QATAR UNIVERSITY

COLLEGE OF HEALTH SCIENCE

CHARACTERIZATION OF GENOMIC ALTERATIONS AND TCRB REPERTOIRE OF TUMOR- INFILTRATING LYMPHOCYTES IN BREAST CANCER

BY

AYEDA ABDULSALAM AHMED

A Thesis Submitted to

the Faculty of the College of Health

Science

in Partial Fulfillment

of the Requirements

for the Degree of

Masters of Science in

Biomedical Sciences

January 2018

COMMITTEE PAGE

The members of the Committee approve the Thesis of Ayeda Abdulsalam Ahmed defended on 26/12/2017.

Ahmed Mohamed Malki Mohamed Youssef
Thesis/Dissertation Supervisor

Dr. Nasser Moustafa Ragheb Rizk
Committee Member

Dr. Joel Malek
Committee Member

Ala-Eddin Al Moustafa
Committee Member

Approved:

Asma Al-Thani, Dean, College of Health Science

ABSTRACT

AHMED, AYEDA,A., Masters of Science:

January : 2018, Biomedical Sciences

Title: Characterization of Genomic Alterations and TCRβ Repertoire of Tumor- Infiltrating Lymphocytes in Breast Cancer

Supervisor of Thesis: Ahmed M. Youssef.

Breast cancer still remains a major cause of morbidity and mortality among women in Qatar and worldwide. More recent studies indicate that the diversity and the composition of the entire set of antigen receptors within tumor-infiltrating lymphocytes (TILs) is strongly correlated with tumor prognosis and therapeutic response with breast cancer. Unfortunately, the relationship between somatic mutational load and TCR diversity of TILs across breast cancer still limited. For this purpose, first we characterized the somatic mutations of Formalin-Fixed Paraffin-Embedded breast cancer samples from 79 patients using NGS of a panel of cancer related genes. Second, we classified and identified the TCRß repertoire for these 11 samples using the ImmunoSEQ platform. Preliminary data demonstrated that the 11 patients had high diversity of TCRß-CDR3 within the tumors. However, there was no statistically significant association between the somatic mutational loads in the gene panels we sequenced and the number of productive TCRß-CDR3 rearrangements.

Table of Contents

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

APC	Antigen presenting cell
CD4	Helper T cell
CD8	Cytotoxic T cells
BCR	B-cell receptor
C	Constant
CCNB1	Cyclin B1
CDR3	Complementarity determining region 3
CK	Cytokeratins
CRC	Colorectal cancer
D	Diversity
DCIS	Ductal Carcinoma in situ
DCs	Dendritic cells
ER	Estrogen receptor
ESR-1	Estrogen Receptor 1
ExAC	Exome Aggregation Consortium
FFPE	Formalin-Fixed Paraffin-Embedded
FOXA1	Forkhead Box A1 Protein
FOXP3	Regulatory T cells
GATA3	GATA-binding protein 3
gDNA	Genomic DNA
H&E	Hematoxylin and eosin
HER1	Human epidermal growth factor receptors 1

HER2	Human epidermal growth factor receptor 2
HER2E	Human epidermal growth factor receptor 2 - enriched
HMC	Hamad Medical Corporation
IDC	Invasive (or infiltrating) ductal carcinoma
IHC	Immunohistochemistry
IL-12	Interleukin-12
ILC	Invasive (or infiltrating) lobular carcinoma
J	Joining
LCIS	Lobular Carcinoma in situ
LPBC	Lymphocyte predominant breast cancer
M	Million
MHC	Major histocompatibility complex
MKI67	Proliferation-Related gene Ki-67
MYBL2	MYB Proto-Oncogene Like 2
NCI	National Cancer Institute
NCI	National Cancer Institute
NGS	Next generation sequencing
NHGRI	National Human Genome Research Institute
OS	Patient overall survival
PR	Progesterone receptor
RAG	Recombination activating gene
SNVs	Single nucleotide variants

TAA	Tumor-associated antigens
TCGA	The Cancer Genome Atlas
TCR	T-cell receptor
TdT	Terminal deoxynucleotidyl transferase
TILs	Tumor-infiltrating lymphocytes
TN	Triple Negative
TSA	Tumor-specific antigen
UMI	Unique molecular indices
V	Variable
WHO	World Health Organization
XBP1	X-Box Binding Protein 1

ACKNOWLEDGMENT

I take this opportunity to express my gratitude to everyone who supported and assist me in making this study possible:

- Dr. Ahmed Malki, Associate Professor of Molecular Oncology in the Biomedical Sciences (Qatar University) for his support, advice and encouragements.
- Dr. Joel Malek, Assistant Professor in Genetic Medicine and Director of Genomics core and Yasmin Mohamoud, manager of the Genomics core (Weill Cornell Medical College- Qatar), for their patience, great support, valuable comments, guidance and supervision during the graduation project.
- Dr. Shahinaz Bedri and Devendra Bansal (Weill Cornell Medical College-Qatar) for their faith on me to process this project.
- Our project collaboration from Hamad Medical Corporation: Dr. Salha Bujassoum, Mahmoud Mohamed, Hekmat Bugrein, Mufeed Almisteri, and Imad Bin Mujeeb for providing us the FFPE samples
- Genomics Core Team at Weill Cornell Medical College- Qatar: Iman Al-Azwani, Lisa Sara Mathew, Eman Al-Dous, Candice Purchase and Shameem Younuskunju for standing beside me in each step of conducting my Master degree.
- My greatest gratitude goes to my friends especially: Rajaa Dalloul and Noor Al-Hajri for their friendship, support and encouragement.
- Last and not least, I would like to thank my Family for their love and emotional support throughout my life and my studies.

CHAPTER 1 : INTRODUCTION

Breast cancer still remains a major cause of morbidity and mortality among women worldwide despite the favorable outcomes provided by early cancer detection and newer therapies (CDC, 2017) . In Qatar, according to the Qatar Cancer Society, breast cancer is the most common cancer among women with incidence rate of about 39% of all cancers in females. Furthermore, 3% of the diagnosed breast cancer cases are in the age group of 15-19 years (Qatar Cancer Society, 2015).

Breast cancer disease is highly complex and heterogeneous, with significant genotypic and epigenetic diversity. Cancergenesis is a result of a multistep accumulation of genetic alterations such as mutations, rearrangements, and copy number alterations that can significantly affect disease progression (Koboldt et al., 2012). From the clinical perspective, different schemes are used to classify breast cancer in order to select the most appropriate treatment. Traditionally, the major categorizations are based on the histological appearance, stage, and grade. Currently, breast cancer subtypes are classified into four groups based on gene expression profiles of hormone receptors (estrogen and progesterone), human epidermal growth factor receptor 2 (HER2) status and Ki67 index. The four subtypes are luminal A, luminal B, HER2 and Triple Negative (TN) (Cheang et al., 2009; Eisenbeisz, 2016).

In the past few years, gene expression and proteomic platforms have been used in an effort to identify and confirm novel diagnostic markers of breast cancer and more importantly to evaluate disease development and treatment response (Criscitiello et al., 2014). However, the capacity of these methods is limited in predicting the length of disease-free survival. These methods showed that there are other factors that malignant cells require for their growth and proliferation that are not directly linked to changes in the genetic elements of the tumor (Criscitiello et al., 2014). To define other potential markers, several investigators have focused on the breast cancer microenvironment. Tumor microenvironment studies have shown that the presence of tumor-infiltrating lymphocytes (TILs) serves as a favorable marker of the disease. Thus, host immune response to the tumor plays a vital role in defining patient prognosis (Asano et al., 2016; Criscitiello et al., 2014). More recent clinical studies indicate that the diversity and the composition of the entire set of antigen receptors within TILs is strongly correlated with tumor prognosis and therapeutic response of a number of solid tumors, including breast cancer (Hadrup et al., 2013). Furthermore, many studies reported that patients with aggressive breast cancer subtypes have high level of TILs (Stanton & Disis, 2016).

Presently, it is unclear why some patients have cancer with higher T cell infiltration and diversity than others, even within the same subtype. Some studies hypothesize that aggressive tumors have a high rate of genomic instability which might increase the chance of presenting mutated proteins at the cell surface of the cancer cells. This stimulates the immune response by increasing the level of T cell diversity within the tumor (Criscitiello et al., 2014; Sherene Loi, 2013).

Unfortunately, the relationship between somatic mutational load and TCR diversity of TILs across breast cancer subtypes remains largely unexplored. With the availability of next generation sequencing (NGS), it has become possible to sequence TCRβ-CDR3 and profile the TILs across patients with different range of mutational load.

1.1 Hypothesis

There is an association between tumor mutational load and T cell diversity within each breast cancer subtype.

1.2 Research Aims and Objectives

The aim of this study is to examine the association between somatic mutation load of and their corresponding to TCR β composition in tumor using the most advanced high throughput DNA sequencing technologies on Formalin-Fixed Paraffin-Embedded (FFPE) samples.

Objective:

- Characterize the somatic mutations of FFPE breast cancer samples from 79 patients within each breast cancer subtype using next generation sequencing of a panel of breast cancer related genes.
- Classify and identify the TCRβ repertoire for a set of these samples using the ImmunoSEQ platform at the survey level.

CHAPTER 2 : REVIEW OF THE LITERATURE

2.1. Breast Cancer

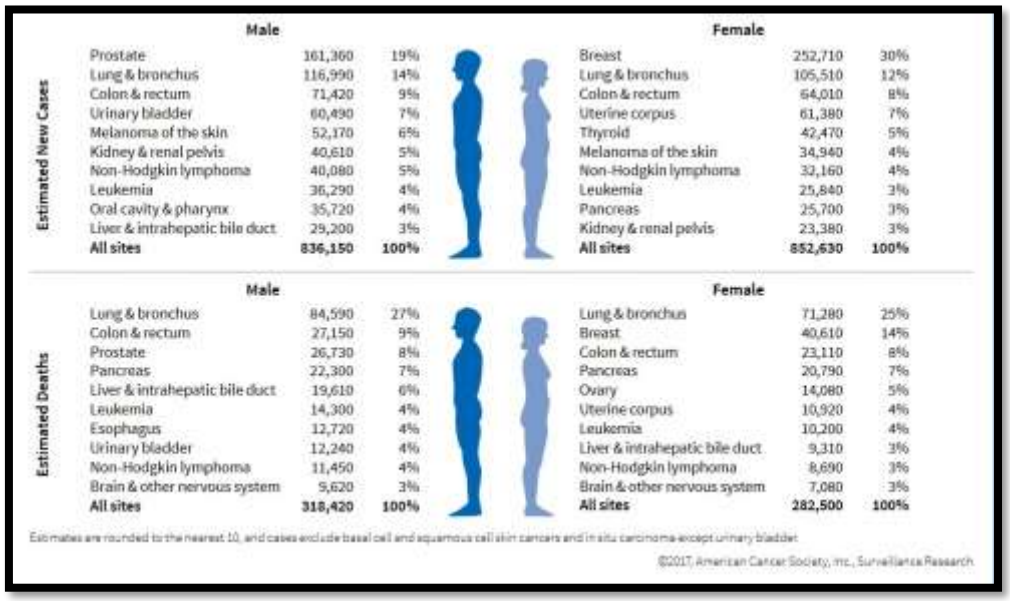
Breast cancer is a disease characterized by uncontrolled growth of abnormally dividing cells beyond their boundaries in the breast tissue. Breast cancer is classified into several forms and each type has its own distinct morphology, behavior and clinical characteristic. According to the World Health Organization (WHO), invasive (or infiltrating) ductal carcinoma (IDC) and lobular carcinoma (ILC) are the most common diagnosed breast cancer types among women at a rate of 70- 80% (Viale, 2012) . Breast cancer can occur at any age but typically older women especially at age 50 years and older are more likely to develop breast cancer disease (Bilimoria & Morrow, 1995; National cancer Institution, 2012).

2.2. Breast Cancer Current Statistics

Breast cancer is one of the most prevalent and deadly malignant diseases in women worldwide. The American Cancer Society's reported that in 2017, breast cancer in women ranked as the most common diagnosed cancer and as the second leading cause of cancer death after lung cancer (Figure 2.1). It is estimated that during 2017, 252,710 new breast cancer cases are expected to be diagnosed. This would account for about 30 % of all female cancers diagnose. Furthermore, it is expected that 40,610 would die of the breast cancer, representing 14% of all cancer cases. Since the last decade, the incidence rates were stable, while mortality trends declined significantly by 38%, likely due to both early diagnosis and improved

treatment efficacy. Breast cancer can also occur in men but is rare. It is about 100 times less common among men compared to the women (American Cancer Society, 2017).

In Qatar, according to the Qatar Cancer Registry, breast cancer is continuing to be the most prevalent cancer among women, accounting for 39% of cancer cases in women. Furthermore, 3% of the diagnosed breast cancer are in the age group of 15-19 years (Qatar Cancer Society, 2015). According to the arab countries registry between 2003-2007, Qatar ranked as one of the top countries for the incidence rate of breast cancer at 45.7 per 100,000 populations (after Bahrain and Kuwait respectively). 32% of the incidences were among the Qatari population and the remaining 68% were among non-Qatari patients (Al Bader et al., 2016).



Estimated New Cases	Male			Female		
	Prostate	161,360	19%	Breast	252,710	30%
	Lung & bronchus	116,990	14%	Lung & bronchus	105,510	12%
	Colon & rectum	71,420	9%	Colon & rectum	64,010	8%
	Urinary bladder	60,490	7%	Uterine corpus	61,380	7%
	Melanoma of the skin	52,170	6%	Thyroid	42,470	5%
	Kidney & renal pelvis	40,610	5%	Melanoma of the skin	34,940	4%
	Non-Hodgkin lymphoma	40,080	5%	Non-Hodgkin lymphoma	32,160	4%
	Leukemia	36,290	4%	Leukemia	25,840	3%
	Oral cavity & pharynx	35,720	4%	Pancreas	25,700	3%
	Liver & intrahepatic bile duct	29,200	3%	Kidney & renal pelvis	23,380	3%
	All sites	**836,150**	**100%**	**All sites**	**852,630**	**100%**

Estimated Deaths	Male			Female		
	Lung & bronchus	84,590	27%	Lung & bronchus	71,280	25%
	Colon & rectum	27,150	9%	Breast	40,610	14%
	Prostate	26,730	8%	Colon & rectum	23,110	8%
	Pancreas	22,300	7%	Pancreas	20,790	7%
	Liver & intrahepatic bile duct	19,610	6%	Ovary	14,080	5%
	Leukemia	14,300	4%	Uterine corpus	10,920	4%
	Esophagus	12,720	4%	Leukemia	10,200	4%
	Urinary bladder	12,240	4%	Liver & intrahepatic bile duct	9,310	3%
	Non-Hodgkin lymphoma	11,450	4%	Non-Hodgkin lymphoma	8,690	3%
	Brain & other nervous system	9,620	3%	Brain & other nervous system	7,080	3%
	All sites	**318,420**	**100%**	**All sites**	**282,500**	**100%**

Estimates are rounded to the nearest 10, and cases exclude basal cell and squamous cell skin cancers and in situ carcinoma except urinary bladder.

©2017, American Cancer Society, Inc., Surveillance Research

Figure 2.1. **Estimation of New Cancer Cases and Deaths in the United States of America by Sex – 2017**. Breast cancer in women ranked as the most common diagnosed cancer and as the second leading cause of cancer death after lung cancer (American Cancer Society, 2017).

2.3. Risk Factors

There are several important factors associated with increased risk of breast cancer (American Cancer Society, 2017). Simply being a woman is the most significant risk factor. This might be because of the activity of the female hormones estrogen and progesterone, which can promote breast cancer cell growth. Age is another crucial factor, where the chance of getting cancer is significantly increased with age. About 5% to 10% of breast cancer cases are assumed to be hereditary, inherited mutations in particular BRCA1, BRCA2, or other breast cancer susceptibility genes are associated with increased risk of breast cancer. Menstrual and reproductive history as well as dense breast tissue can also be associated with increased risk for breast cancer. Additionally, lifestyle risk factors like: lack of physical activity, poor diet, being overweight or obese, drinking alcohol and radiation to the chest are associated with incidences of breast cancer Breast Cancer Classification (American Cancer Society, 2017).

2.4. Breast Cancer Classification

Different schemes are used to classify breast cancer, each of these schemes use various principles and criteria to divide breast cancer into subclasses. The main categorizations are based on the histological appearance, stage, grade and receptor status. These classifications are periodically updated as cancer cell biology knowledge expands. The purpose of the classification is to help in selecting the most appropriate treatment. In fact, this disease is highly complex and heterogeneous and there is no single treatment for all the subclasses. Certain classes

of breast cancer are aggressive and life-threatening, and must be treated aggressively, while others are less aggressive and can be treated with less invasive treatments; such as lumpectomy. Selecting the best treatment is mainly based on the available evidence that is provided by the categorization (Eisenbeisz, 2016).

2.4.1. Grade

Grade is the description of the tumor based on how close cancer cells appear and their growth patterns compared to normal cells. It is an indicator of how rapidly the cancer cells are dividing and spreading. The grade scale system helps physicians to predict prognosis and develop a treatment plan. In general, a "well-differentiated" tumor indicates that the cells and tissue of the tumor are close in appearance to the normal cells and tissue. This type of tumor tends to have well-organized patterns and slower rate of growth. While " undifferentiated " or "poorly differentiated" tumors have abnormal cell appearance and may grow aggressively in disorganized and irregular patterns. This method of classification depends on visual observation of breast cancer cells and tissue under a microscope (National Cancer Institute, 2013).

According to the National Cancer Institute(NCI), most of the cancer cells are rated on a scale from 1 to 3. However, some types of cancers for instance breast and prostate cancers have their own grading systems. The most common grading system for the breast cancer is Nottingham (also called the Elston-Ellis modification of the Scarff-Bloom-Richardson grading system). This system grades tumors based on three features, which are tubule formation, nuclear grade and mitotic rate. Each

feature gets a score from1 to 3, then their scores are summed, yielding a total score between 3 to 9 (National Cancer Institute, 2013). Three possible grades are listed in Table 2.1.

Table 2.1.

Breast Cancer Grading System

Score	Grade	Description
Score 3–5	Grade 1	Low grade or well differentiated
Score 6–7	Grade 2	Intermediate grade or moderately differentiated
Score 8–9	Grade 3	High grade or poorly differentiated

2.4.2. Stage

Staging refers to how widespread and large the cancer is when it is first diagnosed. Knowing the stage helps physicians to determine how serious the cancer is, and how successful the treatment might be. This classification method requires several examinations such as chest x-ray, CT scans, MRI, mammograms and lab tests (National Cancer Institute, 2015).

The TNM system is the most widely used cancer staging system in hospitals and medical centers. This system is applicable to all forms of cancers, including breast cancer. In the TNM system:

- The T refers to the size and extent of the of the breast tumor.
- The N refers to spread of the cancer to nearby lymph nodes.
- The M refers to whether the cancer has metastasized (spread beyond the breast to other parts of the body).

Sometimes stage is expressed as a number on a scale of 0 to IV. As rule, the lower the value is, the less the cancer has spread, and the higher the value is, the more the cancer cells have spread into adjacent tissues and affected lymph nodes (National Cancer Institute, 2015). Stages number described in the Table 2.2.

Table 2.2.

Stages of Breast Cancer

Stage	Description
0	Cancer cells in situ, limited to inner lining surface of the organ and has not spread to nearby tissue.
I, II, and III	Cancer has grown in size and has spread to distant lymph nodes (except stage I).
IV	Cancer has metastasized and spread to distant parts of the body.

2.4.3. Histological Appearance

Histological appearance is the primary method used for diagnosis and classification of cancer (Eisenbeisz, 2016). This method involves direct examination of cancer biopsy by a pathologist to determine the type of cancer whether is in situ or invasive. However, histological analysis does not always explain the differences in breast cancer (Makki, 2015). Currently the WHO classifies breast cancer into 20 major tumors and 18 minor subtypes. This classification has been adopted worldwide. The most common histological types of cancer include; Ductal Carcinoma in situ (DCIS) which is a cancer that is in the ductal system but not spread to the nearby tissues. IDC which originates is in the milk ducts attacks other surrounding tissues and can get to other parts of the body through the lymph nodes. Lobular Carcinoma in situ (LCIS) is another type but very rare. It is not easily detected because there is no formation of palpable mass. ILC originates from lobules, the milk-producing glands at the end of breast ducts. This type is considered the second common type of breast cancer consisting of 10% of cases (American Cancer Society, 2017; Viale, 2012).

2.4.4. Breast Cancer Hormones Status

The most recent breast cancer classification is based mainly on gene expression profiles of hormone receptor (estrogen (ER) and progesterone (PR)) and HER2 status (Zhang et al., 2014). In the last decade, the high-throughput gene-expression platforms such as microarray-based gene-expression methods have been extensively applied for breast cancer studies to identify signatures associated with prognosis and response to therapy (Weigelt & Reis-Filho, 2009). The breast cancer molecular classification was established in 2000 when Perou and his group reported the gene expression profiles of 38 primary invasive breast cancer tissue by using complementary DNA microarrays. Their results exhibited marked variation between cancer tissues and they ended up classifying breast cancer into four molecular subtypes, each one has its own clinical characteristic (Perou et al., 2000; Weigelt & Reis-Filho, 2009)

The four-intrinsic subtype are:

1. Luminal A
2. Luminal B
3. HER2-enriched (HER2E)
4. Triple negative (TN)

2.4.4.1. Luminal A

Luminal A is the most prevalent subtype among women and comprise 50%-60% of all breast cancer cases (Yersal & Barutca, 2014). This subtype is defined as ER and/or PR positive and HER2 negative. It is also characterized by the high expression of luminal epithelial cytokeratins(*CK*) 8 and 18, and other luminal markers such as Estrogen Receptor 1(*ESR-1*), GATA-binding protein 3 (*GATA3*), Forkhead Box A1 Protein (*FOXA1*), X-Box Binding Protein 1 (*XBP1*). Also, Lumina A tumors exhibit a low expression of proliferation-related genes, such as Cyclin B1 (*CCNB1*), Proliferation-Related gene Ki-67 (*MKI67*) and MYB Proto-Oncogene Like 2 (*MYBL2*) gene (Koboldt et al., 2012; Zhang et al., 2014).

Patients with Luminal A type have a good prognosis and are always subjected to endocrine therapy with tamoxifen, to inhibit the functions of ER. The therapeutic strategies of blocking the estrogen signaling pathway have been highly effective for ER^+ subtypes and are currently used as the first-line clinical treatment option (Zhang et al., 2014). The relapse rate is significantly low compared to the other subtypes and recurrence commonly occurs to bone, whereas other organs such as liver, lung and central nervous system occur in less than 10% of patients (Yersal & Barutca, 2014) .

2.4.4.2. Luminal B

The Luminal B subtype is not very common compared to Luminal A and accounts for 15%-20% of total diagnosed cancers. Like the Luminal A subtype, Luminal B is also characterized by the expression of ER, PR genes. In contrast to Luminal A, this subtype is associated with the higher expression of proliferation-related genes, such as *CCNB1, MKI67 , MYBL2* and basal-like gene such as Human epidermal growth factor receptors 1 (*HER1*) & *HER2*. (Yersal & Barutca, 2014; Zhang et al., 2014). Patients with these subtypes have more aggressive tumor phenotype, higher histological grade, proliferative index, worse prognosis, higher recurrence rate and lower survival rates compared to luminal A subtype (Sotiriou & Pusztai, 2009; Zhang et al., 2014). In a clinical practice, Ki67 index is used as a potential marker to distinguish between Luminal subtypes, where it is highly expressed in the luminal B subtype (Cheang et al., 2009). As the Lumina B tumor is highly proliferative and expresses ER and some of basal-like gene patterns, the patients could be subjected to combined therapeutic strategy of chemotherapy and endocrine treatment. (Dai et al., 2015).

2.4.4.3. HER2E

HER2E cancer accounts for 15-20% of breast cancer subtypes. This subtype is characterized by overexpression of HER2 and HER2-associated genes and negative expression for both ER and PR. In addition, HER2E tumor are associated with the higher expression of proliferation-related genes. Patients with HER2E tumor have high histological and nuclear grades. These tumors mostly display aggressive biological and clinical behavior (Yersal & Barutca, 2014). Patients diagnosed with this subtype can benefit from HER2 targeted therapy such as Transtuzumab. The targeted therapies for HER2 subtype have significantly improved overall survival and reduced the risk for recurrence (Petrelli et al., 2008).

2.4.4.4. TN

TN represents 10 % to 20 % of all diagnosed breast cancer. This type does not express ER, PR and HER2, hence referred to as triple-negative (Dillon et al., 2016). However, TN tumors express high levels of basal myoepithelial markers, such as *CK5, CK 14, CK 17*. TN tumor are characterized with high histological and nuclear grade, high mitotic activity, lymphocytic infiltrate and high proliferative rate. Like HER2, TN tumors exhibit aggressive clinical behavior and very high tendency to metastasize to the brain and lung (Yersal & Barutca, 2014). Unlike other breast cancer subtypes, TN is very difficult to treat due to lack of therapeutic target receptors, thus leaving chemotherapeutic treatments the only option for systemic therapy of TN patients (Weisman et al., 2016).

2.5. Breast Cancer New Prognostic Markers

Despite the favorable outcomes provided by earlier cancer detection and newer therapies, breast cancer still remains a major cause of morbidity and mortality among women worldwide. In the past few years, gene expression and proteomics platforms have been used in an effort to identify and confirm novel diagnostic markers of breast cancer and more precisely to evaluate disease development and response to treatment (Criscitiello et al., 2014; Qin & Ling, 2012).

