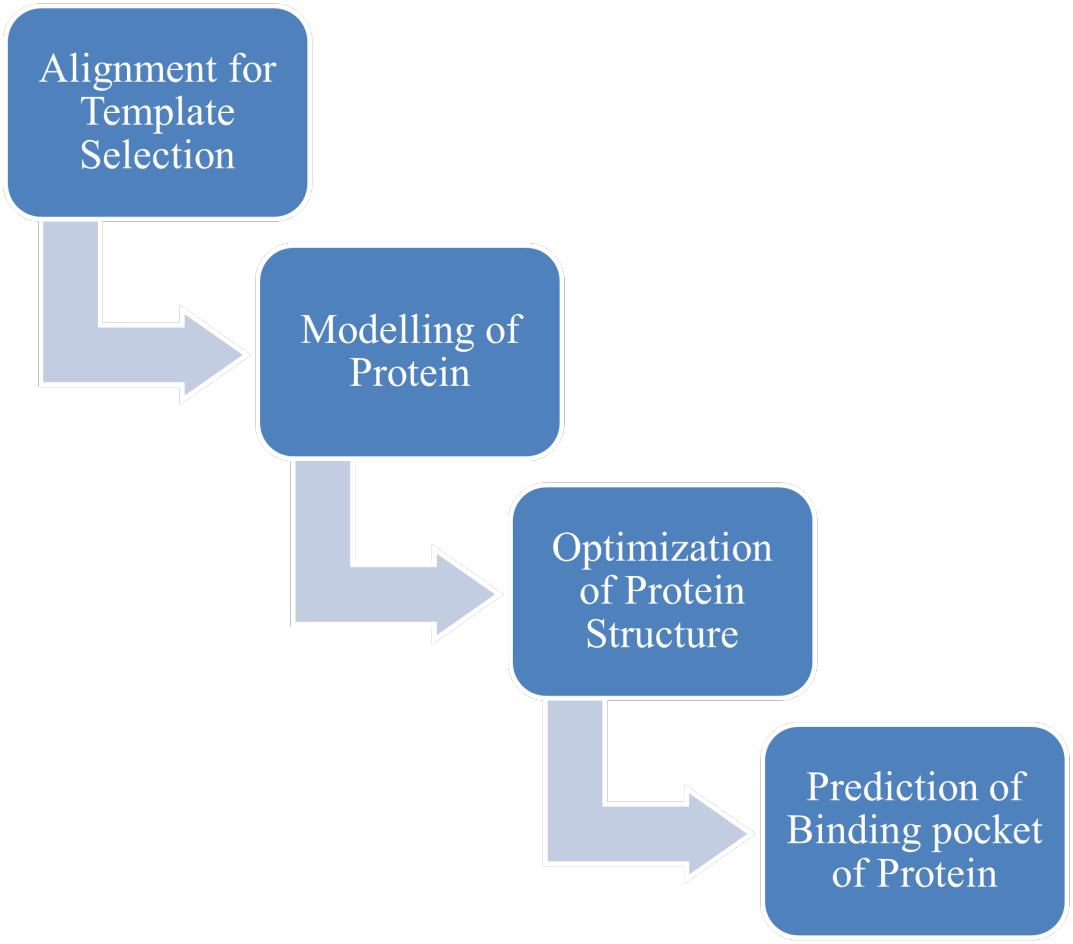

Figure 3.6. Preprocessing of Protein

Alignment of Target for Template Selection

Once the structure for protein is obtained from UniProt the next task is the selection of a suitable template for the target. BLASTp of the protein domain is performed.

BLAST investigates regions of similarity between the biological sequences. The tool compares sequences of protein or nucleotide to sequence database and computes the statistical significance. BLAST searches for a suitable template for the target in the relevant databases and outputs the templates with the maximum identity.

Modelling of Protein

Based upon the template sequence the next task is the modelling the protein based upon the template sequence. Swiss Model will be employed to model the protein.

Swiss Model is an online and automatic homology modelling server accessed through Expasy Web Server. The purpose of Swiss modelling is to make homology modelling of proteins of all life sciences possible.

The link to Swiss Model is:

https://swissmodel.expasy.org/

Optimization of Protein Structure

The model of the protein once generated is optimized. Optimization is done by adjusting a suitable pH of the protein structure and removing water molecules.

Optimization will be done by Play Molecule

Play Molecule is a freely available, online tool which allows target identification and validation, lead discovery and identification.

Link to visit play molecule website: https://www.playmolecule.com/

Prediction of Binding Pocket

Binding pocket is the cavity on the protein surface or the interior of the protein that possess suitable characteristics for binding with ligand. The prediction of the domain of protein where the ligand will interact is a fundamental step in molecular

docking. This step specifies the residues of interaction between the target and the ligand.

DoGSiteScorer will be used for the prediction of binding domain of the protein.

Automated binding domain detection and analysis software that allows the detection of potential pockets in the protein structure. In addition, global properties, shape, size, and chemical properties of the predicted domains are calculated. Drug ability score is calculated based upon linear combination of the descriptors that include hydrophobicity, enclosure and volume.

The link to visit DoGSiteScorer is:

https://bio.tools/dogsitescorer

3.8.3 Preprocessing of Ligands

The preprocessing of ligands involves ligands similes conversion to mol2, 3D coordinates generation and addition of polar hydrogen atoms. Furthermore, from mol2 format the ligands are converted to pdbqt format.

The ligands are downloaded in the form of similies from the databases. They necessarily should be metamorphosed too mol2 format, 3D coordinates should be generated, and hydrogens should be added. So that they can be put forward for molecular docking for their conversion to pdbqt through autodock.

OpenBabel would be employed for the preprocessing of ligands.

OpenBabel is a online and free file format converter that coverts the files from one format to nearly all the formats and jump from one program to another.

The link to visit OpenBabel is:

http://www.cheminfo.org/Chemistry/Cheminformatics/FormatConverter/index.html

AutoDock is a suite that allows automatic docking options. The core task of autodock involves the prediction of small molecules such as drug candidates or substrates, interaction with a receptor of predicted 3d structure. AutoDock distribution

involves two software's AutoDock 4 and AutoDock Vina.

AutoDock 4 is based upon two main tools, AutoDock that allows the docking of the target to a set of ligands to a collection of candidate grids.

Autogrid precalculated the grids where the docking will be performed. In addition to docking the automated affinity grids can be analyzed. AutoDock Vina calculates the grids internally for the atoms that are required and this is performed virtually.

AutoDock tools possess GUI that aids to organize bonds that are treated as rotatable in the docking process.

3.8.4 Docking of Protein and Ligand

Docking involves the interaction of protein and ligand. Where the residues of active site of ligand will interact with the binding pocket of targets. In this operation numerous ligands will be docked to the target. This assessment helps to analyze the binding affinities of various ligands. Docking will be performed with the help of autodock vina.

AutoDock Vina

AutoDock Vina is an open-source tool that performs docking between the target and the ligand. AutoDock Vina usually implements AutoDock 4 tools that are much faster, accurate and easy to use. The depiction the output should not have a statistical bias referring to the conformation of the input. The tools detect the syntactic accuracy of the input and reports the errors. It verifies the invariance of covalent bond lengths automatically in the output. Artificial restrictions like number of atoms in the input, the size of the search space, exhaustiveness of the search and the number of torsions are avoided by the tool.

3.8.5 Boxplot of Binding Affinities

The ligands when docked with the target through autodock vina it yields an output of binding affinities. Lesser the binding affinity of the ligand the better. Foe the analysis of binding affinities of all the ligands visually boxplot is the foremost course of action. R provides a package to construct the boxplot of all the ligands through the package called ggplot2. It takes an xlsx files consisting of the binding affinities of all the ligands and displays it in the form of boxplot.

R language will be used for the construction of boxplot of the binding affinities of the ligands.

Chapter 4

RESULTS

Figure 4.1. Case 2 Petri Net Model: circles represent entities and squares represent transitions

4.1 Results

The therapeutic targets of Nasopharyngeal cancer were identified through microarray and mRNA Seq analysis. The identification of therapeutic target was done through the selection of DEG's. DEG's were identified using algorithm based in microarray and RNA Seq. the target identified through DEG analysis were then docked on the ligands.

4.2 Microarray

Microarray was used for the identification of differentially expressed genes. The outturn of microarray analysis of datasets for Nasopharyngeal cancer is discussed below.

4.2.1 Data Retrieval

The dataset with the GEO accession ID of "GSE53819" was retrieved from GEO database. Table 4.1 shows the parameters of the selected dataset. Table 4.1: Parameters of Dataset of Agilent dataset 'GSE53819'

R-studio is used for microarray analysis. The dominant feature of using R Studio is that it provides multiple features that empowers the user to visualize the

results. Numerous types of graphs and plots are designed to visually observe the results for validation of the results. The resulting plots are discussed below.

Boxplot

Boxplot is s systemized strategy of putting on view the dispersal of data on five feature summaries. The five number summary includes

- Minimum
- Second quartile
- Median
- Third quartile
- Maximum

Figure 4.1 represents the distribution of genes across the samples in the dataset of Nasopharyngeal cancer. Each box represents an individual samples. The dotted line represents the outliers. The division is based on first, second and third quantile

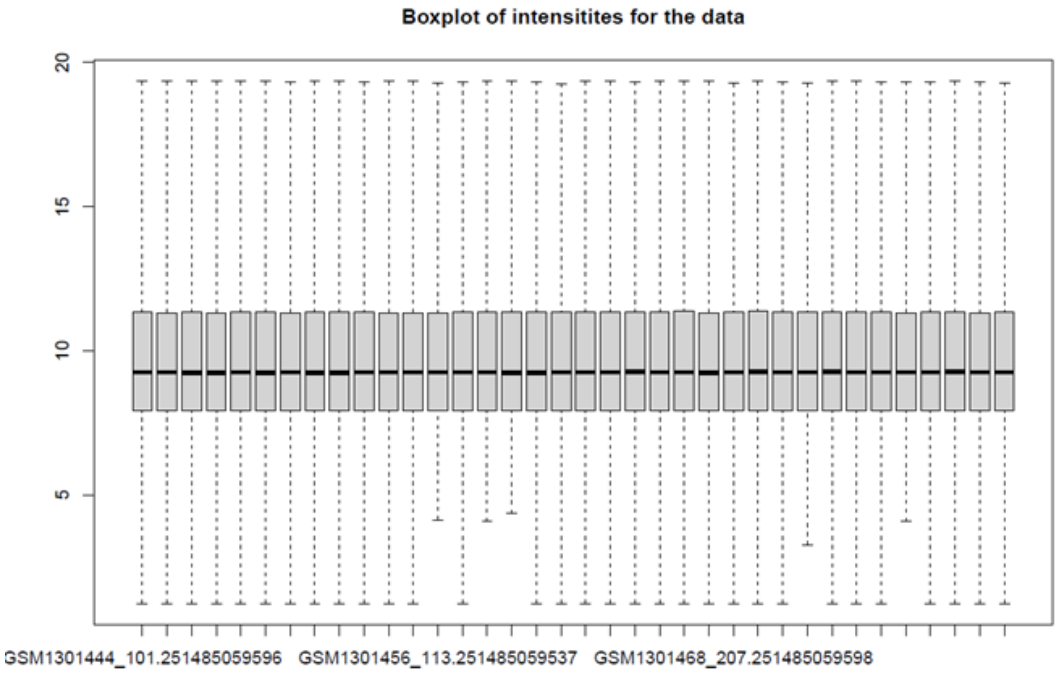

Figure 4.2. BoxPlot

Heatplot

Heatplot is a technique used for visualization of data that represents the immensity of a expression in term of color hues in 2 dimensions (2D). The disparity in color gives obvious hints to the observer about the clustering of data. In figure 4.2 the distribution of genes in their respective samples. Intensity of the gene expression can be noted form the plot. High expression is shown from the intensity of the color. The sharp red tones show high expression of gene. Keep that as a reference the variation of the tones towards slightly or more obvious lighters hues can be termed as lower expression of the genes. The trend in the below plot shows the variation of hue from deep red to dark orange to lighter orangish to yellow.

