DISCOVERY OF POTENTIAL BIOMARKERS IN ORAL SQUAMOUS CELL CARCINOMA USING NEXT GENERATION SEQUENCING AND PROTEOMIC TECHNOLOGIES

JESINDA PAULINE A/P KERISHNAN

FACULTY OF DENTISTRY
UNIVERSITY OF MALAYA
KUALA LUMPUR

2018

DISCOVERY OF POTENTIAL BIOMARKERS IN ORAL SQUAMOUS CELL CARCINOMA USING NEXT GENERATION SEQUENCING AND PROTEOMIC TECHNOLOGIES

JESINDA PAULINE A/P KERISHNAN

THESIS SUBMITTED IN FULFILMENT OF THE REQUIREMENTS FOR THE DEGREE OF DOCTOR OF PHILOSOPHY

DEPARTMENT OF ORAL AND CRANIOFACIAL SCIENCES
FACULTY OF DENTISTRY
UNIVERSITY OF MALAYA
KUALA LUMPUR

2018

UNIVERSITY OF MALAYA

ORIGINAL LITERARY WORK DECLARATION

Name of Candidate: Jesinda Pauline A/P Kerishnan

Matric No: DHA130008

Name of Degree: Doctor of Philosophy

Title of Thesis: Discovery of Potential Biomarkers in Oral Squamous Cell Carcinoma Using Next Generation Sequencing and Proteomic Technologies

Field of Study: Oral Oncology

I do solemnly and sincerely declare that:

(1) I am the sole author/writer of this Work;
(2) This Work is original;
(3) Any use of any work in which copyright exists was done by way of fair dealing and for permitted purposes and any excerpt or extract from, or reference to or reproduction of any copyright work has been disclosed expressly and sufficiently and the title of the Work and its authorship have been acknowledged in this Work;
(4) I do not have any actual knowledge nor do I ought reasonably to know that the making of this work constitutes an infringement of any copyright work;
(5) I hereby assign all and every rights in the copyright to this Work to the University of Malaya ("UM"), who henceforth shall be owner of the copyright in this Work and that any reproduction or use in any form or by any means whatsoever is prohibited without the written consent of UM having been first had and obtained;
(6) I am fully aware that if in the course of making this Work I have infringed any copyright whether intentionally or otherwise, I may be subject to legal action or any other action as may be determined by UM.

Candidate's Signature Date:

Subscribed and solemnly declared before,

Witness's Signature Date:

Name:

Designation:

DISCOVERY OF POTENTIAL BIOMARKERS IN ORAL SQUAMOUS CELL CARCINOMA USING NEXT GENERATION SEQUENCING AND PROTEOMIC TECHNOLOGIES

ABSTRACT

Oral cancer patients have one of the lowest survival rates in the world due to its poor prognosis. Early detection and diagnosis using more reliable biomarkers will improve the current survival rate of OSCC patients. The present study aims to identify potential molecular biomarkers associated to OSCC using a combination of genomics and proteomics technologies. First, the demographic characteristics of patients were analysed to identify and determine the suitable cohort for this study. Demographic study of 208 OSCC patients and 134 non-OSCC controls confirmed that the local female Indian community that practices betel quid chewing is at the highest risk of developing OSCC and past exposure to HPV16 infection could contribute to the onset of OSCC. To identify potential OSCC biomarkers, 12 pairs of gDNA from OSCC and its adjacent non-malignant tissues were subjected to exome sequencing. The sequencing analysis identified 50 somatically mutated genes in OSCC of which *CASP8, USP40, NOTCH1,* and *COL11A1* were further evaluated as candidate biomarkers. These 4 genes were previously reported in various cancers including the head and neck cancer. However, the exact association of *USP40* and *COL11A1* mutations in OSCC were not fully described. Most of the SNVs identified in these genes were found to be novel in OSCC and were characterized as deleterious. Finally, genotyping of the candidate genes using Fluidigm SNP Genotyping in a larger population of 167 OSCC patients successfully identified *USP40* as a promising biomarker for OSCC. In the proteomic study, sera of 25 OSCC patients and 25 healthy controls were subjected to 2-DE and Western blotting, whereas 6 pairs of extracted proteins from OSCC and its adjacent non-malignant tissues were subjected to label free LC-MS. A total of 27 differently expressed proteins in

OSCC were identified. Among these proteins; LRG, A1BG, PRO2044, ACTBM, HBB, CRNS1, HBA, F8WAH6 and SCND3 were found to be uniquely expressed in OSCC when compared with other cancers. These proteins may have the potential as specific biomarkers for OSCC. *SYNE1* (Nesprin-1) was the only biomarker identified by both genomic and proteomic approach. Lastly, functional enrichment and pathway analysis were also performed on these 77 potential biomarkers using ConsensusPathDB, DAVID v6.8 and STRING v10.1 to elucidate the biological function and pathways associated with OSCC. Based on these analyses, the most significant biological function of these biomarkers in OSCC was its involvement with exosomes in the extracellular region. Whereas, the most significant pathway identified was the platelet activation, signalling and aggregation pathway. Findings from both the biological function and pathway analysis indicate that the identified biomarkers play an important role in cancer metastasis. In summary, the study had successfully identified a combination of 13 novel potential biomarkers and further improved our current understanding on the biological functions and pathways associated with OSCC. However, further studies are required to validate these biomarkers in a larger cohort and to fully understand the role of these biomarkers in OSCC.

Keyword: Oral Squamous Cell Carcinoma, Biomarker, Exome sequencing, LC-MS, Functional analysis

PENEMUAN BIOPENANDA YANG BERPOTENSI BAGI KARSINOMA SEL SKUAMUS MULUT DENGAN MENGGUNAKAN TEKNOLOGI PENJUJUKAN GENERASI BARU DAN PROTEOMIK

ABSTRAK

Pesakit kanser mulut mempunyai salah satu daripada kadar kemandirian hidup paling rendah di dunia disebabkan prognosis yang buruk. Pengesanan dan diagnosis awal menggunakan biopenanda yang lebih dipercayai akan meningkatkan kadar kemandirian hidup pesakit karsinoma sel skuamus mulut (OSCC). Kajian ini bertujuan untuk mengenal pasti biopenanda OSCC yang berpotensi dengan menggunakan gabungan teknologi genomik dan proteomik. Pertamanya, ciri-ciri demografi pesakit dianalisis untuk mengenal pasti dan menentukan kohort yang sesuai untuk kajian ini. Kajian demografi dari sejumlah 208 pesakit OSCC dan 134 subjek kawalan bukan OSCC mengesahkan bahawa komuniti wanita India tempatan yang mengamalkan tabiat mengunyah betel quid berada pada tahap risiko tertinggi dalam kejadian OSCC dan pendedahan lampau kepada jangkitan HPV16 dapat menyumbang kepada permulaan OSCC. Untuk mengenal pasti biopenanda OSCC yang berpotensi, 12 pasang gDNA dari tisu OSCC dan tisu bersebelahan OSCC yang tidak malignan dianalisis dengan menggunakan penjujukan exome. Analisis penjujukan ini mengenal pasti 50 gen yang bermutasi somatik pada OSCC, dimana *CASP8*, *USP40*, *NOTCH1*, dan *COL11A1* telah dipilih sebagai gen calon untuk analysis yang selanjutnya. Gen-gen ini pernah dilaporkan dalam pelbagai jenis kanser termasuk kanser kepala dan leher . Tetapi, perkaitan mutasi *USP40* dan *COL11A1* dengan OSCC tidak difahami sepenuhnya. Kebanyakan varian nukleotida tunggal (SNV) yang dikenal pasti dalam gen-gen ini didapati novel dalam OSCC dan dicirikan sebagai mudarat. Akhir sekali, pengenotipan gen calon dengan menggunakan *Fluidigm SNP Genotyping* dalam populasi lebih besar yang sebanyak 167 pesakit OSCC, telah berjaya mengenal pasti USP40 sebagai

biopenanda OSCC. yang berpotensi. Dalam kajian proteomik, sera daripada 25 pesakit OSCC dan 25 subjek kawalan sihat telah dianalisis dengan menggunakan 2-DE and pemblotan Western, manakala 6 pasang protein yang diekstrak dari tisu OSCC dan tisu bersebelahan OSCC yang tidak malignan dianalisis dengan menggunakan kaedah LC-MS yang berlabel bebas. Sejumlah 27 protein yang mempunyai pengekspresan berbeza dalam OSCC telah dikenal pasti. Antara protein-protein ini; LRG, A1BG, PRO2044, ACTBM, HBB, CRNS1, HBA, F8WAH6 dan SCND3 didapati mempunyai pengekspresan yang unik dalam OSCC jika dibandingkan dengan kanser lain. Protein-protein ini mungkin mempunyai potensi sebagai biopenanda khusus untuk OSCC. *SYNE1* (Nesprin-1) adalah satu-satunya biopenanda yang dikenal pasti melalui kedua-dua pendekatan genomik dan proteomik. Akhir sekali, analisis pemerkayaan fungsi dan laluan biologi terhadap 77 biopenanda yang berpotensi dilakukan menggunakan ConsensusPathDB, DAVID v6.8 dan STRING v10.1 untuk menerangkan fungsi biologi dan laluan biologi yang berkaitan dengan OSCC. Berdasarkan analisis ini, fungsi biologi yang paling penting bagi biopenanda OSCC ini adalah penglibatan exosome di kawasan ekstraselular. Manakala laluan biologi yang paling penting ialah laluan pengaktifan platelet, pengisyaratan dan pengagregatan . Kedua-dua penemuan ini menunjukkan peranan penting dalam metastasis kanser. Secara ringkas, kajian kami telah berjaya mengenal pasti gabungan 13 biopenanda novel yang berpotensi serta meningkatkan pemaham semasa kami terhadap fungsi biologi dan laluan biologi yang berkaitan dengan OSCC. Walau bagaimanapun, kajian lanjut diperlukan untuk mengesahkan biopenanda-biopenanda ini dalam kohort yang lebih besar dan memahami sepenuhnya peranan biopenanda-biopenanda ini dalam OSCC.

Kata Kunci: Karsinoma Sel Skuamus Mulut, Biopenanda Biopenanda, Penjujukan Exome, LC-MS, Analisis Fungsi Biologi

ACKNOWLEDGEMENTS

First and foremost, I would like to thank God for giving me the strength, knowledge, ability and blessings to undertake and complete this research.

I have great pleasure in acknowledging my gratitude to my supervisor, Associate Professor Dr. Chen Yeng for providing her heartfelt support, opportunity and guidance. Thank you for your patience and understanding throughout the research and my time in University of Malaya.

I would also like to thank my co-supervisor, Professor Dr. Tang Thean Hock for the opportunity given to me and his guidance during this research. My sincere appreciation also goes to my teammates, collaborator and my peers for their support and assistance.

I would also like to take this opportunity to thank Bright Spark Program (BSP), University of Malaya (UM) for awarding me a scholarship and providing the financial aid throughout my postgraduate program. Not forgetting also University of Malaya, Department of Oral and Craniofacial Science, Oral Cancer Research and Coordinating Center (OCRCC), Dental Research Management Center (DRMC), and the Dean and staff members of Faculty of Dentistry, UM for this opportunity and their countless assistance.

Lastly, I would like to give my special appreciation to my wonderful parents and siblings for their constant love, prayers, support, patience and assistance throughout my time in UM and throughout this research.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF SYMBOLS AND ABBREVIATIONS

ACN	:	Acetonitrile
ACTB	:	Actin, cytoplasmic 1/ Beta-Actin
ACTBM	:	Putative beta-actin-like protein 3
ACTC	:	Actin, alpha cardiac muscle 1
ACTG	:	Actin, cytoplasmic 2
ACTS	:	Actin, alpha skeletal muscle
ALBU	:	Albumin
ANOVA	:	Analysis of variance
ASCII	:	American Standard Code for Information Interchange
ARID2	:	AT-rich interaction domain 2
A1BG	:	alpha-1B-glycoprotein
BAM	:	Binary Alignment/Map
BWA	:	Burrows-Wheeler Aligner
CASP8	:	Caspase 8
CDKN2A	:	Cyclin-dependent kinase inhibitor 2A
CDKI	:	Cyclin dependent kinase inhibitor
CDS	:	Coding DNA sequence
CGH	:	Comparative genomic hybridization
CLU	:	Clusterin
CNS	:	Conserved noncoding sequence
CNV	:	Copy Number Variations
COSMIC	:	Catalogue of Somatic Mutation in Cancer
COL11A1	:	Collagen type XI alpha 1 chain
CRNS1	:	Carnosine synthase 1

C3	:	Complement component 3
DAVID	:	Database for Annotation, Visualization and Integrated Discovery
DNA	:	Deoxyribonucleic acid
DNAH8	:	Dynein axonemal heavy chain 8
dNTP	:	Deoxynucleotide
EASE	:	Expression Analysis Systematic Explorer
EBV	:	Epstein Barr Virus
EF1DL	:	Putative elongation factor 1-delta-like protein
EGFR	:	Epidermal Growth Factor
ELISA	:	Enzyme Immunoassay
E7ENN3	:	Nesprin-1
E7ERU0	:	Dystonin
FAT1	:	Atypical cadherin
FDA	:	Food and Drug Administration
FMN2	:	Formin 2
FOCAD	:	Focadhesin
F8WAH6	:	Elastin
GATK	:	Genome Analysis Toolkit
GC	:	Guanine-Cytosine
gDNA	:	Genomic DNA
GO	:	Gene Ontology
GUI	:	Graphical User Interface
GWAS	:	Genome wide association study
HAP	:	Haptoglobin
HBA	:	Hemoglobin subunit alpha
HBB	:	Hemoglobin subunit beta