However, the capacity of these methods is limited in predicting the length of disease-free survival due to factors not directly linked to changes in genetic elements of tumors, which the malignant cells needed for the growth and proliferation. For example, the prognostic and predictive genetics signatures were mainly limited with TN and HER2E subtypes that are by nature highly proliferative and aggressive compared to luminal subtypes. To overcome this limitation, several investigators have focused on the breast cancer microenvironment to define other promising diagnostic markers. Recent evidence has demonstrated that the host immune response to tumor microenvironment plays a vital role in defining patient prognosis and their response to treatment. Later on, many clinical studies supported the concept that the presence of TILs within the tumor tissue indicates an antitumor cellular immune response, and the degree of infiltration has been strongly correlated with good tumor prognosis and therapeutic response(Asano et al., 2016; Criscitiello et al., 2014).

2.6. Immune System

The immune system consists of two subsystems: innate and adaptive immunity (or acquired immune systems). Innate immune system is composed of physical epithelial barriers, chemicals and cells that protect the host against the invasion of pathogens by providing immediate, non-specific and non-memory response. Unlike the innate immune system, the adaptive immune system relies on specialized cells that are more specific for any individual foreign antigens and their effective responses occur only after several days of exposure to pathogens. Also, the adaptive response produces memory cells that persist in a dormant state for decades after initial contact, but they can rapidly re-express after subsequent exposure to the same target antigen (Janeway et al., 2001). Moreover, the most significant difference between the two systems is that all recognition mechanisms of the innate immune system are encoded in the genes of the host's germ-line. For instance, the innate immune cells rely on the recognition of only specific molecules and molecular patterns which are associated with entire classes of pathogens such as bacteria, viruses and fungi and absent from the host cells; therefore, it is a limited response. On the other hand, the adaptive immune system can specifically recognize an almost infinite diversity of antigens by a process called somatic gene rearrangements of antigen-binding molecules, so each potential foreign antigen can be targeted specifically (Chaplin, 2010). The adaptive system not only works to defend the host against hordes of microorganisms, but also helps to eliminate defected and mutated cells from the body. Impaired adaptive immunity

leads to an increase in the susceptibility to infection, autoimmune diseases and even cancer (Janeway et al., 2001).

2.7. Adaptive System

The adaptive immune system is composed of two major response classes: humoral immunity and cellular immunity, and they are carried out by two distinct types of lymphocytes: B cells and T cells respectively. Each recognizes antigens through the antigen receptors on their surfaces; the B-cell receptor (BCR) on B cells and T-cell receptor (TCR) on T cells. Both T and B lymphocytes play central roles in the adaptive immune system but cell-mediated immunity plays a key role in transforming cell recognition and rejection (Janeway et al., 2001). Indeed, the maturation and activation processes in the adaptive immune system are highly complex**.** Briefly, both lymphocytes cells are generated in the bone marrow; only the B cell matures there whereas the T cell migrates to the thymes to mature (Janeway et al., 2001).

There are two main phases for lymphocyte maturation: first, the lymphocytes programmed to recognize sufficient diversity of target antigens through a process called somatic DNA recombination. This is also known as VDJ recombination; as discussed in more detail in later sections. Second, lymphocytes with receptors which bind to self-antigens and react towards healthy self-tissue are removed during development by inducing apoptosis in the cell (Janeway et al., 2001). After maturation, these cells enter the blood circulation, then the secondary lymphoid organs, such as lymph nodes and spleen where they localize and wait for

exposure to the antigen that they are pre- programmed to recognize. Localized cells are known as naïve lymphocytes. Antigenic activation of these cells requires interaction with antigen presenting cells (APCs) such as B lymphocytes, macrophages and dendritic cells (DCs). The main role of APC in this mechanism is taking up antigens from infected peripheral tissues into the lymph nodes for presentation and activation of the appropriate naïve lymphocyte. After exposure to antigen the selected lymphocytes with specific target receptors to the antigen are proliferated and mature into effector cells (Janeway et al., 2001).

In humoral immunity, activated effector B cells secrete different types of antibody called immunoglobulins, which help in the elimination of extracellular microorganisms and prevent spread of their infections. Naïve B cells are triggered by the encounter with the direct antigens in the lymph nodes and usually require T helper cells to produce antibodies. These antibodies mainly circulate in the bloodstream and body fluids where they can act over long distances. The antibodies are specifically bind to antigens or toxins that stimulate their production then block their ability to bind to the host cell receptors (Janeway et al., 2001).

In the cellular immunity, T lymphocytes are divided into:1) cytotoxic T cells (CD8), directly attack and kill infected cells, 2) helper T cell (CD4), help to activate cells in both humoral and cellular immune responses and 3) regulatory T cells (FOXP3), down-regulate immune responses. Activation of naïve T cells requires at least two independent signals: i) antigen specific signals and ii) co-stimulatory signals (Figure 2.2). The first signal is based on the interaction of TCR with antigen peptides presented by the major histocompatibility complex (MHC) on cell surface

of APC, either class I or class II molecules. Unlike BCR, TCRs do not bind directly to antigens; instead their antigen on APC must be processed to suitable peptides that bind to MHC (Janeway et al., 2001).

The second signal is derived from the interaction of co-stimulatory receptors CD28 on T cell with B7-1 and B7-2 ligand on APC (signal 2), without second signals, T cells cannot be activated, but instead become tolerized or anergic (Iwai et al.,2017; MacLeod, 2015). The effecter T cells then migrate from lymphoid organs to sites of infection where they can act only locally on adjacent cells. Most importantly, once cytotoxic T cells activate, they will kill only the infected cell that bearers the exact antigens. Effecter cytotoxic T cells trigger apoptosis in the targeted infected cells either through secreting perforin and granzymes or displaying Fas ligand on their surface. On the other hand, effecter helper T cells secrete a variety of signal proteins called cytokines. These cytokines act as mediators by influencing the behavior of the numerous cell types that they help (Alberts et al., 2002).

MHC proteins have an important role in the adaptive immune system. The function of MHC molecules is to display non-self-peptide antigens on the cell surface for recognition by the appropriate T cells. MHC class I molecules present peptides derived from intracellular/endogenous proteins to cytotoxic T cells, and they are located on the cell surface of all nucleated cells in the body. While MHC class II display peptides derived from extracellular/exogenous proteins to CD4+ T cells and they are found only on APCs (Janeway et al., 2001).

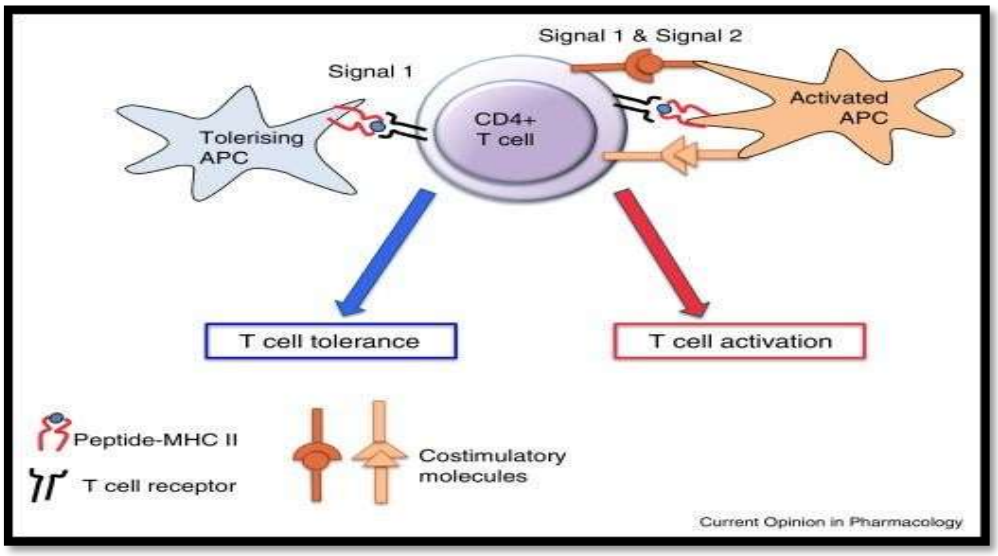


Figure 2.2. **Activation of T Cells Requires Two Signals: Antigen Specific Signals and Co-stimulatory Signals.** The first signal is based on the interaction of TCR with antigen peptides presented by the major histocompatibility complex (MHC) on cell surface of APC, either class I or class II molecules. The second signal is derived from the interaction of co-stimulatory receptors CD28 on T cell with B7-1 and B7-2 ligand on APC. Without second signals, T cells cannot be activated, but instead become tolerized or anergic (MacLeod, 2015).

In this chapter, I will therefore only discuss concepts relevant to T cells and cancer since T cells play a key role in transforming cell recognition and rejection.

2.8. The Development of TCR Diversity

2.8.1. V(D)J Recombination

In the invertebrate immune system, the diversity of the T cell receptors is huge, which enables the immune system to provide broad protection against the vast diversity of nonself-antigens. The TCR protein is encoded by a unique protein-coding system of germline genes. The TCR is composed of two disulfide-linked polypeptide chains (alpha [α] and beta [β], or gamma [γ] and delta [δ]). Approximately 95% of human T cells express α/β and only 5% of circulating T cells express γ/δ (Cruse & Lewis, 2010). Each chain is composed of constant (C) and variable (V) regions. The C-domain is proximal to the cell membrane and the V-domain is distal to the membrane (Figure 2.3). The recombinable genes which rearrange during T cell development and contribute to the great diversity of TCR fall into the V domain. Basically, the TCR β chain V-domain is encoded by three genes segments, the variable (V), diversity (D) and joining (J), while the TCR α chain V-domain is encoded by V and J genes segments, without D (Murphy, 2014).

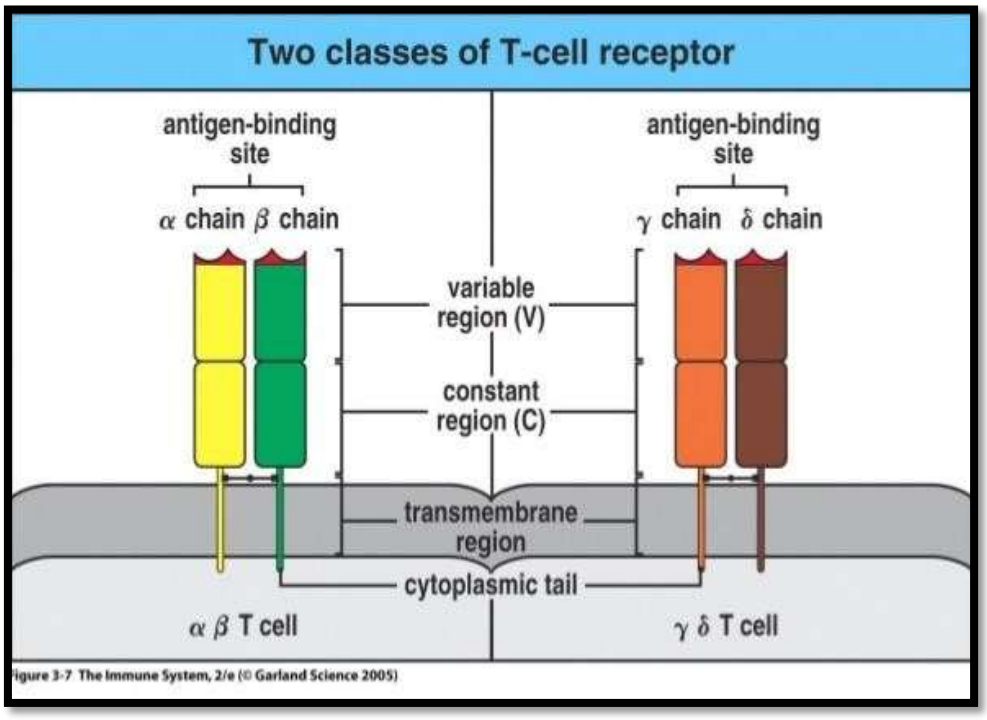


Figure 2.3. **The Basic Structure of Two Classes of TCR**. Each T cells consist of two different polypeptide chains, both of which are heterodimer with constant (C) and variable (V) regions. The C-domain is anchor to the cell membrane and the V-domain is distal to the membrane (Murphy, 2014).

V(D)J recombination or somatic recombination is a process by which T cells randomly assemble different gene segments of V, D, and J in the β chain or V and J genes in the α chain, in order to generate unique antigen receptors. There are multiple copies of V, D, and J gene segments within each receptor gene locus bearing high sequence homology. For instance, the human TCR β-chain locus is located on chromosome 7 and contains 52 V gene, 2 D gene and 13 J gene segments, while the TCR α -chain locus is located on chromosome 14 and consists of ~70 V genes and 61 J genes (Murphy, 2014). In each recombination only one copy of V/(D) /J gene can take part by random selection. Initially one of the D segments is joined with one of the J segments, next is the assembly of the DJ region to one of V segments, yielding the final VDJ region that has a size of ~ 500 bp. Together, this process provides the basis for the TCR diversity (Murphy, 2014; Woodsworth et al., 2013).

Furthermore, along with recombination, the diversity of TCR is significantly increased by the addition and deletion of a set of non-template nucleotides at the joints between the gene segments. This process is termed as junctional diversity. The added nucleotides are known as P-nucleotides and N-nucleotides. P nucleotides are so named because they make up palindromic sequences, which are added at the ends of the gene segments, and N nucleotides are so named because they are non-template encoded (Janeway et al., 2001; Saada et al., 2007).

Briefly, during junctional diversity (Figure 2.4), RAG enzyme cleaved coding segments to create hairpin structures at the end of gene segments. After the formation of the DNA hairpins, RAG catalyzes single-stranded cleavage at a random point within the coding sequences. This lead to the formation of a single-stranded tail from a few nucleotides of the coding sequence along with the complementary bases from the other DNA strand, thus generating the palindromic sequences. Then, random nucleotides are added by terminal deoxynucleotidyl transferase (TdT) enzyme to the ends of single -stranded gene segments. After the addition of the nucleotides, the two single-stranded ends pair over a short region. The unpaired nucleotides are removed by exonucleases and gaps are filled by repair enzymes, which help in creating coding joints between gene segments (Janeway et al., 2001; Saada et al., 2007).

The two processes: V(D)J recombination and junctional diversity generate complementarity determining region 3 (CDR3), which is the main domain of the TCR that is in contact with peptide-MHC antigen and largely determine TCR specificity (Figure 2.5) (Saada et al., 2007; Woodsworth et al., 2013). It is estimated that the adaptive immune system of each person can generate approximately up to 10^{16} distinct αβ pairs of TCR CDR3 (Robins et al., 2009).

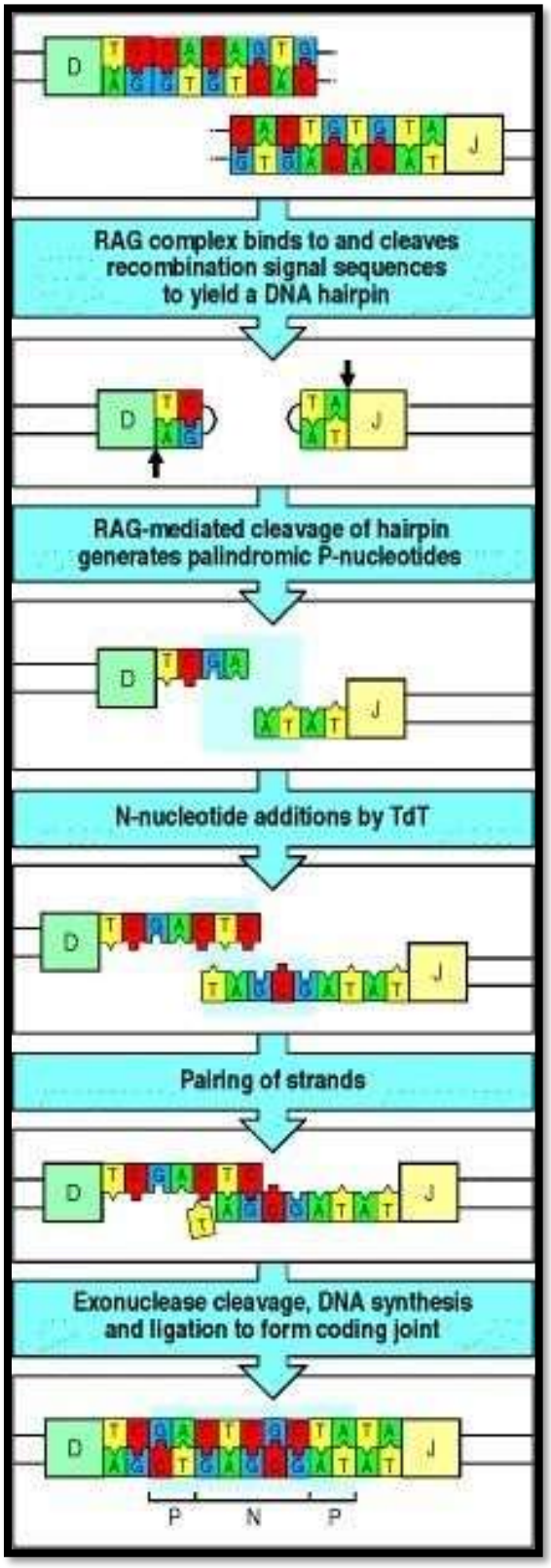


Figure 2.4. **Generation of Junctional Diversity steps on TCR** (Janeway et al., 2001).

2.8.2. The TCR Development in the Thymus

It is good to point out that V(D)J recombination of TCR α and β chains does not happen in parallel. In fact during T cell maturation, TCR β chain gene segments rearrange first. Only when these rearrangement events yield a productive TCRβ chain, the cell will be able to produce a functional pre-T-cell receptor and blocks further gene rearrangement. Following a successful productive rearrangement, the TCR continues to rearrange its TCR α locus until a productive and useful TCR β/α is generated. If the TCR fails to synthesize productive rearrangement the T cell will die (Murphy, 2014).

In addition to V(D)J recombination during TCR development, the β chain of TCR exhibits allelic exclusion. Allelic exclusion is a regulatory mechanism which states that only one of the two alleles of the β chain loci is rearranged and expressed in the T cell, thus ensuring that a given T cell will make TCR molecules with only a single specificity (Murphy, 2014). In the case of the α chain, allelic exclusion may not always be the rule; rearrangement on both alleles is very common and 20-30% of mature T cells could express two productive V_{α}-J_{α} rearrangements on the cell surface. However, the α chains of both rearrangement are regulated by post-translational events (Rybakin et al., 2014). Based on this fact, studying the diversity and the clonotype aspects of T cell receptors rely on sequencing the β chains not the α. Since each T cell only expresses single β chain rearrangement on its surface, the number of β chain sequences is a reflection of the number of TCR clonotypes present in a sample (Woodsworth et al., 2013).

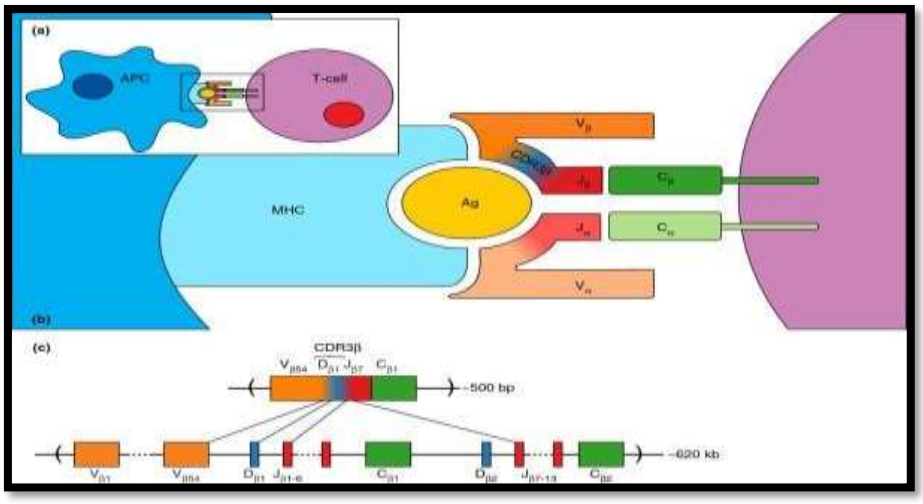


Figure 2.5. **The Interaction of TCR and Antigen Peptide. A)** The T-cell receptor (TCR) encounter with antigen presenting cell (APC) that presents peptide antigen by MHC on its cell surface. **B)** Complementarity determining region 3 (CDR3) is the main domain that directly contacts with peptide-MHC. CDR3 region is unique to each TCR and largely determine TCR specificity to peptide antigen. **C)** Representation of TCR-β V(D) J gene recombination. In each recombination only one copy of each different gene segments of V, D, and J in the β chain can take part in the process by random selection, resulting in TCR diversity (Woodsworth et al., 2013).

2.9. The Concept of Tumor Immune Surveillance

The link between the immune system and cancer has been under investigation for many years and is still ongoing. Paul Ehrlich in 1909 first postulated the idea that nascent transformed cells spontaneously arise in our bodies and that the immune response could effectively eradicate them before they are manifested clinically (R. Kim et al., 2007). However, this idea was not vigorously pursued as at that time the field of immunology was still immature. In the mid-20th century, 50 years later, interest in this area was renewed by Medawar and his colleagues who clarified the significant role of cellular components of the immune system in recognizing and mediating allograft rejection. Their experiments showed strong immune-mediated rejection of transplanted tumors derived from noninbred strains of mice. However, the results were used to argue for the existence of a tumor specific immunity, rather than that the tumor was seen simply as a foreign graft (Dunn et al., 2002; Smyth & Hayakawa, 2004).

Soon after, with the wide availability of inbred strains of mice, it was possible to address whether tumors arising in mice were immunologically distinguishable from normal cells in the same syngeneic mice. The results showed that the mice were able to provoke their immune system and prevent the outgrowth of syngeneic tumors induced by chemical carcinogens or virus. This work formulated the fundamental principle of cancer immune surveillance, which is that immune cells would recognize the presence of transformed tissue in the body if tumor cells express distinctive recognition structures on the surface of cells, as was

postulated by Macfarlane Burnet and Lewis Thomas in 1957. At the core of their hypothesis was the discovery of the existence of tumor specific antigens. Furthermore, they proposed the idea that lymphocytes were mainly responsible for the recognition and elimination of genetically transformed cells (Dunn et al., 2002; Smyth & Hayakawa, 2004).

Years later, a large number of experiments with mouse models harboring mutations in one or several immune response genes have been performed to understand the mechanism of tumor immune surveillance. The accumulated evidence supports that immune surveillance of cancer was dependent on lymphocytes cells. The relative importance of lymphocytes derived from the use of mouse models lacking the recombination activating genes 1 (Rag 1) or Rag 2. Those genes encode proteins that are involved in the initiation of V(D)J recombination during B and T cell development. The loss of Rag genes function in mouse models results in absence of mature B and T cells. In experiments using both wild type mice and mouse models exposed to chemical carcinogens, the frequency of tumor formation and the kinetics of tumor growth were measured. The Rag knockout mice failed to prevent the formation of cancer. They formed tumors earlier and with greater frequency compared to wild type mice that had the same genetic background (R. Kim et al., 2007).

Further analysis has underscored a central role of T lymphocytes in the antitumor immune response by studying mice that lacked T cells. These experimental mice showed rapid formation of large tumors compared to wild-type strains (R. Kim et al., 2007). Other studies have reached similar conclusion using

mice strains lacking IFNγ, interleukin-12 (IL-12), components of the MHC class I antigen processing and presentation pathways, CD8+ T cells, perforin or granzymes and Fas. All these are considered important components of the cell mediated mechanism of tumor recognition and elimination (Dranoff, 2011). These findings indicate a distinct contribution of T cells in the regulation of tumor growth, and inhibition of initial tumor (Dranoff, 2011; R. Kim et al., 2007).

Recently, A large body of evidence from clinical studies and mouse models supported that the immune system plays a dual role in cancer. It can participate both i) in tumor elimination and control by destroying cancer cells or inhibiting their outgrowth (via the actions of the adaptive immune system) or ii) in tumor development by establishing conditions within the tumor microenvironment that facilitate tumor cell survival, outgrowth and spread in the host (via chronic inflammation by the innate immune system). The dual role of the immune system in cancer is known as cancer immunoediting (Criscitiello et al., 2014).

Immunoediting represents a contemporary view of the relationship between the immune system and the tumor. It defines a complex process that leads to the selection of tumors that can evade the immune system as the immune system edits the tumors to kill those that it can recognize. Immunoediting can be divided into three phases: elimination, equilibrium, and escape. The elimination phase represents the classical concept of cancer immunosurveillance, equilibrium is when tumor cells are held in control but are not eradicated by the immune system. The escape phase refers to the final stage when immune cells fail to restrict tumor outgrowth and the tumor becomes clinically detectable. Studies report that even

with advance stages of cancer the immune can directly or indirectly help in defining tumor prognosis and response to treatment (Dushyanthen et al., 2015). In summary, the data obtained from both mouse and human studies provided strong evidence of the complex role of the immune system in tumor initiation, progression, and inhibition.