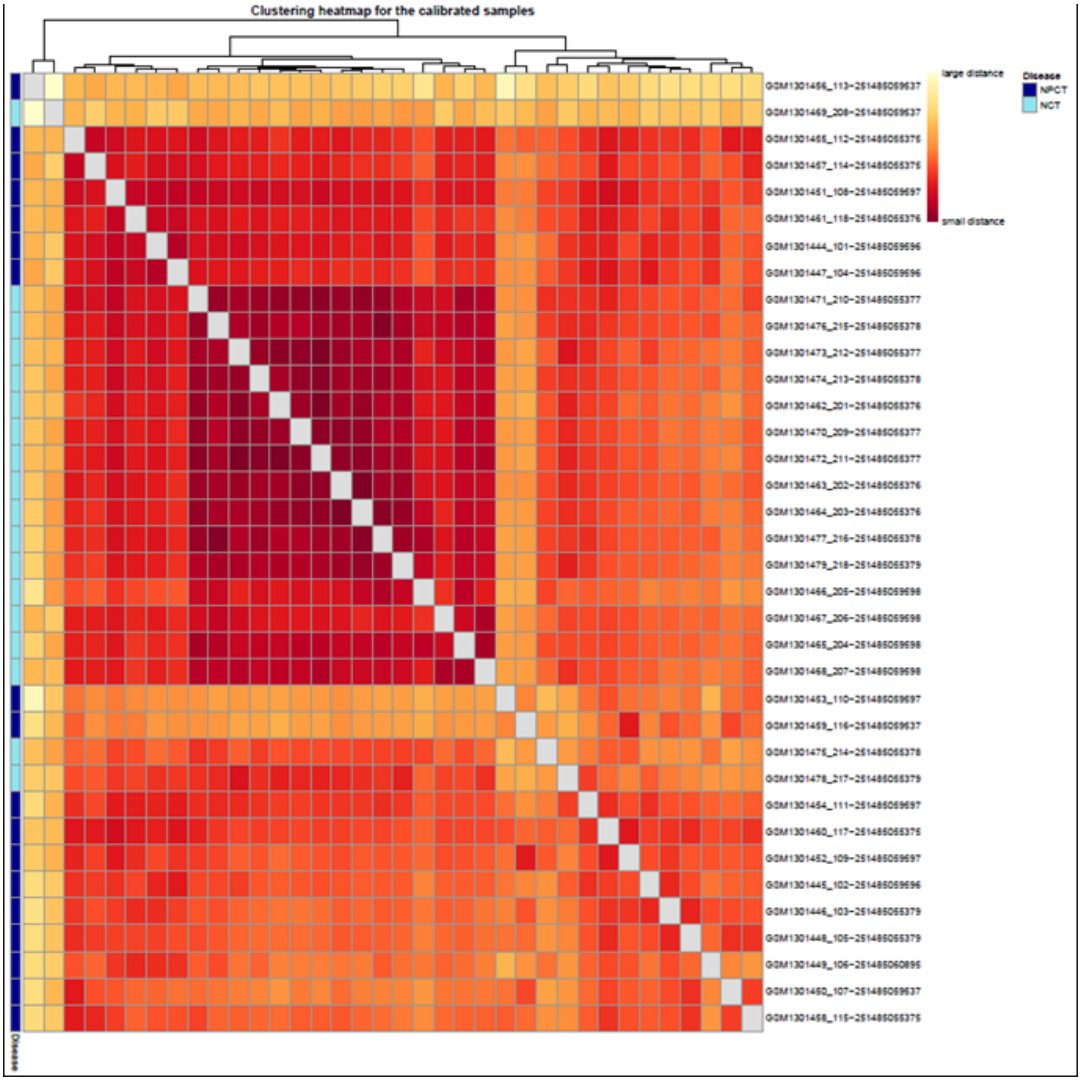

Figure 4.3. HeatPlot

Histogram

Histogram represents the distribution of data where the data is classified in continuous ranges. Below in figure 4.3 the distribution of gene expression across the sample is shown in the form of histogram. The upregulated and down regulated genes can be identified established on the stretch of the bar that depicts the median of gene expression.

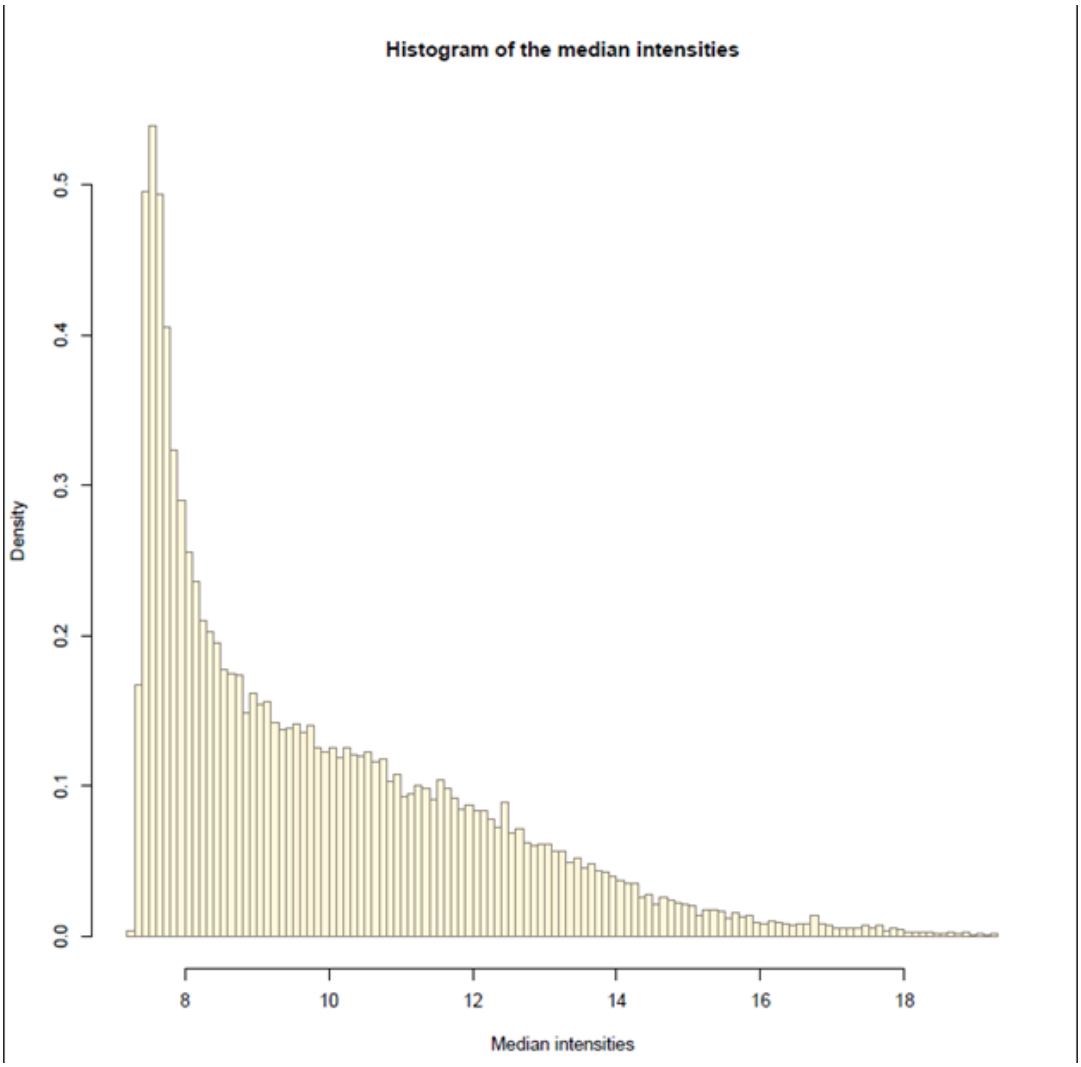

Figure 4.4. Histogram

PCA Plot

Principal Component Analysis portrays the recommendation of dimensional reduction. It transforms huge datasets in smaller sets still calibrating enormous information. In PCA plot the samples that cluster together represents similar characteristics. Figure 4.4 shows samples clustering. The sample close to each other represents similar gene expression.

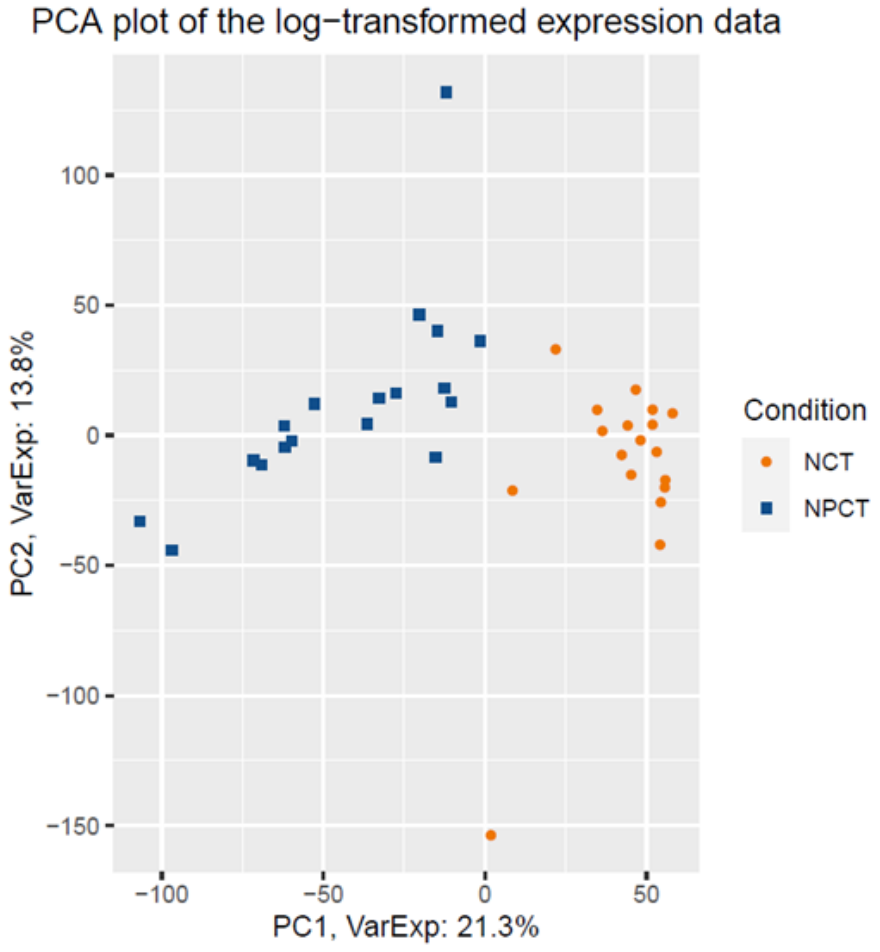

Figure 4.5. PCA Plot

RLE Plot

The Relative Log Expression plots are devised for visualizing outburst of variation in large sets of data of microarray samples. The variation of gene expression across the samples comparable to one another is shown in figure 4.5. The graphs show a diversity in the variation within the dataset. It represents the variation of expression of GSM1301456 is the highest variation to every sample analyzed and GSM1301471 shows the least variation.

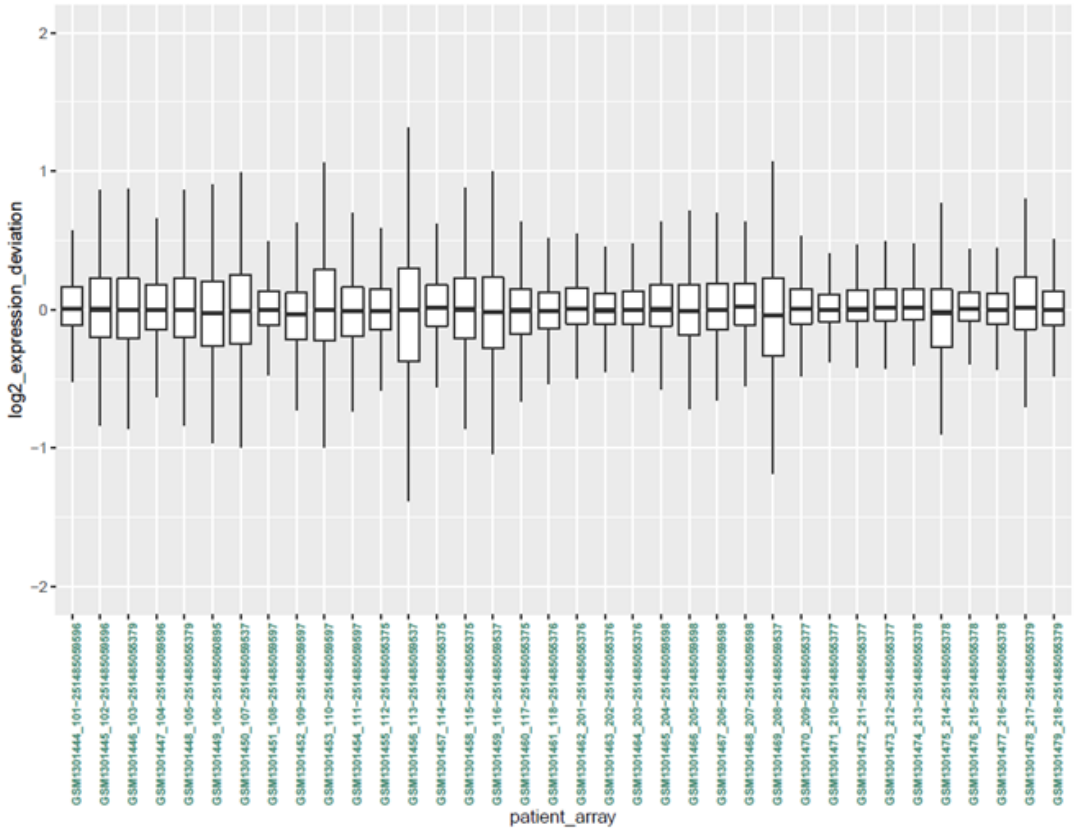

Figure 4.6. RLE Plot

Volcano Plot

Volcano plot is a scatter plot that depicts the statistical significance versus enormity of change termed as fold change. It empowers nimble identification of genes with huge fold changes and statistically significant. Upregulated genes lie in the top right corner whereas downregulated genes originate in the top left corner. The volcano plot in figure 4.6 represents the extent of differentially expression of genes in up and down regulated ends of the spectrum of microarray