HNSCC	:	Head and Neck Squamous Cell Carcinomas
HPLC	:	High-performance liquid chromatography
HPV	:	Human Papillomavirus
HRP	:	Horseradish peroxidase
IFC	:	Integrated Fluidic Circuit
IgG	:	Immunoglobulin G
IgM	:	Immunoglobulin M
IgA	:	Immunoglobulin A
Indels	:	Insertion/Deletion
iTRAQ	:	Isobaric Tags for Relative and Absolute Quantitation
KEGG	:	Kyoto Encyclopedia of Genes and Genomes
LC-MS	:	Liquid Chromatography - Mass Spectrometry
LM-PCR	:	Ligation-Mediated Polymerase Chain Reaction
LOH	:	Loss-of-heterozygosity
LRG	:	Leucine-rich α2-glycoprotein
LRP1B	:	LDL receptor related protein 1B
MgCl2	:	Magnesium Chloride
MOCDTBS	:	Malaysian Oral Cancer Database and Tissue Bank System
MS	:	Mass Spectrometry
NCBI	:	National Center for Biotechnology Information
NCR	:	National Cancer Registry
NGS	:	Next generation sequencing
NOTCH1	:	Notch 1
nPCR	:	Nested PCR
NTC	:	No-Template Control
OCRCC	:	Oral Cancer Research and Coordination Centre

OD	:	Optical Density
OrCGDB	:	Oral Cancer Gene Database
OSCC	:	Oral Squamous Cell Carcinoma
PCR	:	Polymerase Chain Reaction
PIK3CA	:	Phosphatidylinositol-4,5-bisphosphate 3-kinase catalytic subunit alpha
POTEE	:	POTE ankyrin domain family member E
POTEF	:	POTE ankyrin domain family member F
pRb	:	Retinoblastoma protein
proapo-A1	:	Proapolipoprotein A1
PTEN	:	Phosphatase and tensin homolog
QC	:	Quality Control
RAS	:	Renin-Angiotensin System
RBP4	:	Retinol-binding protein 4 precursor
RNA	:	Ribonucleic Acid
SAM	:	Sequence Alignment/Map
SCND3	:	SCAN domain-containing protein 3
SD	:	Standard Deviations
SDS–PAGE	:	Sodium dodecyl sulfate–polyacrylamide gel electrophoresis
SNP	:	Single Nucleotide Polymorphism
SNV	:	Single Nucleotide Variants
SPON1	:	Spondin 1
SPSS	:	Statistical Package for the Social Sciences
STA	:	Specific Target Amplifications
STRING	:	Search Tool for the Retrieval of Interacting Genes
SVEP1	:	Sushi, von Willebrand factor type A, EGF and pentraxin domain

		containing 1
SYNE1	:	Spectrin repeat containing nuclear envelope protein 1
TBE	:	Tris-borate-EDTA
TBS	:	Tris-buffered saline
TERC	:	Telomerase RNA component
TFA	:	Trifluoroacetic
TIL11	:	Tubulin polyglutamylase TTLL11
TITIN	:	Titin
TiTv	:	transition/transversion
TMB	:	tetramethylbenzidine
TP53	:	Tumour protein p53
TP63	:	Tumour protein p63
UCH	:	Ubiquitin carboxyl-terminal hydrolase
UniProtKB	:	The Universal Protein Resource Knowledgebase
UPS	:	Ubiquitin-proteasome system
USP40	:	Ubiquitin specific peptidase 40
VCA	:	Viral Capsid Antigen
VCF	:	Variant Calling Format
WGS	:	Whole genome sequencing
WHO	:	World Health Organization
2-DE	:	Two-Dimensional Electrophoresis

LIST OF APPENDICES

CHAPTER 1: INTRODUCTION

1.1 Introduction

Oral cancer is a cancer of Head and Neck, and remains as a growing problem in many parts of the world. According to Warnakulasuriya (2009), both oral and oropharyngeal cancer represents the sixth most common cancer in the world with an annual estimated incidence of 275,000 and 130,300 respectively. For the past few years, oral cancer has remained with high mortality rate of over 50% (Bagan *et al.*, 2010). Oral cancer incidence rates have been reported with a wide geographical spread. Particularly in south-central Asia, oral cancer was ranked as one of the most common types of cancer (Petersen *et al.*, 2005). According to the National Cancer Registry (NCR) of Malaysia, oral cancer was ranked the fourth most common cancer among Indian females in the year 2006 (Omar, Z. A. *et al.*, 2006).

Oral squamous cell carcinoma (OSCC) accounts for 90% of oral cancers and develops predominantly in the tongue and the floor of the mouth (Bagan *et al.*, 2010). Other less common sites of OSCC are the buccal mucosa, gingiva, soft palate and retromolar trigone (Ogbureke & Bingham, 2012). OSCC is a multifactorial disease caused by long term exposure to carcinogenic chemical, physical or viral stimuli. The most common aetiologies widely associated with OSCC are tobacco, alcohol and betel quid chewing with these factors acting separately or synergistically (Warnakulasuriya, 2010). Asian population has been reported to have higher risk of OSCC, attributed to the life-style of chewing tobacco, areca nut or betel quid (Warnakulasuriya, 2010). It has also been reported that, two-third of OSCC cases occur in the developing countries (Johnson *et al.*, 2011; Scully, Crispian & Bedi, 2000). Furthermore, OSCC occurs mainly in the elderly with the age range between 50 to 70 years. However, in the recent

years occurrence in younger aged patients below 40 years has been widely reported (Rao *et al.*, 2013).

OSCC has high mortality rate as it remains unnoticed by the patient at the early stage. Majority of the cases are usually detected at its advanced stage, despite the easy accessibility of the oral cavity for examination. This has been the significant cause of low improvement on the survival rate over the years (Jemal *et al.*, 2009). OSCC can frequently grow without exhibiting any pain or symptoms that are recognizable. It is characterized by its invasive, frequent perineural growth and lymph nodes metastasis. In many cases, patients tend to develop second primary cancer (Da Silva *et al.*, 2011). Low public awareness plays a significant factor in contributing to the high mortality rate as well. Lack of awareness on the signs and symptoms as well as the risk factors of OSCC is said to be responsible for the delay in diagnosis. Early detection is believed to be the most effective way to improve the devastating effect of the disease. To date, patients with signs and symptoms are usually screened using the conventional method of oral examination. This is followed by diagnosis using microscopic observation on the cells and tissues (Masthan *et al.*, 2012). While these methods are common and practical, they are subjective and are not sensitive to detect OSCC at the early stage. Therefore, there is an urgent need to develop biomarkers to identify high risk individuals, and improve early stage detection and response to therapy (Ralhan, 2007). However, the pathophysiologies of oncogenesis have to be fully understood in order to develop a reliable biomarker.

Oncogenesis is a complex, multi-step process of quantitative or qualitative alteration. It occurs in the genetic events within signal transduction pathways that influence normal cellular physiology (Vogelstein & Kinzler, 1993). Damages to the deoxyribonucleic acid (DNA) located at various points of the short (p) or long (q) arms of chromosome leads to the development of cancer. This happens when the DNA is

exposed to various mutagens through the environment or lifestyle. OSCC develops as a result of multiple molecular events in various genes and chromosomes. The impact of these damages includes cell dysregulations, cell growth cycle or mechanisms to repair cell damage and removal of dysfunctional cells. These impairments or enhancements of the gene function leads to oncogenesis (Scully, C *et al.*, 2000).

Throughout the years, researchers have identified more than 63 karyotypes in oral cancer. Among them, Y deletions and recurrent loss of chromosome 9,13 and 18 are the most commonly reported alteration (Wong *et al.*, 1996). It has also been reported that deleted regions in chromosome 9p21-22 are detected in 2/3 of all head and neck cancers (Ah-See *et al.*, 1994; Nawroz *et al.*, 1994). Researchers have identified several key genes that play a major role in oral cancer including tumour protein p53 (TP53) which is a tumour suppressor gene and epidermal growth factor receptor (EGFR) which is an oncogene (Tsantoulis *et al.*, 2007). Genes encode for 100,000 - 10 million proteins that play important roles in the cellular processes. Therefore, in (pre)malignant cells it can be altered in various ways (Baak *et al.*, 2003). Moreover, the potential involvement of a large number of genes along with cellular and molecular heterogeneity in oral cancer suggest that multiple proteins could also be targeted as an effective strategy to identify novel markers for early diagnosis (Hanash *et al.*, 2002; Mali, 2014).

To fully understand the impact of these alterations, genomic and proteomic analyses can be applied. Genomics is a study of the human genome and proteomics are the analysis of the proteins complement to the genome. Both approaches play a major role in the understanding, diagnosis, prognosis and treatment of cancer (Baak *et al.*, 2003). Studies have shown that genomics along with proteomics allow the identification of a reliable biomarker (Nagaraj, 2009). Biomarker is defined by World Health Organization (WHO) as substance measured in the body or its product that would influence or predict the incidence of a disease (Strimbu & Tavel, 2010). It functions as a

marker for normal biological activity, pathogenic activity or response to a therapy. In cancer, these markers are either produced by the cancer cell itself or by the body in response to cancer. Over the years, the US Food and Drug Administration (FDA) has approved around 20 protein biomarkers that have been used in variety of cancer (Li, D. & Chan, 2014).

The current genomic and proteomic advancements promises the identification and development of a new pool of biomarkers. Because of the unique association between genomic changes with cancer development, biomarkers have been the highlight in cancer research throughout the years (Hartwell *et al.*, 2006). With the availability of the advanced high throughput genomic and proteomic technologies, researchers have a better understanding on the molecular pathogenesis of cancer. This further allows the identification of reliable candidate biomarkers (Ralhan, 2007; Wu, J.-Y. *et al.*, 2010).

Throughout the years, many emerging technologies have been utilized in cancer research. These technologies have become the major driving force of cancer research. It further opens new avenues to identify causal factors and to understand the mechanism or cancer progression. The newly emerged technologies include the next generation sequencing (NGS) platform which is an alternative strategy to Sanger sequencing. The major advances of NGS are the ability to produce a large plethora of data with low cost (Metzker, 2010). It is also capable of producing millions of sequence reads in parallel. In general, NGS facilitates the discovery of mutations that determine phenotypes which is fundamental in genetic research. The broadest application of NGS would be the human genome re-sequencing project. This project improved the current understanding on the effect of genetic variation towards the health and disease (Metzker, 2010). Over the years, NGS has been considered as the most technically feasible method to catalogue mutations in multiple cancer types world-wide (Meldrum *et al.*, 2011). Furthermore, NGS has been reported to open opportunities to identify a large number of

potential biomarkers in various diseases (Manne *et al.*, 2005). One of the most common NGS platforms to date is exome sequencing. This platform produces sequencing data from the functionally relevant genome.

Although NGS stands as the ideal platform for genome analysis, other approaches such as proteomics are required for better understanding of the multifactorial diseases such as cancers. Proteomics represents the study on the complete protein complement which allows system-level biomarker discovery (Altelaar *et al.*, 2013). Protein is the biological endpoint of a living organism. During the transformation of a healthy cell into a neoplastic cell, apparent changes take place that may affect the cellular function. This includes differential protein modification, altered expression and aberrant localization (Srinivas *et al.*, 2002). Determining these changes is the key focus of proteomics. Proteomic biomarker discovery involves the combination of two-dimensional electrophoresis (2-DE) for protein separation coupled with mass spectrometric analysis for protein sequencing. Since these approaches do not identify low abundance proteins, technology such as liquid chromatography-mass spectrometry (LC-MS) is included as a complement to 2-DE (Tainsky, 2009; Wulfkuhle *et al.*, 2003). Throughout the years, proteomics has grown at an astonishing rate and has been extensively applied in various researches (Mesri, 2014; Wilkins *et al.*, 2006). Therefore, in addition to the genomic approach, proteomic approach was applied in this study.

Overall, oral cancer has become one of the lowest survival rate cancers worldwide with poor prognosis despite recent advancement in therapeutic intervention (Gómez *et al.*, 2010). To curve this problem, minimizing diagnostic delay to achieve early detection is a necessity and a key foundation to improve the current survival rate of OSCC. The best method for early cancer diagnosis, prognosis or therapeutic response prediction is the use of serum or tissue biomarkers (Kulasingam & Diamandis, 2008). In this study, exome sequencing was applied to study the genomic changes in OSCC and

to further identify the potential biomarkers. To further enhance the data obtained from NGS, various proteomic technologies were applied. It is hypothesized that potential genomic and proteomic based OSCC biomarkers can be identified using next generation sequencing and proteomic technologies.

1.2 Research Objectives

The aim of this study is to identify potential biomarkers and pathways that may assist in the early detection of oral squamous cell carcinoma (OSCC) using next generation technology. The present study was initiated to test the hypothesis that with the combination of genomic and proteomic approaches, molecular changes in OSCC that may serve as biomarkers for early detection can be identified. To achieve these goals, the following objectives were constructed;

a. To determine the characteristic and population distribution of OSCC patient cohort.
b. To identify and elucidate genetic variations and the somatic mutation underlying OSCC using exome sequencing.
c. To identify potential genomic biomarkers by evaluating and validating candidate mutated genes that contributes to OSCC.
d. To identify potential proteomic biomarker using 2-DE, immunoblotting and label free LC-MS.
e. To underline possible biological functions and pathways that may attribute to oral cancer based on the identified biomarkers.

CHAPTER 2: LITERATURE REVIEW

2.1 Overview of Oral Cancer

2.1.1 Cancer of the Head and Neck

Cancer is an abnormal growth of cells characterized by multiple changes in gene expression. It has been a major hindrance to the well-being of humans for years (Ruddon, 2007). In 2012, the world cancer burden rose to an estimation of 14 million new cases per year and was expected to rise further (Torre *et al.*, 2015). Cancer has a broad range of causal risk factors, including physical agents, infectious agents, genetic mutations and life style-related exposure such as alcohol and tobacco consumption (Schottenfeld, 2006). There are over 100 distinct types of cancers and subtypes of tumours reported worldwide, such as breast, colon, cervix and prostate cancer (Hanahan & Weinberg, 2000). Among them is oral cancer, a type of cancer under the head and neck cancer group and one of the most prominent cancer in Malaysia (Zain, R. & Ghazali, 2001).

Head and neck cancer includes squamous cell carcinoma in the mucosal surface of the larynx (including nasopharynx and oropharynx), pharynx, salivary gland, nasal cavity and the oral cavity (Vokes *et al.*, 1993; Walden & Aygun, 2013). Oral cavity cancer (OCC) and oropharyngeal cancer (OPC) are among the most common cancers of the head and neck (Chaturvedi *et al.*, 2013). To date, there are several types of cancer; however 90% of them are squamous cell carcinoma which is a well-known malignancy (Bagan *et al.*, 2010; Ogbureke *et al.*, 2012). Oral squamous cell carcinoma (OSCC) develops in the oral cavity. It is differentiated by their histological type which includes the mucosal lining of the lip, the floor of the mouth, the buccal mucosa and the hard palate; however, it is commonly occurred in the tongue region (Bagan *et al.*, 2010; Lambert *et al.*, 2011). As for oropharyngeal squamous cell carcinoma (OPSCC), it usually occurs at the base of the tongue, tonsil and oropharynx (Hussein *et al.*, 2017).