2.10. TIL

In clinical practice, the presence of T cells in tumors and their potential impact have been studied over the past years. Studies provide compiling evidence for a positive correlation between the presence of high density of T cells at the tumor site and improved patient overall survival (OS) (Hadrup et al., 2013). Most of the convincing evidence originates from studying patients with cutaneous melanomas. TIL patterns were classified into three categories as brisk, non-brisk and absent according to criteria suggested by Clark et al. and Elder et al. "Brisk" was defined as the presence or infiltration of lymphocytes within tumor, the "non-brisk" category is when lymphocytes are present in one or more focal location of the tumor, and "absent" are cases in which no lymphocytes were present in the tumor. Early data proposed that a brisk TIL in melanoma lesions was a positive prognostic factor (Hadrup et al., 2013; Mihm et al., 2015). In cohort study, of 5-10 year follow-ups for more than 500 patients with melanoma, the results indicated that patients in the brisk TIL category had highly significant survival advantage compared to patients in non-brisk and absent TIL groups. The result showed that the brisk TIL group tended to live 1-1.5 to 3 times longer than those with in the absent group and non-brisk patients had intermediate survival (Dunn et al., 2002).

Similar methodological approaches for evaluating TIL have been used in several other solid tumor studies. Same positive correlation between the presence of TILs and improved patient survival have been found in breast, ovarian, bladder, prostate, colon and other cancers. This data suggested the use of TIL as a diagnostic marker in routine clinical practice for the predication of patient survival outcome (Hadrup et al., 2013).

2.11. Tumor Mutation Burden and TIL

At present, there is limited knowledge as to why some patients have cancer with heavy T cell infiltration than others, even within patients that have similar cancer subtypes. Some studies report that tumorigenesis can induce T-cell-mediated immune response against malignant cells because the genetic alterations in the tumor lead to profound changes in the cells. The genetic alterations can either lead to the presentation of new antigens known as tumor-specific antigens (TSA), antigens only expressed on tumor cells. Or they can cause change in gene expression level on the cell surface known as tumor-associated antigens (TAA), antigens expressed on both normal and tumor cells but the quantity is significantly higher in tumor cells. As suggested these tumor antigens can trigger the immune system to recognize and destroy cancer cells (Escors & David, 2014).

Several studies elucidated that not all intracellular foreign antigens will have a chance to be recognized by CD8 T cells. Studies clarified that for a mutation to give rise to a foreign antigen (called neoantigens; an antigens that is recognized

with high specificity by patient T cells), three criteria must be met: a) the mutated protein must be processed and presented as a peptide; derived from the intracellular proteolysis machinery ii) the mutant peptide must be loaded with high affinity on MHC class I molecules on the surface of cancer cells iii) the patient's T cells repertoire must contain a TCR with an optimal binding affinity and specificity for the mutant epitope (Martin et al., 2016).

The vaccine studies are the most appropriate examples for the effect of these restrictive criteria to the immune response. Not all viral proteins given in a vaccine can be naturally processed, presented on MHC class I, or/ and recognized by CD8 T cells (Martin et al., 2016). For instance, in one study that focused on influenza A virus epitopes, in total of 180 peptides used in this experiment, predicted from protein sequence of several H1N1 viruses, only 120 (66.6%) were found to be binding to the MHC molecule, of those only 13 (7.2 %) were verified as actual strong CD8 cytotoxicity epitopes (Lundegaard et al., 2010). Indeed, this data illustrated that only a small proportion of mutations can become neoantigens and a small subset of those can become strong CD8 cytotoxicity epitopes.

Furthermore, many studies underscore the significant role of mutational load in increased T cell infiltration and diversity in human cancer. The Cancer Genome Atlas (TCGA); which is an important database created by both National Cancer Institute (NCI) and the National Human Genome Research Institute (NHGRI), contain a comprehensive catalog of somatic tumor mutations of 33 types of cancers identified by deep sequencing. The TCGA data has revealed that human tumors are heterogeneous and the spectrum of somatic mutations are extremely

variable within and across tumor types. In addition, the analysis of cancers genomes has highlighted the positive association between total tumor mutational load or neoantigen burdens with increased cytotoxic T cell infiltration within samples (Brown et al., 2014; Roszik et al., 2016).

2.12. Tumor Mutation Burden and TCR Diversity

It has been reported that the diversity of T cell receptor (TCR) repertoire within the TIL is a fundamental property of an effective immune defense system and is closely related to the capacity of T cells to recognize cancer antigens (Li et al., 2016). Several studies proposed that neoantigens derived from somatic mutations might shape and modulate the diversity of infiltrating T cells in human cancer (Savage et al., 2014). Recently, with the availability of high throughput approaches, it became possible to address this issue. One important study used deep sequencing data to analyze the TCR repertoire of the tumor microenvironment in different types of tumors. This study reported that tumors that showed high levels of T cell infiltration had more TCR reads compared to tumors that had low TIL infiltration. As each T cell clone possesses a unique TCR, the study confirmed positive correlation between somatic mutational load and T cell clonotype diversity (Li et al., 2016).

2.13. Breast Cancer and TIL

Breast cancer has not traditionally been considered a typical immunogenic tumor, perhaps owing to its low mutational rate when compared to melanomas and renal cell carcinomas. However, TIL is observed in breast cancer and has been reported to be a good prognosis feature for subsets of the disease (Criscitiello et al., 2014; Dushyanthen et al., 2015). Clinical trials reported that TIL located in the surrounding stroma contributes to treatment efficacy and survival rate in all subtypes of breast cancer. Higher TIL counts are significantly observed with ER negative, larger tumors, higher histologic grade, high HER2 amplification and more involved lymph-nodes tumors. All these are mostly feature of HER2 and TN subtypes (S. Loi et al., 2014).

Notably, a lot of observational studies reported that TN patients are more likely to have tumors with >50% lymphocytic infiltrate, which is called lymphocyte predominant breast cancer (LPBC). Every 10 % increase in TIL was associated with increased survival benefit in LPBC patients. However, HER2 patients mostly had similar TIL infiltration as TN patients but have not shown the same survival benefits (Stanton & Disis, 2016).

Some studies have attempted to understand why TN and HER2 patients have high TIL infiltration. These studies suggested this might be because both of TN and $HER2^{+}$ subtypes exhibit high proliferation rates and high genomic instability compared to luminal breast cancer subsets. Genomic instability, as suggested by the available data, can promote antitumor immune responses through

inducing and presenting large number of tumor antigens and therefore promotes increased level of T cells infiltration within the tumor (Criscitiello et al., 2014; Sherene Loi, 2013).

In addition, the high throughput sequencing data provides a deeper understanding of tumor molecular biology for each breast cancer subtypes. It has been observed that TN patients had the highest mutation rates among all the breast cancer subtypes, whereas luminal subsets had the lowest mutation rates (Wang et al., 2017). The TCGA breast cancer dataset that consists of 762 invasive breast cancers confirmed that ER negative subtypes have a higher spectrum of mutations compared to the ER positive samples (Haricharan et al., 2014). Unfortunately, the relationship between the load of somatic mutation and TCR diversity of infiltrating T lymphocytes across breast cancer subtypes remains limited.

2.14. Approaches for TILs Evaluation in the Breast Cancer

Recently TIL has emerged in clinical practice as a predictive and prognostic biomarker of long term breast cancer disease control. In December 2013, a group of breast cancer researchers from around the world discussed the important need to consider methodological and criteria for evaluating and scoring the degree of lymphocytic infiltration. The International TILs Group recommended using hematoxylin and eosin (H&E)- stained slides of tumor sections as a standard method. Scoring the percentage of lymphocytes would be done by selecting areas with average density of stromal lymphocytes (Figure 2.6) (Salgado et al., 2015). This semiquantitive method appears to be sufficient for pathologist-based

assessment. However, this method is limited in providing detailed information about the immune subpopulations and cannot discriminate between TIL subsets. Another major issue with this method is the heterogeneity of lymphocyte distributions within some tumors as this might yield results that are not representative of the entire tumor (Lee et al., 2016; Salgado et al., 2015) .

Immunohistochemistry (IHC) assay is another method utilized in clinical practice and research studies for quantification of T cell population. This method is dependent on using antibodies against human CD3, CD4, and CD8 on FFPE tissue. Then the stained section is scored and evaluated either by automated digital image analysis or by visual scoring of a pathologist (Metzger et al., 2012; Rathore et al., 2014).

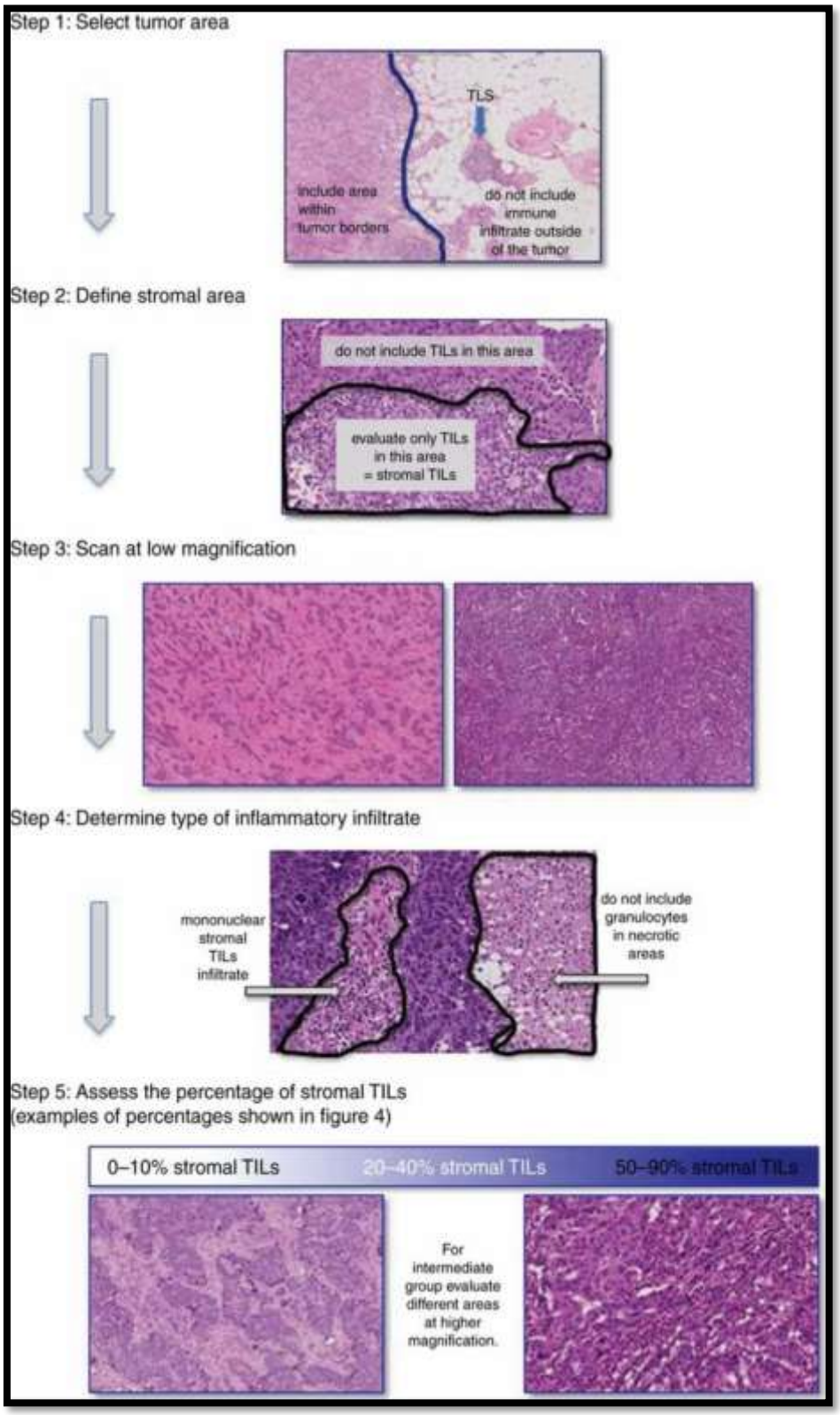


Figure 2.6. **Standard Method for Tumor-Infiltrating Lymphocytes (TILs) Evaluation of the Breast Cancer** (Salgado et al., 2015).

2.15. Methodologies for TCR Repertoire Analysis

Due to the importance of TCR repertoire diversity in understanding the adaptive immune system, new molecular platforms have been developed to monitor the TCR within cancer and other immunological diseases. These new approaches enabled deeper analyses of the degree of heterogeneity of T lymphocyte in the specimens. Before the emergence the high throughput sequencing, several low-throughput techniques had been used to assess the diversity of TCR repertoire. The most important methods were sanger sequencing and gel electrophoresis based methods known as immunoscope or spectratyping (Ciupe et al., 2013; Dziubianau et al., 2013).

Sanger sequencing was used to read nucleotide sequences of the rearranged DNA but it was labor-intensive and generated limited data the TCR repertoire (Ciupe et al., 2013; Dziubianau et al., 2013). Spectratyping approach was also used for analyzing the diversity and the spread of clonalities in the samples. This method depends on the principle that each of T-cell clone have particular sequences or lengths of CDR3 products. The spectratyping focused on generating information about the sizes of CDR3 in T lymphocytes and their pattern of distribution in the specimen (Ciupe et al., 2013).

Recently, next-generation sequencing (NGS) technologies have provided high resolution analysis of numerous areas of biology and medicine. Because of the high throughput nature of these technologies they have provided deep insights into properties and behaviors of the adaptive immune system. These methods are able to profile, monitor lymphocytes and capture accurate quantitative measurements of TCR sequences in the sample from the first time (Dziubianau et al., 2013).

CHAPTER 3 : MATERIALS AND METHODS

3.1. Materials and Reagents

Table 3.1-3.3 contains a list of all materials and resources used in the project.

Table 3.1.

List of Reagents

Item	Company	Part Number
Qubit® dsDNA BR Assay Kits	Life Technologies	Q32850
ImmunoSEQ hsTCRB KIT	Adaptive Biotechnologies	ISK10001
Agilent High Sensitivity DNA Kit	Agilent Technologies	5067-4626
Aglilent DNA 1000 Kit	Agilent Technologies	5067-1504
Agencourt AMPure XP	BECKMAN COULTER	A63881
MiSeq Reagent Kit v2 (300 Cycles)	Illumina	MS-102-2002
PhiX Control v3	Illumina	FC-110-3001
MiSeq Reagent Kit v3 (150 cycle)	Illumina	MS-102-3001
KAPA Library Quantification Kit Illumina® platforms	Kapa Biosystems	KK4824
Qubit® dsDNA HS Assay Kits	Life Technologies	Q32854

SMRTbell Damage repair kit	Pacific Biosciences	100-465-900
QIAseq DNA QuantiMIZE Assay KiT	QIAGEN	333414
GeneRead DNAseq Targeted Panels V2 (Human Breast Cancer Panel)	QIAGEN	181900
GeneRead DNAseq Panel PCR Kit V2 (96)	QIAGEN	181942
DNQC-100Y-R GeneRead TM DNA Library I Core Kit(12)	QIAGEN	180432
GeneRead TM Adapter I Set 12-plex(72)	QIAGEN	180984
GeneRead TM DNA Librart I Amp Kit (100)	QIAGEN	180455
GeneRead TM Size Selection Kit (50)	QIAGEN	180514
QIAquick R PCR Purification Kit (50)	QIAGEN	28104
QIAGEN Multiplex PCR plus KIT	QIAGEN	206152

Table 3.2.

List of Instruments

Item	Company	Part Number
2100 Bioanalyzer system	Agilent	G2943CA
Mi-seq Systems	Illumina	SY-411-9001DOC
7500 Fast Real-Time PCR System, laptop	Applied Biosystems	4351106
Veriti™ 96-Well Thermal Cycler	Applied Biosystems	4375786
Qubit® 2.0 Fluorometer	Life Technologies	Q32866

Table 3.3.

List of Software

Item	Company
Immunoseq analyzer 3.0	Adaptive biotechnologies
Ingenuity Variant Analysis	QIAGEN
Biomedical Genomics Workbench version 4	QIAGEN
Rstudio V.3.3.2.	Rstudio
Mutagen	NCBI

3.2. Study Design

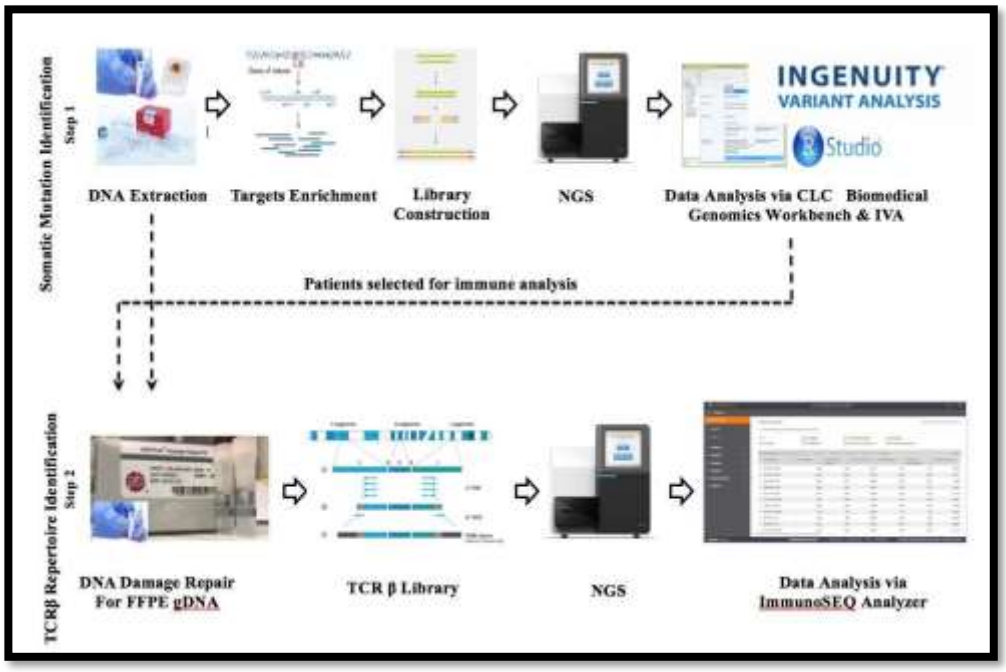


Figure 3.1. **Study Design Workflow.** The workflow illustrates the steps of somatic mutation identification and TCRB repertoir identification. First, somatic mutations were identified from FFPE breast cancer clinical biopsies from 79 patients. Based on the results of the somatic mutational load analysis, 11 patients were selected for the immune analysis.

3.3. Methods

3.3.1. Tumor Sample Collection

For this study, we included a total of 182 breast cancer samples provided by Hamad Medical Corporation (HMC) collaborators. This retrospective study was approved by the Medical Research Center institutional review board of HMC with a waiver of the informed consent requirements. All the FFPE biopsies were obtained from patients with no history of any cancer therapy. These clinical samples were collected between 2004 and 2012. Sample selection was based on the high quantification level of lymphocytes in the tumor tissue. Lymphocyte quantification was determined by the evaluation of hematoxylin and eosin (H&E) stained tissue slides, which is the standard method at HMC. The molecular subtypes of breast cancer have been determined based on expression level of ER, PR, HER2, and Ki-67, as follows (Table 3.4):

Table 3.4.

Molecular Characteristics of Breast Cancer

Molecular Subtypes	**Luminal A**	**Luminal B**	**HER2E**	**TN**
Characteristics	ER+ PR+ HER2- Ki-67 -Low	ER+ PR+ HER2 +/- Ki-67 -High	ER- PR- HER2 +	ER- PR- HER2 -

3.3.2. Macrodissection of Tissue Samples

All FFPE samples were manually macrodissected according to specific instructions of the pathologists, to ensure the presence of maximum cancer cells in the study. Briefly, first the pathologist reviewed the H&E stained slides of FFPE using the light microscope to select and mark areas of cancer cells for dissection. Next, the laboratory technician used these marked H&E slides as guide to dissect FFPE blocks. Furthermore, the exact areas when the FFPE block was taken from in the tumor (acinar, lobular, ductal..etc) were not provided to our study.

3.3.3. DNA Extraction and Quantification

DNA and RNA were extracted from dissected FFPE using the AllPrep DNA/RNA FFPE Kit (Qiagen, USA) according to the manufacturer's instructions. DNA concentrations were determined by Qubit 2.0 Fluorometer dsDNA HS assay kit (lifeTechnologies, USA). Because the genomic DNA (gDNA) extracted from FFPE could been have damaged and fragmented as a result of the fixation and storage conditions, the DNA was qualified and quantified using the QIAseq^{TM} DNA QuantiMIZE assay (Qiagen, USA). This assay utilizes a qPCR-based approach to calculate the amount of amplifiable DNA in each sample. Following the manufacture recommended protocol, 2 PCR components were prepared to amplify more than 20 genomic loci distributed across the human genome. The first mix was SYBR Green PCR master mix with primer assay 100 and the second mix was SYBR Green PCR master mix with primer assay 200, which generate 100 and 200 bp amplicons respectively. 4 ul of each gDNA sample along with the control sample were interrogated by 6 reactions; 3 reactions using the 100 assay and 3 using the 200 assay. After the qPCR run was performed, the C_T values of both assays were exported from the ABI 7500 real –time PCR instrument to an Excel data analysis sheet. This Excel sheet was provided by the manufacture to assist in calculating the integrity of the amplifiable nucleic acid in each sample. RNA samples were stored in -80C for future studies, and were not used in this project.

3.3.4. Breast Cancer Target Enrichment

GeneRead DNAseq Targeted Panels V2; Human Breast Cancer Panel (Qiagen, USA) in combination with GeneRead DNAseq Panel PCR Kit V2 (Qiagen, USA) were used to perform target enrichment by multiplex PCR. The breast cancer panel consists of four primer pools yielding 2,915 amplicons. These amplicons cover mutational hotspot regions located in 46 genes, which are known to be related to breast cancer. This genes list in Table 3.5. Briefly, gDNA samples were amplified using PCR reagents with 4 primer pool mixes (each primer pool in a separate well), and PCR was performed in a standard thermocycler following the protocol recommendation and the number of cycles calculated by QuantiMIZE analysis.

After the completion of the 4 PCR reactions, the products for each sample were combined into one 1.5ml LoBind tube and the enriched DNA was purified using Agencourt® AMPure® XP beads (Beckman Coulter, USA). This step helps exclude large DNA fragments that could contaminate downstream steps. The concentration and the size of the purified amplicons were determined using Qubit 2.0 Fluorometer dsDNA BR assay kit (lifeTechnologies, USA) and Agilent BioAnalyzer 2100 High-Sensitivity DNA kit (Agilent Technologies, USA).

Table 3.5.

Human Breast Cancer Genes Panel (QIAGEN)

Genes	
ACVR1B	KMT2C
AKT1	MAP2K4
ATM	MAP3K1
BAP1	MDM2
BRCA1	MUC16
BRCA2	MYC
CBFB	NCOR1
CDH1	NEK2
CDKN2A	PBRM1
CISD3	PCGF2
EGFR	PHF7
EP300	PIK3CA
ERBB2	PIK3R1
ERBB3	PPM1L
ESR1	PTEN
EXOC2	PTGFR
EXT2	RB1
FBXO32	RET
FGFR1	SEPT9
FGFR2	TP53
GATA3	TRAF5
IRAK4	WEE1
ITCH	ZBED4

3.3.5. Library Preparation and Sequencing

A total amount of 80 -160 ng of purified enriched DNA was used as template to generate NGS libraries. The NGS libraries were prepared using the DNQC-100Y-R GeneRead TM DNA Library I Core Kit(12) (Qiagen, USA) and GeneRead TM Adapter I Set 12-plex(72) (Qiagen, USA). All library preparation steps; End repair, A-tail, adapter ligation, size selection and PCR amplification were performed according to the manufacturer's protocol. The size of the final PCR products was analyzed using Agilent BioAnalyzer 2100 with 1000 DNA kit (Agilent Technologies, USA).

To avoid over-clustering during sequencing, an accurate concentration of the libraries were estimated using Qubit and KAPA Library Quantification Kits for Illumina® platforms (Kapa Biosystems, USA). The quantified samples were then normalized to 2 nM to ensure equal representation of each library within the pool. After normalization, each 12 libraries were pooled together and spiked with 5% of PhiX control. Finally, using MiSeq Reagent Kit v2, 300 cycle, a paired-end 260 bp sequencing run of each pool was carried out on the MiSeq instrument (Illumina, USA).

3.3.6. DNA Damage Repair

SMRTbell Damage repair kit (Pacific Biosciences, USA) was used to perform DNA damage repair for 20 samples, thus to ensure good quality VDJ sequencing results from FFPE samples. According to the kit guidelines, up to 2 ug of gDNA were used. Repaired samples were then purified using Agencourt® AMPure® XP beads (Beckman Coulter, USA).

3.3.7. TCR β CDR3 Amplification and Sequencing

The CDR3 region of TCRβ was amplified and sequenced for 12 samples using the ImmunoSEQ profiling system at the survey level (Adaptive Biotechnologies, USA). This assay utilized a two-step PCR reaction to amplify the TCRβ immune repertoire. For the survey level means, replicate reactions were required for each sample. Briefly, total of 150 ng of purified damage repaired gDNA was used as template. Typically, this concentration will yield the targeted number of T cells (~ 4,000- 30,000 T cells) for non-lymphoid tissue.