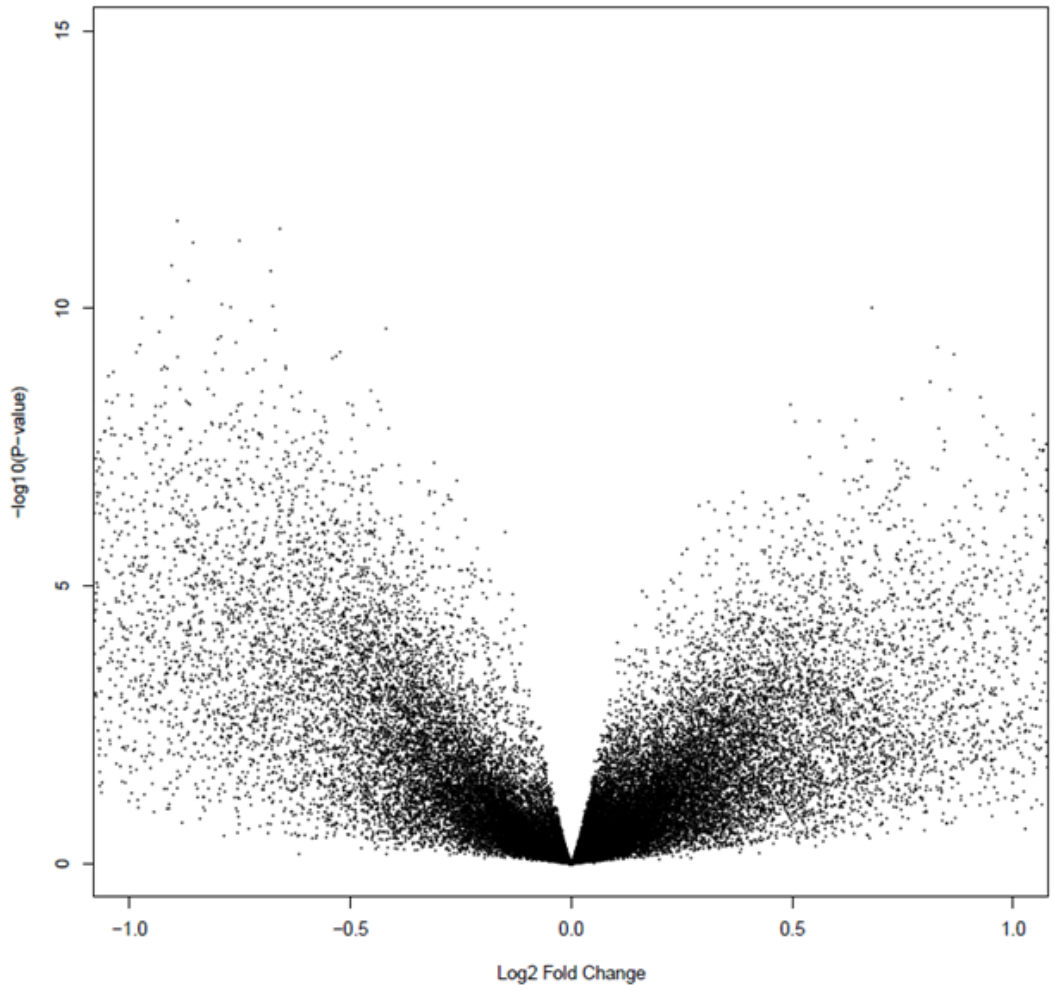

Figure 4.7. Volcano Plot

Enhanced volcano plot

The figure 4.7 shows an enhanced volcano plot. The spread of values from zero on both sides of the graph shows the transposing of genes expression towards abnormal expression. The left side of the plot indicates down regulated genes and right side points up regulated genes. As for the colors; the grey values on the plot show genes that didn't pass the threshold of FC. Blue values passed P-value threshold and the green values passed FC threshold. The most significant are seen in red which passed both the applied thresholds.

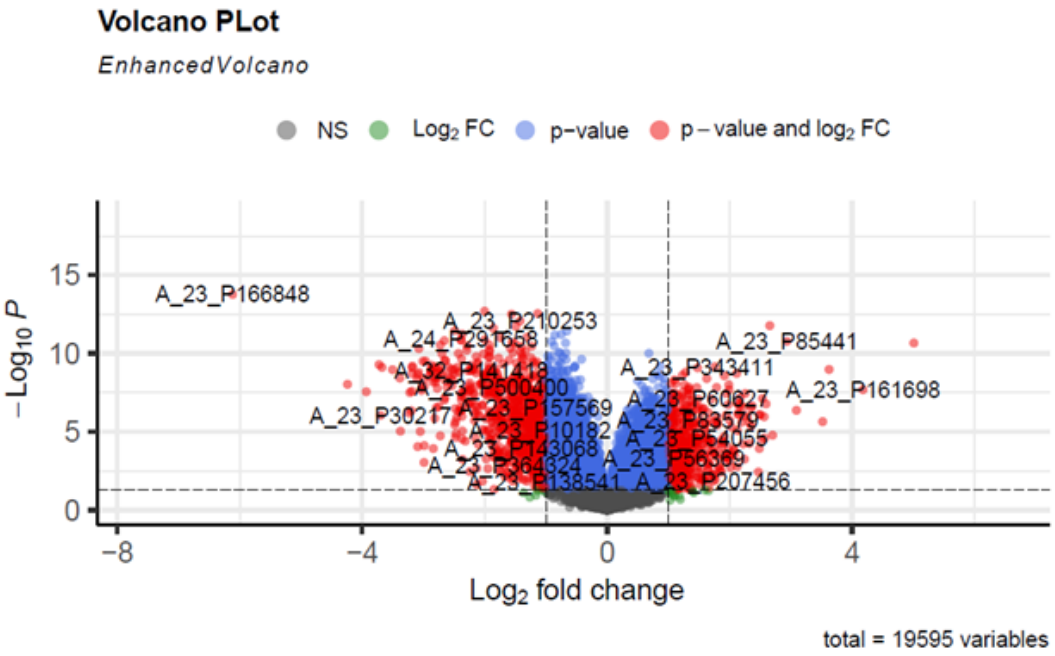

Figure 4.8. Enhanced Volcano Plot

4.3 mRNA Seq

mRNA seq is a procedure for the identification of differentially expressed genes. In addition to microarray mRNA seq was also protruded for the identification of therapeutic targets of Nasopharyngeal cancer. Two datasets RNA seq were manipulated for the identification of therapeutic targets.

4.4 mRNA Seq 1

4.4.1 Data Retrieval

GEO database was used to obtain secondary dataset for Nasopharyngeal cancer. The dataset with the GEO accession ID "GSE118719" was employed for analysis. The attributes of the data are shown in table 4.3

The DEGs were identified with the help of R package ballgown. As this dataset has 2 phenotypes. Ballgown is a package of R language that calculates the differential expression of both phenotypes.

BoxPlot

Boxplot of all the samples is represented in figure 4.7. The graph is between the samples and log2 of FPKM values. Each box corresponds to individual samples. The middle layers represent the medians that are not normalized. The start of the box represents minimum values and end of the box speaks for maximum values and the dots represent outliers.

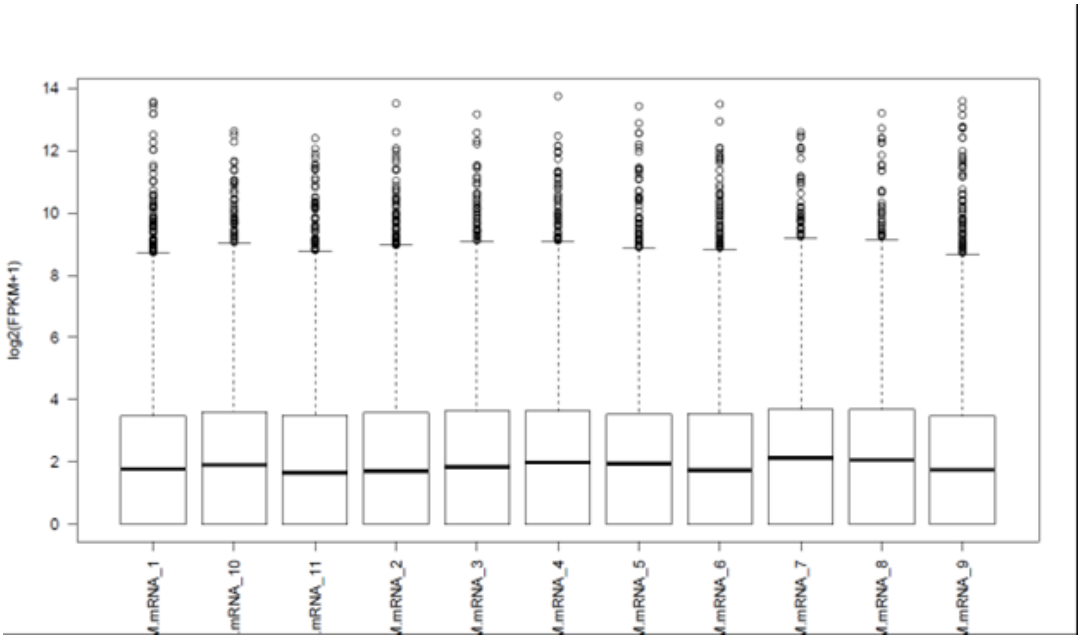

Figure 4.9. Box Plot

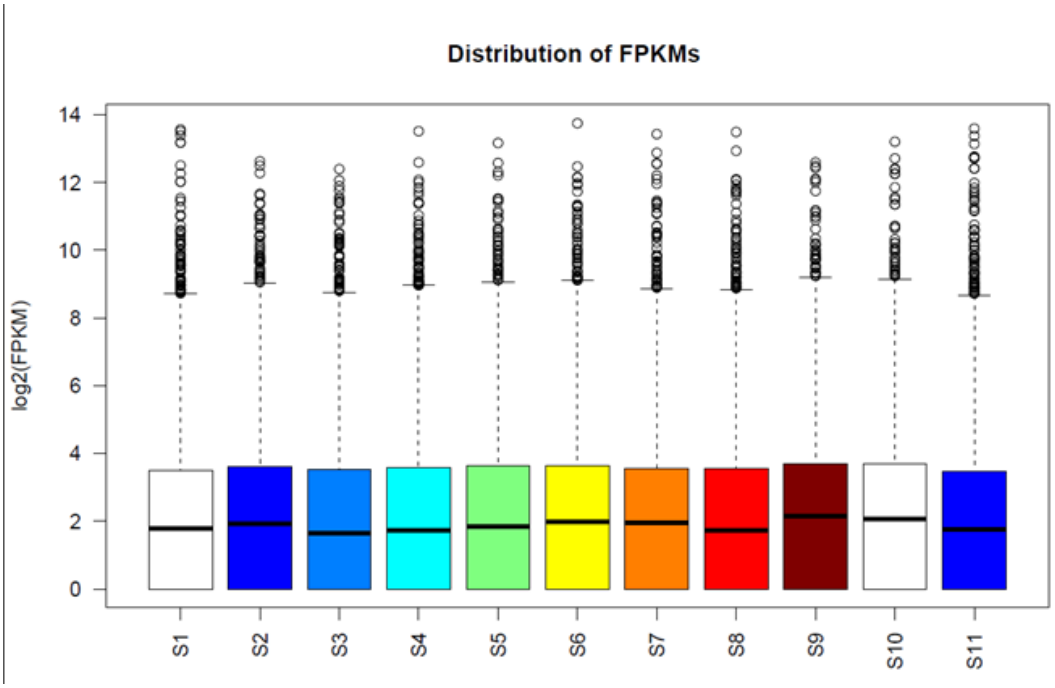

Figure 4.10. Box Plot

Bar Chart

Figure 4.9 represents the bar chart of differential expression and their frequency. In this bar chart a threshold value is chosen for up and down regulated genes. A threshold value of 2 was applied on both sides of the spectrum. The genes on the left side of the graph falling below the threshold line indicates down regulated genes and right side of the plot above the cut offline represents up regulated genes.