Anatomically, both OSCC and OPSCC occurs at different locations that borders each other without any overlaying (Chi *et al.*, 2015). However, it has been a challenge to define both oral cancer and oropharyngeal cancer since oral cavity boarders is not accurately defined (Lambert *et al.*, 2011).

2.1.2 Epidemiology

Oral cancer is one of the most common types of cancer with increasing new reported cases especially in the developing countries (Eskiizmir *et al.*, 2017; Petersen *et al.*, 2005). In recent WHO release, oral cancer was ranked as the 11th most common cancer worldwide with the range of 1 to 10 cases per 100, 000 people in most countries (Benzian, 2013). Thus, oral cancer is classified as one of the most prevalent cancers worldwide with higher incidence and mortality rate in men (Petti & Scully, 2010).

The total global oral cancer incidence rate stands at 170,903 cases with mortality rate of 83,254 cases; in which India represent 41% of the global oral cancer burden with 68% of mortality (Bhatnagar *et al.*, 2012). Oral cancer was reported to be more prevalent in countries across South-East Asia due to the betel quid chewing practices (Cheong, S. *et al.*, 2009). In Malaysia, oral cancer is one of the top ten cancers with high prevalence among Indian and indigenous ethnics (Zain, R. *et al.*, 2001).

2.1.3 Aetiology

Oral cancer has a multi-factorial aetiology which could be grouped as non-modifiable, modifiable or risky life-style (Warnakulasuriya, 2010). It has been reported that alcohol consumption and tobacco smoking are few of the major risk factors of oral cancer worldwide (Ogbureke *et al.*, 2012). However, in certain parts of Asia, the use of smokeless tobacco or betel quid chewing has been identified as the major risk factors for oral cancer (Neville, B. W. & Day, T. A., 2002). In Malaysia, betel quid chewing habits is seen common among the Indian ethnic group from the plantation sector, elderly

Malays living in the rural areas and among the indigenous people of Sabah and Sarawak (Zain, R. *et al.*, 2001).

Although the high incidence of oral cancer is generally associated with tobacco reported that oral cancer may also develop without any exposure to these risk factors (Scully, C & Bagan, 2009). This suggests that other factors such as diet, ionising radiation, genetic predisposition and viruses may play a role in oral cancer (Neville, B.W & Day, T.A, 2002; Scully, C *et al.*, 2009). Human Papillomavirus (HPV) and Epstein Barr Virus (EBV) are the most significant cause of viral associated oral cancer (Jalouli *et al.*, 2010; Meurman, 2010).

HPV can be classified as low-risk or high risk sub-types. There are around 200 HPV sub-types recorded and 20 of these sub-types are cancer risk factors. High risk HPV sub-types such as HPV16 and HPV18 were reported to be involved in epithelial carcinogenesis (Arbyn *et al.*, 2014; Ono *et al.*, 2014). According to the International Agency for Research on Cancer (IARC), HPV16 has been classified as carcinogenic to humans, and is responsible for the development of various cancers including uterine cervix cancer (Bouvard *et al.*, 2009; Termine *et al.*, 2012). HPV16 has also been suggested as an etiological factor for the head and neck squamous cell carcinomas (HNSCC), tonsillar cancers, OPSCC and OSCC (Chai *et al.*, 2016; Termine *et al.*, 2012). Because of its significant association with OSCC, majority of oral cancer studies focus on HPV16 (Hu *et al.*, 2016; Saini *et al.*, 2011).

EBV belongs to the herpes virus family which infects approximately 90% of the world's adult population (Ono *et al.*, 2014). Studies have shown that EBV infected individuals usually carry the virus throughout their whole life without any symptoms and present elevated EBV VCA-IgG (Jenson, 2011). EBV infection is associated with various cancer such as infectious mononucleosis, Burkitt's lymphoma, nasopharyngeal

carcinoma, oral hairy leukoplakia and OSCC (Higa *et al.*, 2003; Thompson & Kurzrock, 2004). In OSCC, EBV infection was reported to be correspondent to the increased risk of OSCC (She *et al.*, 2017). It has been reported that the role of EBV in OSCC is associated with the geographical regions and habits (Polz-Gruszka *et al.*, 2014). For example, EBV prevalence has been reported to increase in OSCC with the influence of betel quid chewing (Acharya *et al.*, 2015). Although there are numerous studies reported the association of EBV with OSCC development, there were no safe conclusion were drawn from it (Sand & Jalouli, 2014). Further studies on a larger population representing various countries and habits are required to support this association (She *et al.*, 2017).

2.1.4 Diagnosis and Prognosis of Oral Cancer

Early diagnosis of oral cancer is significantly low compared to other cancers such as breast cancer (Mashberg, 2000). Usually, different microscopic methods that involve tissue biopsy and staining are used to visualize malignant lesions in the oral cavity. Often times, biopsies from oral cancer patients may include soft tissues that surround the cancer tissue which may not exhibit malignant tissue (Connolly *et al.*, 2003). Nevertheless, a successful diagnosis through tissue biopsy is highly dependent on collecting whole and complete tissue samples from patients. Therefore, the identification of suitable and reliable OSCC biomarkers is essential to develop a reliable early detection tool (Masthan *et al.*, 2012).

2.2 Molecular Basis of Oral Cancer

OSCC is a multi-step process with the involvement of multiple genetic alterations influenced by factors such as exposure to environmental carcinogens and genetic predisposition of an individual (Choi, S. & Myers, 2008). If the damaged or altered cells are not identified by the DNA repairing mechanism, progressive development of tumour occurs (Malarkey *et al.*, 2013). However, the exact mechanism

involved in oral carcinogenesis remains unknown (Karsani *et al.*, 2014). The formation of a cancer cell from normal cell requires multiple genetic and epigenetic alterations. These alterations include inheritance mutation, somatic mutation and methylation of cell's DNA (Riley & Desai, 2009). Genetic mutations in cancer cells usually consist of two major families of genes; oncogenes and tumour suppressor genes (Macdonald *et al.*, 2004; Schulz, 2005).

2.2.1 Oncogene

Oncogenes or proto-oncogenes represent genes that have potential to initiate cancer when mutated or expressed in high levels (Macdonald *et al.*, 2004; Schulz, 2005). The proteins encoded by these oncogenes are usually involved in cell proliferation, cell differentiation and cell death (Futreal *et al.*, 2004). Oncogenes are usually activated through several mechanisms that involve gene amplifications and/or mutation, such as chromosomal translocation, gene amplifications or subtle intragenic mutations (Murugan, A. K *et al.*, 2012; Vogelstein & Kinzler, 2004). Activation of these oncogenes prompts cell growth and proliferation cycle, resulting in uncontrolled growth of tumour cells (Macdonald *et al.*, 2004; Schulz, 2005). Mutations may also occur in genes that are involved in the DNA-repair process, resulting in high somatic mutation rate, which in turn may increase the probability of mutation in a growth-control gene (Futreal *et al.*, 2004). To date, there are several oncogenes reported to be involved in the development of OSCC. These oncogenes are Renin-Angiotensin System (*RAS*) (Tsantoulis *et al.*, 2007), epidermal growth factor receptor (*EGFR*) (Mendes, 2013) and phosphatidylinositol-4,5-bisphosphate 3-kinase catalytic subunit alpha (*PIK3CA*) (Kozaki *et al.*, 2006).

2.2.1.1 *RAS* Gene

RAS, an OSCC oncogene is known to be genetically deregulated in more than 20% of the disease (Murugan, A. K *et al.*, 2012). The *ras* gene family associated with

OSCC includes *H-*, *K-*, and *N-ras* oncogene (Tsantoulis *et al.*, 2007). The frequency of *RAS* mutation in OSCC was found to be low in countries such as Europe and USA. However, it was noted to be high in Asian countries such as India (Sakai *et al.*, 1992; Tsantoulis *et al.*, 2007). The activation of *RAS* is in response to numerous extracellular stimuli such as growth factors. It transduces its signal across the cell membranes and controls the cellular pathways involved in growth, migration, adhesion, cytoskeletal integrity, differentiation and survival of a cell (Murugan, A. K *et al.*, 2012).

2.2.1.2 *EGFR* Gene

One of the most extensively studied oncogenes in OSCC is *EGFR* (Mendes, 2013; Nagpal & Das, 2003). *EGFR* works through the tyrosine kinase cascade; therefore it plays a crucial role in the control of cellular proliferation, apoptosis, invasion, angiogenesis, and metastasis (Mendes, 2013). *EGFR* in carcinogenesis is known to be activated by 3 mechanisms; a) N-terminal ligand-binding domain deletions/mutations, b) EGFR gene over-expression concurrent with the continuous presence of EGF and/or TGF-a and c) deletion in the C-terminus of the receptor (Kalyankrishna & Grandis, 2006; Wong *et al.*, 1996).

2.2.1.3 *PI3KCA* Gene

PI3KCA gene from the phosphatidylinositol 3-kinase (PI3K) signalling pathway has been previously reported as an oncogene of OSCC (Kozaki *et al.*, 2006; Murugan, A.K *et al.*, 2013). *PI3K* regulates signalling pathways that are important in cell proliferation, adhesion, survival, and motility (Samuels *et al.*, 2004). Previous studies have shown that *PIK3CA* plays an important role in the initiation of squamous cell carcinoma and metastasis (Kozaki *et al.*, 2006; Murugan, A.K *et al.*, 2013). *PI3KCA* oncogene is located on chromosome 3g26, where other oncogenes such as tumour protein p63 (*TP63*) and *tel*omerase RNA component (*TERC*) are also located (Tu *et al.*, 2011).

2.2.2 Tumour Suppressor Genes

Tumour suppressor genes represent genes that prevent the progression of normal cells into cancer (Schulz, 2005). These genes are responsible for regulating division of the cells and maintaining the genome stability and integrity (Macleod, 2000; Sherr, 2004). Loss or inactivation of the genetic elements in these genes contributes to the uncontrolled growth of tumour cells (Schulz, 2005; Weinberg, 1991). One of the most commonly mutated tumour suppressor genes in cancer is the *p53* gene. Abnormalities in this gene may have diagnostic, prognostic, and therapeutic impact (Greenblatt *et al.*, 1994; Muller & Vousden, 2014; Whibley *et al.*, 2009). There are several tumour suppressor genes that are involved in the development of oral cancer, including *p53* (Scully, C *et al.*, 2009), phosphatase and tensin homolog (PTEN) (Mendes, 2013) and cyclin-dependent kinase inhibitor 2A (CDKN2A) (Mascolo *et al.*, 2012).

2.2.2.1 *p53* Gene

The *p53 (or TP53)* gene located on chromosome 17p13.1, is the most important tumour suppressor gene that maintains genome stability and is important in cell cycle progression, cellular differentiation, DNA repair and apoptosis (Nagpal *et al.*, 2003; Perez-Sayans *et al.*, 2009). This tumour suppressor gene is a transcription factor that inhibits cell growth and stimulates the death of a cell during cellular stress (Vogelstein *et al.*, 2004). The deregulation of tumour suppressor network in *p53* is seen in many cancer types including OSCC and most often through the loss-of-heterozygosity (LOH) or point mutation in the 17p13 region (Scully, C *et al.*, 2009; Tsantoulis *et al.*, 2007).

2.2.2.2 *CDKN2A* Gene

Alteration in *CDKN2A* gene located on chromosome 9p21 has been widely investigated and reported in oral cancer (Mascolo *et al.*, 2012; Tsantoulis *et al.*, 2007). This tumour suppressor gene encodes cell cycle regulatory protein $p16^{INK4A}$ that functions as a negative regulator by inhibiting the cyclin-dependent kinase 4 and 6

activity which further stimulate cell-cycle arrest in the G1 phase (Mascolo *et al.*, 2012; Sailasree *et al.*, 2008). The cyclin dependent kinase inhibitor (CDKIs) is highly targeted in OSCC because of its ability to prevent retinoblastoma protein (pRb) phosphorylation (Tsantoulis *et al.*, 2007). It has been reported that inactivation of *p16* occurs in the early stage of OSCC that leads to the loss of *p16* expression (Buajeeb *et al.*, 2009).

2.2.2.3 *PTEN* Gene

PTEN a tumour suppressor gene and a product of PI3K has been previously reported to be mutated or epigenetically inactivated in various cancer types including breast, brain, melanoma and oral cancer (Mendes, 2013; Rahmani *et al.*, 2012). *PTEN* is located on chromosome 10q23.3 and is thought to be involved in cellular processes such as survival, differentiation, proliferation, apoptosis and invasion (Mascolo *et al.*, 2012). *PTEN* expression was previously reported to be down-regulated in many cancers types including oral cancer where it occurs in 5-10% of OSCC lesion (Mascolo *et al.*, 2012; Mendes, 2013).

2.2.3 Genetic Polymorphism in Oral Cancer

A detailed study to identify targeted specific gene changes specifically in oral cancer is essential for better understanding of the molecular events that underlies the progression and development of oral disease (Kuo *et al.*, 2003). Based on previous studies on the genetic alteration during the development of squamous cell carcinoma, researchers found that the most common genetic alteration is the loss of chromosomal region 9p2, suggesting that this is an early oral carcinogenesis (Jurel *et al.*, 2014; Perez-Sayans *et al.*, 2009). It was also reported that the loss of chromosome 3p region is also a common early genetic alteration in oral cancer (Garnis *et al.*, 2003; Hogg *et al.*, 2002). Apart from 9p and 3p, loss of chromosomal material in 17p has also known to be high in dysplastic lesions, indicating that these events are early markers of oral carcinogenesis. However, loss of 13q and 8p regions is more frequent in carcinoma

compared to dysplasia and is associated with the later stages of cancer development (Nagpal *et al.*, 2003; Perez-Sayans *et al.*, 2009).