For the first round of PCR, DNA was amplified using the QIAGEN Multiplex PCR plus KIT (Qiagen, USA) with mix of multiplexed V-and J gene primers. The forward primers annealed to the(V) region and the reverse primers annealed to the junction (J) region of TCRβ as provided by ImmunoSEQ hsTCRB KIT (Adaptive Biotechnologies, USA). The reaction cycling conditions were: 95°C 15 min, 30 cycles of [30s at 95°C, 90s at 62°C, and 90s at 72°C], plus a final extension of 3 min at 72°C. After amplification, the PCR products were purified using immunoSeq PCR Cleanup reagent following the manufacturer protocol. For

quality control, the samples were run on Agilent BioAnalyzer 2100 with 1000 DNA kit (Agilent Technologies, USA).

For the second round of PCR, Illumina adapters and unique DNA barcodes were introduced to each PCR replicate. Briefly, 2 ul of each purified first PCR product was mixed into a multiplex PCR reaction with unique barcodes supplied by ImmunoSEQ hsTCRB KIT. The thermocycling conditions consisted of: an initial denaturation step at 95°C for 15 min followed by 7cycles of [30s at 94°C, 40s at 68°C and 1 min at 72 °C], and the final extension of 10 min at 72 °C. The second PCR products were purified according to the protocol recommendations. In this step, each replicate samples has become uniquely identifiable by the additional of barcodes.

Following the two step PCR reactions, equal volumes of each 14-barcoded sample [including positive and negative controls], were pooled together in one tube without normalization. In total we had 2 pooled samples in this study each containing 14 barcodes. The final concentration of the pooled samples was quantified using KAPA Library Quantification Kits for the Illumina® platforms (Kapa Biosystems, USA). Following quantification, the pooled immunoSEQ libraries were spiked with 5% PhiX control and sequenced on the MiSeq platform using MiSeq Reagent Kit v3, 150-cycle (Illumina, USA) as recommended by Adaptive Biotechnologies.

3.3.8. Data Analysis

3.3.8.1. Somatic Mutation Analysis

The output raw data (.fastq files) from the MiSeq instrument were analyzed using Biomedical Genomics Workbench V 4 (CLC Bio, QIAGEN) following QIAGEN GeneRead Panel Analysis workflow, which was designed for the GeneRead DNAseq Targeted Panel users. The workflow starts with mapping sequencing data to human genome reference (hg19). A local realignment was performed to improve the quality of variant detection. The variants were directly called after trimming sequencing primers and dimers. All reads that non-specifically mapped or mapped outside of the targeted regions were excluded. Reads that aligned within the targeted regions were annotated with gene names, amino acid changes, exon numbers and chromosomal numbers. These annotations were obtained from different databases like, ClinVar dbSNP, HapMap, Cosmic and 1000 Genomes project.

Additional filtering steps were performed to remove variants present in the human reference and reported in dbSNP, 1000 Genomes Project, HapMap, Exome Sequencing Projects and Exome Aggregation Consortium (ExAC) databases. Also, variants that were present in more than 3% of the studied population were excluded as they could be due to sequencing artifacts or as-yet undiscovered common variants. For the remaining variants, the selection parameters for candidate variants were Quality >=50, Allele fraction > 2, Read depth >= 30. The final list only included the variants that changed native protein function as predicted by SIFT and PolyPhen-2 Functional Predictions. We excluded variant with allele fraction < 2

because the NGS kit not optimize to detect variants with low allele fraction from FFPE samples. The last filtering steps were done through Ingenuity Variant Analysis (QIAGEN). Finally, the somatic mutation data were summarized and plotted using R Program version 3.3.2.

3.3.8.2. Derivation of Breast Cancer Mutational Signatures

To identify the most likely mutagenic processes underling somatic mutations of our breast cancer samples in Qatar, the mutational profile was analyzed for the 79 patients using an online freely available computational framework MutaGene; https://www.ncbi.nlm.nih.gov/research/mutagene/. To start this analysis, we uploaded 79 VCF files of filtered germline variants to the MutaGene system.

Briefly, MutaGene reported the mutational profile according to six base substitutions which are: C→A, C→G, C→T, T→A, T→C and T→G; all substitutions are referred to by the pyrimidine of the mutated Watson–Crick base pair. Further, each of these substitutions represent information of the neighboring nucleotides in 5' (C, A, G, and T) and 3' (C, A, G, and T) directions from the mutated base, so under each of 6 mutation types there are 16 possible 5’3’ sequence contexts, this mades up in total 96 context-dependent mutation types. The given set of mutations for each patient were displayed as histogram on the basis of the trinucleotide frequency across the GRCh37 (hg19) reference human genome; regardless of their genomic locations; this represents each patient’s mutational fingerprint.

Then, these fingerprints were annotated into mutational signatures and mutagenic components using the MutaGene database to identify the most likely mutagenic processes that may have lead to somatic mutational profiles of each of the breast cancer patients. The MutaGene database includes mutational profiles for more than 9000 genomes and exomes from 37 different cancer types.

3.3.8.3. TCR β Repertoire Analysis

Raw data (.fastq files) from the MiSeq instrument were transferred to Adaptive Biotechnology. The sequencing data for each sample was processed and uploaded to the ImmunoSEQ Analyzer web-based analysis tool. Data processing steps included annotation, TCRB quantification and clonality assessment. Only samples with at least 10X amplification factor and high quality sequence data were included in the study. The analysis pipeline is designed to count both productive and nonproductive TCR rearrangements. Non-productive templates contain stop codons or frameshifts that are unlikely to produce a functional protein receptor. In this study, only the productive TCR rearrangements were reported in this study.

CHAPTER 4 : RESULTS

4.1. FFPE Clinical Samples

FFPE samples are routine diagnostic specimens used in HMC. The pathological laboratories in HMC use standard protocols for the fixation process and subtyping of breast cancer. The molecular characterization of the breast cancer patients are listed in Table 4.1. Most of the provided samples were of the luminal A subtypes, which is the most prevalent subtype in women in Qatar and worldwide. The FFPE samples were isolated from primary breast cancer tissues without lymph nodes and archived up to 144 months before the study started. The process of the sample collection starting from the surgery through fixation were not provided. Additionally, information related to patients age, lymph node status, tumor stage and size were not available for this project.

Table 4.1.

Moleculer Subtypes of Breast Cancer Patients Included in this Study

Breast cancer subtypes	**Number of sample**
Luminal A	108
TN	30
Luminal B	23
HER2	21
Total	**182**

4.2. DNA Quality and Quantity from FFPE

The NGS workflows are highly sensitive to quality and quantity of the input nucleic acid. The integrity of the extracted gDNA was measured by QIAseq™ DNA QuantiMIZE assays to ensure that the fragmented DNA samples yielded enough material to preform NGS library preparation. In general, the quality and quantity of the FFPE extracted gDNA were barely suitable for the downstream analysis. 46 out of 182 samples failed quality check and were excluded from further analysis. For the remaining 136 samples, 2 to 6 PCR additional cycles were needed depending on the assays quality scores for each sample. The additional cycles were recommended by the assay to compensate for the differences in gDNA quantity during the enrichment steps. Because of budget limitations, only 79 FFPE samples were used for breast cancer somatic mutations detection. Table 4.2 summarizes the number of samples in each breast cancer subtypes and its status in this study.

Table 4.2.

Overview of the Number of Each Breast Cancer Subtypes That Passed QC, Failed QC and Included in the Study

Subtypes	Passed QC	Failed QC	Included in the study
Luminal A	84	24	31
Luminal B	18	5	16
HER2E	18	3	16
TN	16	14	16
Total	**136**	**46**	**79**

4.3. Target Enrichment and Library Construction for NGS

The 2100 Bioanalyzer instrument was used to check fragment size after the target enrichment and library NGS construction processes. As shown in the Figure 4.1-A & B, single peaks around 280 bp represent the size of amplicons as described in the manufacturer protocol. The Bioanalyzer results also show that the NGS libraries are clean, and do not contain adapter dimer, which it present should be 120 bp peak. To avoid over-clustering the MiSeq flow cell, NGS libraries were quantified by qPCR and fluorometric quantification method. The results from both approaches gave similar measurement of the DNA concentration for each sample. This data was used in the preparation of the final sequence loading concentration of 2 nM. As a result, the MiSeq sequences generated maximum cluster density, with high Q30 scores as shown in Figure 4.2-A, indicating high read quality.

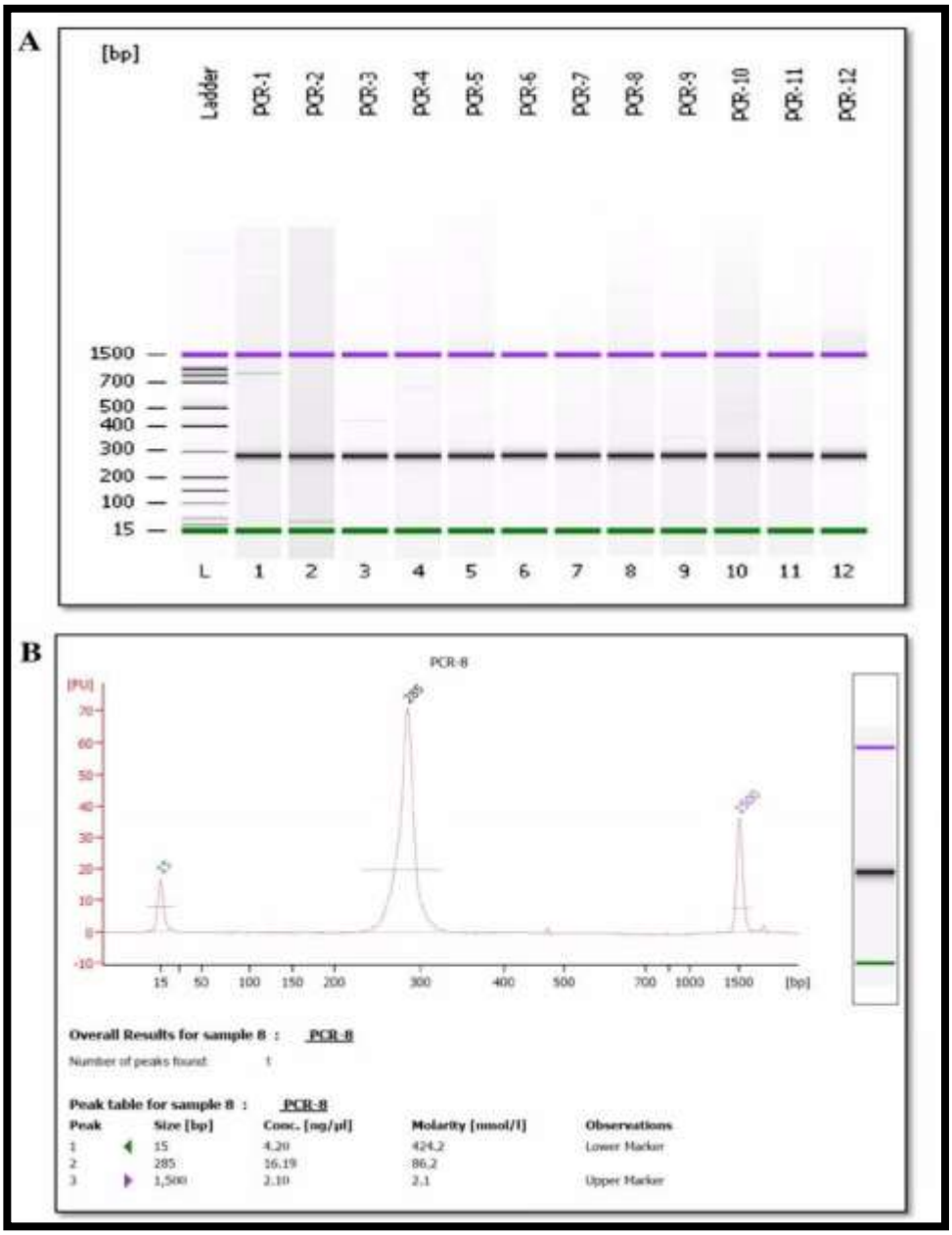


Figure 4.1. **Agilent 2100 BioAnalyzer Results of Final NGS Libraries**. **A)** Bioanalyzer image of 12 different libraries with ladder. Single band around 280 bp obsorved from each library as protcol recommended. **B)** Library product of sample PCR-8 shown as an example, peak size: 285 bp, concentration: 16.19 ng and molority: 86.2 nM.

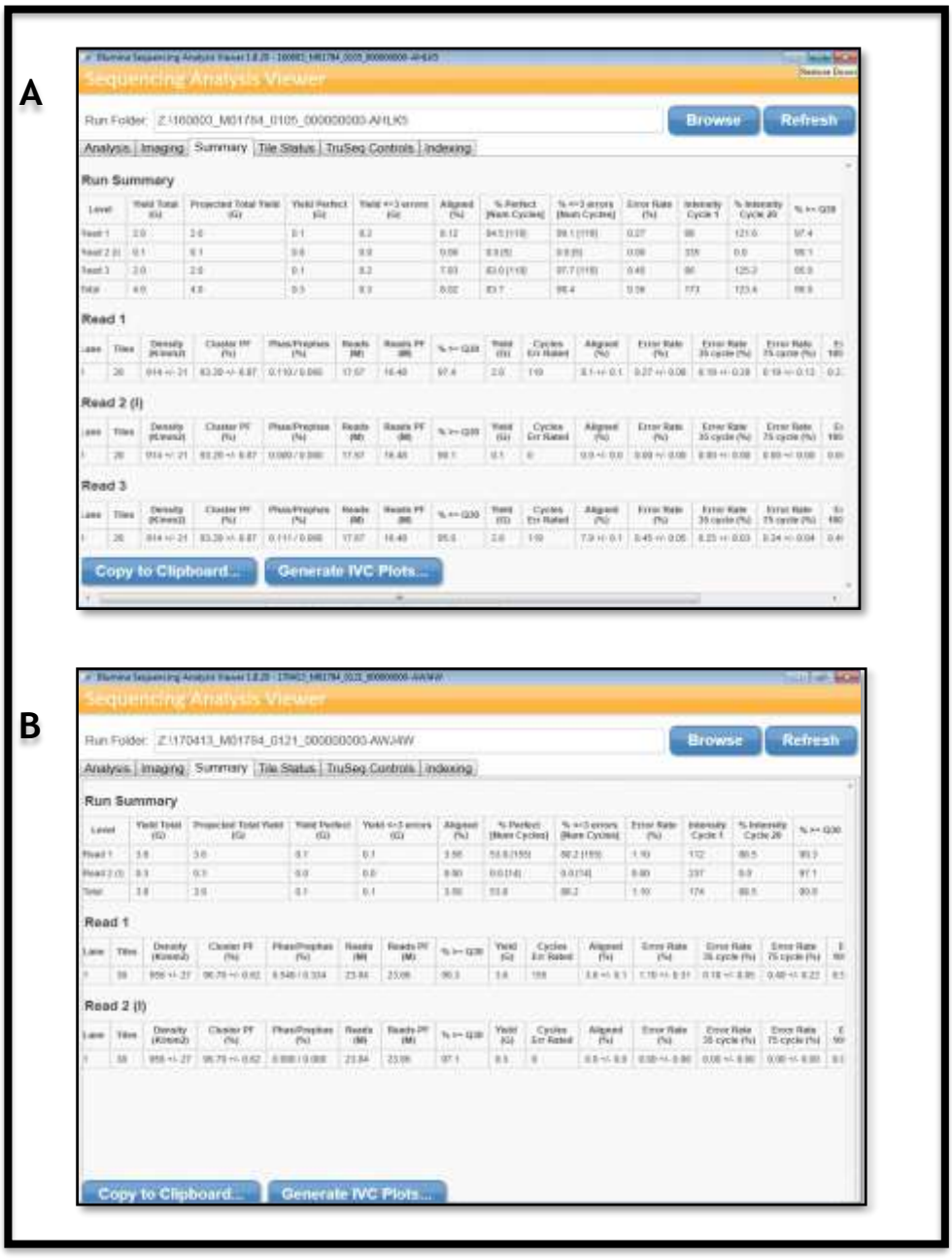


Figure 4.2. **MiSeq Run Summary Report. A)** Shreenshoot of MiSeq run summary report of one of 300bp paired-end runs for targeted breast cancer. Cluster Density: 914 k/mm2), Reads (M)= 17.67.48, Reads PF (M)= 16.48, Read 1 Q30 = 90.1 and Read 2 Q30 = 95. **B)** MiSeq summary report of one of the 150 bp single- end runs for TCR β-CDR3 sequencing. Cluster Density=956 k/mm2, Reads (M)=23.84, Reads PF (M)=23.06 and Read 1 Q30 = 90.3. (M=Million).

4.4. Breast Cancer Somatic Mutations

2,813 somatic mutations were identified from all of the 79 samples compiled in 46 genes. Using SIFT and PolyPhen, 1,557 were predicted to affect protein sequence; 588 were in Luminal A, 341 in Luminal B, 402 in TN and 226 in HER2. Only the variants that affect the protein sequence were included for further downstream analysis. Out of the 1,557 variants, 1378 variants were single nucleotide variants (SNVs) and 179 were dinucleotide variants. The translational impacts of these variants were as following: 1,372 missense, 177 in-frame, 5 frameshift, 3 stop codon.

Overall, 21 of the 46 genes tested were mutated in more than 10% of cases (Figure 4.3-A). *MUC16* was the most common altered gene identified. Variants in *MUC16* were present in 98.7% of patients, followed by *KMT2C, TP53, ZEBED4* and *ERBB2* which were present in 41%, 39.2%, 22.8% and 22.8% of patients, respectively.

The first four most frequently mutated genes in each molecular subtype are listed in Table 4.3. Among the 46 breast cancer genes studied. *MUC16, TP53, KMT2C* are the most frequent to contain mutations in all of the subtypes. *ERBB2* and *RET* mutations were enriched in Luminal A and Luminal B respectively, where each was present in >37 % of the patients.

Figure 4.3-B shows in detail the mutational frequency of the genes in each of breast cancer subtypes. HER2E samples harbored the least number of mutated genes compared to other subtypes, 54 % (24/46) of the breast cancer genes were

mutated in this group. Also, as shown in Figure 4.3-A &B no mutation was detected in *MYC, PTEN CBFB and PHF7* genes in any of the subtypes.

Table 4.3.

The Four Most Frequently Mutated Genes in Each Molecular

Luminal A		**Luminal B**		**HER2E**		**TN**	
(N=31)		(N=16)		(N=16)		(N=16)	
Mutated genes	Frequency (%)	Mutated genes	Frequency (%)	Mutated genes	Frequency (%)	Mutated genes	Frequency (%)
MUC16	96.7%	*MUC16*	100%	*MUC16*	100%	*MUC16*	100%
KMT2C	38.7%	*KMT2C*	50%	*TP53*	62.5%	*KMT2C*	43.7%
ERBB2	35.4%	*RET*	37.5%	*KMT2C*	31.25%	*TP53*	37.5%
TP53	32.2%	*TP53*	31.25%	*FGFR1*	18.75%	*FGFR1*	37.5%

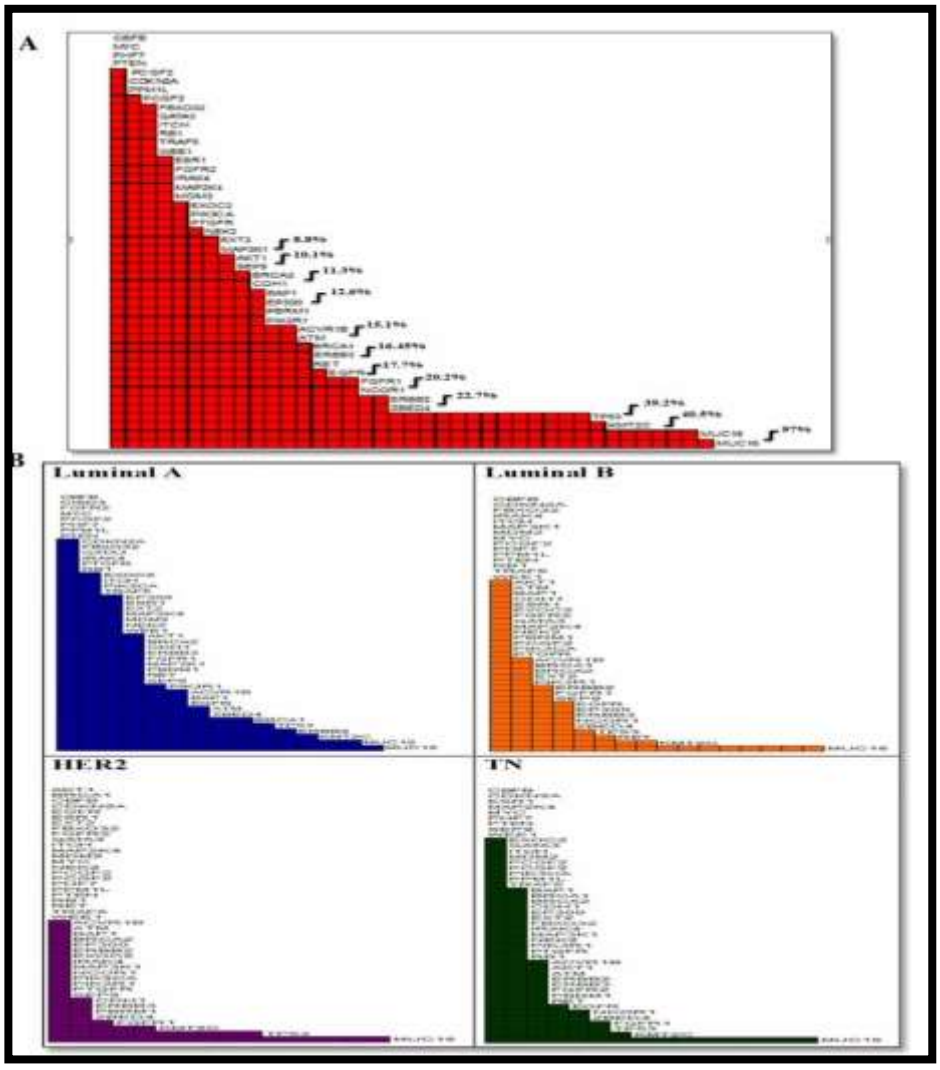


***Figure 4.3*. Spectrum of Mutations of 46 Genes Among 79 Patients. A)** squares represent the number of patients with mutations per gene. *MUC16* present in 78 patients and only split into 2 lines (for the figure resolution) **B)** The mutational frequency in all of 79 patients per gene in each of the four molecular breast cancer subtypes: Lumnal A (n=31), Luminal B (n=16), HER2E (n=16) and TN (n=16).

4.5. Gene Mutation Frequency Per Patient

The number of altered genes in each patient is detailed in Figure 4.4. The results showed that 67% (53/79) of breast cancer patients in this study had at least >=3 mutated genes. Most notably, 8 patients have high rate of genes alterations (>10 genes). Interestingly, only one patient (LA9) in Luminal A subtypes, did not show any mutated genes.

.

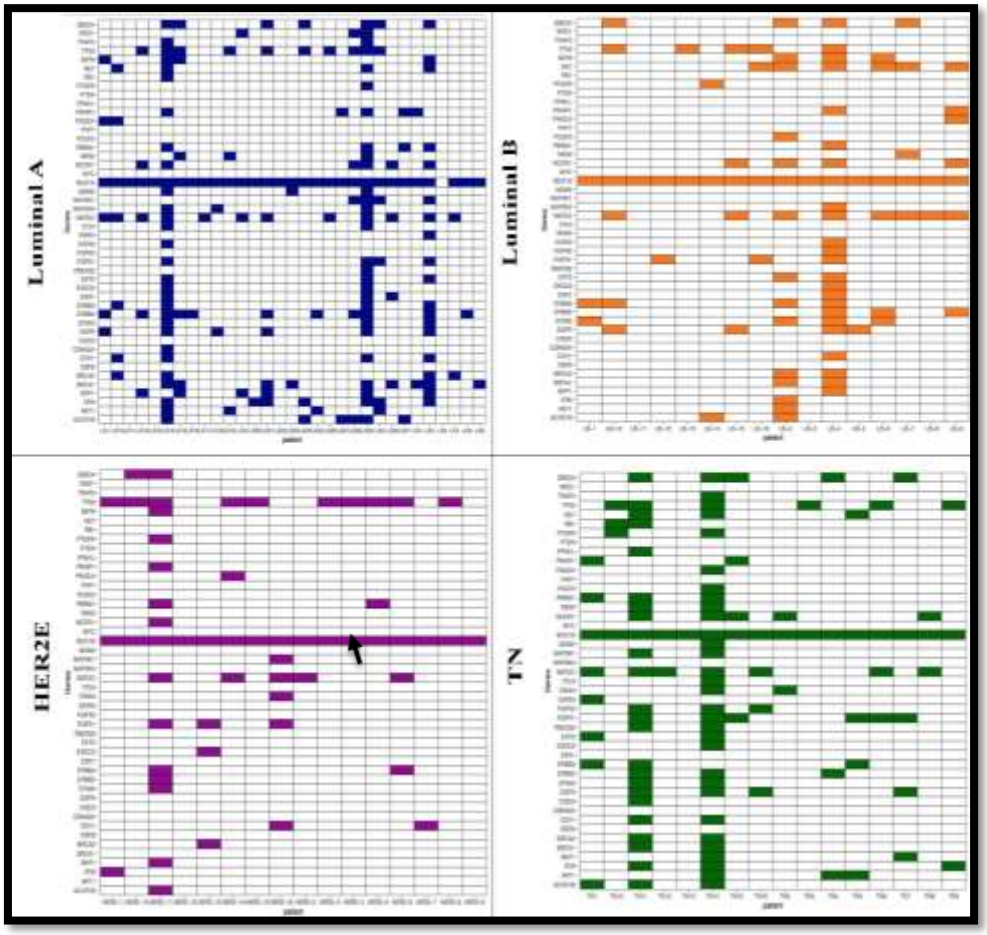


Figure 4.4. **Gene Mutation for Each Patient in the Four Molecular Subtypes.** Lumnal A (n=31), Luminal B (n=16), HER2E (n=16) and TN (n=16). X-axis represent 79 paients , Y-axis represent 46 genes.