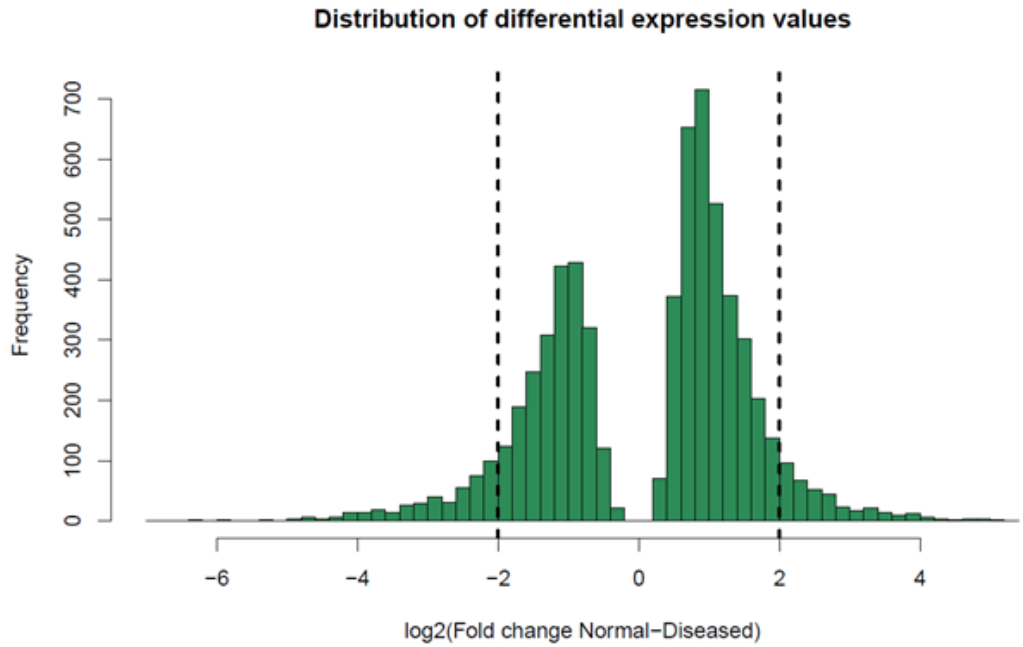

Figure 4.11. Bar Chart of Differential Expression

The distribution of transcript length versus their frequency is shown in figure 4.10. The first bar has the highest frequency and remaining transcripts are of shorter length depicting lower frequency.

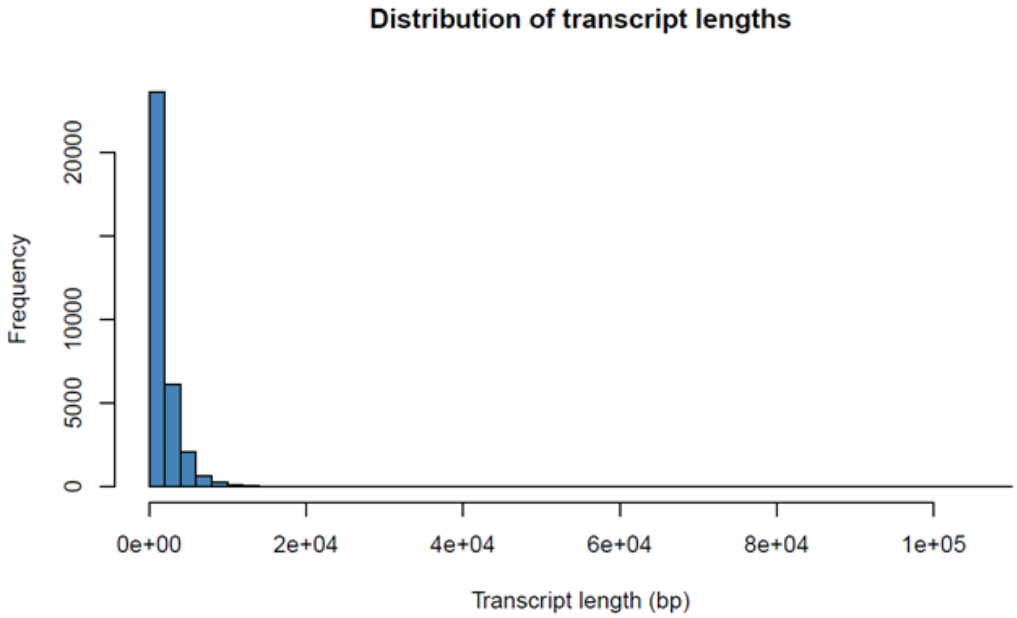

Figure 4.12. Distirbution of Transcript Length

Figure 4.11 shows the transcript distribution count per gene. It can be seen from the bars within the graph that most of the transcripts are extending to 5. Highest frequency of transcript goes as high as 15000>. The legend shows there are 16146 genes with one transcript, 5399 genes with more than 1 transcript and the maximum number of transcripts for 1 gene is 28.

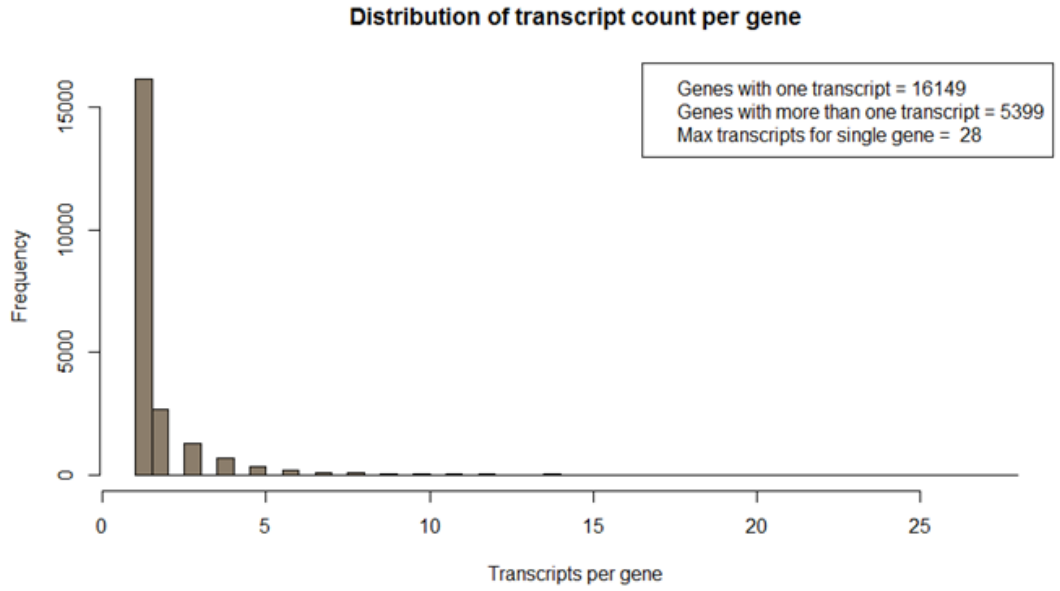

Figure 4.13. Transcript Distirbution per Gene

Enhanced Volcano Plot

Figure 4.12 shows the enhanced volcano plot of Nasopharyngeal cancer. As we move away from zero on both sides of the graph the expression of genes transposes towards abnormal expression. The left side of the plot indicates down regulated genes and right side points up regulated genes. As for the colors; the grey values on the plot show genes that didn't pass the threshold of FC. Blue values passed P-value threshold and the green values passed FC threshold. The most significant are seen in red which passed both the applied thresholds.

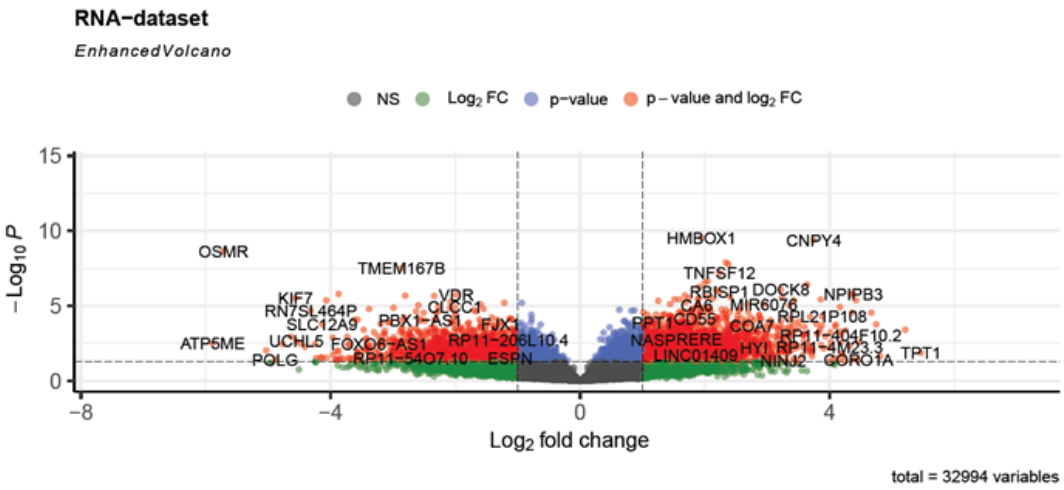

Figure 4.14. Enhanced Volcano Plot

4.5 mRNA Seq 2

Two datasets of mRNA seq were manipulated.

4.5.1 Data Retrieval

The second dataset of mRNA seq was also extracted from GEO database the acessinon Id "GSE68799". The parameters of the dataset are given in the table below

The retrieved data after progressing through the series of quality control procedures were aligned to the reference genome of Homo Sapiens that is hg38. The alignment rate of every sample is shown in table 4.8. Table 4.8: Alignment rate of samples to hg38

4.5.2 Identification of DEGs

To identify differentially expressed genes is the core objective of RNA seq. Ballgown performs a remarkable job of identifying DEGs.

BoxPlot

The boxplot in figure 4.24 represents the samples and log2 of their FPKM values. Each box represents the minimum, median and maximum values by its respective layers of each sample. The outliers are shown by the dots above the box. Figure 4.24: is a boxplot of all the samples FPKM values versus their log2 values. Each individual plot represents a single sample and the dost represent the outliers.

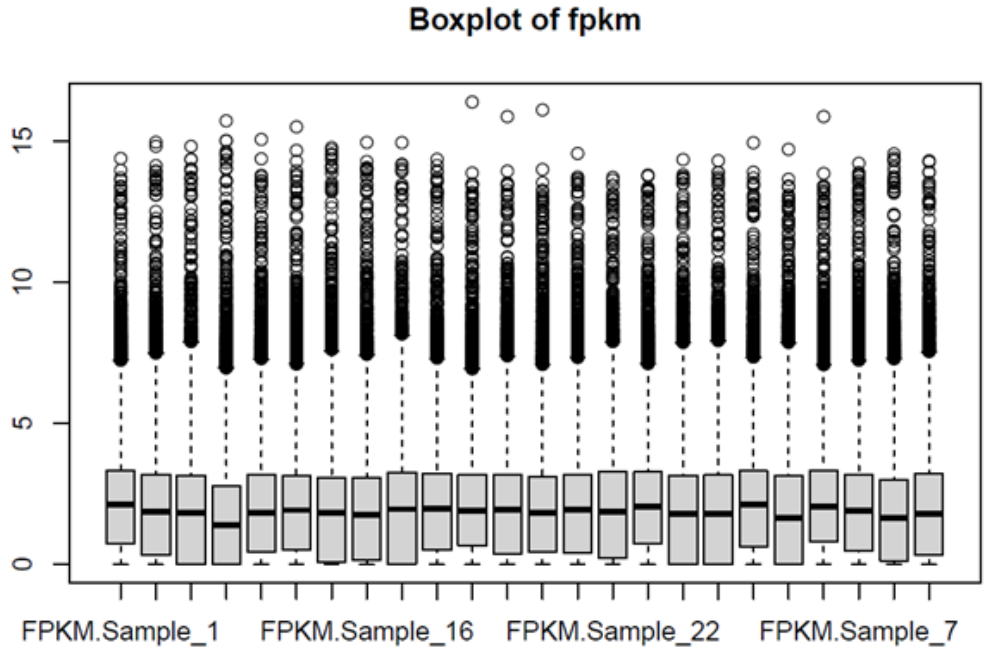

Figure 4.15. Box Plot

Figure 4.24 is a boxplot of all the samples FPKM values versus their log2 values. Each individual plot represents a single sample and the dost represent the outliers