On the other hand, molecular profiles of oral cancer vary throughout the world and are influenced by both etiological factors and ethnicity (Shah & Singh, 2006). For example in India, chewing betel, paan and Areca which is known to be high risk factors of oral cancer, is a common practice. Therefore, oral cancer accounts a higher percentage of all cancers in India, compared to countries such as United Kingdom which has low use of betel, paan and Areca. Most genome-wide studies on oral cancer have been carried out on various intra-oral sites that are associated with different etiological agents (Ambatipudi *et al.*, 2011).

Understanding the genetic mechanism in the development of oral cancer is crucial in the clinical perspective of OSCC. This knowledge may assist in diagnosing, treating and preventing oral cancer and to further prevent the complication of the serious adverse effects resulting from the cancer and its treatment (Agrawal, N. *et al.*, 2011).

2.3 Cancer Biomarker Discovery

2.3.1 Principle of Biomarker

Biomarkers are biological molecules such as DNA and proteins that are measured and evaluated to detect the presence of a disease or to determine the progression and response of a treatment (Mäbert *et al.*, 2014). It is defined by WHO as a substance, structure or process measured in the body, that would influence or predict the incidence of a disease (Strimbu *et al.*, 2010). It is often independently measured or evaluated as a marker to indicate the biological or pathogenic process, or the pharmacological responses (Manne *et al.*, 2005).

In cancer, biomarkers are usually molecules that are produced by the cancer cell itself or the host in response to the disease. It is a biological material used to determine the risk or to detect, diagnose and classify cancer (Tainsky, 2009). To date, over 20 cancer biomarkers has been approved by the FDA (Li, D. *et al.*, 2014). The use of biomarkers in cancer diagnosis is highly applicable due to the unique association of genomic alterations with disease process (Hartwell *et al.*, 2006). Biomarkers can measured through genomics, proteomics or cellular/molecular substance detected at a higher level in cancer patient's blood, urine or body tissue compared to a normal patient (Mäbert *et al.*, 2014; Manne *et al.*, 2005). In recent years, the advancement in -omics technologies such as genomics (genetic profiling) and proteomics (protein profiling) has provided techniques to identify reliable biomarkers which further improve the understanding of the pathways and interactions involved in carcinogenesis (Ralhan, 2007). These robust technologies open new means in biomarker discovery to allow early detection of cancers (Ribeiro *et al.*, 2016).

2.3.2 Genomic and Proteomic Application in Biomarker Discovery

To fully understand the complex process of neoplastic development, a comprehensive and systematic approach such as the application of genomics to cancer biology is essential (Martin & Nelson, 2001; Tainsky, 2009). Although genomic profiling offers a remarkable opportunity to understand and identify molecular alterations in cancer, ultimately to fully understand the most functional level of cancer, proteomic profiling is greatly needed (Hanash *et al.*, 2002; Martin *et al.*, 2001). This is because proteins are responsible for the cellular function and various protein modifications are only apparent at proteomic level (Baer & Millar, 2016; Martin *et al.*, 2001). The outcome of combining both applications allows a better understanding of the mechanisms of cancer, through the identification of previously unknown patterns of expression and cellular responses to environmental factors, which further assist in

identifying potential and reliable biomarkers (Hanash *et al.*, 2002; Mcdermott *et al.*, 2013). The use of proteomics as a downstream analysis of genomic application, functions as a unified analysis which allows the translation of genomic data to molecular functions and phenotypes (Ellis *et al.*, 2013). Therefore, the application of genomic and proteomic approaches in cancer research allows the advancement of the current knowledge on the basic mechanism involved in carcinogenesis, further improving the diagnostic and therapeutic strategies (Martin *et al.*, 2001; Tainsky, 2009).

2.3.3 Genomics in Biomarker Discovery

Genomic biomarkers are tools used to aid disease diagnosis, prognosis and therapeutic intervention and further providing insight on disease aetiology (Ginsburg & Haga, 2006). Furthermore, these genomic based biomarkers are able to foster the current approaches in routine disease management (Hartwell *et al.*, 2006). Mutations in genes, gene transcriptions and translation alterations or the final protein product alteration has the potential to be used as a specific biomarkers for disease (Wulfkuhle *et al.*, 2003). For example, mutations in cancer associated genes such as tumour suppressor genes and oncogenes can be used as genomic biomarkers (Ludwig & Weinstein, 2005). To increase the current understanding of cancer biology and to identify genomic biomarkers, technologies such as whole genome sequencing (WGS), targeted sequencing or next generation sequencing (NGS), genotyping and bioinformatics have been fundamentally applied (Tran *et al.*, 2012). These genomic based biomarkers include, single nucleotide variation (SNVs), chromosomal aberrations, copy number alterations, microsatellite instability and differential promoter-region methylation (Ludwig *et al.*, 2005).

2.3.3.1 High Throughput Next Generation Technology Platform

Because of the current limited available biomarkers, new researches have emerged to identify large numbers of potential biomarkers using high throughput next

generation technologies (Manne *et al.*, 2005). These technologies offer new opportunities in developing accurate and cheaper biomarker based tests (Etzioni *et al.*, 2003). Advancement in these genomic technologies allows rapid screening of specimens to identify specific changes in the gene and its expression, exclusively in a cancer cell (Wulfkuhle *et al.*, 2003). This next generation technologies which include next generation sequencing (NGS), "omics' profiling, SNP analysis and microarray-based comparative genomic hybridization (CGH) are able to cross-examine the cancer genome and provide a significant insight into cancer biology (Mäbert *et al.*, 2014).

2.3.3.2 Genotyping

Cancer mutation genotyping was introduced as a single base assay, given that some cancer mutations are known to occur at similar bases in the tumour genome from different patients (i.e. recurrent mutations) (Tran *et al.*, 2012). In the recent genotyping studies, it has been revealed that these genetic mutations are identified in cancers such as oral squamous cell carcinomas or adenocarcinomas, giving opportunities in developing molecularly targeted biomarkers (Li, T. *et al.*, 2013; Pickering *et al.*, 2013). Moreover, clinical application of this single base assay as a molecular biomarker has been proven successful in non-small-cell lung cancer (Li, T. *et al.*, 2013). In the recent years, NGS technologies such as exome sequencing offers rapid and novel approach in identifying single base mutations (Li, T. *et al.*, 2013; Mäbert *et al.*, 2014).

2.3.3.3 Bioinformatics

Bioinformatics represents a statistical and scientific computational approach by analysing large and complex data sets to answer biological questions (Baxevanis & Ouellette, 2004; Tran *et al.*, 2012). Bioinformatics comprises of information management and algorithm development mainly in assembling, annotating and comparing genomes which are important in cancer genomics (Tran *et al.*, 2012). Bioinformatics tools applied with the advanced genomic technologies will have a

significant impact on cancer biomarker discovery studies (Kulasingam *et al.,* 2008). This further introduces opportunities to unravel interaction networks and molecular pathways, further characterizing the fundamental molecular mechanism involved in carcinogenesis (Mäbert *et al.*, 2014).

2.3.4 Proteomics in Biomarker Discovery

Proteomic technologies functions as analogous to technologies such as genomic platform by providing large accumulation of various sequences and variants produced along with quantitative data on biological expression (Ellis *et al.,* 2013). The proteomic analysis allows the identification and quantification of proteins and peptides in a biological sample (Arnott & Emmert-Buck, 2010). The expression level of these proteins provides the most reliable data characterizing the disease and normal biological system (Cox & Mann, 2007) . Proteins or peptides are known to be a reliable biomarker for cancer diagnosis. Targeting proteins as cancer biomarkers has become widely applied since proteomic approach characterizes the proteins in cancer, whether these proteins are modified or unmodified (Srinivas *et al.,* 2002). Proteomics, together with genomic allow a reliable molecular characterization of OSCC as well as in determining diagnostic and novel therapeutic targets (Rezende, T. M. B. *et al.,* 2010).

2.3.4.1 2-Dimensional Electrophoresis and Mass Spectrometry

Proteomics was initially pursued with 2-dimensional gel electrophoresis (2D-E) in the mid-1970s and has been the most widely used tool for separating proteins (Baak *et al.*, 2003; Cox *et al.*, 2007; Srinivas *et al.*, 2002). 2-DE separates proteins in the gel in first dimension based on the isoelectric points and then in a second dimension based on the molecular masses (Marouga *et al.,* 2005; Srinivas *et al.*, 2002). In the first dimension, the use of narrow immobilized pH gradients increases resolving power and detect low-abundance proteins. Finally, radioactive/fluorescent labelling or

silver/coomassie blue staining allows the visualization of the proteins in a single gel (Srinivas *et al.*, 2002). The relative quantities can be determined when the ratio of the spot intensities are quantified, thus the differences between samples can be compared (Hood *et al.*, 2004; Zhu *et al.*, 2009).

However, since 2-DE does not fully deliver on proteomics analysis, biological mass spectrometry (MS) is applied (Cox *et al.*, 2007). 2-DE coupled MS has increased the power of proteomics. Additionally, MS becomes a powerful tool to detect protein interaction and posttranslational modification (Aebersold & Mann, 2003; Cox *et al.*, 2007). Furthermore, mass spectrometry platform such as the Matrix-assisted desorption/ionization time-of-flight mass spectrometry (MALDI-TOF MS) allows the identification and analysis of small amounts of proteins that are isolated from the gel (Baak *et al.*, 2003; Srinivas *et al.*, 2002). MS is applied by measuring the relative abundance of heavy and light peptide formed in each sample. Each peptide is identified by generating and analysing peptide fingerprints (Srinivas *et al.*, 2002). Therefore, the protein abundance in tissues or cells in two different states can be determined (Hood *et al.*, 2004; Srinivas *et al.*, 2002). Finally, peptide spectra obtained from MS can be annotated to protein sequence database such as National Center for Biotechnology Information (NCBI) or The Universal Protein Resource Knowledgebase (UniProtKB) to identify the protein (Baak *et al.*, 2003).

2.3.4.2 Immunoblotting

Although 2-DE identifies proteins in biological samples, numerous post-translational forms of protein regulation, such as regulating enzymes and low abundance proteins remain undetected (Goldszmid *et al.*, 2014). To address this, immunoproteomic approach with pooled human antibodies were employed to detect host-specific response proteins (Mu *et al.*, 2014) . This immunoproteomics approach can be used to identify antigens targeted by the immune system, in patients' sera during the cancer progression.

Furthermore, immune responses are also known to be involved in the mechanism of carcinogenesis (Goldszmid *et al.*, 2014).

Immunobloting or western blotting is a technique used in immunoproteomics to separate and identify proteins. Proteins are separated based on the molecular weight through gel electrophoresis and are transferred onto a membrane to produce spots representing a protein. Subsequently, the membrane is incubated with labelled antibodies specific to the targeted proteins and the bound antibodies are then detected through the development of image on film (Mahmood & Yang, 2012).

2.3.4.3 Label Free LC-MS

Usually, small proteins do not generate sufficient digested products for MS identification through sequencing of peptides (Srinivas *et al.*, 2002). Therefore, to address this, detection of proteins through liquid chromatography coupled tandem liquid-chromatography (LC-MS) such as label free LC-MS was introduced (Milac *et al.*, 2012; Srinivas *et al.*, 2002). Label free LC-MS is a proteomic technique that does not require isotope labelling such as those used in Isobaric Tags for Relative and Absolute Quantitation (iTRAQ) analysis. This technique was introduced to address some of the issues in labelling methods and to achieve faster, simpler and cleaner protein quantification results (Zhu *et al.*, 2009).

Samples for LC-MS are digested using trypsin or other photolytic enzymes prior to the analysis (Milac *et al.*, 2012). Samples are separated using liquid chromatography (LC) and further analyzed using mass spectrometry (MS). The obtained data are usually analyzed through identification of peptide, quantification of peptide and statistical analysis (Zhu *et al.*, 2009). Protein quantifications are usually analyzed based on the measurement of ion intensity changes such as peptide peak area or peak heights in

chromatography and based on the spectral count of the identified proteins after MS analysis (Clough *et al.*, 2009; Zhu *et al.*, 2009).

In the LC-MS system, conventional high-performance liquid chromatography (HPLC) pumps and columns coupled tandem mass spectrometer are integrated. Since the pumps and the mass spectrometer are controlled by the same software, ion detection and coupling of chromatography works efficiently (Srinivas *et al.*, 2002). Therefore, single LC-MS analysis allows the isolation and sequencing of numerous selected peptides from one sample exposing new avenues for novel biomarker discovery and further assisting researchers to understand the mechanisms in carcinogenesis (Huang, S. K. *et al.*, 2009; Srinivas *et al.*, 2002).

2.3.5 Functional Enrichment Analysis

Functional enrichment methods allow the understanding of the intracellular signalling pathways that underlie the development of cancers (Lin, J. *et al.*, 2007). Along with genomic and proteomic study, it can significantly contribute in discovering molecular markers. Functional enrichment integrates the identified molecules to biological annotations, such as gene ontology, functional classifications and metabolic pathways (Curtis *et al.*, 2005). In general, with functional enrichment analysis, the biological processes, component or structures in which the individual genes or proteins are involved can be described. This includes on how or where these genes or proteins interact with each other (Khatri *et al.*, 2012). There are currently various high-throughput enrichment tools available, including Onto-Express, GoMiner, MAPPFinder, The Database for Annotation, Visualization and Integrated Discovery (DAVID), Expression Analysis Systematic Explorer (EASE), FuncAssociate, Search Tool for the Retrieval of Interacting Genes/Proteins (STRING) and GeneMerge (Huang, D. W. *et al.*, 2009).

2.3.6 Cancer Biomarkers in OSCC

Throughout the years, researchers have continually searched for OSCC biomarker panels to assist in early detection and diagnosis. Saliva was identified has the best matrix for biomarker discovery as it is non-invasive and easy to be collected and the idea of using salivary biomarkers for OSCC early detection has attracted much interest over the years (Kawahara *et al.*, 2016). Over 100 potential OSCC salivary biomarkers have been reported in the past; however not all has shown significant differences when compared between OSCC patients and the controls (Cheng *et al.*, 2014). Despite of the numerous studies on salivary biomarkers for OSCC, none of the identified candidate biomarkers have progressed through the development and validation phase to be used clinically (Kawahara *et al.*, 2016). Genomic and proteomic biomarkers have also been identified in the past; however, these biomarkers are only applicable in the cancer monitoring and therapeutic intervention and not for early detection or screening of OSCC (Fernández-Olavarría *et al.*, 2016; Santosh *et al.*, 2016). Although detection of biomarkers for OSCC has progressed greatly in the past years, there are still several aspects of OSCC biomarkers that are not fully understood (Ni *et al.*, 2015). Therefore, a profound understanding on the molecular biology of OSCC is important to identify potential biomarkers that are applicable for OSCC screening (Ni *et al.*, 2015).