4.6. Total Mutational Load

To test the hypothesis that higher mutational load will increase the likelihood of creating more neoantigens and this will lead to increased T cell quantity and diversity within the tumor microenvironment, the total number of mutations (variants) present in specimens of each subtypes were analyzed. Most notably, no specific patterns were observed in any subtypes as some studies reported with ER negatives (Haricharan et al., 2014). There were marked differences in mutational burden among patients. The median mutational load was TN:11, Luminal B: 9, Luminal A:9, HER2E:7 (Figure 4.5).

Regardless of the tumor subtypes, we selected 20 samples for further immune analysis. Three criteria were considered when the samples were selected. There are: the mutational load, QC call score (obtained from QIAseq^{TM} DNA QuantiMIZE assay which indicated the degree of DNA fragmentation) and the amount of gDNA left.

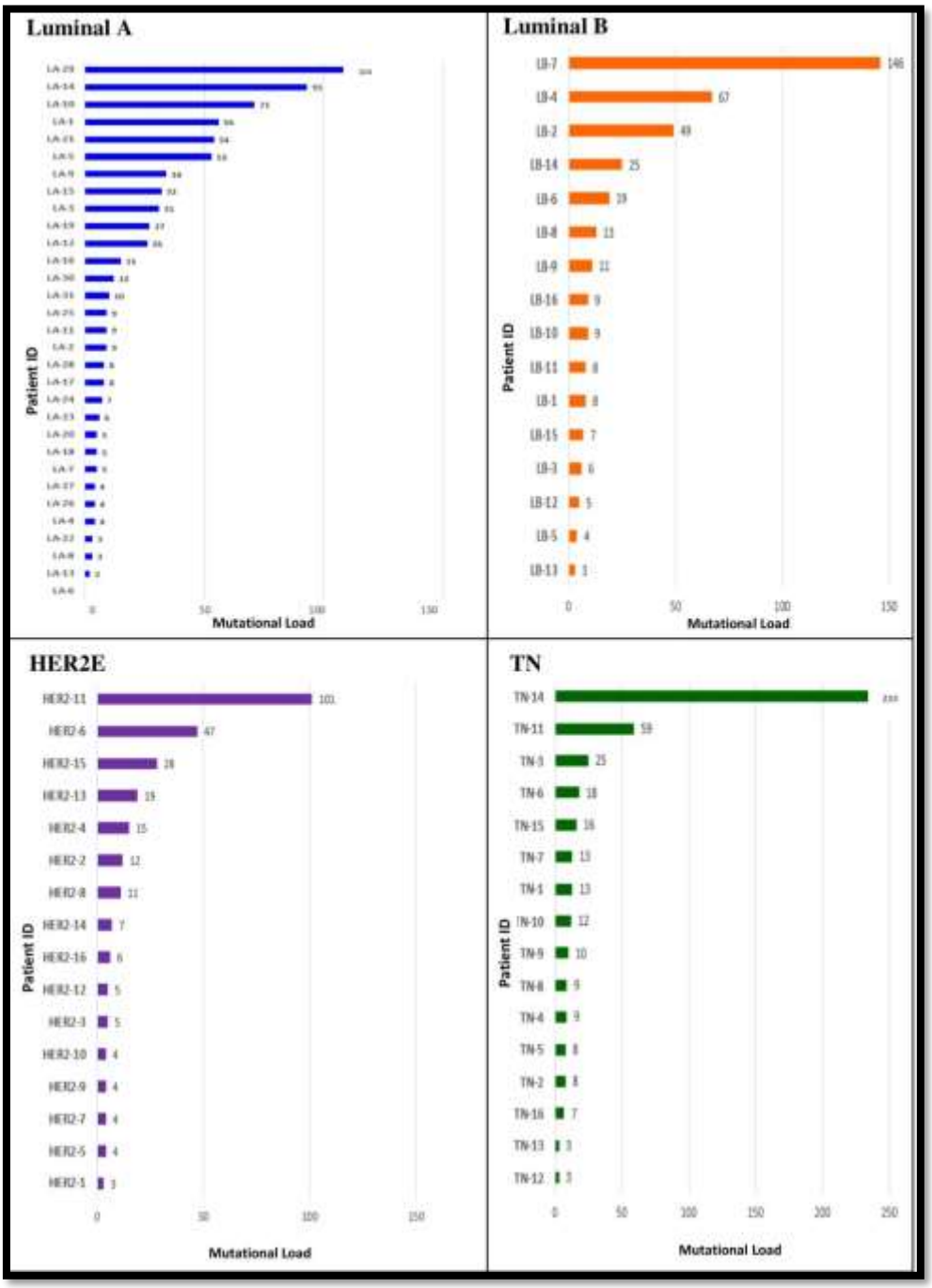


***Figure 4.5*. Somatic Mutaional Load in Each Patients of the Four Molecular Subtypes.** Luminal A (n=31), Luminal B (n=16), HER2E (n=16) and TN (n=16). The identified 1,557 variants were distributed by patient.

4.7. Mutational Signatures

All cancers are caused by somatically acquired mutations. In some cancer types, a proportion of somatic mutations are known to be generated by defects in the DNA repair process or due to exposure to certain carcengenesis; such as tobacco or ultraviolet light. These mutagenic processes have been previously reported and often produce distinctive mutational patterns in the cancer tissue (Alexandrov et al., 2013). In this study, the somatic mutational profile for each patient is analyzed using MutaGene to examine the most likely mutagenic signature behind the mutational pattern in each subtype (APPENDIX: B).

Most remarkably, no specific trends were observed in any of the subtypes. Most patients displayed large peaks under the substitution mutations C→T and T→C classes, which corresponded to UV radiation or unknown etiology (Figure 4.6-A). Also, the results showed that in 13 patients the deamination of methylcytosine was largely responsible for more than 55% of the patient's mutations (Figure 4.6-B). Tobacco or aflatoxin were other common mutagenics observed in 65 patients and contributed to 5% to 21% (mean 11%) of total patient mutations. AID/APOBEC contributed to the mutation burden in 32 patients by 5% to 18% (mean 7%).

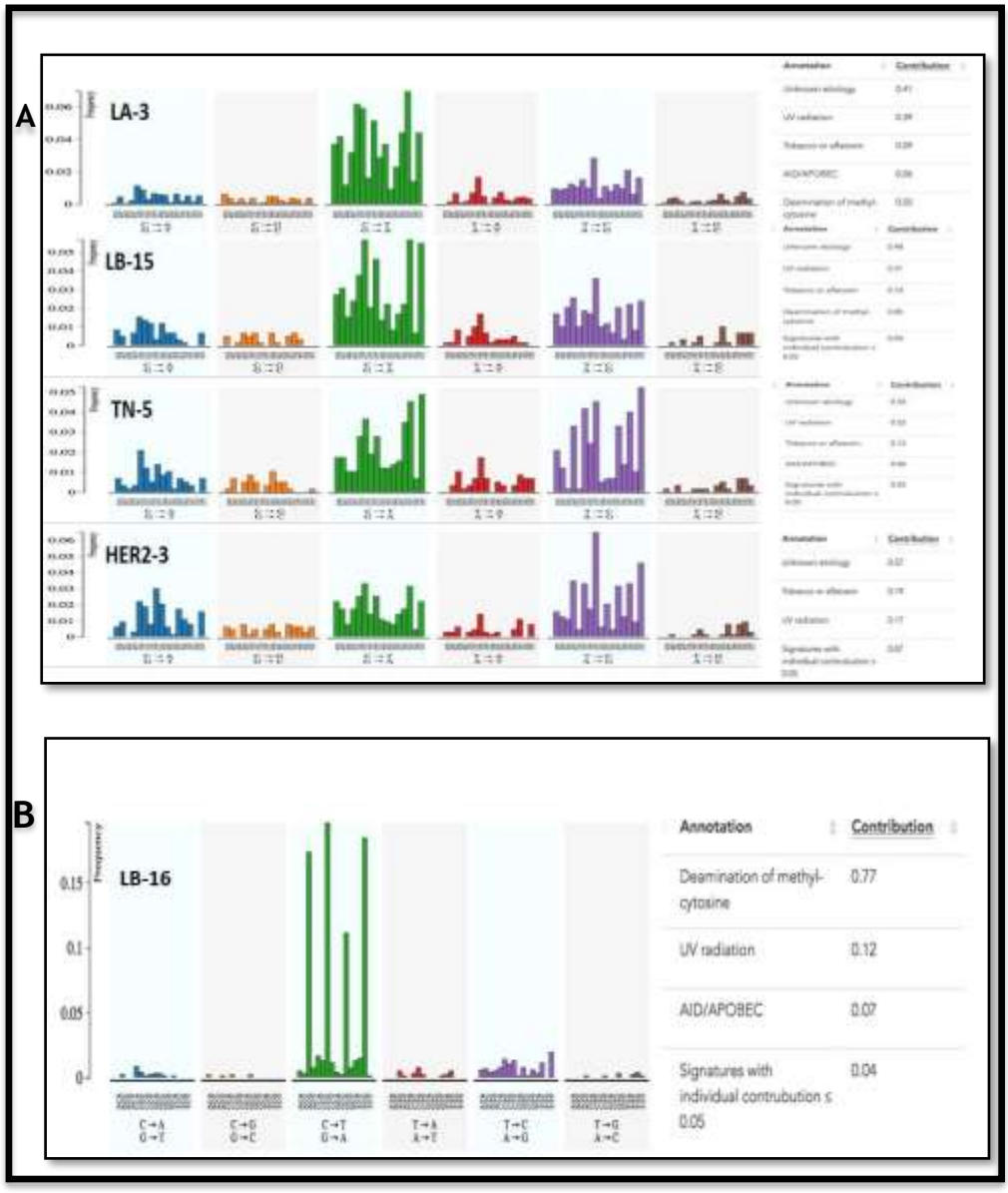


Figure 4.6. **Contribution of Mutational Signatures to Mutational Profile of the four molecular breast cancer subtypes**. **A)** The four patients shown as an example from each subtype**,** no specific trends were observed and most of the patients displayed large peaks under the substitution mutations C→T. **B**) LB-16 shown as example out of 13 patients of the studied patients displayed that deamination of methylcytosine was largely responsible for more than 55% of the patient's mutations.

4.8. Quality Control of TCR β CDR3 Library and Sequencing

Prior to sequencing, DNA damage repair kit was used for the selected 20 samples to improve the FFPE gDNA by repair any nicks on the DNA strands and fill in any gaps. Only 12 samples had sufficient starting input to be used for the ImmunoSEQ profiling assay at the survey level, listed in Table 4.4**.** Those samples had different spectrum of mutational load; some with low mutation and some with high mutational burden**.** Figure 4.7-A & B shows the TCR β-CDR3 amplification product in both the first and second PCR steps. The first PCR yielded a peak around 200 bp in size. After the addition of the NGS adapters and DNA barcodes during the second PCR, the final product at approximately 400 bp (as recommended by Adaptive Biotechnologies). Good quality immunoSEQ sequencing data was generated from the MiSeq instrument as shown in the run summary in Figure 4.2-B. Using The immunoSEQ analyzer, 11 of the samples passed the coverage and read quality thresholds required. The sample that did not pass the quality thresholds was excluded from further analysis.

Table 4.4.

Selected Samples for ImmunoSEQ Profiling Assay at the Survey Level and their Mutational Load

Patient ID	Mutational Load
LB-13	3
TN-12	3
TN-13	3
HER2-7	4
HER2-9	4
HER2-16	6
TN-1	13
TN-15	16
LB-14	25
HER2-15	28
HER2-6	47
LB-7	146

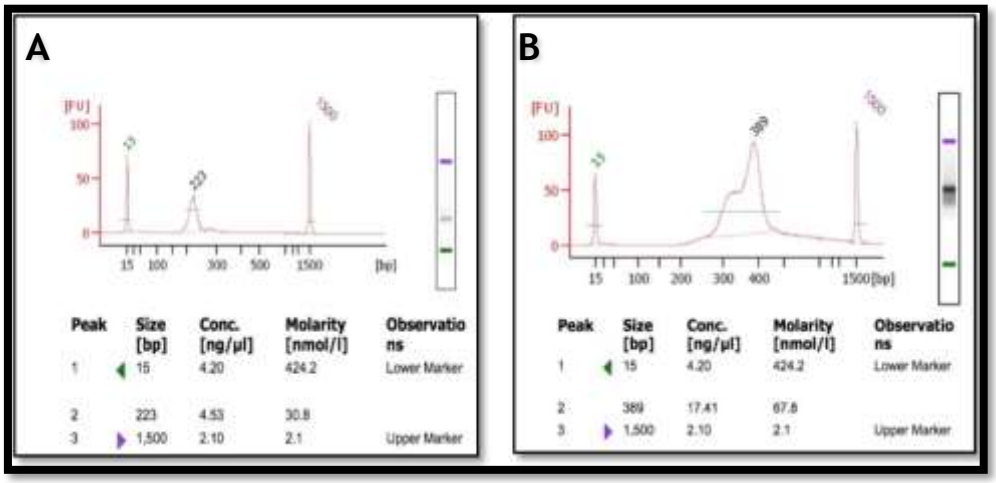


Figure 4.7. **Quality Check of TCR β-CDR3 Library from Both 1st and 2nd PCR Steps**. **A)** 1st PCR: product of amplifying CDR3 region exhibited a band at approximately 200 bp. **B)** 2nd PCR: final product after introducing NGS adapters and DNA barcodes yelided one band of approximately 400 bp.

4.9. TCRβ Rearrangement and Diversity

The total number of TCRβ rearrangements detected from the 11 samples was 22,960. 18533 of them were productive rearrangements that are in-frame and do not contain a stop codon. Of the productive rearrangements, the total number of unique TCRβ-CDR3 reads identified was 16570, distributed between 411 to 3206 per patient (median 1629). Most of Productive TCR-CDR3 numbers obtained from each sample were unique reads, indicating samples have high diversity level of TCRβ-CDR3 within tumor tissue studied (Figure 4.8).

Furthermore, the diversity of T clones in each sample was assessed. The score of the clonality was calculated based on the Shannon diversity index. The value of clonality was ranged from 0 to 1. Values close to 1 represent monoclonal distribution and values close to 0 represent polyclonal distribution. Using this metric, clonality of the TCRβ repertoires of the 11 tumor tissues ranged from 0.0032 to 0.0243 (median 0.0095). These results indicate that the TCRβ-CDR3 in these patient's tumors were more polyclonal compared to the median clonality of the ImmunoSEQ hsTCRB assay (Adaptive Biotechnologies, USA) of an adult T-cell repertoire in blood which is about 0.075.

Figure 4.9 shows the total count of unique TCRB-CDR3 and their level of diversity for each patient. Overall, HER2-16 and HER2-9 patients had the most abundant unique TCRB CDR3 rearrangements. Patient TN-15 with 16 mutational loads had the lowest total diversity of the TCR repertoire in the group; with only 411 unique CDR3. Remarkably, patients carrying only three somatic mutation had

> 1000 unique TCRβ sequence, while in other samples that had only one or three extra mutations, the diversity of TCR almost doubled (see patients HER2-16 and HER2-9). On the other hand, patients HER2-15 and HER2-6 with high mutational load, did not exhibit a remarkable increase in TCR diversity.

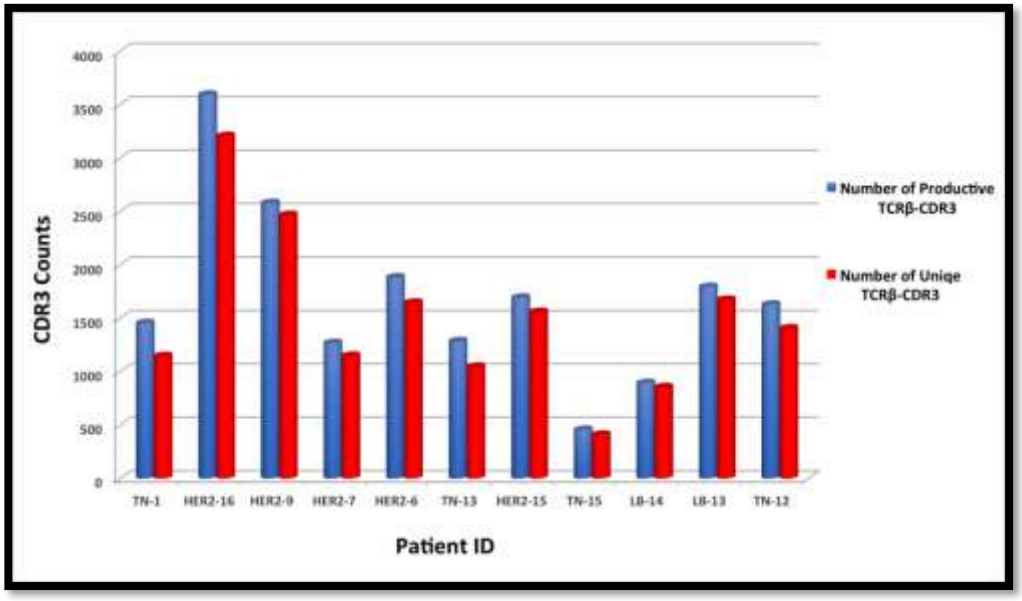


Figure 4.8. **Number of Productive and Unique TCRβ-CDR3 in Each Sample**. Most of Productive TCR-CDR3 numbers were unique reads, indicating high diversity level of TCRβ-CDR3 within tumor tissue studied.

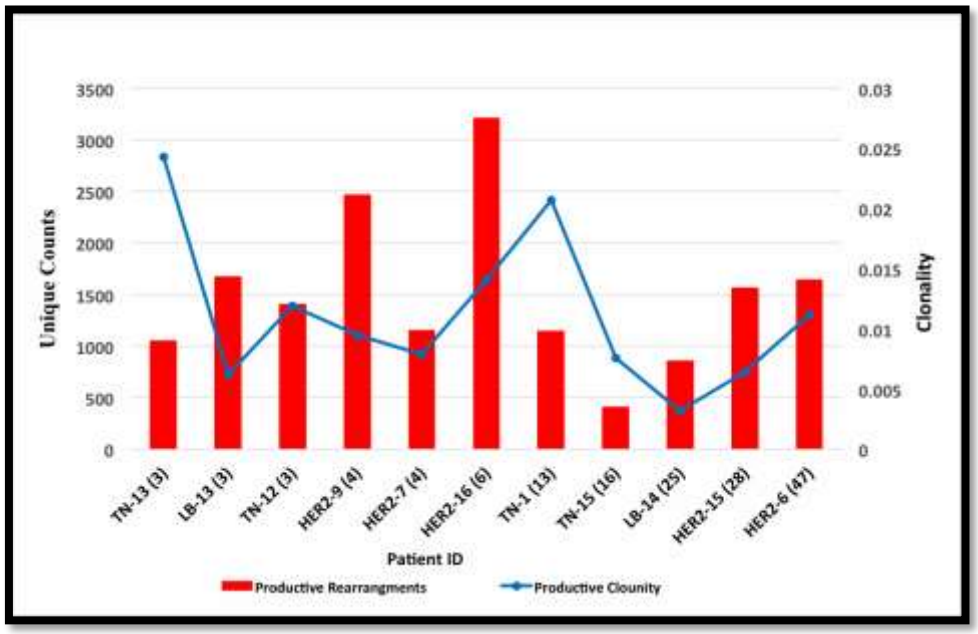


Figure 4.9. **TCRB CDR3 Diversity of 11 Patients.** The red bar represent the total count of unique TCRβ-CDR3 rearrangements for each patient. Blue line represents the value of clonality, which ranges from 0 (polyclonal distribution) to 1 (monoclonal distribution). The somatic mutational load of each patient stated next to the patient ID in the brackets.

4.10. Association of Mutational Load with TCRβ-CDR3 Diversity

The association between the mutational load and the clinical outcome in several cancers, including the breast have been previously reported (Haricharan et al., 2014), but there have been few reports on the effect of mutational load on TCRβ diversity in breast cancer and clonality level . In the study, association between T cell diversity and load of mutation were first tested (Figure 4.10-A). According to the p-value =0.61 the correlation between mutational load and the number of productive unique TCRβ-CDR3 in the studied patient is not statistically significant. Additionally, there was no significant relationship between the load of mutation and clonal diversity p = 0.43 (Figure 4.10 -B).

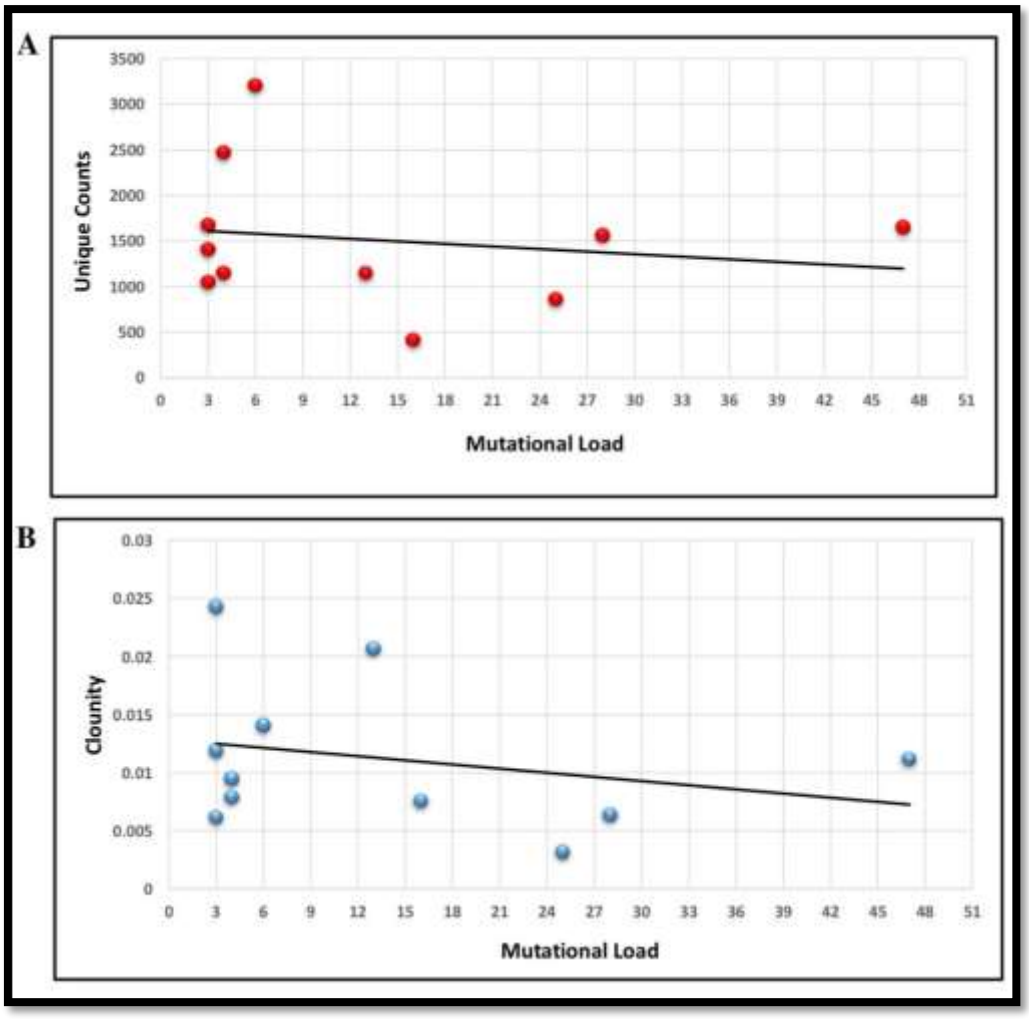


Figure 4.10. **Assocation Betweern Mutational Load and TCR β Diversity Within Tumor Tissues Studied. A)** The patients exhibited no signifigant assocation between mutational load and the number of productive TCRβ-CDR3 rearrangements (r = -0.172, p = 0.61). **B)** No significat relationship was displayed between mutaional load and clonal frequency (r = -0.2, p = 0.43).

CHAPTER 5 : DISCUSSION

5.1 Discussion

The relationship between the immune system and cancer has occupied the research community for many years now. The critical function of the TCR in the recognition of cancer cells has been well explained. The role of T cell in the inhibition of tumor formation and progression is also demonstrated in many studies (Dranoff, 2011; R. Kim et al., 2007). Additionally clinical studies of solid tumors including, breast cancer, confirmed the positive correlation between the presence of TILs and improved patient survival (Hadrup et al., 2013). However, there is limited knowledge about why some patients have increased TIL infiltration and diversity than others even within the same cancer subtypes.

A large number of studies have attempted to understand the mechanisms that are involved in regulating T cell infiltration and their clonal diversity within the tumor microenvironment. Such studies suggested that the high level of genomic instability in cancer cells might promote antitumor immune responses and increase the level of TILs through inducing and presenting a large number of tumor antigens (Criscitiello et al., 2014; Sherene Loi, 2013). The TCGA has also reported evidence showing a correlation between mutation burdens or neoantigen load with increased T cell infiltration within tumors (Brown et al., 2014; Roszik et al., 2016). Unfortunately, these relationships across breast cancers subtypes remains limited.