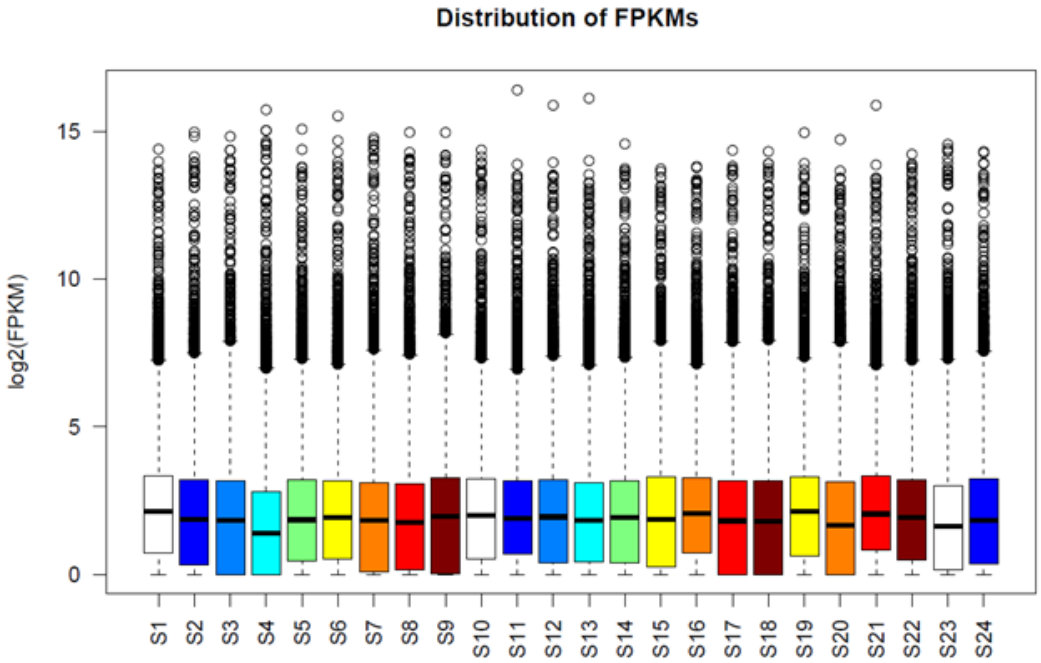

Figure 4.16. Box Plot

BarChart

Figure 4.25 represents the distribution of transcript count per gene. The transcript distribution count per gene. It can be seen from the bars within the graph that most of the transcripts are extending to 5. Highest frequency of transcript goes as high as 15000>. The legend shows there are 10871 genes with one transcript, 4515 genes with more than 1 transcript and the maximum number of transcripts for 1 gene is 24.

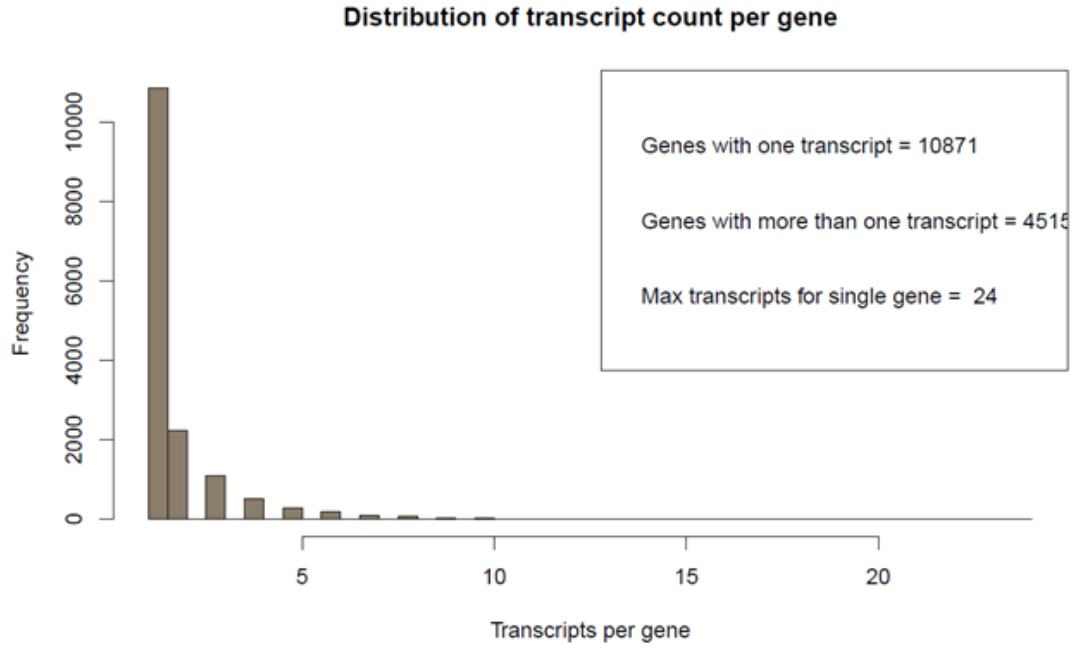

Figure 4.17. Transcript Count Per Gene

The graph of differential expression is shown in figure 4.26. A threshold value of fold change is selected for up and downregulated genes. +2 and -2 is the threshold value for up and down regulated genes respectively. The left side of the graph represents down and right side of the graph represents upregulated gene.

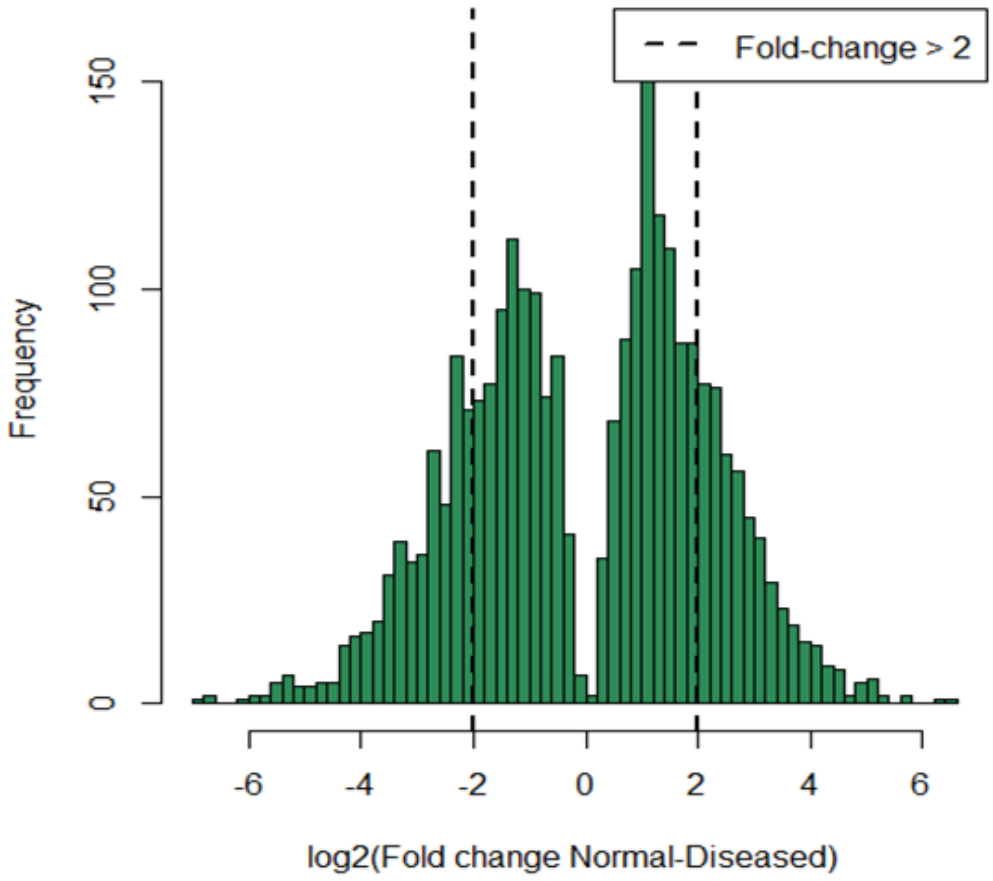

Figure 4.18. Bar Chart of Differential Expression

The distribution of transcript length versus their frequency is shown in figure 4.10. The first bar has the highest frequency and remaining transcripts are of shorter length depicting lower frequency.

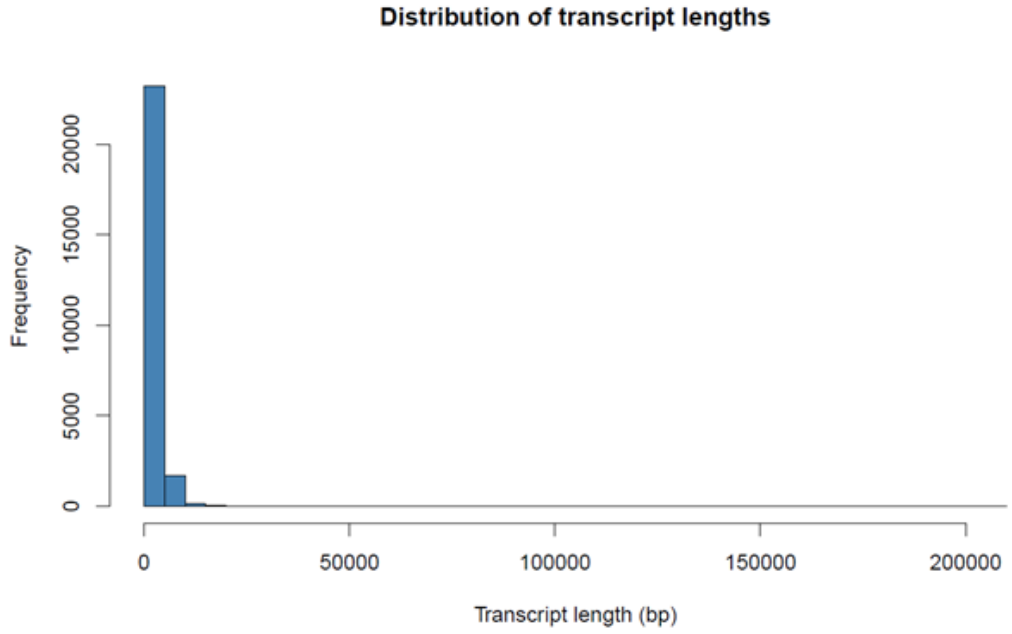

Figure 4.19. Distirbution of Transcript Lengtj

Figure 4.28 shows the enhanced volcano plot. The left and right extremes of the graph depict the down and upregulated genes respectively shown in red color.

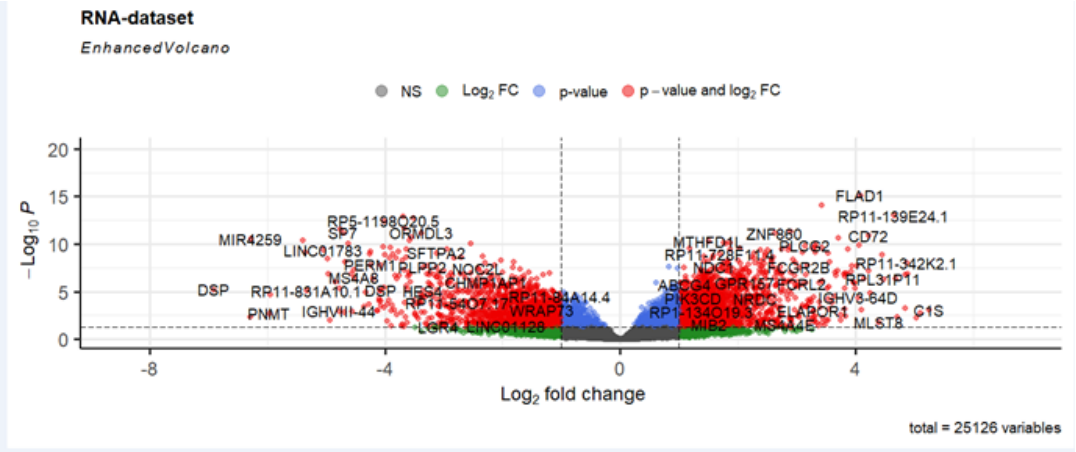

Figure 4.20. Enhanced Volcano Plot

4.6 Comparative Analysis

After the identification of DEGs in both the dataset of RNASeq and microarray comparative analysis was performed for the identification of common genes in all the datasets. 'Draw Venn Diagram" was used for this purpose. 3 different lists of DEGs were given as input. A total of 19 genes were found to be common among the 3 datasets. 285 genes were found between datasets 'GSE68799' and 'GSE118719'. 92 genes were found between datasets 'GSE68799' and 'GSE 53819'. 94 genes were found common between 'GSE118719' and 'GSE 53819'.