2.4 Next Generation Sequencing

2.4.1 Next Generation Sequencing Technology

Cancer association studies comparing frequencies of genetic polymorphisms between cases and controls, offer a powerful approach in identifying variants. This approach allows the use of high-throughput technology in identifying novel regions and pathways associated with carcinogenesis by screening through these regions in the sample across the entire human genome (Chung *et al.*, 2009).

Previously, Sanger Sequencing and genome wide association study (GWAS) was widely used in cancer genomic studies. More than 20 novel disease loci have been identified in breast cancer, prostate cancer, colorectal cancer, lung cancer, and melanoma using GWAS (Easton & Eeles, 2008). However, higher cost and excessive data production by GWAS, have made genomic profiling in cancer research to be less feasible (Biesecker *et al.*, 2011). On the other hand, this technology serves as a paradigm shift from the conventional method such as Sanger Sequencing and increase the scale of sequencing in a massive magnitude (Ku *et al.*, 2012; Majewski *et al.*, 2011). In addition, this targeted sequencing approach has the advantages of increasing the sequence coverage on the region of interest such as exons, consequently reducing the cost and producing high throughput compared to the conventional platform (Majewski *et al.*, 2011).

Therefore, researchers have begun to explore the use of targeted sequencing or next generation sequencing such as exome sequencing that targets only the certain regions of the genome (Bamshad *et al.*, 2011). This technology has become a major driver for various genetic researches using genomic samples (Mertes *et al.*, 2011). In time, these revolutionizing technologies will replace the use of conventional methods in exploring genetic alteration in diseases (Rabbani *et al.*, 2014).

2.4.2 Target Capture System

The major highlight of next generation technology is the strategies involved in the direct selection of genomic regions to sequence. This is called target enrichment or target capture strategy. By selectively sequencing the genomic loci of interest such as the exonic region, both the cost and efforts are significantly reduced compared to whole genome sequencing (Mamanova *et al.*, 2010; Mertes *et al.*, 2011). The need for this strategy arises from the current capacity and capability of NGS platform to sequence complex samples (Summerer, 2009). Apart from this, targeted re-sequencing of multiple

causative genes has become a widely explored area for the diagnostic development (Voelkerding *et al.*, 2010).

Target capture techniques are usually characterized based on a range of technical consideration on the ease of use and its performance. However, the practical importance of target capture relies on the aim of a research and the methodologies that are applied (Mertes *et al.*, 2011). These capture strategies are categorized either as oligonucleotide array or amplification-based capture (Voelkerding *et al.*, 2010). Currently several techniques are available according to the nature of the research such as hybrid capture, selective circulation and polymerase chain reaction (PCR) amplification (Mertes *et al.*, 2011). Some examples of hybrid capture platforms include as Agilent SureSelect Human All Exon, Roche /NimleGen's SeqCap EZ Exome Library and Illumina's TruSeq Exome Enrichment platform (Clark *et al.*, 2011; Mertes *et al.*, 2011). Example of a PCR based captured technique include Fluidigm Access Array platform which uses microfluidic chip containing reaction chamber in nanolitter scale separated by valves (Voelkerding *et al.*, 2010).

2.4.3 Exome Sequencing

With the recent developments in genomic technologies, exome sequencing has become a feasible method to explain the genetic alteration in chronic disorder, such as cancer. With the previous use of exome sequencing in many researches, it is known that the success of exome sequencing in the discovery of novel causal mutations for rare diseases is well established, and in addition it is increasingly being used as a diagnostic tool (Ku *et al.*, 2012).

Exomes/exons that are targeted in exome sequencing, are important sequences of DNA representing the regions in the genes (Figure 2.1) that play an important function in the translation of proteins (Ng *et al.*, 2010). Exomes constitute

approximately 1% of the human genome and are estimated to harbour 85% of disease-causing mutations (Ku *et al.*, 2012; Majewski *et al.*, 2011). Therefore, making these regions as an ultimate parameter that can be used to study the relationship of such variation to health and disease (Majewski *et al.*, 2011).

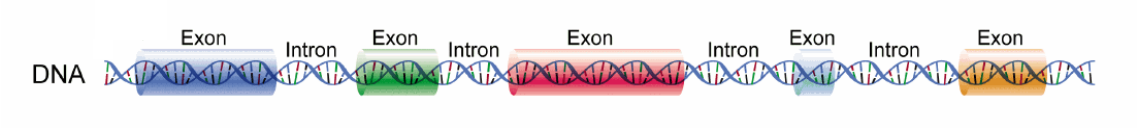


Figure 2.1: Exon region in a DNA.
Exon is a coding region in a DNA interspersed with the non-coding Introns. Exon contains information that are needed to encode proteins (Saberkari *et al.*, 2013).

Exome sequencing has become a feasible method to explain the genetic alteration in chronic disorder such as cancer (Choi, M. *et al.*, 2009). Since this technology targets the coding region of the genome and because most high-penetrance variations are mediated by variations such as non-synonymous, frame shifting and canonical splice variation, exomes sequencing is an ideal platform to study these variations (Biesecker *et al.*, 2011).

2.4.4 Exome Sequencing in Cancer Research

Exome sequencing was first clinically introduced to diagnose a rare form of inflammatory bowel disease in an infant. This technology successfully identified the causal variant of this disease (Warr *et al.*, 2015; Worthey *et al.*, 2011). Following this, exome sequencing has been applied to various cancer researches such as the head and neck cancer (Stransky *et al.*, 2011), oesophagogastric cancer (Chong *et al.*, 2013), prostate cancer (Barbieri *et al.*, 2012), breast cancer (Banerji *et al.*, 2012), renal carcinoma (Varela *et al.*, 2011), and melanoma (Wei, X. *et al.*, 2011).

The application of NGS parallel sequencing technology allows researchers to feasibly catalogue various classes of somatically-acquired mutations in cancer

(Majewski *et al.*, 2011; Shyr & Liu, 2013). Exome sequencing is able to identify complex cancer genomic alterations, such as point mutation, small insertion or deletion, and copy number mutation (Shyr *et al.*, 2013). Furthermore, this technology also introduces new avenues in understanding the molecular pathogenesis that underlies cancers (Majewski *et al.*, 2011). Additionally, this technology is capable of sequencing specific oncogenes and/or tumour suppressor genes with high coverage in samples that has low percentage of tumour cell (Majewski *et al.*, 2011).

Furthermore, cancer association studies using NGS have successfully identified alterations between tumours and matched normal samples (controls). This offers a powerful approach in identifying and distinguishing variants from somatic cells (Chung *et al.*, 2009; Koboldt *et al.*, 2012; Shyr *et al.*, 2013). In addition, this technology allow the identification of novel regions and pathways that are associated with carcinogenesis (Chung *et al.*, 2009)**.**

Previously, the application of next generation sequencing technology in OSCC particularly exome sequencing is limited. However, in the recent years exome sequencing has been frequently used in cancer research offering a wealth of information in the genomic level (Rizzo *et al.*, 2015). In a study by Hayes *et al.* (2016), the authors exhibit the use of exome sequencing on OSCC cell lines to demonstrate molecular changes of *FAT1* and *CASP8*. In a separate study by Al-Hebshi *et al.* (2016), exome sequencing on shammah-associated OSCC identified novel driver events and pathways demonstrating genetic heterogeneity. Likewise, exome sequencing was able to identify both known and novel genes that were significantly and recurrently mutated in OSCC (India Project Team of the International & Consortium, 2013; Su *et al.*, 2017). As a result, exome sequencing has become an effective and popular approach to genetically profile various cancers (Shen *et al.*, 2015).

CHAPTER 3: MATERIALS AND METHODS

3.1 Clinical Samples and Study Design

Serum samples collected in this study comprised of 231 OSCC sera and 159 normal healthy controls. In addition, a total of 167 extracted genomic DNA (gDNA) from OSCC tissues were included in the genomic analysis. Finally, 12 pairs of OSCC tissue with matched non-malignant (normal) adjacent tissues corresponding to the gDNA were obtained for both genomic and proteomic analysis.

All of these samples and data were obtained from the Malaysian Oral Cancer Database and Tissue Bank System (MOCDTBS) (Zain, R. B. *et al.*, 2013), under the Oral Cancer Research and Coordination Centre (OCRCC), University of Malaya (UM). Ethical approval for the study was obtained from Medical Ethics Committee of Faculty of Dentistry, UM. Samples were collected with written consent and were in accordance with the Medical Ethics Committee.

Overall, the study was conducted in 4 phases; which were Demographic Profiling, Genomic Study, Proteomic Study and Functional Annotation Analysis (Figure 3.1).

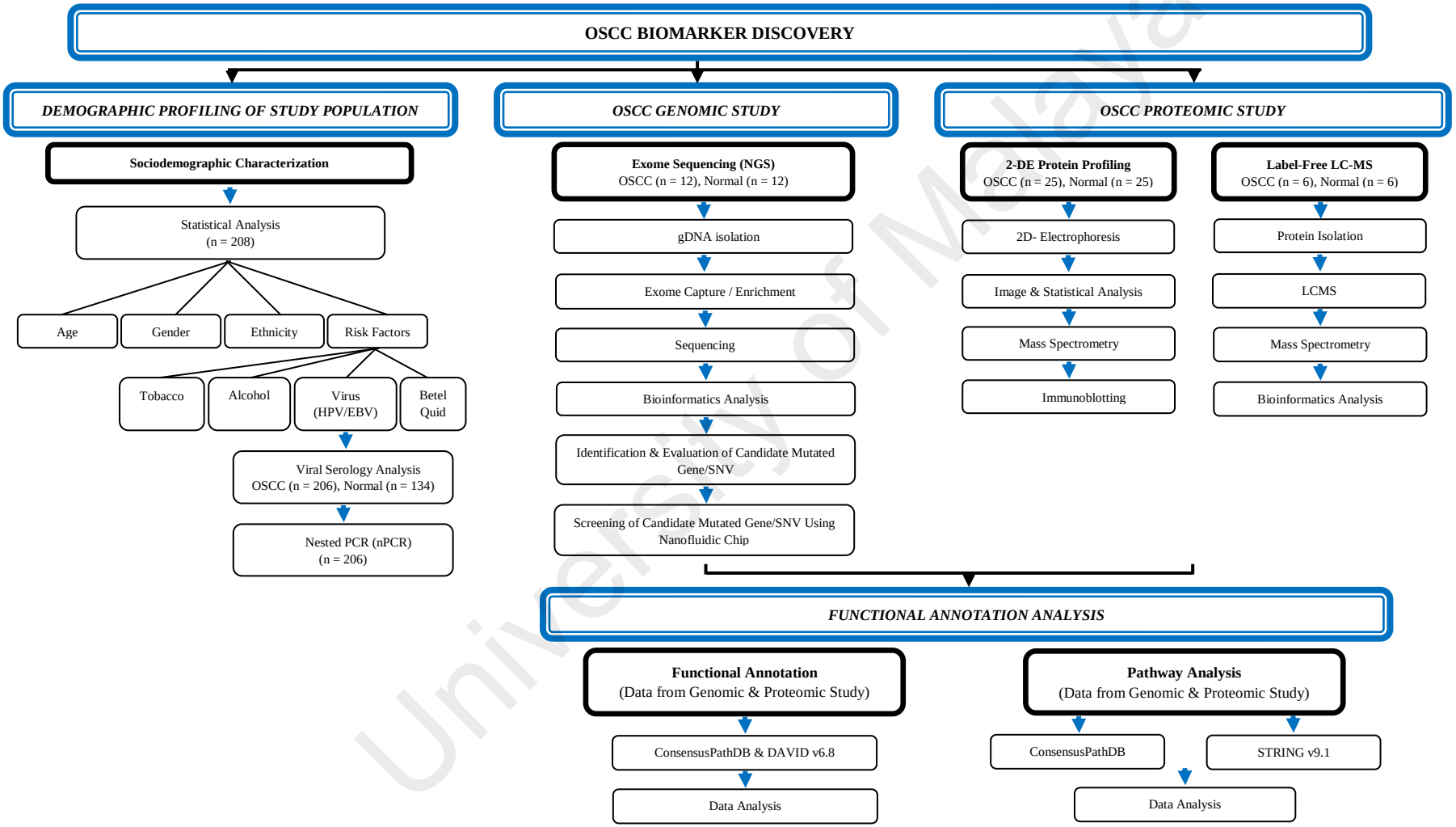


Figure 3.1: OSCC biomarker discovery workflow.

3.2 Demographic Profiling of the Study Cohort

Demographic and clinical data of 206 OSCC patients and 134 normal healthy controls were obtained from MOCDTBS to identify the demographic profile of the study cohort. The demographic and clinical data included age, gender, ethnicity, viral infection, risk habits, cancer stage, treatment and indication of other cancer/disease.

To determine the HPV and EBV infection status of these patients and controls, serum and tissue samples were collected and screened for the presence of these viruses. Therefore, a total of 206 OSCC sera and 134 sera from healthy individuals (control group) were used in the viral serological analysis. In addition, 84 gDNA samples extracted from fresh frozen OSCC tissue corresponding to the samples included in the serological analysis were used for HPV validation through nested PCR (nPCR). Approval for this study was obtained from the Medical Ethics Committee for the Faculty of Dentistry, University of Malaya (reference no: DF DR1307/0077(U)). This demographic profiling was conducted to identify and to set the inclusion/exclusion criteria for the genomic and proteomic study as well as to be further used throughout the study.

3.2.1 Socio-demographic Characterization

To identify and to determine the socio-demographic characteristic of the study cohort, statistical analysis was conducted on 206 OSCC patients and 134 normal healthy controls. Statistical analysis was done on patients' age, gender, ethnicity, viral infection and risk habits. Common risk habits of OSCC patients in Malaysia included in this characterization were tobacco smoking, alcohol drinking and betel quid chewing. The viral infection included in this characterization was HPV and EBV.