The traditional approaches of TIL characterization such as H&E staining and IHC are limited to providing qualitative measurements of TILs and cannot address the heterogeneity of the TCR diversity within a tumor. With the availability of NGS based methods, it has become possible to analyze T cell infiltration at the genetic level by sequencing the TCR rearrangements. This method provides deeper understanding of the complex relationship between somatic mutations and the diversity of TCRs within a tumor. Hence in this study we characterized somatic mutations of 79 FFPE breast cancer samples from patients within each breast cancer subtype using NGS sequencing of a panel of breast cancer associated genes. To understand the relationship between mutational load and T cell diversity, we sequenced TCRβ-CDR3 for 11 patients with different levels of mutational load.

In general, archived FFPE tissue is a highly valuable source of DNA, RNA and protein for molecular analysis of cancer. Hospitals are routinely creating several FFPE blocks from cancer patient biopsies and preserving for decades (Al-Attas et al., 2016). However, the biological material in FFPE could be degraded due to the process of fixation and embedding as well as the storage period and conditions. These factors might limit the FFPE usefulness in molecular studies. FFPE samples are mostly used as a reasonably reliable material in identifying cancer somatic mutations through targeted NGS approaches. However, the heavily degradation of isolated DNA could have negative impact on the results if not take into consideration prior to the sequencing process (De Leng et al., 2016).

In light of the negative impact of FFPE on the gDNA, we took several precautions to ensure producing high quality results from the NGS workflow, especially because our FFPE samples aged 4 to 13 years and were stored at room temperature. For example, the DNA extraction kit that we used (AllPreppa DNA/RNA FFPE Kit) helps in reversing formaldehyde modification that were induced during fixation process. Also important was selecting suitable samples for targeted NGS sequencing with good gDNA using the QIAseq^{TM} DNA QuantiMIZE assay. For the downstream process, the QIAGEN GeneReader NGS workflows compatible with FFPE material were used for the enrichment of 2915 amplicons of breast cancer genes, library preparation and somatic mutation analysis. These steps ensured that the selected 79 out 136 yielded high quality sequencing results and amplicons coverage. Moreover, along with these steps for reduce false positive results we only were called variants with high quality reads, deep depth and Allele fraction > 2.

The second objective of our study; the assessment of the TCRβ repertoires from long period archived FFPE specimen, was much more challenging. The available protocols for sequencing TCRβ- CDR3 from non-lymphoid tissue required high quality and quantity of gDNA. None of these protocols supported FFPE samples as input material. The justification that the gDNA in FFPE could be highly damaged and fragmented and thus may compromise the results of TCRβ-CDR3 identification.

The main issue in the use of damaged gDNA is that damaged bases could interfere with the DNA polymerase function that is to synthesize complementary

DNA strand from the damaged template during the PCR reaction. Furthermore, increased damage may artificially increase the number of TCRB clones that are reported. The DNA damage can lead to misincorporation, transient stalling and termination of DNA polymerization (Clark et al., 2011). To avoid these negative effects, several studies started to use the SMRTbell Damage repair kit to identify and fix DNA damage before high-throughput sequencing (Clark et al., 2011; K. E. Kim et al., 2014; Kong et al., 2017). Mostly these studied used the SMRTbell Damage repair kit to treat high molecular weight gDNA for denovo genome assembly application.

In accordance with these observations, we decided to use the SMRTbell Damage repair kit in our study to compensate for the poor gDNA quality from FFPE prior to immune sequencing. Before using the kit, we selected 20 FFPE samples with good gDNA quality. After treatment only 12 sample had sufficient starting material to proceed with immune sequencing. Based on ImmunoSEQ Analyzer web-based analysis for TCR beta database, we were successfully able to sequence 11 samples with good quality sequencing reads and output data.

For the data analysis, the ideal method for identification of somatic mutations in tumor is done by comparing the genomes of the tumor to the normal tissue derived from the same patient. In our case, the matched normal tissues were not available, so we relied on five publicly available databases to identify potential cancer somatic variants. This method is considered by several studies as an alternative pipeline for cancer variant identification whenever normal sample from the same patient is absent (Kalatskaya et al., 2017; Liu et al., 2015). Software such

as Biomedical Genomics workbench (CLC Bio, Qiagen) provides such a tool for these unpaired tissue samples. It incorporates the publicly available databases [dbSNP, 1000 Genomes Project, HapMap, Exome sequencing Projects and ExAC databases] that are rich with common germline variants, some of these databases included variants from Arab population (Koshy et al., 2017) . Furthermore, we also considered the variants that were present in more than 3% of our study as common variants in the population according to the definition of genetic polymorphism of the common allele or sequencing artifact as common results of PCR duplicates of NGS pipelines issues (Ebbert et al., 2016; Keats & Sherman, 2013).

In the breast cancer gene panels sequenced in this study, we evaluated the frequency of somatic mutations as well as the mutational load in 79 breast cancer patients. The most common genetic alteration identified in all of the subtypes was the *MUC16* gene also known as the CA125 gene, present in 98.7% of all patients. *MUC16* encodes proteins that paly an essential role in forming the protective mucous barrier. The products of the *MUC16* gene is mostly used as a marker for ovarian cancer, with higher expression levels correlated with poorer outcomes (NCBI, 2017). Similar findings has been reported in several other studies; the COSMIC database classified *MUC16* gene as one of the top frequently mutated genes in general cancers (Tan et al., 2015). Moreover, a meta-analysis study generated from a total of 602 breast cancer samples the *MUC16* was reported to be one of most frequently mutated genes in breast cancer but its mutational frequency ranged only between 4% to 14% across molecular subtypes (Cornen et al., 2014). The possible explanation for our observation of higher than usual mutational

frequency is because *MUC16* gene is the largest gene in the whole panel we sequenced, therefore more likely to have a mutation compared to rest of the genes in the panel.

Our result showed that *TP53* was also largely mutated in the breast cancer patients, this is in agreement with previous studies (Cornen et al., 2014; Liu et al., 2015; Tan et al., 2015). In addition, our data demonstrated that the mutational genes frequency were almost similar across subtypes. For example, *MUC16, KMT2C* and *TP53* were at the top of the list in terms of most frequently mutated genes across all molecular subtype, in addition the frequency of mutation on other genes were almost alike. Furthermore among our small sample size (79 patients) no mutation was detected in *MYC, PTEN, CBFB* and *PHF7* genes. Those genes also were not included in both lists of commonly mutated genes in general cancer of COSMIC databases and in breast cancer meta-analysis study that included huge number of patients in their analysis (Cornen et al., 2014; Tan et al., 2015).

In one of most comprehensive molecular studies done in breast cancer, somatic mutations of primary tumors were identified in 825 people using exome the sequencing approach (Koboldt et al., 2012). The study data exhibited that *TP53* was highly mutated in breast cancer, present in 37 % of the total patients. The aggressive clinical and biological types had higher *TP53* alteration ratios; 80% and 72 % in TN, HER2 respectively, followed by 29% and 12 % in Luminal B and Luminal A respectively. On our results also showed that the TN group also had the highest mutational frequency in *TP53* compared to other breast cancer groups. The TP53 gene was mutated in 62.5%, 37.5%, 32.2% and 31.25% in TN, HER2E,

Luminal A and Luminal B patients respectively. Taken together, with the absence of normal tissue pairs, our results revealed that mutations commonly found in breast cancer were at frequencies similar to those previously reported in other breast cancer molecular subtypes, thus indicating that FFPE mutational profiles are reliable and gDNAs extracted from FFPE samples were not significantly damaged. Furthermore. The results demonstrated the reliability of using QIAGEN GeneReader NGS workflows for breast cancer panel on FFPE samples.

In regards of mutational load, unlike the TCGA data set, that included 762 invasive breast cancer and showed significant correlation between ER negative subtypes and high somatic mutational burden (Haricharan et al., 2014), we identified no specific patterns of mutational load among any breast cancer groups. Mutational loads ranged from high to low in each subtype. These differences may be attributed to the fact that we were restricted to 46 genes while the TCGA data study used exome sequencing to cover all possible somatic mutations. In addition, we were so conservative when we called patient variants, thus may have lost some important variants in each subtype.

According to Alexandrov et al. there are ~30 genome wide mutational signatures across human cancers (Figure 5.1). Some of these signatures are very common and almost present in every cancer type, while others are rare and only found in specific types of cancers. C> T substitutions are frequently present in every identified mutational signature. Aetiologies of some signatures are associated with defective DNA mismatch repair or known mutagenic exposures, while others still remain unknown (COSMIC, 2017). In breast cancer disease, there are several

signatures that contribute to the somatic mutations. For example, signatures 8, 17, 18 and 30 have unknown aetiology, signatures 6, 20 and 26 are associated with defective DNA mismatch repair. Signatures 2 and 13 has been attributed to activity of the AID/APOBEC. Previous studies reported that deamination of 5-methylcytosines, an endogenous mutational process, modifies CpG sequence and occurs so frequently throughout evolution. This mutagenic process has been observed in all cancer types and samples (COSMIC, 2017; Nik-Zainal et al., 2016). In our samples, the mutational signatures were mostly attributed to unknown atetiology. Moreover, UV radiation, tobacco, aflatoxin and AID/APOBEC were the main underling somatic mutational profiles identified in our samples. 13 of our patients showed that deamination of 5-methylcytosines was behind > 55 % of the mutation pattern profiles. As been well known that formalin fixation process causes random artificial C > T mutation on FFPE samples (Munchel et al., 2015). This raise an important point for the possibilities of that mutational profiles for 13 patients affected from the fixation issue.

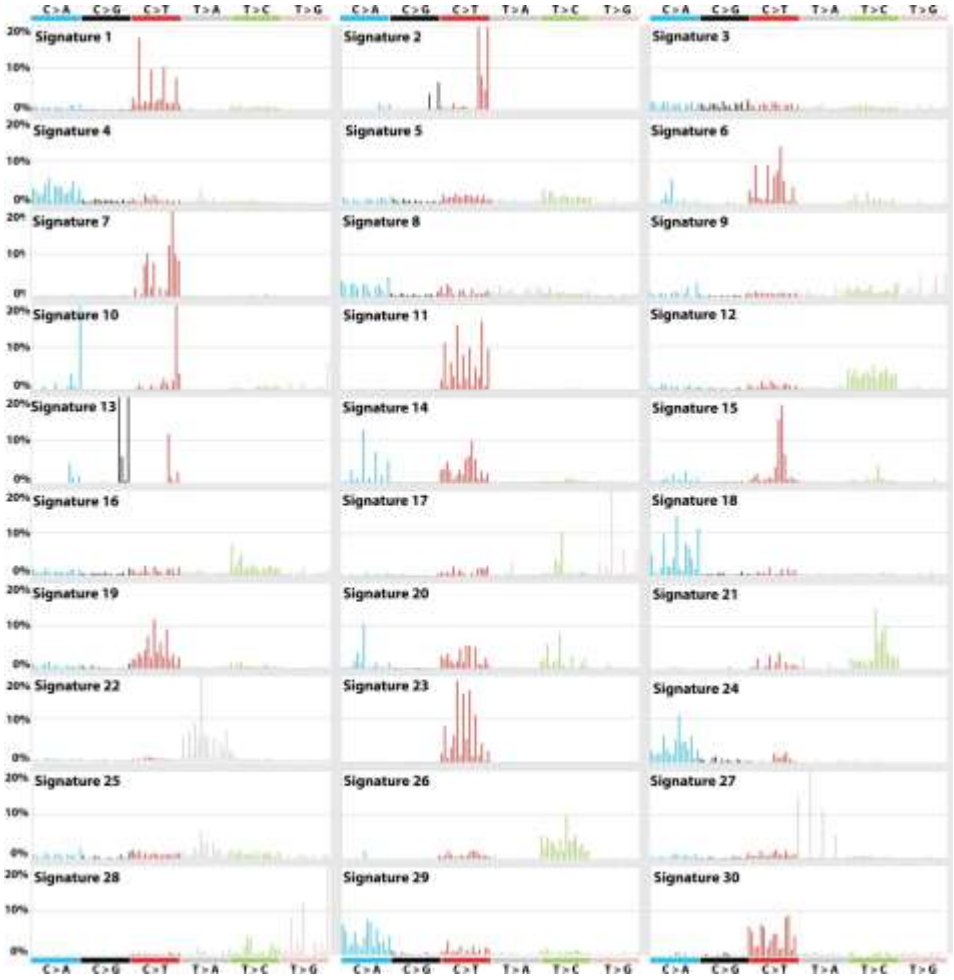


Figure 5.1. **Patterns of Mutational Signatures Found in Human Cancer.**X-axis displayed the 96 substitution classfications of mutational type difined by clasess and sequence context of the neighboring nucleotides in 5' and 3' directions from the mutated base. Y-axis represent the percentage of mutations attributed to specific mutational type (COSMIC, 2017).

In the present study, we demonstrated the possibility of using archived FFPE to evaluate and characterize TCR-CDR3 in the tumor microenvironment of several patients. Our data showed high diversity level of the TCRβ repertoire within each tumor. However, because of our limited sample size we failed to identify statistically significant relationship between somatic mutational load in the sequenced breast cancer genes and the diversity level of the TCRβ repertoire.

Other studies had extracted TCR sequences from 9142 RNA-seq samples across 29 cancer types including breast cancer. The study found that breast cancer samples displayed positive association between diversity of T cell clonotypes and somatic mutational load (Li et al., 2016). In addition, a study in breast cancer confirmed correlation between the composition of the TCR repertoire in tumors with somatic mutation patterns by assessing multiple regions of the tumors from 5 patients (Kato et al., 2017) . In our study, further effort is needed by increasing the number of samples to improve the statistical power of the relationship between mutational load and T-cell clonal diversity in the breast cancer.

5.2 Limitations and Prospective

There are several limitations in our study. First, sample size was small and this provided insufficient power for statistical test of association between the T cell diversity with the mutational load. Second, absence of paired normal tissue for each sample lead us to only rely on publicly available databases and stringent cut-off values for somatic variants. This may have led to excluding some important variants in our patients. Third, sample selection was based on the presence of high quantification level of lymphocytes in the tumor tissue and thus might have led to masking the real differences in diversity of TCR across 11 sequenced samples. Finally, the exact areas when the FFPE block was taken from in the tumor were not provided. As breast cancer known to be heterogeneous and TCR diversity is reported to change accordingly. This may have introduced some kind of sampling bias if selection was solely based on TIL infiltration without stander criteria being applied consistently on all the samples.

5.3 Future Direction

- Recently, Qiagen released advance digital DNA sequencing approach for use with FFPE samples. This technology depends on sequencing unique molecular indices (UMI) to allow deeper sequencing of PCR products and detects low-frequency variants with high confidence. Using a breast cancer panel of these technology would help to remove PCR duplicates and minimizes PCR amplification error.
- Sequencing TCRβ-CDR3 for more patients with different range of mutational load will helps on improve the statistical power of the association between mutational load and the TCRβ repertoire.
- Assessment of somatic mutational load and TCR from multiple regions in breast cancer of each samples would provide deeper understanding of the heterogeneity of TCRβ repertoire within patient samples and across other patients. Also, it would minimize sampling bias.
- Characterization of TCR-CDR3 from FFPE samples were successfully evaluated in this present study. Thus, should provide frame-work for analysis of large-scale of FFPE samples in Qatar in the future.

CHAPTER 6 : CONCLUSION AND RECOMMENDATION

In this study, we demonstrated that gDNA from FFPE samples could be used as a resource to identify somatic mutations of breast cancer using NGS based methods. Moreover, in the present study we also showed that the TCRβ repertoire could successfully be evaluated from the FFPE samples. However, the use of FFPE samples presents challenges in sample processing and correcting for false positive variant calls. These challenges are caused by formalin fixation process and storage period and conditions. These factors mostly introduce C>T artifact mutations and should be taken into account during analysis. Current studies in cancer are mostly using NGS high throughput technology to detect low allele frequency somatic variants. As the mutational signature studies showed there are many mutagenic processes that cause C>T mutation in human cancer genomes (Alexandrov et al., 2013; COSMIC, 2017), this limits the FFPE usefulness in low allele frequency molecular genetic analysis. This raises an important need for pathological laboratories in Qatar to change the collection and storage process of cancer samples from FFPE to frozen biospecimens. Frozen tissue storage method has already been initiated in pathology departments and biobanks in United Kingdom since the last few years (Shabihkhani et al., 2014).

REFERENCES

Al-Attas, A., Assidi, M., Al-Maghrabi, J., Dallol, A., Schulten, H. J., Abu-Elmagd, M., … Al-Qahtani, M. (2016). Enhancement of pathologist's routine practice: Reuse of DNA extracted from immunostained formalin-fixed paraffin-embedded (FFPE) slides in downstream molecular analysis of cancer. *Cancer Genomics and Proteomics*, *13*(5), 399–406.

Al Bader, S. B., Bugrein, H., Elmistiri, M., Elmistiri, M., & Alassam, R. (2016). The Development of Breast Cancer Screening in Qatar (January 2008 – April 2015). *Evidence Based Medicine and Practice*, *2*(2), 1–6.

Alberts, B., Johnson, A., Lewis, J., Raff, M., Roberts, K., & Walter, P. (2002). *Molecular biology of the cell*. Garland Science.

Alexandrov, L. B., Nik-Zainal, S., Wedge, D. C., Aparicio, S. A. J. R., Behjati, S., Biankin, A. V., … Stratton, M. R. (2013). Signatures of mutational processes in human cancer. *Nature*, *500*(7463), 415–421.

American Cancer Society. (2017a). Cancer Facts & Figures 2017.

Asano, Y., Kashiwagi, S., Goto, W., Kurata, K., Noda, S., Takashima, T., … Hirakawa, K. (2016). Tumour-infiltrating CD8 to FOXP3 lymphocyte ratio in predicting treatment responses to neoadjuvant chemotherapy of aggressive breast cancer. *British Journal of Surgery*, *103*(7), 845–854.

Bilimoria, M. M., & Morrow, M. (1995). The woman at increased risk for breast cancer: evaluation and management strategies. *CA: A Cancer Journal for Clinicians*, *45*(5), 263–78.

Brown, S. D., Warren, R. L., Gibb, E. A., Martin, S. D., Spinelli, J. J., Nelson, B. H., & Holt, R. A. (2014). Neo-antigens predicted by tumor genome meta-analysis correlate with increased patient survival. *Genome Research*, *24*(5), 743–750.

CDC. (2017). CDC - Breast Cancer Statistics. Retrieved August 25, 2017, from https://www.cdc.gov/cancer/breast/statistics/index.htm

Chaplin, D. D. (2010). Overview of the immune response. *The Journal of Allergy and Clinical Immunology*, *125*(2 Suppl 2), S3-23.

Cheang, M. C. U., Chia, S. K., Voduc, D., Gao, D., Leung, S., Snider, J., … Nielsen, T. O. (2009). Ki67 index, HER2 status, and prognosis of patients with luminal B breast cancer. *Journal of the National Cancer Institute*, *101*(10), 736–50.

Ciupe, S. M., Devlin, B. H., Markert, M., & Kepler, T. B. (2013). Quantification of total T-cell receptor diversity by flow cytometry and spectratyping. *BMC Immunology*, *14*(1), 35.

Clark, T. A., Spittle, K. E., Turner, S. W., & Korlach, J. (2011). Direct Detection and Sequencing of Damaged DNA Bases. *Genome Integrity*, *2*(1), 10.

Cornen, S., Guille, A., Adélaïde, J., Addou-Klouche, L., Finetti, P., Saade, M. R., … Chaffanet, M. (2014). Candidate luminal B breast cancer genes identified by genome, gene expression and DNA methylation profiling. *PLoS ONE*, *9*(1).

COSMIC. (2017). Signatures of Mutational Processes in Human Cancer. Retrieved November 18, 2017, from

http://cancer.sanger.ac.uk/cosmic/signatures

Criscitiello, C., Esposito, A., Gelao, L., Fumagalli, L., Locatelli, M., Minchella, I., … Curigliano, G. (2014). Immune approaches to the treatment of breast cancer, around the corner? *Breast Cancer Research : BCR*, *16*(1), 204.

Cruse, J. M., & Lewis, R. E. (Robert E. (2010). *Atlas of immunology*. CRC Press/Taylor & Francis.

Dai, X., Li, T., Bai, Z., Yang, Y., Liu, X., Zhan, J., & Shi, B. (2015). Breast cancer intrinsic subtype classification , clinical use and future trends, *5*(10), 2929–2943.

De Leng, W. W. J., Gadellaa-Van Hooijdonk, C. G., Barendregt-Smouter, F. A. S., Koudijs, M. J., Nijman, I., Hinrichs, J. W. J., … Lolkema, M. P. (2016). Targeted next generation sequencing as a reliable diagnostic assay for the detection of somatic mutations in tumours using minimal DNA amounts from formalin fixed paraffin embedded material. *PLoS ONE*, *11*(2), 1–18.

Dillon, J. L., Mockus, S. M., Ananda, G., Spotlow, V., Wells, W. A., Tsongalis, G. J., & Marotti, J. D. (2016). Somatic gene mutation analysis of triple negative breast cancers. *The Breast*, *29*, 202–207.

Dranoff, G. (2011). Experimental mouse tumour models: what can be learnt about human cancer immunology? *Nature Reviews Immunology*, *12*(1), 61.

Dunn, G. P., Bruce, A. T., Ikeda, H., Old, L. J., & Schreiber, R. D. (2002). Cancer immunoediting: from immunosurveillance to tumor escape. *Nature Immunology*, *3*(11), 991–998.

Dushyanthen, S., Beavis, P. A., Savas, P., Teo, Z. L., Zhou, C., Mansour, M., …

Loi, S. (2015). Relevance of tumor-infiltrating lymphocytes in breast cancer. *BMC Medicine*, *13*(1), 202.

Dziubianau, M., Hecht, J., Kuchenbecker, L., Sattler, A., Stervbo, U., Rödelsperger, C., … Babel, N. (2013). TCR repertoire analysis by next generation sequencing allows complex differential diagnosis of T cell-related pathology. *American Journal of Transplantation*, *13*(11), 2842–2854.

Ebbert, M. T. W., Wadsworth, M. E., Staley, L. A., Hoyt, K. L., Pickett, B., Miller, J., … Ridge, P. G. (2016). Evaluating the necessity of PCR duplicate removal from next-generation sequencing data and a comparison of approaches. *BMC Bioinformatics*, *17*(Suppl 7).

Eisenbeisz, J. (2016). *Mammography Review: for Technologists*. XRC LLC; 1 edition (July 24, 2016). Retrieved from https://books.google.com.qa/books?id=og04BAAAQBAJ&printsec=frontcover#v=onepage&q&f=false

Escors, D., & David. (2014). Tumour immunogenicity, antigen presentation and immunological barriers in cancer immunotherapy. *New Journal of Science*, *2014*.

Hadrup, S., Donia, M., & Thor Straten, P. (2013). Effector CD4 and CD8 T cells and their role in the tumor microenvironment. *Cancer Microenvironment : Official Journal of the International Cancer Microenvironment Society*, *6*(2), 123–33.

Haricharan, S., Bainbridge, M. N., Scheet, P., & Brown, P. H. (2014). Somatic mutation load of estrogen receptor-positive breast tumors predicts overall

survival: an analysis of genome sequence data. *Breast Cancer Research and Treatment*, *146*(1), 211–20.

Iwai, Y., Hamanishi, J., Chamoto, K., & Honjo, T. (2017). Cancer immunotherapies targeting the PD-1 signaling pathway. *Journal of Biomedical Science*, *24*(1), 26.

Janeway, C., Travers, P., Walport, M., & Shlomchik, M. (2001). *Immunobiology 5 : the immune system in health and disease*. Garland Pub.

Kalatskaya, I., Trinh, Q. M., Spears, M., McPherson, J. D., Bartlett, J. M. S., & Stein, L. (2017). ISOWN: accurate somatic mutation identification in the absence of normal tissue controls. *Genome Medicine*, *9*(1), 59.

Kato, T., Park, J.-H., Kiyotani, K., Ikeda, Y., Miyoshi, Y., & Nakamura, Y. (2017). Integrated analysis of somatic mutations and immune microenvironment of multiple regions in breast cancers. *Oncotarget*, *8*(37), 62029–62038.

Keats, B. J. B., & Sherman, S. L. (2013). Population Genetics. In *Emery and Rimoin's Principles and Practice of Medical Genetics* (pp. 1–12). Elsevier.

Kim, K. E., Peluso, P., Babayan, P., Yeadon, P. J., Yu, C., Fisher, W. W., … Landolin, J. M. (2014). Long-read, whole-genome shotgun sequence data for five model organisms. *Scientific Data*, *1*, 140045.

Kim, R., Emi, M., & Tanabe, K. (2007). Cancer immunoediting from immune surveillance to immune escape. *Immunology*, *121*(1), 1–14.

Koboldt, D. C., Fulton, R. S., McLellan, M. D., Schmidt, H., Kalicki-Veizer, J., McMichael, J. F., … Palchik, J. D. (2012). Comprehensive molecular

portraits of human breast tumours. *Nature*, *490*(7418), 61–70.

Kong, N., Ng, W., Thao, K., Agulto, R., Weis, A., Kim, K. S., … Weimer, B. C. (2017). Automation of PacBio SMRTbell NGS library preparation for bacterial genome sequencing. *Standards in Genomic Sciences*, *12*(1), 27.