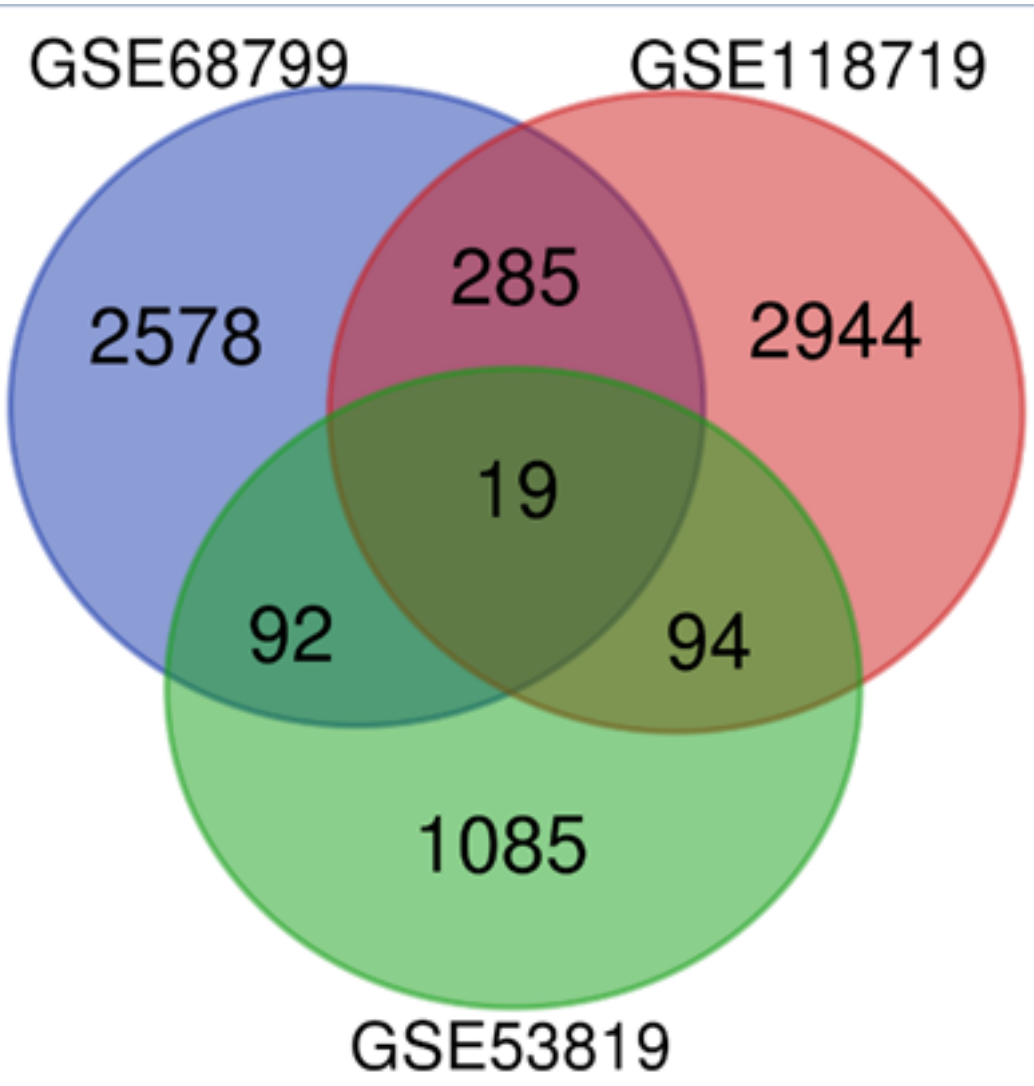

Figure 4.21. Common Gene

4.7 Pathway Analysis

Pathway analysis was performed on 19 common genes obtained because of comparative analysis. This was done to find out the pathways in which these genes are involved. The topmost significant pathways with respect to p-value and entities found are shown in table 4.8

4.8 Protein Modelling

The most significant pathways with the maximum number of entities and p-value less than 0.05 were shortlisted. The entities (proteins) involved in these pathways were validated from literature for their role in Nasopharyngeal cancer. The final gene should be disease causing and should be up regulated. Based upon these requirements Insulin-like growth factor 2 mRNA-binding protein (IGF2BP3) was finalized.

4.8.1 IGF2BP3

Insulin-like growth factor 2 mRNA-binding protein 3 is a protein that in humans is encoded by the IGF2BP3 gene. Multidomain RNA-binding protein, IMP3 (also called IGF2BP3), which contains six RNA-binding domains (RBDs): four KH and two RRM domains. IGF2BP3 mRNA levels were differentially expressed between NPC and adjacent normal tissues After the selection of protein and investigating the best suitable template for protein modelling through BLASTp “6GQE” was finalized with 100

4.8.2 Non-Covalent Interactions

The top 10 ligands that manifested paramount binding affinity with the protein were shortlisted. PLIP was used to discern the non-covalent interaction between the

ligand and the protein. The non-covalent interaction between the protein and the short-listed ligands is discussed below.

In the figure below the interactions are shown between IGF2BP3 and BDBM50128432 are shown. In the figure, different colors depict different aspects of the exchanges. Protein is represented through blue, ligand through orange, water through lilac, charge center through yellow, aromatic ring center through white, hydrophobic interactions through dotted lines and hydrogen bonds through sticks.

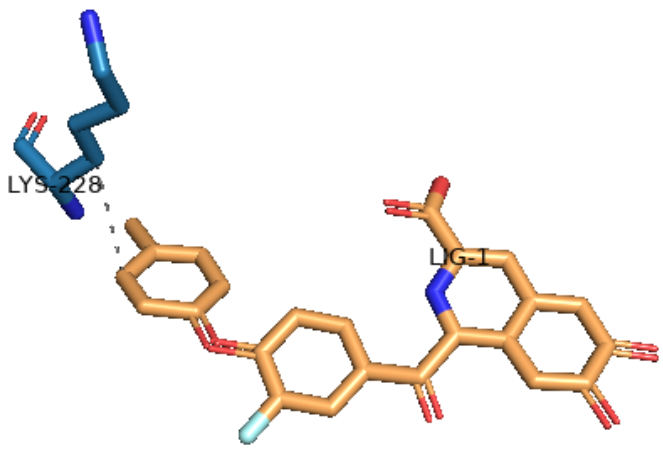

Figure 4.22. Non-Covalent Interactions with BDBM50128432

In the figure below the interactions are shown between IGF2BP3 and BDBM50128454 are shown. In the figure, different colors depict different aspects of the exchanges. Protein is represented through blue, ligand through orange, water through lilac, charge center through yellow, aromatic ring center through white, hydrophobic interactions through dotted lines and hydrogen bonds through sticks.

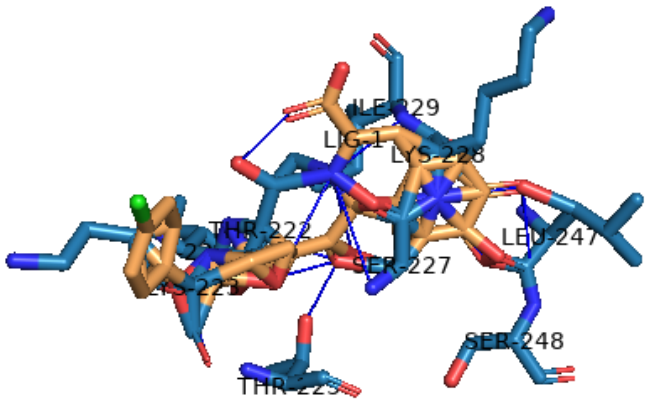

Figure 4.23. Non-Covalent Interactions with BDBM50128454

In the figure below the interactions are shown between IGF2BP3 and BDBM50106439 are shown. In the figure, different colors depict different aspects of the exchanges. Protein is represented through blue, ligand through orange, water through lilac, charge center through yellow, aromatic ring center through white, hydrophobic interactions through dotted lines and hydrogen bonds through sticks.

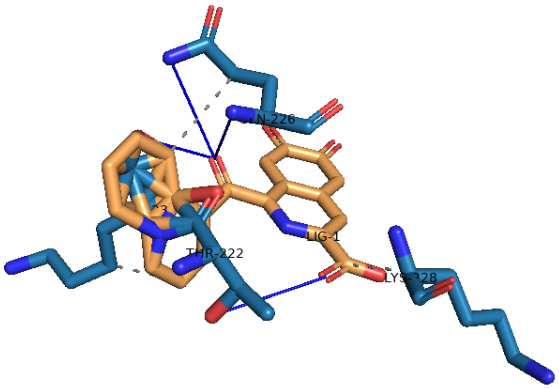

Figure 4.24. Non-Covalent Interactions with BDBM50106439

In the figure below the interactions are shown between IGF2BP3 and BDBM50128431 are shown. In the figure, different colors depict different aspects of the exchanges. Protein is represented through blue, ligand through orange, water through lilac, charge center through yellow, aromatic ring center through white, hydrophobic interactions through dotted lines and hydrogen bonds through sticks.

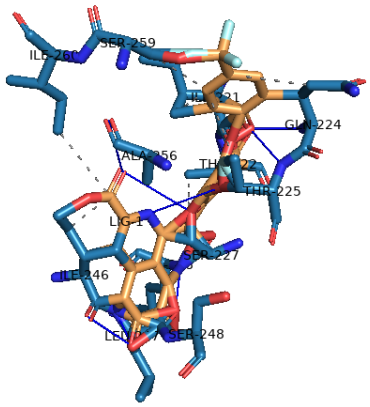

Figure 4.25. Non-Covalent Interactions with BDBM50128431

In the figure below the interactions are shown between IGF2BP3 and DB00619 are shown. In the figure, different colors depict different aspects of the exchanges. Protein is represented through blue, ligand through orange, water through lilac, charge center through yellow, aromatic ring center through white, hydrophobic interactions through dotted lines and hydrogen bonds through sticks.

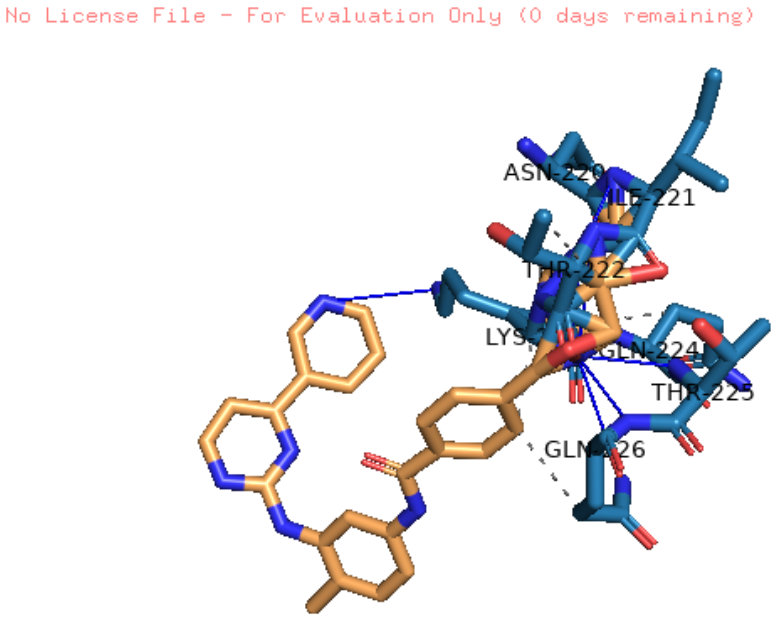

Figure 4.26. Non-Covalent Interactions with DB00619

In the figure below the interactions are shown between IGF2BP3 and DB11978 are shown. In the figure, different colors depict different aspects of the exchanges. Protein is represented through blue, ligand through orange, water through lilac, charge center through yellow, aromatic ring center through white, hydrophobic interactions through dotted lines and hydrogen bonds through sticks.

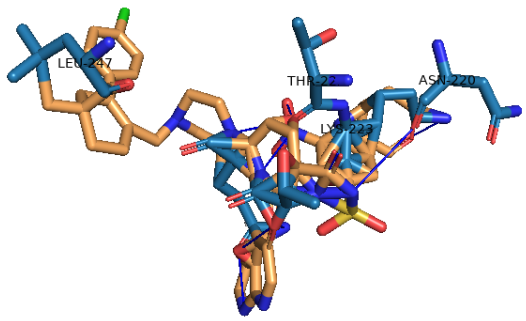

Figure 4.27. Non-Covalent Interactions with DB11978

In the figure below the interactions are shown between IGF2BP3 and BDBM50458517 are shown. In the figure, different colors depict different aspects of the exchanges. Protein is represented through blue, ligand through orange, water through lilac, charge center through yellow, aromatic ring center through white, hydrophobic interactions through dotted lines and hydrogen bonds through sticks.