Furthermore, to further evaluate whether the predictors of OSCC (i.e., HPV, EBV, gender, race, and age) were significantly associated, logistic regression analysis

was conducted. The logistic regression analysis determines whether a variable (risk factor) could predict the likelihood of OSCC. R^2 in the logistic regression model equals to 0.561 (Cox and Snell) and 0.774 (Nagalkerke), whilst χ^2 model was equal to 254.3. The assumptions related to normality, multicollinearity, nomoscedasticity, independent errors and outliers were checked and met. A *p*-value < 0.05 was considered as statistically significant. All statistical analyses in this study were conducted using the Statistical Package for the Social Sciences (SPSS) software version 12.0.1.

3.2.2 Viral Serology

To determine the study cohort's HPV and EBV status, serum samples were screened for the presence of IgG and IgM antibodies to HPV type 16 and also IgG and IgM antibodies against the viral capsid antigen (VCA) of EBV using Enzyme Immunoassay (ELISA). A similar assay was also performed on a normal population to obtain a comparable data.

3.2.2.1 HPV16 Serology

Both IgG and IgM antibodies against HPV16 antigens were detected in patient sera using HPV16 antibody ELISA kits based on the manufacturer's protocols (Cusabio, Wuhan, China). The microplates from this ELISA kit were pre-coated with HPV-specific antigens. Serum samples were diluted 1000-fold prior experimentation using the provided sample diluent. The diluted samples were then pipetted into the microtiter wells and further incubated for 30 min at 37°C. The wells were then washed, and were added with horseradish peroxidase (HRP)-conjugated anti-human IgG/IgM. It was then incubated again for 30 min at 37°C. After the second wash, 3,3',5,5' tetramethylbenzidine (TMB) substrate solution was pipetted into each well, followed by sulphuric acid solution to terminate the enzyme substrate reaction. The colour changes were spectrophotometrically measured at 450 nm using a microplate reader (Tecan Infinite m200 Pro, Tecan Group Ltd., Mannedorf). Based on the manufacturer's

protocol, the samples' HPV16 antibody (IgG or IgM) valence was detected through optical density (OD). The cut-off level for seropositivity was determined according to the manufacturer's guidelines. The positive and negative controls were provided with the kit.

3.2.2.2 EBV VCA Serology

Viral capsid antigens (VCA) of EBV IgG and IgM antibodies in patient sera were detected using EBV VCA IgG and IgM ELISA kits based on the manufacturer's protocols (Diagnostic Automation Inc, CA, USA). This ELISA micro-well strips was pre-sensitized with EBV antigen by passive absorption. All sera samples were diluted to obtain a uniform concentration using the provided sample diluents. The sample diluents for EBV VCA IgM ELISA contains antihuman IgG. This antihuman IgG precipitated and removed IgG and rheumatoid factor leaving only IgM in the samples to react with the immobilized antigen. The diluted sera samples were then added to the strips and incubated for 25 min at 25°C. After thorough washing, Peroxidase Conjugates goat anti-human IgG (γ chain specific) / IgM (μ chain specific) was added in and the strips was further incubated for 25 min at 25°C. The wells were subsequently washed and TMB substrate solution was pipetted into each well and incubated for 15 min at 25°C. Peroxidase Substrate Solution was added to terminate the enzyme substrate reaction. These changes were then measured using 450 nm using a microplate reader (Tecan Infinite m200 Pro, Tecan Group Ltd., Mannedorf) spectrophotometrically. EBV antibody's (IgG or IgM) valence in the samples was detected through optical density (OD) according to the manufacturer's protocol and was further correlated to the Calibrator. The cut-off level for seropositivity was determined according to the manufacturer's guidelines. The positive and negative controls were provided with the kit.

3.2.3 HPV Validation Using Nested PCR

To verify the presence of HPV, samples that were serologically screened for HPV were further analysed through nested polymerase chain reaction (PCR) assay. Consequently, a total of 84 genomic DNA samples corresponding to the samples used in the serological assay were obtained from the MOCDTBS. In order to carry out a thorough and accurate PCR assay, genomic DNA sample preparation, PCR reagent preparation/setup, and PCR product analysis were conducted using dedicated instruments in three separate rooms. To prevent false negative results all gDNA sample were tested for amplification of a 268-bp region of human β-globin. This was conducted by analysing 5 μL of each DNA sample in a PCR assay which targets the 268-bp region of the β-globin-specific gene using the primers: PC04 and GH20 (Table 3.1) (Zehbe & Wilander, 1996). To visualize the PCR products, 2% agarose gel electrophoresis was used. All β-globin-positive DNA samples were then subjected to DNA amplification via nPCR to detect HPV.

Table 3.1: Primers used to detect HPV in clinical samples.

Primer Set	Primer Name	5'-3' sequence
β-globin	GH2O	GAA GAG CCA AGG ACA GGT AC
	PCO4	CAA CTT CAT CCA CGT TCA CC
PGMY09/11	PGMY11-A	GCA CAG GGA CAT AAC AAT GG
	PGMY11-B	GCG CAG GGC CAC AAT AAT GG
	PGMY11-C	GCA CAG GGA CAT AAT AAT GG
	PGMY11-D	GCC CAG GGC CAC AAC AAT GG
	PGMY11-E	GCT CAG GGT TTA AAC AAT GG
	PGMY09-F	CGT CCC AAA GGA AAC TGA TC
	PGMY09-G	CGA CCT AAA GGA AAC TGA TC
	PGMY09-H	CGT CCA AAA GGA AAC TGA TC
	PGMY09-I	G CCA AGG GGA AAC TGA TC
	PGMY09-J	CGT CCC AAA GGA TAC TGA TC
	PGMY09-K	CGT CCA AGG GGA TAC TGA TC
	PGMY09-L	CGA CCT AAA GGG AAT TGA TC
	PGMY09-M	CGA CCT AGT GGA AAT TGA TC
	PGMY09-N	CGA CCA AGG GGA TAT TGA TC
	PGMY09-P	G CCC AAC GGA AAC TGA TC
	PGMY09-Q	CGA CCC AAG GGA AAC TGG TC
	PGMY09-R	CGT CCT AAA GGA AAC TGG TC
	HMB01	GCG ACC CAA TGC AAA TTG GT
GP5+/GP6+	GP5+	TTT GTT ACT GTG GTA GAT ACT AC
	GP6+	GAA AAA TAA ACT GTA AAT CAT ATT C

First, to screen for the presence of HPV in the OSCC gDNA samples, primary PCR assay was performed using the L1 consensus PCR primer pools PGMY09/11 (primary PCR) and primer HMB01 (Table 3.1) which targeted the 450-bp region. This procedure was adopted from previously described protocols (Erhart *et al.*, 2016; Gravitt *et al.*, 2000; Winder *et al.*, 2009), which were further modified and optimized. Briefly, 1

μL (30 - 50 ng) of OSCC DNA sample was amplified with an equimolar mixture of the primary PCR primers (i.e., PGMY09 and PGMY11; final concentration of 10 pmol for each). The DNA sample and primer mixture were combined with PCR Buffer containing 2 mM $MgCl_2$, PCR grade deoxyribonucleoside triphosphates / deoxynucleotide (dNTP) mix (10 mM of each nucleotide), 2U of FastStart Taq DNA Polymerase (Roche, Germany), and nuclease free water. Applied Biosystems® Veriti® 96-Well Thermal Cycler (USA) was used to perform the amplification. The PCR cycling conditions applied for the primary PCR primers (PGMY09/11) were as follows: 95°C for 5 min (initial denaturing), followed by 40 cycles of 95°C for 1 min (denaturing), 60°C for 1 min (annealing), and 72°C for 1 min (elongation). This was followed by a final extension period of 10 min at 72°C and storage at 4°C.

The secondary PCR included amplification of the primary PCR product using the general consensus primer $GP5^+/6^+$ (Table 3.1), targeting a 150-bp region based on a previously modified protocol (Haws *et al.*, 2004; Van Den Brule *et al.*, 2002). The buffer, reagents, and instrument used in this PCR amplification were similar to the primary PCR protocol. The PCR cycling conditions for the secondary PCR primers were as follows: 94°C for 120 s (initial denaturing); followed by 40 cycles of 94°C for 45 s (denaturing); 48°C for 4 s, 38°C for 30 s, 42°C for 5 s, 66°C for 5 s (annealing); and 71°C for 90 s (elongation). This was then followed by a final extension period of 10 min at 72°C and storage at 4°C. A positive control (DNA from HPV-positive HeLa cells) was included in all PCR analyses (Rocha-Zavaleta *et al.*, 2004). A non-template control to evaluate contamination and accuracy (negative control) of the analyses was also included.

Products for all PCR reactions were electrophoresed using 2% low melting point agarose gels (Vivantis Inc., USA) for 35 min at 110V, in 1X Tris-borate-EDTA (TBE) buffer. The agarose gels were then stained with ethidium bromide (OmniPur® Ethidium

Bromide, EMD, USA), visualized, and photographed (MultiDoc-It Imaging System, UVP, Upland, CA). HPV nPCR was randomly repeated on several samples and compared with the initial results, as a measure of quality control.

3.3 Genomic Profiling of OSCC Samples (Exome Sequencing)

In this study, 12 pairs of fresh-frozen oral cancer tissue with matched non-malignant (histopathologically confirmed normal) adjacent tissues were selected. The oral cancer tissues were obtained from different intra-oral sites including the buccal, tongue, palate, lower lip and floor of the mouth. The inclusion and exclusion criteria for sample selection were determined based on the data obtained from demographic profiling study. All samples were from patients free of alcohol use or tobacco use and have no history of other types of cancer or chronic diseases. In addition, samples were collected from patients who have not undergone any type of cancer treatment (i.e. chemotherapy, radiation) prior to biopsy in order to obtain a natural state of the tumour. Approval for this study was obtained from the Medical Ethics Committee for the Faculty of Dentistry, University of Malaya (reference no: DF OB1505/0069(P)).

3.3.1 Sample Processing

All oral cancers and matched non-malignant (normal) adjacent tissues were sectioned to achieve a neoplastic cellularity of >60% and were reviewed histologically by a OCRCC assigned pathologist. The genomic DNA was isolated from the sectioned tissue using the DNeasy Blood and Tissue kit (Qiagen, Germany) according to the manufacturer's protocol. The purity and concentration of the extracted gDNA were determined using the NanoDrop 2000 UV-Vis Spectrophotometer (Thermo Fisher Scientific, Wilmington, DE, USA). The gDNA was further analysed through gel-electrophoresis to determine the quality of the DNA. A good quality, purity and concentration were achieved in order to generate suitable libraries for exome sequencing.

The quantity of the samples was ≥ 6ug with the concentrations of ≥30 ng/µl. The purity of the samples (A260/A280) was between 1.8~2.0. The gDNA purity of the samples used in this study had an average ratio of 1.88. The samples were then sent to Beijing Genomic Institute (BGI-Hong Kong Co., Limited), Hong Kong, China, to perform exome sequencing.

3.3.2 Exome Capture / Target Enrichment

Prior to sequencing, the genomic DNA was randomly fragmented using Covaris focused-ultrasonicator system (Covaris, MA, USA), resulting in DNA fragments with a base pair peak at 200 to 300b. This was followed by ligation of adapters to both ends of the resulting fragments. The adapter-ligated templates were then purified by the AgencourtAMPure SPRI beads (Agencourt, Beverly, MA, USA) and fragments with insert size about 200bp were excised. These extracted DNAs were then amplified by ligation-mediated polymerase chain reaction (LM-PCR), purified, and hybridized to the SureSelect Biotinylated RNA Library (BAITS) (Agilent Technologies Inc. USA) for exome capture.

Exome DNA capture was performed to enrich exonic sequencing (Gnirke *et al.*, 2009) using Agilent SureSelect Human All Exome V4+UTRs-71M (6GB sequence) which provides a comprehensive coverage of the coding regions and UTRs. This target capture design has the target size of 71Mb with the flanking region of 20,965 genes and 335, 765 targeted exons. The probes/baits used in this target enrichment system were 120nt cRNAs which were designed to be specific to the exonic sequence. The custom probes designed for this experiment was generated using the online Agilent SureDesign software (https://earray.chem.agilent.com/suredesign) to detect human exon, based on the available database. The design can be reviewed on the probe position and target

coverage by uploading the .bed format file generated by SureDesign to the UCSC Genome Browser (http://genome.ucsc.edu/) (Dorschner, 2014).

Hybridized fragments were bound to the streptavidin-labelled magnetic beads whereas non-hybridized fragments were washed out after 24h incubation. The RNA baits were further digested to obtain the targeted DNA of interest (Ernani & Leproust, 2013). Captured LM-PCR products were finally subjected to Agilent 2100 Bioanalyzer (Agilent Technologies Inc. USA) to estimate the magnitude of enrichment (Figure 3.2).

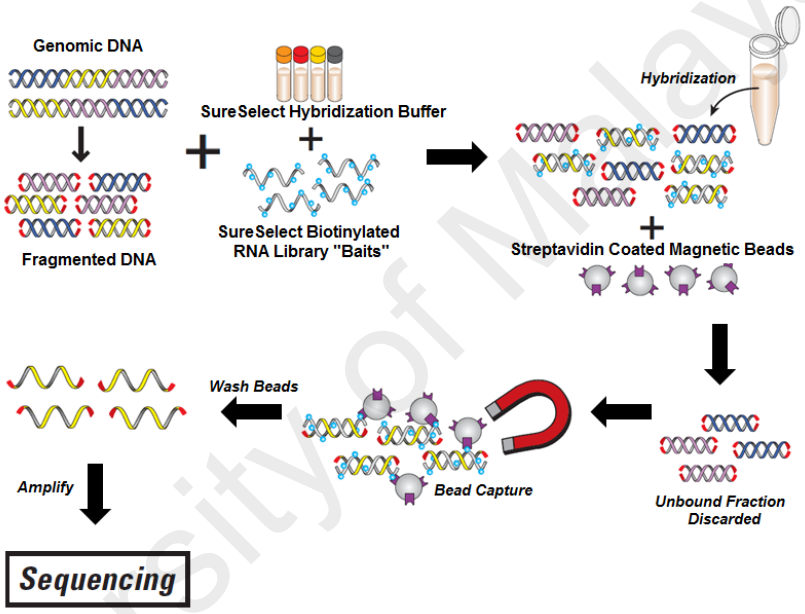


Figure 3.2: SureSelect target enrichment system workflow.
The system uses biotinylated baits targeting exon through hybridization. Magnetic beads are used to capture the hybridized products to be sequenced (Ernani *et al.*, 2013).