Koshy, R., Ranawat, A., & Scaria, V. (2017). al mena: a comprehensive resource of human genetic variants integrating genomes and exomes from Arab, Middle Eastern and North African populations. *Journal of Human Genetics*, *62*(10), 889–894.

Lee, H. J., Park, I. A., Song, I. H., Shin, S.-J., Kim, J. Y., Yu, J. H., … Penault-Llorca, F. (2016). Tertiary lymphoid structures: prognostic significance and relationship with tumour-infiltrating lymphocytes in triple-negative breast cancer. *Journal of Clinical Pathology*, *69*(5), 422–430.

Li, B., Li, T., Pignon, J.-C., Wang, B., Wang, J., Shukla, S. A., … Liu, X. S. (2016). Landscape of tumor-infiltrating T cell repertoire of human cancers. *Nature Genetics*, *48*(7), 725–732.

Liu, S., Wang, H., Zhang, L., Tang, C., Jones, L., Ye, H., … Zhou, T. (2015). Rapid detection of genetic mutations in individual breast cancer patients by next-generation DNA sequencing. *Human Genomics*, *9*(1), 2.

Loi, S. (2013). Tumor-infiltrating lymphocytes, breast cancer subtypes and therapeutic efficacy. *Oncoimmunology*, *2*(7), e24720.

Loi, S., Michiels, S., Salgado, R., Sirtaine, N., Jose, V., Fumagalli, D., … Sotiriou, C. (2014). Tumor infiltrating lymphocytes are prognostic in triple negative breast cancer and predictive for trastuzumab benefit in early breast

cancer: results from the FinHER trial. *Annals of Oncology*, *25*(8), 1544–1550.

Lundegaard, C., Lund, O., Buus, S., & Nielsen, M. (2010). Major histocompatibility complex class I binding predictions as a tool in epitope discovery. *Immunology*, *130*(3), 309–18.

MacLeod, M. K. (2015). Antigen-based immunotherapy (AIT) for autoimmune and allergic disease. *Current Opinion in Pharmacology*, *23*, 11–16.

Makki, J. (2015). Diversity of Breast Carcinoma: Histological Subtypes and Clinical Relevance. *Clinical Medicine Insights. Pathology*, *8*, 23–31.

Martin, S. D., Brown, S. D., Wick, D. A., Nielsen, J. S., Kroeger, D. R., Twumasi-Boateng, K., … Nelson, B. H. (2016). Low Mutation Burden in Ovarian Cancer May Limit the Utility of Neoantigen-Targeted Vaccines. *PLOS ONE*, *11*(5), e0155189.

Metzger, G. J., Dankbar, S. C., Henriksen, J., Rizzardi, A. E., Rosener, N. K., Schmechel, S. C., … Molnar, B. (2012). Development of Multigene Expression Signature Maps at the Protein Level from Digitized Immunohistochemistry Slides. *PLoS ONE*, *7*(3), e33520.

Mihm, M. C., Mulé, J. J., & Mulé, D. J. J. (2015). Reflections on the Histopathology of Tumor-Infiltrating Lymphocytes in Melanoma and the Host Immune Response. *Cancer Immunology Research*, *3*(8), 827–35.

Munchel, S., Hoang, Y., Zhao, Y., Cottrell, J., Klotzle, B., Godwin, A. K., … Chien, J. (2015). Targeted or whole genome sequencing of formalin fixed tissue samples: potential applications in cancer genomics. *Oncotarget*, *6*(28).

Retrieved from www.impactjournals.com/oncotarget

Murphy, K. (2014). *Janeway's immunobiology.* Garland Science.

National Cancer Institute. (2013). Tumor Grade Fact Sheet - National Cancer Institute. Retrieved October 8, 2017, from https://www.cancer.gov/about-cancer/diagnosis-staging/prognosis/tumor-grade-fact-sheet

National Cancer Institute. (2015). Cancer Staging - National Cancer Institute. Retrieved October 8, 2017, from https://www.cancer.gov/about-cancer/diagnosis-staging/staging

National cancer Institution. (2012). Breast Cancer Risk in American Women - National Cancer Institute. Retrieved October 8, 2017, from https://www.cancer.gov/types/breast/risk-fact-sheet

NCBI. (2017, October). MUC16 mucin 16, cell surface associated [Homo sapiens (human)]. Retrieved October 25, 2017, from http://www.ncbi.nlm.nih.gov/pubmed/94025

Nik-Zainal, S., Davies, H., Staaf, J., Ramakrishna, M., Glodzik, D., Zou, X., … Stratton, M. R. (2016). Landscape of somatic mutations in 560 breast cancer whole-genome sequences. *Nature*, *534*(7605), 47–54.

Perou, C. M., Sørlie, T., Eisen, M. B., van de Rijn, M., Jeffrey, S. S., Rees, C. A., … Botstein, D. (2000). Molecular portraits of human breast tumours. *Nature*, *406*(6797), 747–752.

Petrelli, F., Cabiddu, M., Cazzaniga, M. E., Cremonesi, M., & Barni, S. (2008). Targeted therapies for the treatment of breast cancer in the post-trastuzumab era. *The Oncologist*, *13*(4), 373–81.

Qatar Cancer Society. (2015). Breast Cancer. Retrieved November 3, 2017, from http://www.qcs.qa/?page_id=284

Qin, X. J., & Ling, B. X. (2012). Proteomic studies in breast cancer (Review). *Oncology Letters*, *3*(4), 735–743.

Rathore, A. S., Kumar, S., Konwar, R., Makker, A., Negi, M. P. S., & Goel, M. M. (2014). CD3+, CD4+ & CD8+ tumour infiltrating lymphocytes (TILs) are predictors of favourable survival outcome in infiltrating ductal carcinoma of breast. *The Indian Journal of Medical Research*, *140*(3), 361–9. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/25366203

Robins, H. S., Campregher, P. V, Srivastava, S. K., Wacher, A., Turtle, C. J., Kahsai, O., … Carlson, C. S. (2009). Comprehensive assessment of T-cell receptor beta-chain diversity in alphabeta T cells. *Blood*, *114*(19), 4099–107.

Roszik, J., Haydu, L. E., Hess, K. R., Oba, J., Joon, A. Y., Siroy, A. E., … Woodman, S. E. (2016). Novel algorithmic approach predicts tumor mutation load and correlates with immunotherapy clinical outcomes using a defined gene mutation set. *BMC Medicine*, *14*(1), 168.

Rybakin, V., Westernberg, L., Fu, G., Kim, H.-O., Ampudia, J., Sauer, K., & Gascoigne, N. R. J. (2014). Allelic Exclusion of TCR α-Chains upon Severe Restriction of Vα Repertoire. *PLoS ONE*, *9*(12), e114320.

Saada, R., Weinberger, M., Shahaf, G., & Mehr, R. (2007). Models for antigen receptor gene rearrangement: CDR3 length. *Immunology and Cell Biology*, *85*(April), 323–332.

Salgado, R., Denkert, C., Demaria, S., Sirtaine, N., Klauschen, F., Pruneri, G., …

Loi, S. (2015). The evaluation of tumor-infiltrating lymphocytes (TILs) in breast cancer: recommendations by an International TILs Working Group 2014. *Annals of Oncology*, *26*(2), 259–271.

Savage, P. A., Leventhal, D. S., & Malchow, S. (2014). Shaping the repertoire of tumor-infiltrating effector and regulatory T cells. *Immunological Reviews*, *259*(1), 245–58.

Shabihkhani, M., Lucey, G. M., Wei, B., Mareninov, S., Lou, J. J., Vinters, H. V, … Edu, G. U. (2014). The procurement, storage, and quality assurance of frozen blood and tissue biospecimens in pathology, biorepository, and biobank settings. *Clin Biochem*, *47*(0), 258–266.

Smyth, M. J., & Hayakawa, Y. (2004). EVIDENCE FOR THE EXISTENCE OF CANCER IMMUNOSURVEILLANCE. *Annals of Cancer Research and Therapy*, *12*(1/2), 9–32.

Sotiriou, C., & Pusztai, L. (2009). Gene-Expression Signatures in Breast Cancer. *New England Journal of Medicine*, *360*(8), 790–800.

Stanton, S. E., & Disis, M. L. (2016). Clinical significance of tumor-infiltrating lymphocytes in breast cancer. *Journal for Immunotherapy of Cancer*, *4*, 59.

Tan, H., Bao, J., & Zhou, X. (2015). Genome-wide mutational spectra analysis reveals significant cancer-specific heterogeneity. *Scientific Reports*, *5*(1), 12566.

Viale, G. (2012). The current state of breast cancer classification. *Annals of Oncology*, *23*(suppl 10), x207–x210.

Wang, T., Wang, C., Wu, J., He, C., Zhang, W., Liu, J., … Liu, X. (2017). The

Different T-cell Receptor Repertoires in Breast Cancer Tumors, Draining Lymph Nodes, and Adjacent Tissues. *Cancer Immunology Research*, *5*(2). Retrieved from http://0-cancerimmunolres.aacrjournals.org.

Weigelt, B., & Reis-Filho, J. S. (2009). Histological and molecular types of breast cancer: is there a unifying taxonomy? *Nature Reviews. Clinical Oncology*, *6*(12), 718–30.

Weisman, P. S., Ng, C. K. Y., Brogi, E., Eisenberg, R. E., Won, H. H., Piscuoglio, S., … Wen, H. Y. (2016). Genetic alterations of triple negative breast cancer by targeted next-generation sequencing and correlation with tumor morphology. *Modern Pathology*, *29*(5), 476–488.

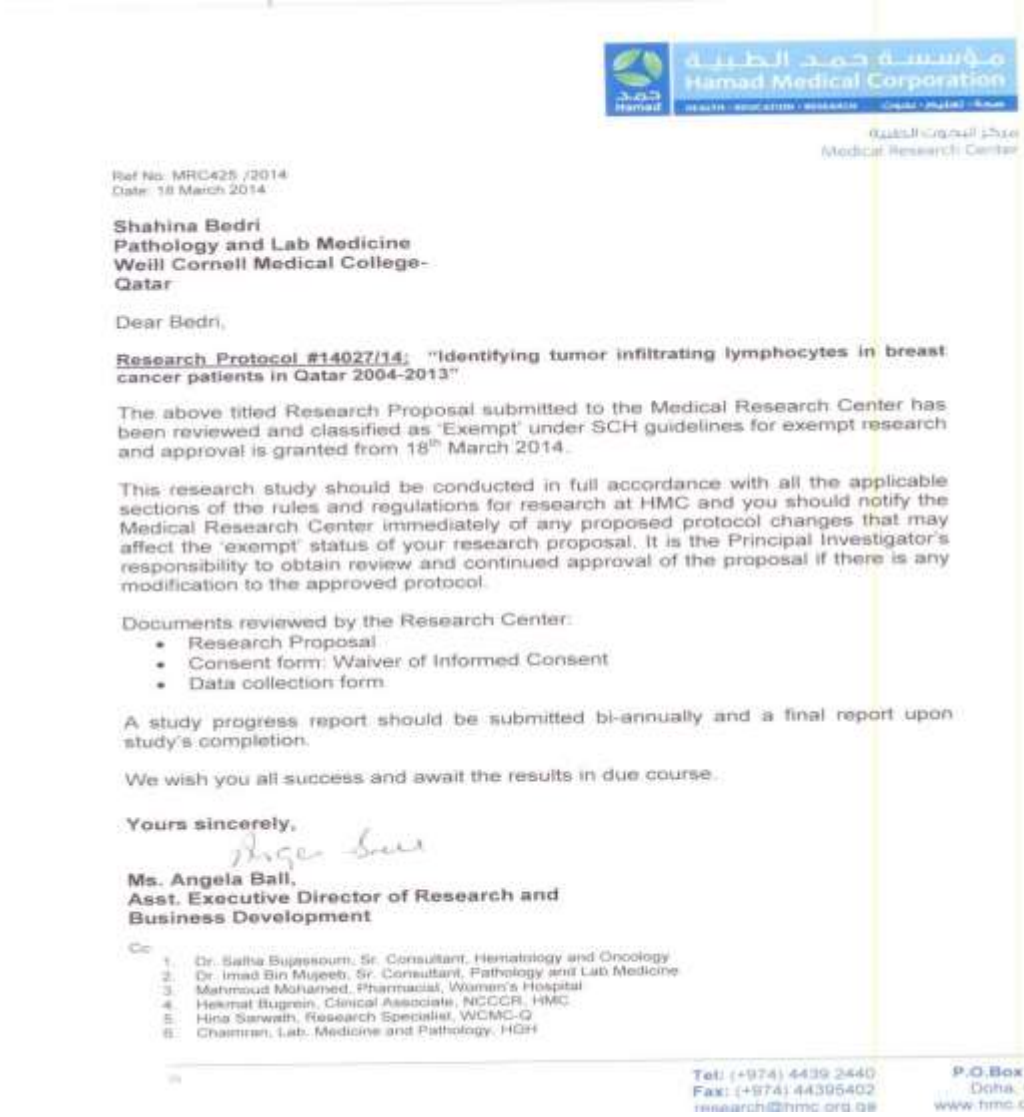
Woodsworth, D. J., Castellarin, M., & Holt, R. A. (2013). Sequence analysis of T-cell repertoires in health and disease. *Genome Medicine*, *5*(10), 98.

Yersal, O., & Barutca, S. (2014). Biological subtypes of breast cancer: Prognostic and therapeutic implications. *World Journal of Clinical Oncology*, *5*(3), 412–24.

Zhang, M. H., Man, H. T., Zhao, X. D., Dong, N., & Ma, S. L. (2014). Estrogen receptor-positive breast cancer molecular signatures and therapeutic potentials (Review). *Biomedical Reports*, *2*(1), 41–52.

APPENDICES

Appendix A: Ethical Approval



مؤسسة حمد الطبية
Hamad Medical Corporation

مركز البحوث الطبية
Medical Research Center

Ref No: MRC425 /2014
Date: 18 March 2014

Shahina Bedri
Pathology and Lab Medicine
Weill Cornell Medical College-
Qatar

Dear Bedri,

Research Protocol #14027/14: "Identifying tumor infiltrating lymphocytes in breast cancer patients in Qatar 2004-2013"

The above titled Research Proposal submitted to the Medical Research Center has been reviewed and classified as 'Exempt' under SCH guidelines for exempt research and approval is granted from 18th March 2014.

This research study should be conducted in full accordance with all the applicable sections of the rules and regulations for research at HMC and you should notify the Medical Research Center immediately of any proposed protocol changes that may affect the 'exempt' status of your research proposal. It is the Principal Investigator's responsibility to obtain review and continued approval of the proposal if there is any modification to the approved protocol.

Documents reviewed by the Research Center:

- Research Proposal
- Consent form: Waiver of Informed Consent
- Data collection form

A study progress report should be submitted bi-annually and a final report upon study's completion.

We wish you all success and await the results in due course.

Yours sincerely,

Ms. Angela Ball,
Asst. Executive Director of Research and
Business Development

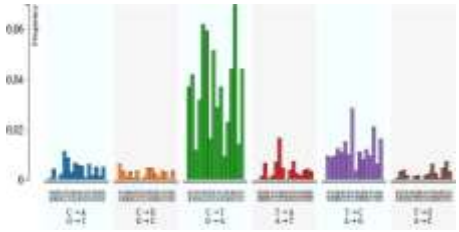
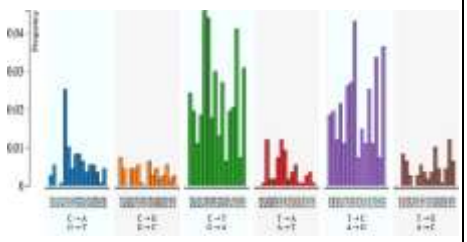
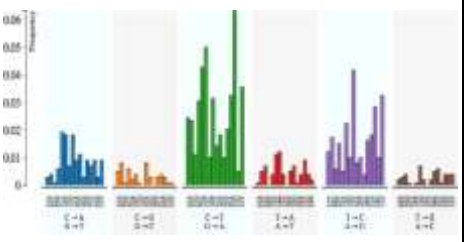
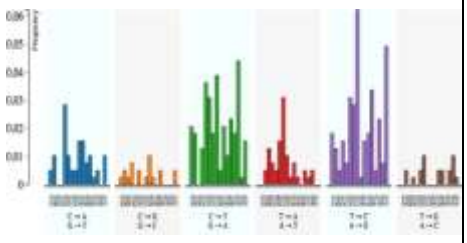
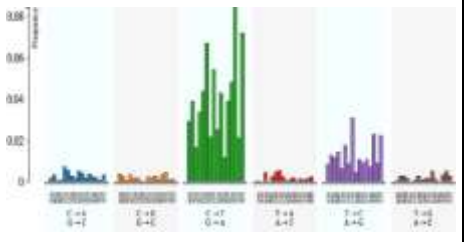
Cc

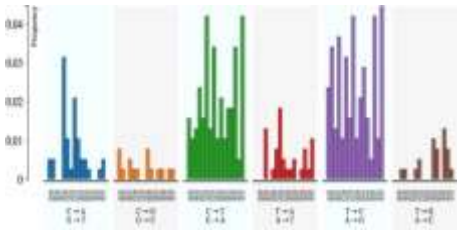
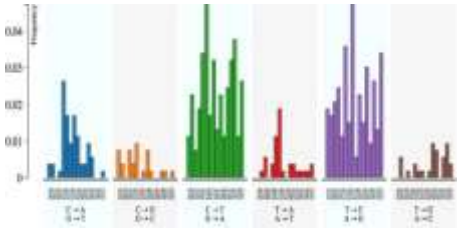
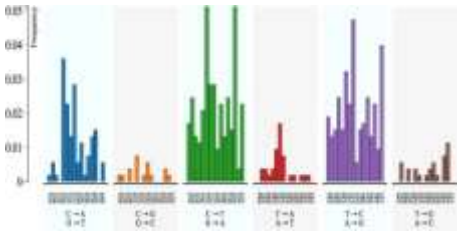
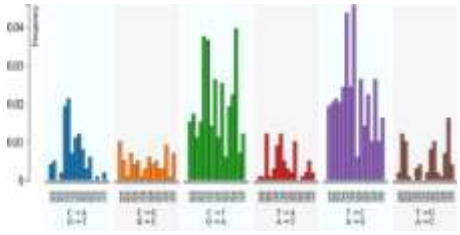
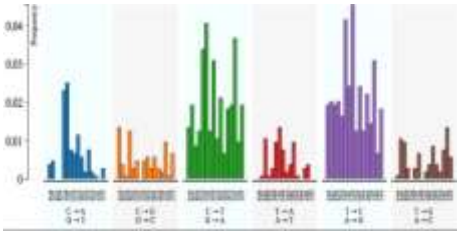
1. Dr. Sajna Bujassoum, Sr. Consultant, Hematology and Oncology
2. Dr. Imad Bin Mujeeb, Sr. Consultant, Pathology and Lab Medicine
3. Mahmoud Mohamed, Pharmacist, Women's Hospital
4. Hekmat Bugrein, Clinical Associate, NCCCR, HMC
5. Hina Sarwath, Research Specialist, WCMC-Q
6. Chairman, Lab. Medicine and Pathology, HGH

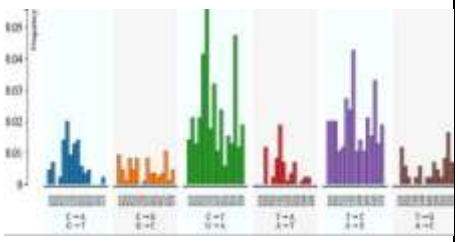
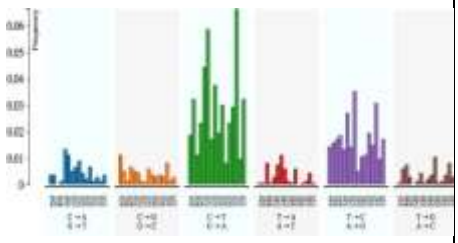
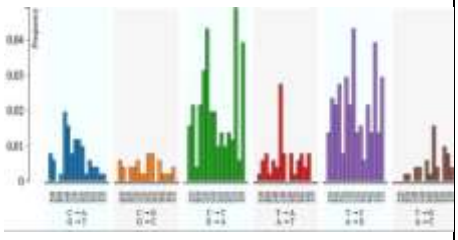
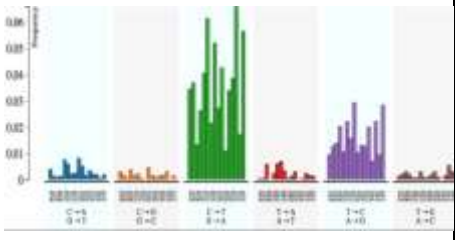
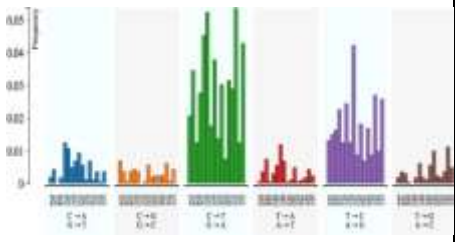
Tel: (+974) 4439 2440
Fax: (+974) 44395402
research@hmc.org.qa

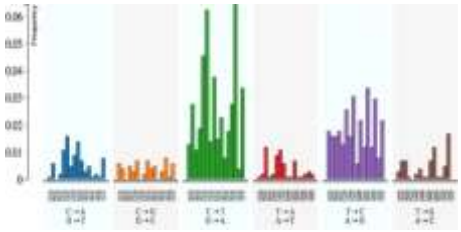
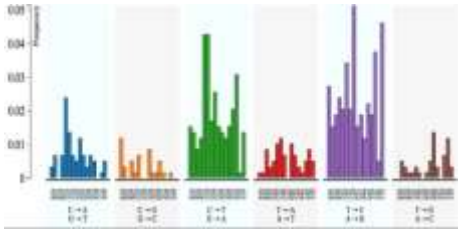
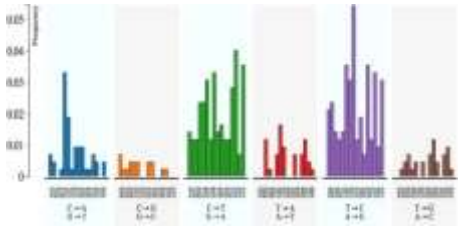
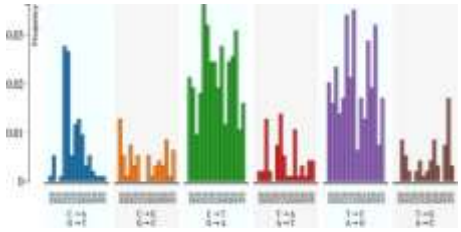
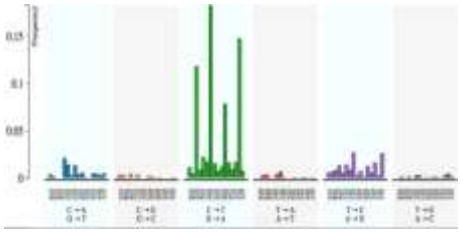
P.O.Box 3050
Doha, Qatar
www.hmc.org.qa

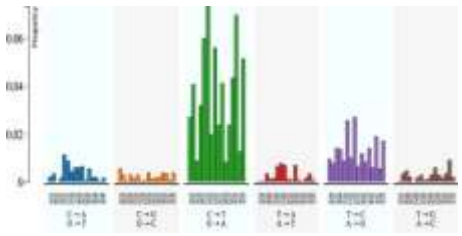
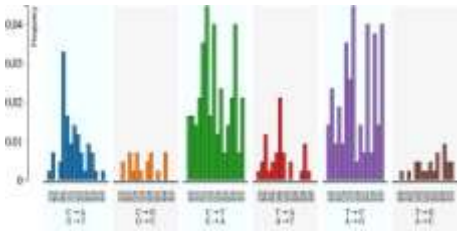
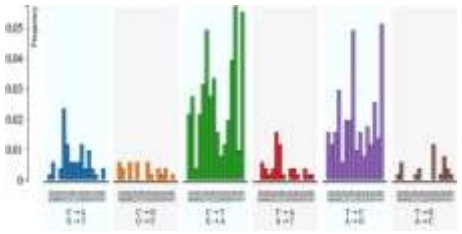
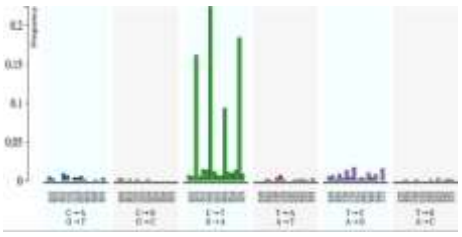
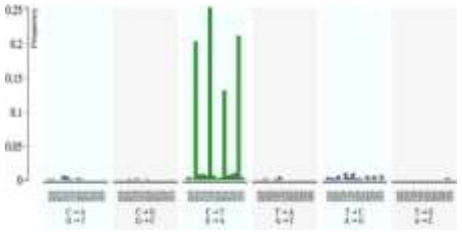
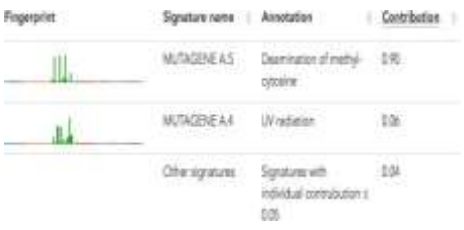
Appendix B: Results of Mutational Signatures for 79 patients