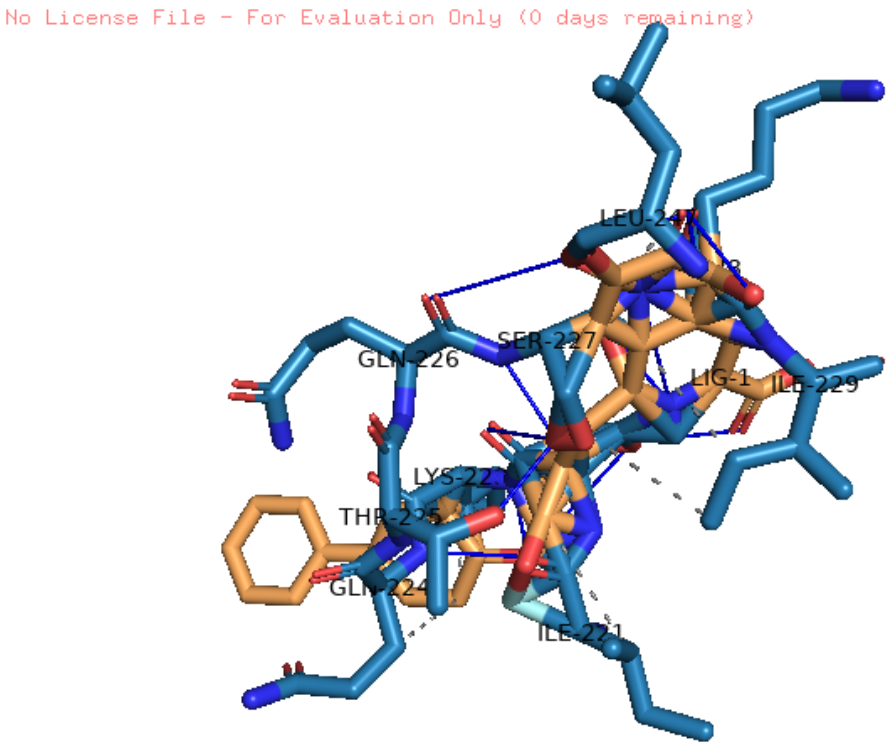

Figure 4.28. Non-Covalent Interactions with BDBM50128436

In the figure below the interactions are shown between IGF2BP3 and BDBM50128431 are shown. In the figure, different colors depict different aspects of the exchanges. Protein is represented through blue, ligand through orange, water through lilac, charge center through yellow, aromatic ring center through white, hydrophobic interactions through dotted lines and hydrogen bonds through sticks.

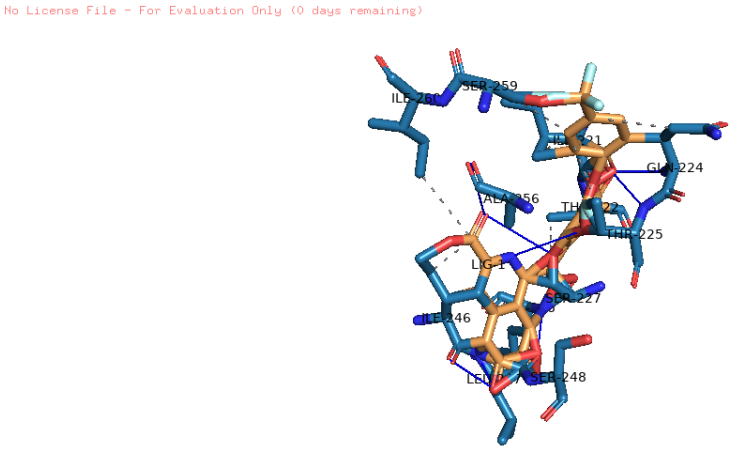

Figure 4.29. Non-Covalent Interactions with BDBM50128431

In the figure below the interactions are shown between IGF2BP3 and BDBM50128436 are shown. In the figure, different colors depict different aspects of the exchanges. Protein is represented through blue, ligand through orange, water through lilac, charge center through yellow, aromatic ring center through white, hydrophobic interactions through dotted lines and hydrogen bonds through sticks.

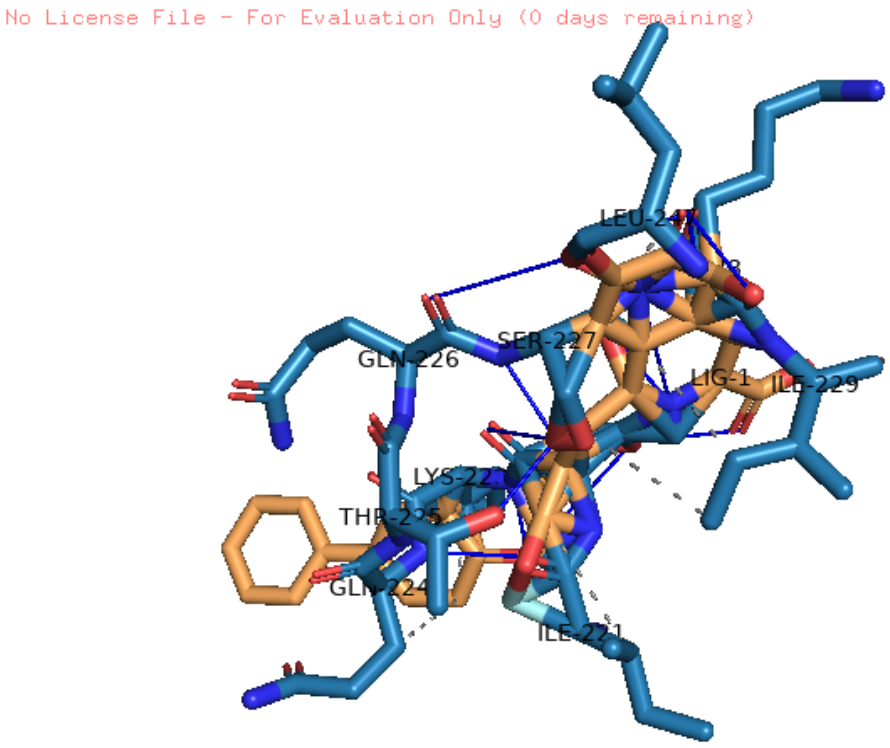

Figure 4.30. Non-Covalent Interactions with BDBM50128436

In the figure below the interactions are shown between IGF2BP3 and DB11952 are shown. In the figure, different colors depict different aspects of the exchanges. Protein is represented through blue, ligand through orange, water through lilac, charge center through yellow, aromatic ring center through white, hydrophobic interactions through dotted lines and hydrogen bonds through sticks.

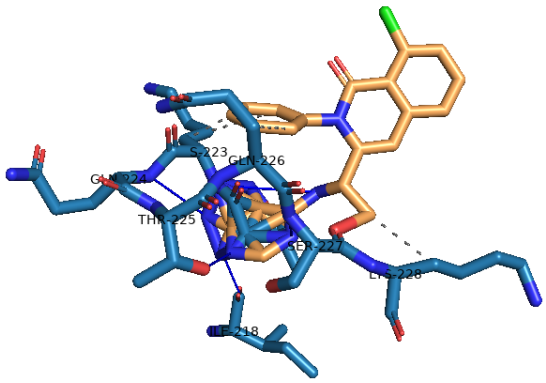

Figure 4.31. Non-Covalent Interactions with DB11952

Chapter 5

DISCUSSION

Discussions Nasopharyngeal cancer falls into the category of head and neck cancers. It originates in nasopharynx. Nasopharynx is a region that originates from the upper region of the throat and behind the nasal cavity close to Eustachian tubes. There are many causes of nasopharyngeal cancer but for the most part it is associated to Epstein Bar Virus. Epstein Bar Virus also known as Human Herpes Virus 4 that possess a liner doble stranded DNA of 172kb in length. Cancer is a genetic disease and identification of therapeutic targets is of paramount significance. The therapeutic targets are upregulated and down regulated genes whose expression level is beyond normal. Such genes can be targeted for the treatment of any disease or any type of cancer. Next Generation sequencing furnished a platform for identification of therapeutic targets. NGS has totally restructured cancer research at genetic and transcriptomic level by identification of abnormal genes whose expression level is above or below normal. Microarray and RNA seq were employed for the identification of therapeutic targets.

The core unit of this research was the identification of therapeutic targets of nasopharyngeal carcinoma and the identification if the most effective therapeutics for nasopharyngeal carcinoma, One dataset of microarray and two datasets of RNA seq were analyzed for the identification of differentially expressed genes.

Comparative analysis of the differentially expressed genes among all the datasets yielded 19 common genes. KLHL14, CD72, TREM2, RASGRP3, PTPN6, TFEB, CBFA2F3,SOX4,NAV1,IGF2BP3,PHACRR1, OSBPL10,ANGPTL1,RALGPS2, AFAP1-AS1, RAI14,S1PR4 are the common differential expressed genes in all the three datasets that can be termed as therapeutic targets. Out of all the genes IGF2BP3 was brought forward in analysis as therapeutic target. In our analysis for the gene that

is brought forward should possess experimentally determined 3D structure, binding domain, and relevance in disease.

IGF2BRP3 possessed all the characteristics to be termed as therapeutic target for nasopharyngeal carcinoma. IGF2BRP3 termed as insulin like growth factor 2 mRNA binding protein 3 belongs to the class of RNA binding proteins that perform a pivotal role in RNA life cycle regulation. RNA binding proteins bind to RNA and direct nuclear export, intracellular localization of targets, translation rates. Mutations is RNA binding protein results in aberrant translation of RNA and cancer progression.

The next task was the search the most effective therapeutics to inhibit IGF2BRP3 expression in cancer. For this purpose, BindingDB and DrugBank were used. They are the two databases that have all the ligands available. A total of 180 drugs were assessed against IGF2BRP3 for the most effective therapeutic The similes of the ligands were downloaded that were converted to mol2 format and further to pdbqt. These ligands were docked to IGF2BRP3 in AutoDock Vina that yielded the binding affinity of every drug against the target. Out of 180 drugs top 10 drugs were shortlisted in terms of maximum binding affinity and least hindrance between the protein and ligand.

The shortlisted ligand candidates include BDMB50128436, BDBM50128454, DB00619, DB01590, DB11730, BDBM50128432, DB11987, BDBM50128431, DB11952, BDBM50106439 with .8.8kcal/mol to -7.7kcal/mol of binding energy. The best ligand against IGF2BRP3 was BDBM50128436 with -8.8kcal/mol of binding energy.

Chapter 6

CONCLUSION AND FUTURE PERSPECTIVES

Nasopharyngeal carcinoma is aggressive cancer of nasopharynx region. Out of all neck and head cancers nasopharyngeal cancer is the is accountable for 18% of the cases. There are many causes for the onset of this cancer and causes millions of deaths year around. The 5-year survival rate for nasopharyngeal cancer is 85%. To date, there is very limited treatment available given for the nasopharyngeal cancer.

In this project IGF2BP3 was modified as to be a suitable therapeutic target for nasopharyngeal cancer. High levels of IGF2BP3 were observed in all cases of nasopharyngeal cancer. RNA-binding proteins have an important role in messenger RNA (mRNA) regulation during tumour development and carcinogenesis.

180 ligands were docked to IGF2BP3, the out file of the binding affinities was generated. Ligands and several inhibitors were docked the selected protein of interest, some being FDA approved drugs for nasopharyngeal cancer and others were anti-cancer drugs. The maximum binding affinity observed was –8.8 kcal/mol by BDBM50128436.

IGF2BP3 marks itself as a primary contributor to tumorigenesis by upregulation in nasopharyngeal cancer. It can be further targeted to achiever positive results for tumor inhibition.

REFERENCES

[1] R. Warrington, W. Watson, H. L. Kim, and F. R. Antonetti, "An introduction to immunology and immunopathology," *Allergy, Asthma & Clinical Immunology*, vol. 7, no. 1, pp. 1–8, 2011.

[2] S. E. Turvey and D. H. Broide, "Innate immunity," *Journal of Allergy and Clinical Immunology*, vol. 125, no. 2, pp. S24–S32, 2010.

[3] G. Zaitseva, A. Kiseleva, and I. Paramonov, "Major histocompatibility complex class i chain-related gene a (mica) and b (micb) polymorphism," *Russian journal of hematology and transfusiology*, vol. 61, no. 2, pp. 100–104, 2019.

[4] J. Ferlay, I. Soerjomataram, R. Dikshit, S. Eser, C. Mathers, M. Rebelo, D. M. Parkin, D. Forman, and F. Bray, "Cancer incidence and mortality worldwide: sources, methods and major patterns in globocan 2012," *International journal of cancer*, vol. 136, no. 5, pp. E359–E386, 2015.

[5] K. D. Miller, R. L. Siegel, C. C. Lin, A. B. Mariotto, J. L. Kramer, J. H. Rowland, K. D. Stein, R. Alteri, and A. Jemal, "Cancer treatment and survivorship statistics, 2016," *CA: a cancer journal for clinicians*, vol. 66, no. 4, pp. 271–289, 2016.

[6] H. Hosseini, M. Obradović, M. Hoffmann, K. L. Harper, M. S. Sosa, M. Werner-Klein, L. K. Nanduri, C. Werno, C. Ehrl, M. Maneck *et al.*, "Early dissemination seeds metastasis in breast cancer," *Nature*, vol. 540, no. 7634, pp. 552–558, 2016.