3.3.3 Sequencing

The PCR product from target enrichment was then loaded on Illumina HiSeq™ 2000 Sequencing Platform, to perform high-throughput sequencing for each captured products respectively to ensure that each samples meet the desired average fold-coverage. The sequencing generated at least 200 effective mean depths per sample. This 200 effective mean depth which represents the total amount of 14G of clean data (data

mapped to the reference genome, duplication removed and fully mapped to the target region) was equivalent to 200 times of the whole target region length covered by the probes.

Raw image obtained post-sequencing were processed using Illumina Base Calling Software 1.7 (CASAVA1.7.0) for base calling with default parameters and the sequences of each library were generate as 90bp paired-end reads. The software performed alignment of the sequence to the reference genome, variant analysis and read counting subsequently. The raw data obtained from the sequencing were further analysed through bioinformatics.

3.3.4 Bioinformatics Analysis

3.3.4.1 Raw Data

The bioinformatics analysis began with sequencing data (raw data) which were generated from the Illumina pipeline. The raw reads contained sequence of the adapter, high content of unknown based and low quality reads which were removed before data analysis from the raw FASTQ data. Low quality reads were defined if more than half of the bases in a read are low-quality base (base quality ≤ 5). Reads with unknown bases of more than 10% were removed as well. The original image data were transferred into sequence data via base calling, which was defined as raw data or raw reads and saved as FASTQ file. These FASTQ files contained the detailed read sequences and the read quality information. In each FASTQ file, every read was described by four lines (Figure 3.3).

```
Line 1 : @A80GVTABXX:4:1:2587:1979#ACAGTGAT/1
Line 2 : NTTTGATATGTGTGAGGACGTCTGCAGCGTCACCTTTATCGGCCATGGT
Line 3 : +
Line 4 : BTTMKZXUUUddddddddddddddddddddddddddddaddddddd^WYYU
```

Figure 3.3: FASTQ format.
The raw sequencing data is generated in FASTQ format. FASTQ contains read sequence details and quality information.

The first and third lines of the FASTQ file represents sequences names generated by the sequence analyser, while the second line represents the sequence obtained post sequencing and the fourth line represents the sequencing quality value, in which each letter corresponds to the base in line 2. The base quality is equal to the American Standard Code for Information Interchange (ASCII) value of the character in line 4 minus 64 (e.g. the ASCII value of c is 99, then its base quality value is 35). ASCII is a character encoding standard used in electronic communication. Starting from the Illumina GA Pipeline v1.5, the range of base quality values is from 2 to 41. Table 3.2 demonstrates the relationship between Illumina HiSeqTM 2000 sequencing error rate and the sequencing quality value. Specifically, if the sequencing error rate is denoted as E, Illunima HiSeqTM 2000 base quality value is denoted as Q. This relationship was calculated based on the following formula;

$$sQ = -10 \log_{10} E$$

Table 3.2: Relationship between Illumina HiSeqTM 2000 error rate and sequencing quality.

Sequencing error rate (%)	Sequencing quality value	Character
5	13	M
1	20	T
0.1	30	^

Finally, to obtain a clean data, the adapter sequence in the raw data was removed, and low quality reads which had too many Ns and low base quality bases were discarded.

3.3.4.2 Data Alignment and Quality Control

Once a clean data is obtained through filtering and QC, the sequence reads were mapped or aligned to the reference genome. Burrows-Wheeler Aligner (BWA) was used to align the clean data to the reference database. BWA was chosen as alignment tool due to its efficiency to align relatively short nucleotide sequences against a long reference by producing accurate and fast results with low error rates. BWA provided a flexible parameter setup. The output of BWA alignment was presented in Sequence Alignment/Map (SAM) format. The sequence reads were aligned to the human genome build37 (*hg19*) (http://www.ncbi.nlm.nih.gov/assembly/2758/). *Picard* was used to mark duplicated reads. Duplicated reads were information, redundantly produced by PCR. The data obtained through alignment was in Binary Alignment/Map (BAM) format files; a compressed form of SAM file. These BAM format files were used to fix mate information of the alignment, add read group information, to mark duplicate reads caused by PCR and for variant calling. Finally, the alignment results were combined with the BAM format file. Both SAM and BAM file contains sequences of the reads, position of reads, mapping quality and several other statistics on alignment.

3.3.4.3 Variant Analysis

Once the alignment of the sequencing reads to the reference genome was completed, the reads were subjected to variant calling. Variant calling was done to identify Single Nucleotide Polymorphism (SNPs), Single Nucleotide Variations (SNVs), indels, and Copy Number Variations (CNVs) by comparing the aligned reads to the reference genome. In this study, somatic mutations were emphasized as it plays a

distinct role in tumour development. Somatic variants are tissue-specific mutations that are present only in somatic cells. SNPs identification was performed using the SOAPsnp software, Genotype with the highest probability at a given locus was identified for each individual sequencing sample and the consensus sequence of the sample was assembled and saved in the conserved noncoding sequence (CNS) format. Using the consensus sequence, the polymorphic loci between the identified genotype and the reference can be filtered and highlighted; this will constitute the high confident SNV dataset. The dataset is saved as tab-separated file in text format. Once the SNVs were identified, ANNOVAR was used for annotation and classification. ANNOVAR represents a software tool that employs up-to-date information from the publically available database to functionally annotate genetic variants from sequencing.

Single Nucleotide Variants (SNVs) were identified using the Varscan software. Varscan was applied to identify normal (adjacent) sample 1 and tumour sample 2 specific SNVs by simultaneously comparing reads counts, base quality, and allele frequency between the normal tissue (N) and tumour tissue (T). VarScan has the ability to identify tumour specific somatic substitutions by comparing tumour (T) and normal (N) tissue in a pair. It identified specific SNVs and insertion/deletion (Indels) by simultaneously comparing reads counts, base quality, and allele frequency between the normal and tumour tissue and classifies the variations by somatic status. Finally, once the SNVs were identified, ANNOVAR was used for annotation and classification.

Filters were applied throughout the analysis to obtain more confident variant results. These pipelines also include the purity estimation. Finally, ANNOVAR was used to annotate the confident variant results and for classification. The final variants were then fed to the downstream advanced analysis pipeline. Quality Control (QC) was

applied throughout pipeline to obtain clean data, alignment, and variant calling. The standard analysis pipeline is described in Figure 3.4.

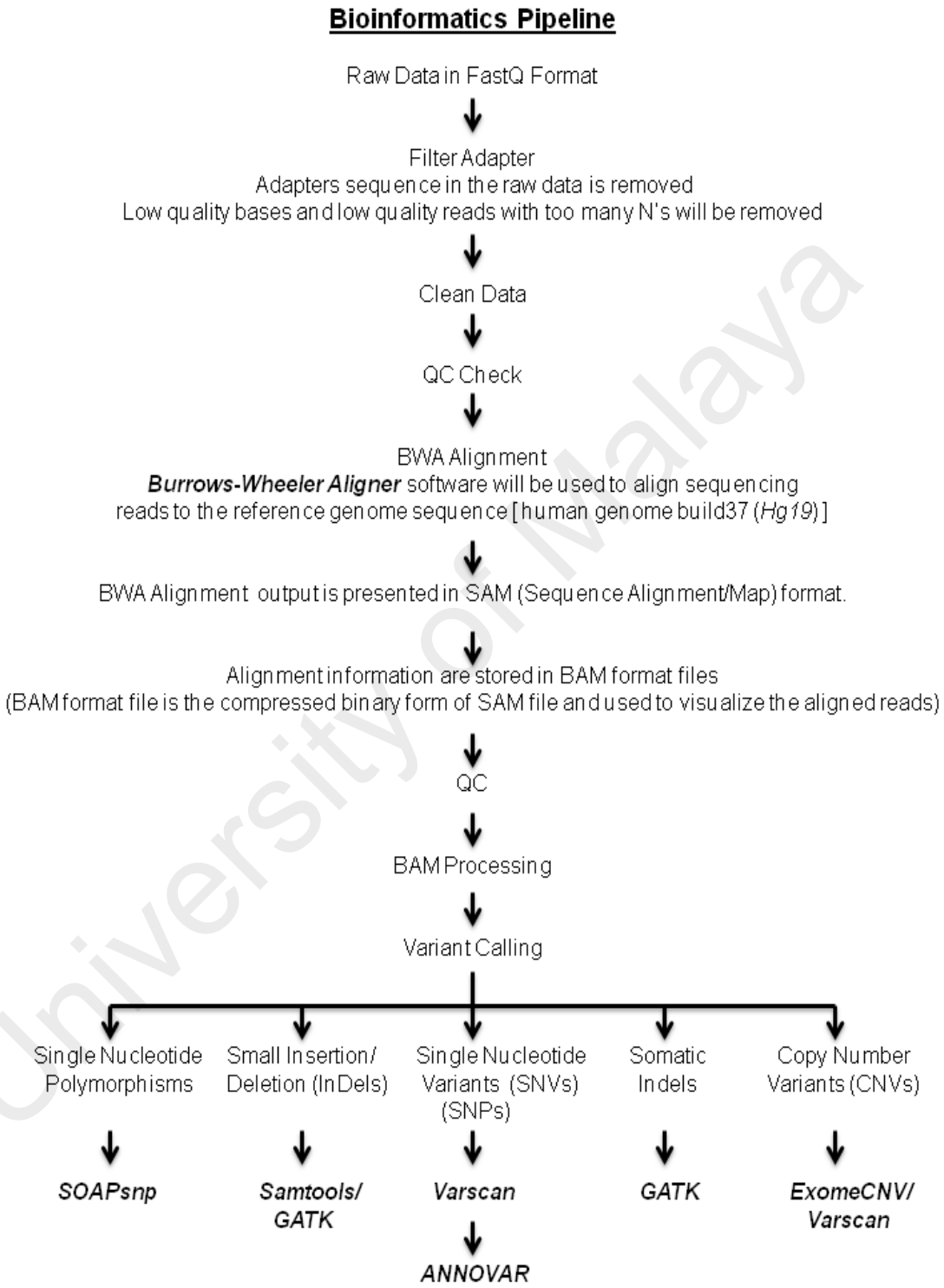


Figure 3.4: Bioinformatics workflow.
The workflow represents a systematic overview of the bioinformatics pipeline applied to analyze the exome sequencing data and to further identify somatic mutations.

3.3.5 Identification and Evaluation of Candidate Mutated Genes and SNVs

The data obtained through exome sequencing should be clinically correlated to OSCC and must be statistically significant to the research. This is because exome sequencing produces a plateau of data which may or may not be relevant to OSCC. In other words, not all mutations detected by exome sequencing are likely to be pathologically relevant. Thus, the non-relevant or germline mutations were excluded to identify those that have a probability to destroy protein function or affect highly conserved amino acids. To remove all of these common variants, any potential somatic mutations that were observed in the public-domain database on human genome were removed. The database used in this analysis were, 1000 genome database (http://phase1browser.1000genomes.org/index.html) and dbSNP135 database (https://www.ncbi.nlm.nih.gov/SNP/).

3.3.5.1 Identifying Candidate Mutated Gene and Data Visualization

To filter and analyse the final clean data from exome sequencing, software R i386 version 3.2.2 and version 3.3.0 was employed. R is a mathematical software with open-source environment for statistical computing and visualization based on the S (statistical programming) language (Dessau & Pipper, 2008; Rossiter, 2012). The analysis was done with the R console using Graphical User Interface (GUI) for Windows (Figure 3.5).

Figure 3.5: Display of R console graphical user interface (GUI) for Windows. The R GUI was used to analyse the exome sequencing data. The left window represents the console window on which the commands (R language) are keyed in to produce the graphics in the right window.

To identify the candidate somatic mutation through mutation burden, GenVisR waterfall function was applied using the software R i386 version 3.2.2. The GenVisR waterfall function; a Bioconductor package was used to calculate and illustrate the mutational burden of variants and further differentiate the variant types (Team, 2013). The command line script that was designed and inputted in the R GUI console to generate the mutation burden was as below;

> library(GenVisR)

> set.seed(426)

> mutation_data=read.table(file.choose(), header=TRUE, sep="\t")

> most_deleterious<-

*c("missense","nonstop","nonsense","frame_shift_del","frame_shift_ins","splice_*

site_del","splice_site_ins","splice_site","in_frame_del","in_frame_ins","splice_region_del","splice_region_ins","splice_region","5_prime_flanking_region","3_prime_flanking_region","3_prime_untranslated_region","5_prime_untranslated_region","rna","intronic","silent")

> waterfall(mutation_data, fileType="Custom",variant_class_order = most_deleterious, mainXlabel=TRUE, maxGenes=50, mainLabelCol="amino_acid_change", mainLabelSize=2)

The spectrum of the mutation were also observed using GenVisR TiTv (transition/transversion) graph. Transition - transversion ratio was commonly compared in new and previously described variants (Challis *et al.*, 2012). Following was the command line script designed and inputted in the R GUI console to generate the mutation spectrum.

> library(GenVisR)

> set.seed(426)

> Trans_data=read.table(file.choose(), header=TRUE, sep="\t")

> TvTi(Trans_data, lab_txtAngle=75, fileType="MGI")

> TvTi(Trans_data, type = "Frequency", palette = c("#77C55D", "#A461B4", "#C1524B","#93B5BB", "#4F433F", "#BFA753"), lab_txtAngle = 75, fileType = "MGI")

To identify SNVs of interest from the data obtained through exome sequencing, qqman Manhattan plot was generated using R i386 version 3.3.0. Manhattan plot refers as a plot of the $-log_{10}$ (p value) of the association statistic on the *y*-axis versus the chromosomal position of the SNV on the *x*-axis (Turner, 2014). Manhattan plot was

generated using the below designed command line script, inputted in the R GUI console.