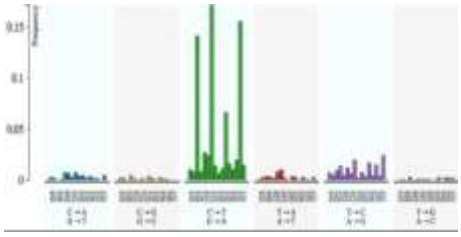
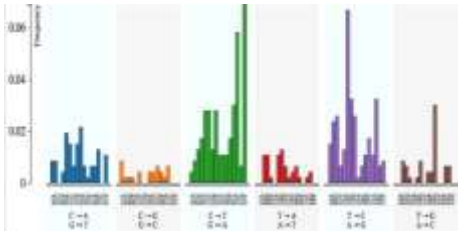
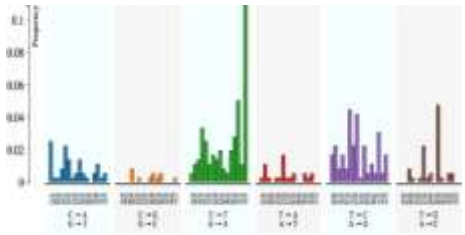
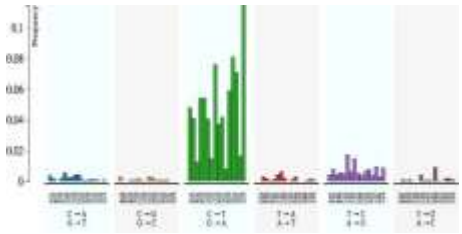
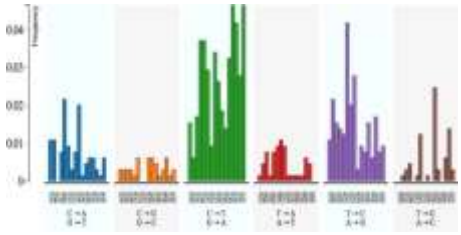
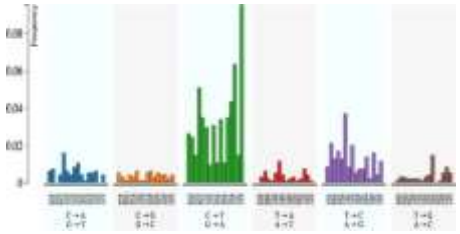
Patient ID	Mutational Profile / Fingerprint	Contribution of Mutational Signatures to Mutational Profile
LA-1		
LA-2		
LA-3		
LA-4		
LA-5		

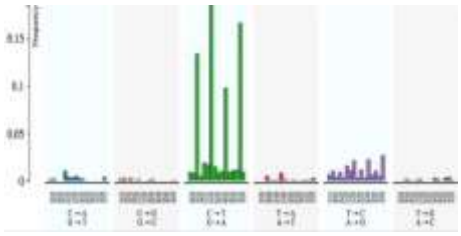
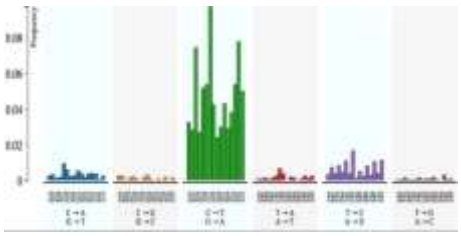
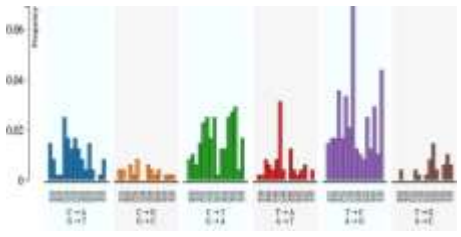
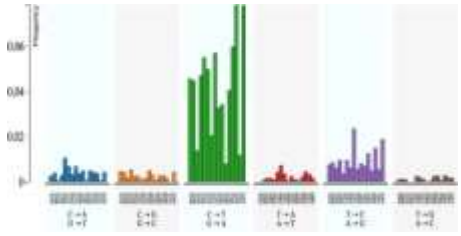
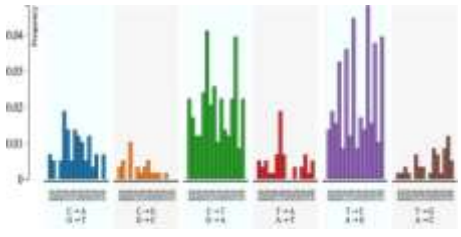
Patient ID	Mutational Profile / Fingerprint	Contribution of Mutational Signatures to Mutational Profile
LA-6		
LA-7		
LA-8		
LA-9		
LA-10		

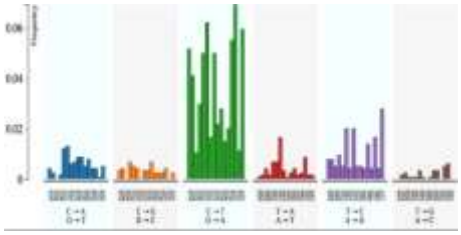
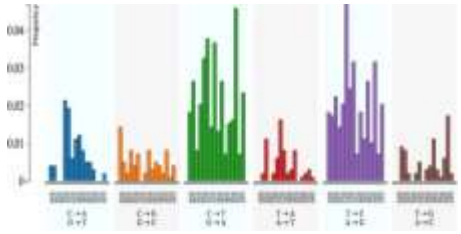
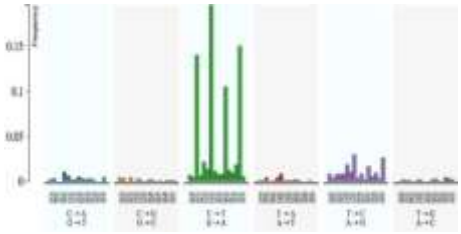
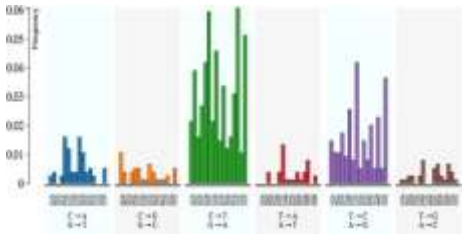
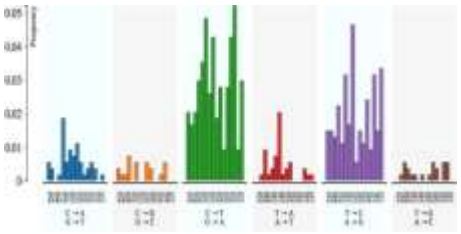
Patient ID	Mutational Profile / Fingerprint	Contribution of Mutational Signatures to Mutational Profile
LA-11		
LA-12		
LA-13		
LA-14		
LA-15		

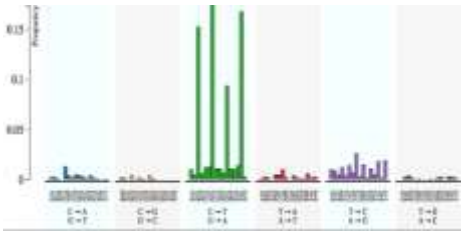
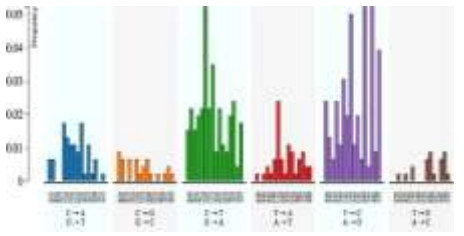
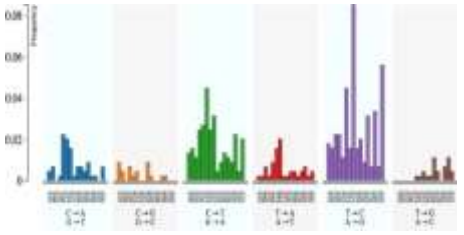
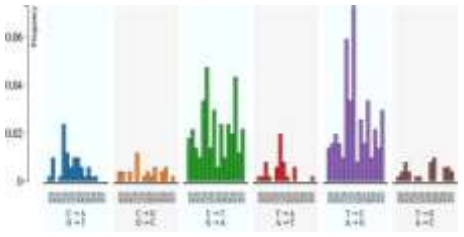
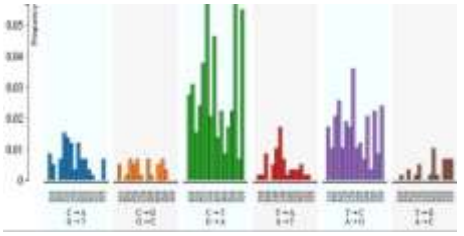
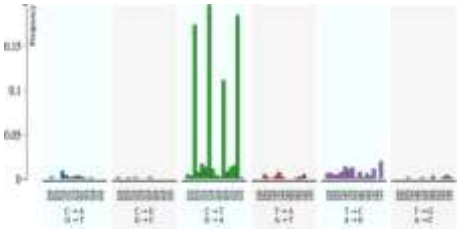
Patient ID	Mutational Profile / Fingerprint	Contribution of Mutational Signatures to Mutational Profile
LA-16		
LA-17		
LA-18		
LA-19		
LA-20		

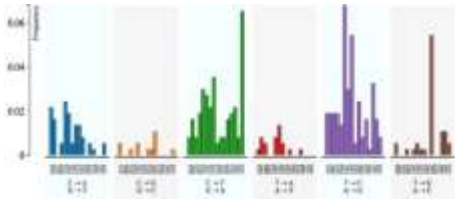
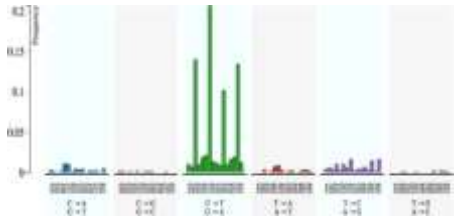
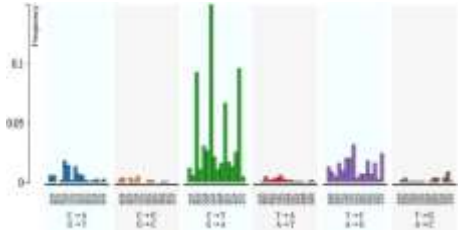
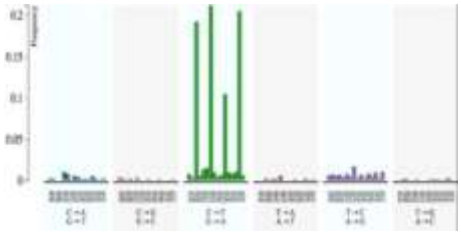
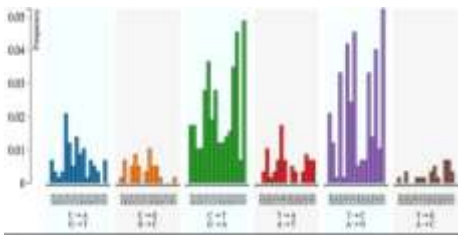
Patient ID	Mutational Profile / Fingerprint	Contribution of Mutational Signatures to Mutational Profile
LA-21		
LA-22		
LA-23		
LA-24		
LA-25		

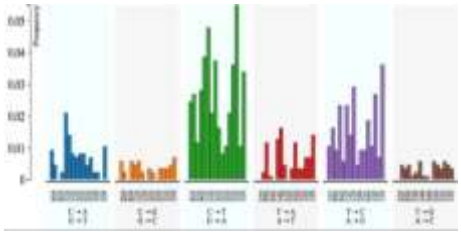
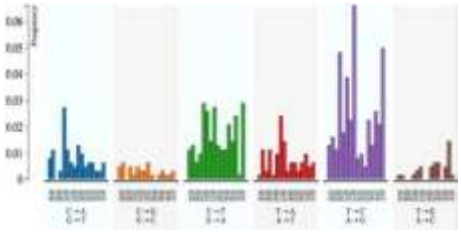
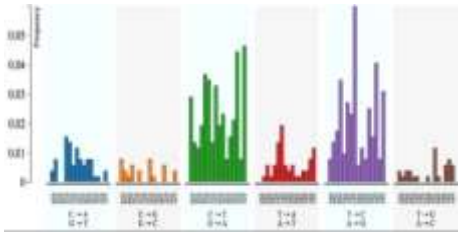
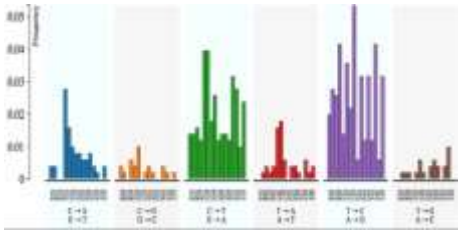
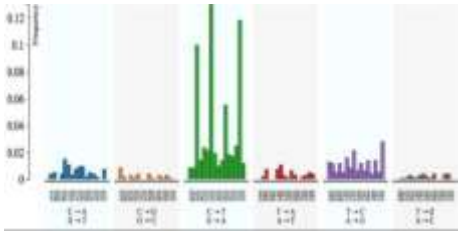
Patient ID	Mutational Profile / Fingerprint	Contribution of Mutational Signatures to Mutational Profile
LA-26		
LA-27		
LA-28		
LA-29		
LA-30		
LA-31		

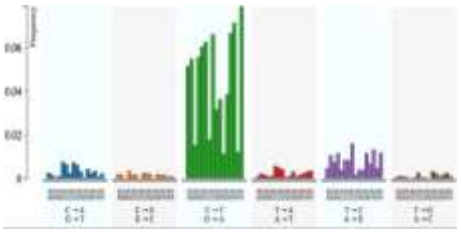
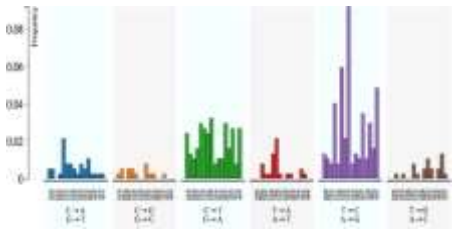
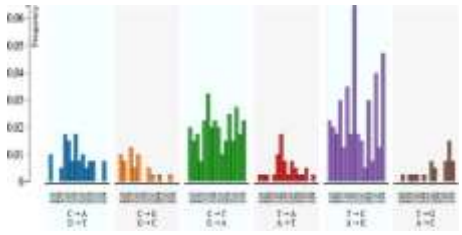
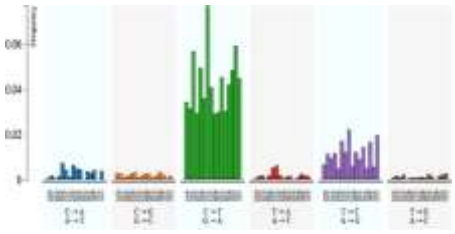
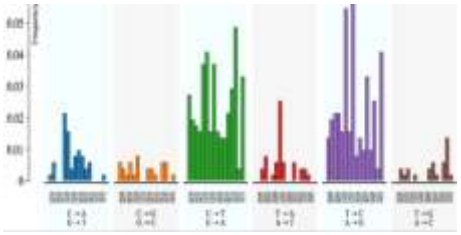
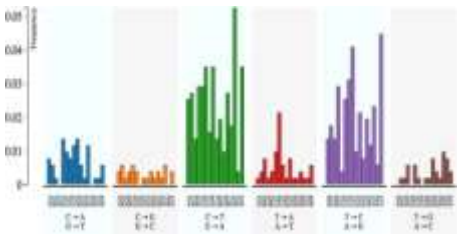
Patient ID	Mutational Profile / Fingerprint	Contribution of Mutational Signatures to Mutational Profile
LB-1		
LB-2		
LB-3		
LB-4		
LB-5		

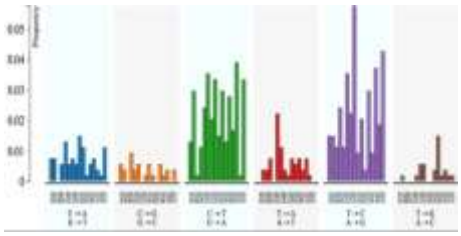
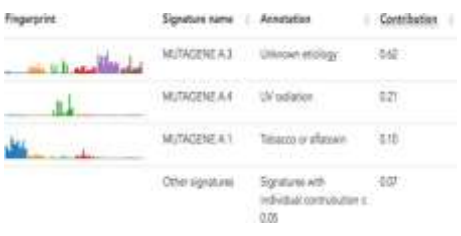
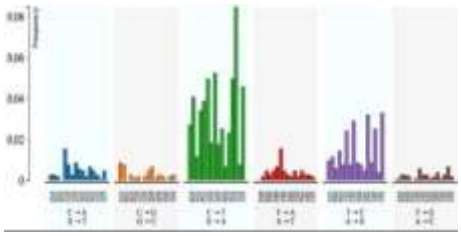
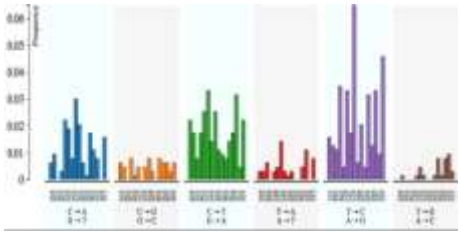
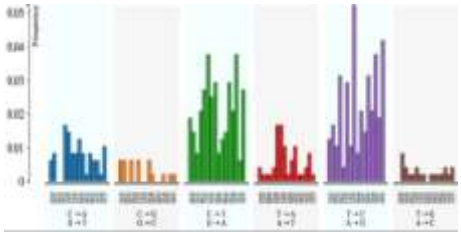
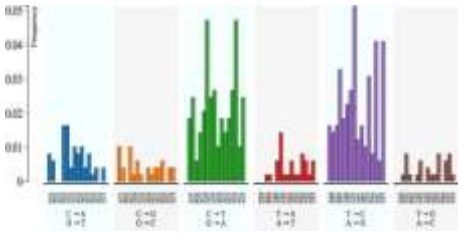
Patient ID	Mutational Profile / Fingerprint	Contribution of Mutational Signatures to Mutational Profile
LB-6		
LB-7		
LB-8		
LB-9		
LB-10		

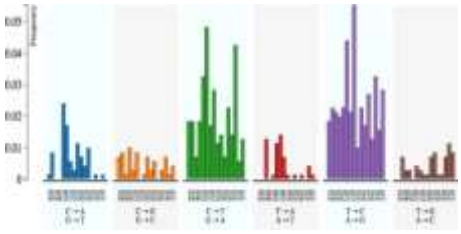
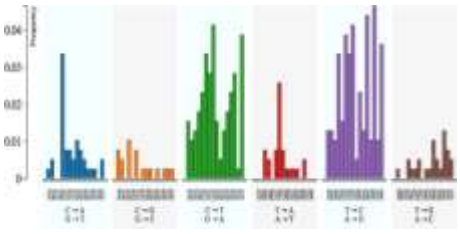
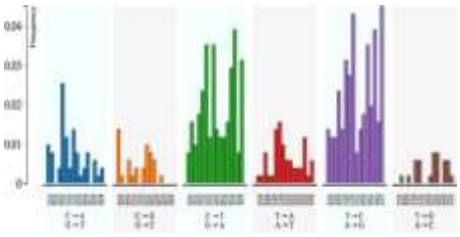
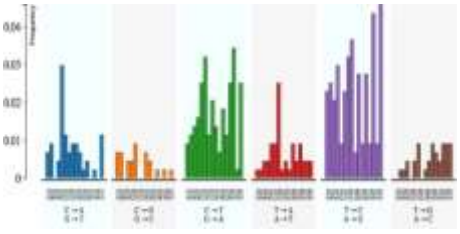
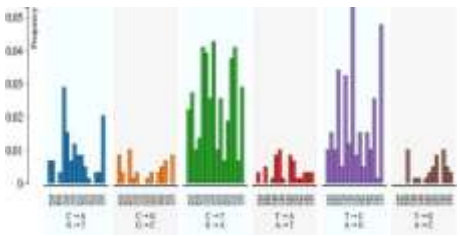
Patient ID	Mutational Profile / Fingerprint	Contribution of Mutational Signatures to Mutational Profile
LB-11		
LB-12		
LB-13		
LB-14		
LB-15		
LB-16		

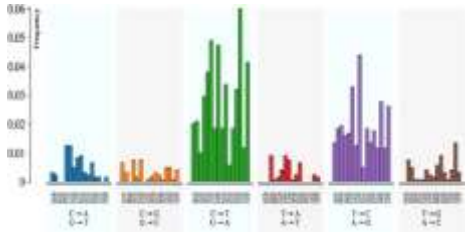
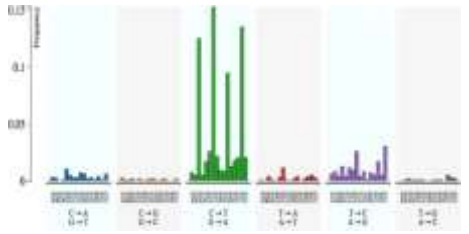
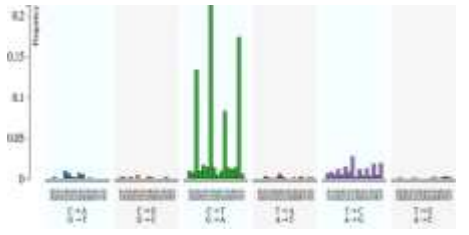
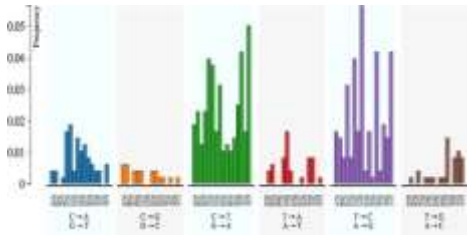
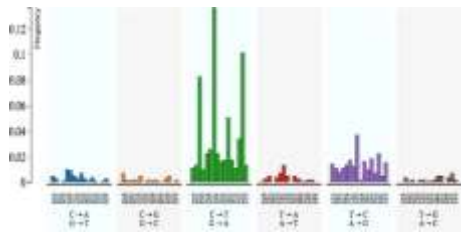
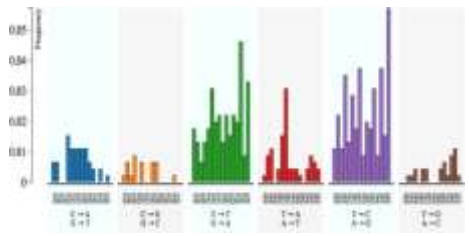
Patient ID	Mutational Profile / Fingerprint	Contribution of Mutational Signatures to Mutational Profile
TN-1		
TN-2		
TN-3		
TN-4		
TN-5		

Patient ID	Mutational Profile / Fingerprint	Contribution of Mutational Signatures to Mutational Profile
TN-6		
TN-7		
TN-8		
TN-9		
TN-10		

Patient ID	Mutational Profile / Fingerprint	Contribution of Mutational Signatures to Mutational Profile
TN-11		
TN-12		
TN-13		
TN-14		
TN-15		
TN-16		

Patient ID	Mutational Profile / Fingerprint	Contribution of Mutational Signatures to Mutational Profile
HER2-1		
HER2-2		
HER2-3		
HER2-4		
HER2-5		

Patient ID	Mutational Profile / Fingerprint	Contribution of Mutational Signatures to Mutational Profile
HER2-6		
HER2-7		
HER2-8		
HER2-9		
HER2-10		

Patient ID	Mutational Profile / Fingerprint	Contribution of Mutational Signatures to Mutational Profile
HER2-11		
HER2-12		
HER2-13		
HER2-14		
HER2-15		
HER2-16		

Appendix C: Abstract in Arabic

التركيبة التسلسلية للتغيرات الجينية لكل من أنماط سرطان الثدي ومجموعة مستقبالت خاليات تاء- بيتا للخاليا اللمفاوية المتسلله للورم

المقدمه : مرض سرطان الثدي هو من اكثر الامراض الخبيثة المنتشره والمسببه للوفاه لدى النساء فـي قطر وفـي جميع أنحاء العالم. تشير العديد من الدراسات الحديثه لسرطان الثدي إلى وجود علاقة قوية بين وجود تحسن حالة المريض و استجابته للعالج. فـي هذا الصدد نان الى و زيادة تنوع مجموعة مستقبالت خاليات خاليا تاء الليمفاوية فـي نسيج الورم وفرصة التسلسلية للتغيرات الان حقيقة علاقة مستوى عبء التغيرات الجينية لسرطان الثدي الليمفاوية محدوده وغير مكتملة. من أجل ذلك فـي هذه الدراسة أول حددنا التركيبة بسرطان الثدي بمختلف أنماطه بستخدام عينات البرافين الثابت مع الفورمالين تاء الجينية وعينها ل 79 مريضا

التي تتضمن من الانسجة ويطلق عليها (FFPE) تم ذلك من خالل استخدام NGS. وقد مجموعه بسرطان الثدي بعد ذلك تم اختيار 11 مريضا منهم يحملون مستوى أعباء مختلفة لفحص انسجة سرطان الثدي قمنا بذلك من خالل تصنيف وتحديد التركيبة من الجينات المتعلقة الانتاج أن 11 مريضا كان لديهم تنوع عالي نسبة تنوع وتواجد خاليا تاء الليمفاوية فـي عبء التغيرات الجينية التسلسلية الجينية لمستقبالت خاليات تاء- بيتا. **النتيجة** أظهرت من خاليات تاء- بيتا فـي نسيج الورم. ومع ذلك، لم تكن هناك علاقة بين ارتفاع بالسرطان الثدي وعدد خاليا اللمفاوية تاء- بيتا وتنوعها فـي عينات المرضى. المستهدفه و