[7] H. Gonzalez, I. Robles, and Z. Werb, "Innate and acquired immune surveillance in the postdissemination phase of metastasis," *The FEBS journal*, vol. 285, no. 4, pp. 654–664, 2018.

[8] D. Pardoll, "Cancer and the immune system: basic concepts and targets for intervention," in *Seminars in oncology*, vol. 42, no. 4. Elsevier, 2015, pp. 523–538.

[9] A. W. Lambert, D. R. Pattabiraman, and R. A. Weinberg, "Emerging biological principles of metastasis," *Cell*, vol. 168, no. 4, pp. 670–691, 2017.

[10] F. Pezzuto, L. Buonaguro, F. Caponigro, F. Ionna, N. Starita, C. Annunziata, F. M. Buonaguro, and M. L. Tornesello, "Update on head and neck cancer: current knowledge on epidemiology, risk factors, molecular features and novel therapies," *Oncology*, vol. 89, no. 3, pp. 125–136, 2015.

[11] L. Q. Chow, "Head and neck cancer," *New England Journal of Medicine*, vol. 382, no. 1, pp. 60–72, 2020.

[12] G. M. Suresh, R. Koppad, B. Prakash, K. Sabitha, and P. Dhara, "Prognostic indicators of oral squamous cell carcinoma," *Annals of maxillofacial surgery*, vol. 9, no. 2, p. 364, 2019.

[13] D. E. Johnson, B. Burtness, C. R. Leemans, V. W. Y. Lui, J. E. Bauman, and J. R. Grandis, "Head and neck squamous cell carcinoma," *Nature reviews Disease primers*, vol. 6, no. 1, pp. 1–22, 2020.

[14] J. J. Yang, D. Yu, W. Wen, X.-O. Shu, E. Saito, S. Rahman, P. C. Gupta, J. He, S. Tsugane, Y.-B. Xiang *et al.*, "Tobacco smoking and mortality in asia: a pooled meta-analysis," *JAMA network open*, vol. 2, no. 3, pp. e191 474–e191 474, 2019.

[15] A. K. Dhull, R. Atri, R. Dhankhar, A. K. Chauhan, and V. Kaushal, "Major risk factors in head and neck cancer: a retrospective analysis of 12-year experiences," *World journal of oncology*, vol. 9, no. 3, p. 80, 2018.

[16] A. A. MUHAMMAD, "Histopathological study of nasopharyngeal cancer in aminu kano teaching hosital, kano, nigeria: A ten-year retrospective review (january 2005 to december 2014)," *Faculty of Pathology*, 2017.

[17] A. Jemal, F. Bray, M. M. Center, J. Ferlay, E. Ward, and D. Forman, "Global cancer statistics," *CA: a cancer journal for clinicians*, vol. 61, no. 2, pp. 69–90, 2011.

[18] B. Ahmad and U. Pindiga, "Malignant neoplasms of the ear, nose and throat in north eastern nigeria," *Highland Medical Research Journal*, vol. 2, no. 1, pp. 45–48, 2004.

[19] J. S. Low, E.-T. Chua, F. Gao, and J. T. Wee, "Stereotactic radiosurgery plus intracavitary irradiation in the salvage of nasopharyngeal carcinoma," *Head & neck*, vol. 28, no. 4, pp. 321–329, 2006.

[20] W. Liao, M. Tian, and N. Chen, "Characteristic and novel therapeutic strategies of nasopharyngeal carcinoma with synchronous metastasis," *Cancer Management and Research*, vol. 11, p. 8431, 2019.

[21] M. Draganescu, L. Baroiu, A. Iancu, C. Dumitru, D. Radaschin, E. D. Polea, C. Bobeica, A. L. Tatu, E. Niculet, and G. L. Fekete, "Perspectives on skin disorder diagnosis among people living with hiv in southeastern romania," *Experimental and Therapeutic Medicine*, vol. 21, no. 1, pp. 1–1, 2021.

[22] W. Liu, G. Chen, X. Gong, Y. Wang, Y. Zheng, X. Liao, W. Liao, L. Song, J. Xu, and X. Zhang, "The diagnostic value of ebv-dna and ebv-related antibodies detection for nasopharyngeal carcinoma: a meta-analysis," *Cancer cell international*, vol. 21, no. 1, pp. 1–13, 2021.

[23] N. Lee, J. Harris, A. S. Garden, W. Straube, B. Glisson, P. Xia, W. Bosch, W. H. Morrison, J. Quivey, W. Thorstad *et al.*, "Intensity-modulated radiation therapy with or without chemotherapy for nasopharyngeal carcinoma: radiation therapy oncology group phase ii trial 0225," *Journal of clinical oncology*, vol. 27, no. 22, p. 3684, 2009.

[24] A. A. Adoga, D. D. Kokong, N. D. Ma'an, O. A. Silas, A. M. Dauda, J. P. Yaro, J. G. Mugu, C. J. Mgbachi, and C. J. Yabak, "The epidemiology, treatment, and

determinants of outcome of primary head and neck cancers at the jos university teaching hospital," *South Asian journal of cancer*, vol. 7, no. 03, pp. 183–187, 2018.

[25] A. S. Evans, "The spectrum of infections with epstein-barr virus: a hypothesis," *The Journal of Infectious Diseases*, vol. 124, no. 3, pp. 330–337, 1971.

[26] H. Williams, "Dorothy h. crawford.","*Epstein-Barr Virus: The Impact Of Scientific Advances On Clinical Practice". Blood107 (3)*, pp. 862–869, 2006.

[27] C. M. Borza and L. M. Hutt-Fletcher, "Alternate replication in b cells and epithelial cells switches tropism of epstein–barr virus," *Nature medicine*, vol. 8, no. 6, pp. 594–599, 2002.

[28] Z. Li, X. Zhang, L. Dong, J. Pang, M. Xu, Q. Zhong, M.-S. Zeng, and X. Yu, "Cryoem structure of the tegumented capsid of epstein-barr virus," *Cell research*, vol. 30, no. 10, pp. 873–884, 2020.

[29] L. Frappier, "Epstein-barr virus: Current questions and challenges," *Tumour Virus Research*, vol. 12, p. 200218, 2021.

[30] G. Klein, "Viral latency and transformation: the strategy of epstein-barr virus," *Cell*, vol. 58, no. 1, pp. 5–8, 1989.

[31] P. Gerber, S. Lucas, M. Nonoyama, E. Perlin, and L. Goldstein, "Oral excretion of epstein-barr virus by healthy subjects and patients with infectious mononucleosis," *The lancet*, vol. 300, no. 7785, pp. 988–989, 1972.

[32] N. Goossens, S. Nakagawa, X. Sun, and Y. Hoshida, "Cancer biomarker discovery and validation," *Translational cancer research*, vol. 4, no. 3, p. 256, 2015.

[33] K. T.-W. Lee, J.-K. Tan, A. K.-y. Lam, and S.-Y. Gan, "Micrornas serving as potential biomarkers and therapeutic targets in nasopharyngeal carcinoma: A critical review," *Critical reviews in oncology/hematology*, vol. 103, pp. 1–9, 2016.

[34] S. Behjati and P. S. Tarpey, "What is next generation sequencing?" *Archives of Disease in Childhood-Education and Practice*, vol. 98, no. 6, pp. 236–238, 2013.

[35] S. Marguerat and J. Bähler, "Rna-seq: from technology to biology," *Cellular and molecular life sciences*, vol. 67, no. 4, pp. 569–579, 2010.

[36] R. Govindarajan, J. Duraiyan, K. Kaliyappan, and M. Palanisamy, "Microarray and its applications," *Journal of pharmacy & bioallied sciences*, vol. 4, no. Suppl 2, p. S310, 2012.

[37] K. M. Dnyandev, G. V. Babasaheb, K. V. Chandrashekhar, M. A. Chandrakant, and O. K. Vasant, "A review on molecular docking."

[38] A. J. Davison, R. Eberle, B. Ehlers, G. S. Hayward, D. J. McGeoch, A. C. Minson, P. E. Pellett, B. Roizman, M. J. Studdert, and E. Thiry, "The order herpesvirales," *Archives of virology*, vol. 154, no. 1, pp. 171–177, 2009.

[39] K. A. Chan, J. K. Woo, A. King, B. C. Zee, W. J. Lam, S. L. Chan, S. W. Chu, C. Mak, I. O. Tse, S. Y. Leung *et al.*, "Analysis of plasma epstein–barr virus dna to screen for nasopharyngeal cancer," *New England Journal of Medicine*, vol. 377, no. 6, pp. 513–522, 2017.

[40] A. Jain, W. K. Chia, and H. C. Toh, "Immunotherapy for nasopharyngeal cancer-a review," *Chin Clin Oncol*, vol. 5, no. 2, p. 22, 2016.

[41] S. Wang, H. Xiong, S. Yan, N. Wu, and Z. Lu, "Identification and characterization of epstein-barr virus genomes in lung carcinoma biopsy samples by next-generation sequencing technology," *Scientific reports*, vol. 6, no. 1, pp. 1–8, 2016.

[42] C. De Martel, J. Ferlay, S. Franceschi, J. Vignat, F. Bray, D. Forman, and M. Plummer, "Global burden of cancers attributable to infections in 2008: a review and synthetic analysis," *The lancet oncology*, vol. 13, no. 6, pp. 607–615, 2012.

[43] G. Khan and M. J. Hashim, "Global burden of deaths from epstein-barr virus attributable malignancies 1990-2010," *Infectious agents and cancer*, vol. 9, no. 1, pp. 1–11, 2014.

[44] Y.-P. Chen, A. T. Chan, Q.-T. Le, P. Blanchard, Y. Sun, and J. Ma, "Nasopharyngeal carcinoma," *The Lancet*, vol. 394, no. 10192, pp. 64–80, 2019.

[45] D. Wang and S. Bodovitz, "Single cell analysis: the new frontier in 'omics'," *Trends in biotechnology*, vol. 28, no. 6, pp. 281–290, 2010.

[46] Y. Wu, F. Wei, L. Tang, Q. Liao, H. Wang, L. Shi, Z. Gong, W. Zhang, M. Zhou, B. Xiang *et al.*, "Herpesvirus acts with the cytoskeleton and promotes cancer progression," *Journal of Cancer*, vol. 10, no. 10, p. 2185, 2019.

[47] B. He, J. Zeng, W. Chao, X. Chen, Y. Huang, K. Deng, Z. Huang, J. Li, M. Dai, S. Chen *et al.*, "Serum long non-coding rnas malat1, afap1-as1 and al359062 as diagnostic and prognostic biomarkers for nasopharyngeal carcinoma," *Oncotarget*, vol. 8, no. 25, p. 41166, 2017.

[48] J. Zhao, C. Guo, F. Xiong, J. Yu, J. Ge, H. Wang, Q. Liao, Y. Zhou, Q. Gong, B. Xiang *et al.*, "Single cell rna-seq reveals the landscape of tumor and infiltrating immune cells in nasopharyngeal carcinoma," *Cancer letters*, vol. 477, pp. 131–143, 2020.

[49] Z. Ye, F. Wang, F. Yan, L. Wang, B. Li, T. Liu, F. Hu, M. Jiang, W. Li, and Z. Fu, "Bioinformatic identification of candidate biomarkers and related transcription factors in nasopharyngeal carcinoma," *World journal of surgical oncology*, vol. 17, no. 1, pp. 1–10, 2019.

[50] A. Carbone, "Kshv/hhv-8 associated kaposi's sarcoma in lymph nodes concurrent with epstein-barr virus associated hodgkin lymphoma," *Journal of clinical pathology*, vol. 58, no. 6, pp. 626–628, 2005.

[51] C. Fan, C. Tu, P. Qi, C. Guo, B. Xiang, M. Zhou, X. Li, X. Wu, X. Li, G. Li *et al.*, "Gpc6 promotes cell proliferation, migration, and invasion in nasopharyngeal carcinoma," *Journal of Cancer*, vol. 10, no. 17, p. 3926, 2019.

[52] Y. Mo, Y. Wang, L. Zhang, L. Yang, M. Zhou, X. Li, Y. Li, G. Li, Z. Zeng, W. Xiong *et al.*, "The role of wnt signaling pathway in tumor metabolic reprogramming," *Journal of Cancer*, vol. 10, no. 16, p. 3789, 2019.

[53] C. Virtanen and J. Woodgett, "Clinical uses of microarrays in cancer research," *Clinical Bioinformatics*, pp. 87–113, 2008.

[54] D. Boda, A. O. Docea, D. Calina, M. A. Ilie, C. Caruntu, S. Zurac, M. Neagu, C. Constantin, D. E. Branisteanu, V. Voiculescu *et al.*, "Human papilloma virus: Apprehending the link with carcinogenesis and unveiling new research avenues," *International journal of oncology*, vol. 52, no. 3, pp. 637–655, 2018.