> library(qqman)

> read.csv(file.choose(),header=TRUE)->Exome.Sequencing

> ls()

> str(Exome.Sequencing)

> head(Exome.Sequencing)

> tail(Exome.Sequencing)

> as.data.frame(table(Exome.Sequencing$CHR))

> manhattan(Exome.Sequencing)

> manhattan(Exome.Sequencing, suggestiveline=F, genomewideline=F, cex.axis = 0.7, cex=1.5)

>manhattan(Exome.Sequencing, suggestiveline=F, genomewideline=F, cex.axis=0.7,cex=1.5,col=c("aquamarine4","blue4","forestgreen","darkorchid4","brown","cyan4","darkmagenta","chartreuse4","gold2","blueviolet","azure4","firebrick4","black","hotpink1","salmon","olivedrab3","powderblue","orangered","seagreen","cyan","yellow","violet"))

>snpsOfInterest<-c("233","234","235","236","729","731","800","801","802","2275","2278","2282")

>manhattan(Exome.Sequencing, suggestiveline=F, genomewideline=F, highlight = snpsOfInterest, cex.axis = 0.7, cex=1.5)

3.3.5.2 Evaluation of Candidate Mutated Gene

To identify oral cancer genes that are less commonly mutated, the somatic mutation information obtained from this study was compared to the Sanger COSMIC (Catalogue of Somatic Mutation in Cancer) database (http://www.sanger.ac.uk/cosmic), Sanger COSMIC is a comprehensive database that curates and organize information on somatic mutations in human cancer (Forbes *et al.*, 2010). The candidate somatic mutations were also compared to OrCGDB (Oral Cancer Gene Database), an easily accessible source (http://www.tumor-gene.org/Oral/oral.html) of information on genes that are involved in oral cancer.

In order to understand the association of this mutation with its associated protein and to contextualize this mutation within a structurally related family of proteins, lollipop plot was generated using MutationMapper at cBioPortal (www.cBioPortal.org). A computational algorithm was used to identify the correct protein or protein domain for the mutations that were inputted to the system. The curetted protein database used in this mapping was UniProt (www.uniprot.org).

To further predict whether the SNVs identified and filtered in this study was neutral or deleterious, consensus classifier software called PredictSNP (http://loschmidt.chemi.muni.cz/predictsnp/) was used. PredictSNP is computational tool consisting of 8 established prediction tools (MAPP, nsSNPAnalyzer, PANTHER, PhD-SNP, PolyPhen-1, PolyPhen-2, SIFT and SNAP) to predict the effect of SNVs or mutations on protein function (Bendl *et al.*, 2014). The amino acid sequence of the 13 candidate somatic mutated genes that were identified in this study was then blast to PredictSNP. The input was submitted in FASTA format and was translated into an

interactive sequence. The position and the mutations of the candidate genes were defined to allow the software to further predict the potential pathogenicity of the mutation. The mutations were determined neutral or deleterious based on continuous interval score (-1,+1) of the output value. The mutations are considered neutral if the values in the interval are -1 or 0 and it is considered deleterious if the values in the interval are 0 or +1.

3.3.6 Screening of Candidate Mutated Genes and SNVs

The identified targeted SNVs were validated with SNPtype Assay for SNP Genotyping on the 192.24 Dynamic Array Integrated Fluidic Circuit (IFC) (Fluidigm Corporation, USA). The system used pressure controlled valves in the designated chip. Therefore, the samples and genotyping assay reagents were loaded into the reaction chambers using pressure.

A total of 167 OSCC gDNA were obtained from the MOCDTBS as the validation cohort. Approval for the study was obtained from the Medical Ethics Committee for the Faculty of Dentistry, University of Malaya (ref no: DF OB1505/0069(P)).

3.3.6.1 Specific Target Amplification (STA)

Sequences of the targeted SNV were submitted to Fluidigm D3 Assay Design System (https://d3.fluidigm.com) for the development of biallelic SNPType assay. Specific target amplifications (STA) were designed to detect 12 SNVs in the CASP8, USP40, NOTCH1 and COL11A1 gene (2, 3, 3 and 4 SNVs respectively). The STA was performed by preparing primer pool which consisted of 100 μM SNPtype Assay STA Primer and 100 μM SNPtype Assay LSP for 12 assays with the final concentration of 500.0 nM each. STA was performed using 2.5 μL Qiagen 2x Multiplex PCR Master Mix (PN 206143, Qiagen, Hilden, Germany), 0.5 μL of 10X SNPtype STA primer pool,

0.75 µL of PCR- certified water and 1.25 µL of genomic DNA per sample. STA was then performed with the PCR cycling condition of 95°C for 15 min (initial denaturing), followed by 14 cycles of 95°C for 15 sec (denaturing) and 60°C for 4 min (annealing). Amplification was performed with an Applied Biosystems® Veriti® 96-Well Thermal Cycler (USA). The final product of STA was then further diluted 1:100 in DNA suspension buffer and stored at -20°C prior genotyping.

3.3.6.2 SNPtype Assay for SNV Genotyping on the 192.24 Dynamic Array IFC

Prior genotyping, each 12 SNPtype assay mixes were prepared with 1.5 µL of SNPtype Assay ASP1/ASAP2 (100 µM each) and 4.0 µL SNPtype Assay LSP (100 µM) per STA product with a final concentration of 7.5 µM and 20 µM respectively. Finally, 14.5 µL DNA Suspension Buffer was added to the SNPtype assay mixes prior analysis.

In a DNA-free hood, aliquots of 10X assays for a total of 12 assays were prepared. 12.5 µL 2X Assay Loading Reagent (Fluidigm PN 100-7611) was combined with 7.5 µL PCR-certified water to create the assay pre-mix. 20.0 µL of the assay pre-mix was then further combined with 5 µL of each individual SNPtype assay mix as prepared earlier for a total of 25 µL of 10X assay mix

Lastly, the sample pre-mix was prepared by combining 270 µL of Biotium Fast Probe Master Mix (PN 31005, Biotium Inc, Hayward, CA, USA), 27 µL of 20X SNPType sample loading reagent (Fluidigm PN 100-7608), 9 µL of SNPType reagent (Fluidigm PN 100-7607), 3.24 µL 50 X ROX (PN 12223-012, Life Technologies, Rockville, MD) and 5.76 µL of PCR-certified water. Finally, 2.6 µL of the sample pre-mix was combined with 1.9 µL of each gDNA (STA products) to make a total of 4.525 µL of sample mix solution.

Prior SNP genotyping, the IFC (Figure 3.6) was injected with control line fluid into accumulator 2 (Acc2) using 150 µL syringes. The blue protective film was removed from the bottom of IFC. Next, 3 µL of each assay and 3 µL of each sample were pipetted into the respective inlets on the IFC. 150 µL of pressure fluid were pipette into P1, P2 and P3 wells. 20 µL of pressure fluid were pipette in to the P4 and P5 wells. The carrier surface of the IFC was blotted with a dry lint-free cloth. And finally, the IFC was loaded in the BioMark HD system (Fluidigm, CA, USA) and the sample SNP genotyping was conducted using the SNPtype 192.21 v1 configuration.

Prior to amplification, the IFC was placed on the NanoFlex™ 4-IFC Controller (Fluidigm, CA, USA) for loading and mixing of the sample and the reagent. PCR was performed using the BioMark unit with the cycling condition of 2 minutes at 50 °C and 10 minutes at 95 °C (initial denaturing), followed by 40 cycles of denaturation at 95 °C for 15 s and 1 minute of annealing and extension at 60 °C. Finally, the BioMark Real-Time PCR System was used to acquire endpoint fluorescent image data and the image was analyzed using the Fluidigm SNP Genotyping Analysis software (Fluidigm, CA, USA).

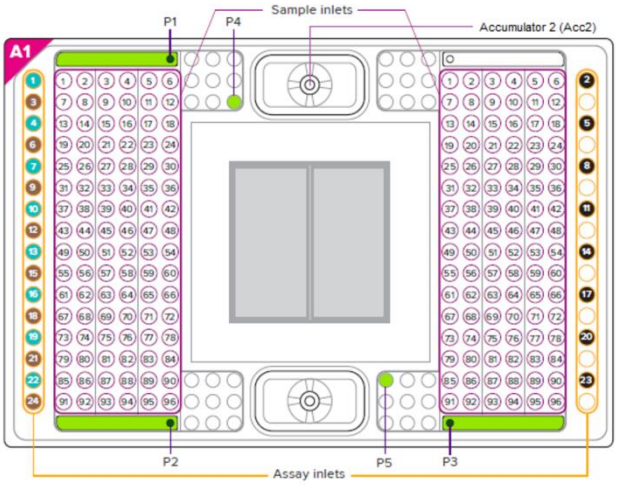


Figure 3.6: 192.24 Dynamic Array Integrated Fluidic Circuit (IFC).
IFC was used to perform OSCC SNP genotyping. Pressure fluids, samples and assay reagents were pipetted into the respective inlets on the IFC, prior loading to the BioMark HD system.

3.4 Proteomic Profiling of OSCC Samples

To further understand the mechanism of cancer development and to improve the genomic biomarker findings, proteomic analysis was incorporated in this study. Consequently, 25 OSCC sera and 25 control sera from healthy individuals were obtained to conduct 2D-electrophoresis and immunoblotting. To perform label-free LCMS, 6 pairs of fresh frozen OSCC tissue with matched non-malignant (normal) adjacent tissues corresponding to the samples used in the NGS analysis was obtained. Approval for this study was obtained from the Medical Ethics Committee for the Faculty of Dentistry, UM (reference no: DF OB1505/0069(P)).

3.4.1 2-Dimensional Electrophoresis Serum Protein Profile

3.4.1.1 2-Dimensional Electrophoresis

Two dimensional electrophoresis (2-DE) was performed as previously described (Chen, Y. *et al.*, 2008). Unfractionated serum samples of patient (10 μl) were lysed, rehydrated in lysis buffer (2M thiourea, 8M urea, 4% CHAPS, 1% dithreitol, and 2% pharmalyte), and further subjected to isoelectric focusing in 13-cm rehydrated precast immobilized dry strips (pH 4–7; GE Healthcare, Sweden) as the first dimension separation. For the second dimension separation, using 8–18% gradient polyacrylamide gels, sodium dodecyl sulfate–polyacrylamide gel electrophoresis (SDS–PAGE) was performed. Gel silver staining was performed as previously described by Heukeshoven and Dernick (1988). Using slightly modified published methods, silver staining and Coomassie Brilliant Blue staining was conducted for MS (Shevchenko *et al.*, 1996).

3.4.1.2 Image and Statistical Analysis

To capture and store the 2-DE gel images, the gels were subjected to the ImageScanner III (GE Healthcare, Sweden). Further on, to evaluate the differentially

expressed protein spots PD-Quest™ 2-D gel analysis software (version 8.0.1, Bio-Rad) was employed. In the serial gels, identical spots were matched and further normalized by correcting for spot quantification values and also gel-to-gel variation which were unrelated to the expression changes. Total densities captured from the gel images (i.e., the raw quantity of each gel spot was divided by the total quantity of all spots within the gel) was used for the normalization method. Protein concentration values were shown as means of percentage volume (% volume) ± standard deviations (SD). To analyse differences between patients and controls, one-way analysis of variance (ANOVA) and the student's t-test were employed. *P*-values less than 0.05 ($p<0.05$) were regarded as statistically significant.

3.4.1.3 Mass Spectrometry Analysis and Database Search

To perform protein identification, spots of interest were excised and subjected to in-gel tryptic digestion. Both MS analysis and database searches were performed at the Proteomic Center, Faculty of Biological Sciences, National University of Singapore. The trypsin digested peptides were then mixed with 1.2 μl of CHCA matrix solution (5 mg/ml of cyano-4-hydroxy-cinnamic acid in 0.1% trifluoroacetic acid [TFA] and 50% acetonitrile [ACN]) and further spotted onto MALDI target plates. Finally, ABI 4800 Proteomics Analyzer MALDI-TOF/TOF Mass Spectrometer was employed for spectra analysis (Applied Biosystems, USA), and the MASCOT search engine (version 2.1; Matrix Science, UK) was utilized for the database searches. In addition, to identify peptides and proteins, the GPS Explorer™ software (version 3.6; Applied Biosystems, USA) was employed along with the MASCOT. Search parameters allowed for C-terminal carbamidomethylation of cysteine (fixed modification), N-terminal acetylation, and methionine oxidation (variable modification). Mass tolerances were set to 100 ppm and ± 0.2 Da, for peptide and fragment respectively. Parameters for the peptide mass fingerprinting (PMF) were set as follows: monoisotropic mass value; ± 0.1 Da peptide

mass tolerance; one missed cleavage allowed in trypsin digest; and 1+ peptide charge state.

Initially peptides were identified with the application of ProteinPilot proteomics software on the Mass Spectrometer (Applied Biosystems, USA). ProteinPilot was further used to calculate and assign score that reflects the relationship of theoretically determined masses and experimentally determined masses. NCBI Unigene human databases (version 3.38) and the International Protein Index (http://www.ebi.ac.uk/IPI) were used to analyse the MS output. A score of > 82 was considered as significant in a total of 100,907 entries that was searched in the MASCOT NCBI database.

3.4.1.4 Immunoblotting

To further improve the biomarker discovery in OSCC, an immunoproteomic approach was applied. The approach utilizes 2-DE immunoblotting assay using OSCC patient and control sera. A total of four categories were assigned as the 2-DE gel immunoblotting protocol. These four categories include; (1) normal sera probed with normal sera; (2) normal sera probed with OSCC sera; (3) OSCC sera probed with normal sera; (4) OSCC sera probed with OSCC sera. To conduct the immunobloting assay, the gels from the 2-DE were transferred onto nitrocellulose membranes using the Multiphor II Novablot semi-dry system (GE Healthcare, Sweden). The membranes were then blocked with SuperBlock (Pierce, USA). After three consecutive wash with Tris-buffered saline (TBS)–Tween-20, the membranes were subsequently incubated with pooled sera from patients or healthy subjects overnight (4°C). The pooled sera contain primary antibodies against various targets (1:200 dilutions). Following the second washing, membranes were further incubated at room temperature for 1 hour with horseradish peroxidase (HRP)-linked monoclonal anti-human IgM (1:5000; Invitrogen, USA). After the final wash, the membranes were visualized using chemiluminescence

substrate (Pierce, USA) and the image was captured on an 18 cm x 24 cm films (Kodak, USA).

3.4.2 Label Free LC-MS Relative Protein Quantitation

Label free LC-MS is a technique that combines both liquid chromatography (LC) and mass spectrometry (MS) without the presence of peptide labels. This technique allows analyte profiling in large numbers by comparing complex mixtures from different biological conditions; i.e normal and cancer (Wiener *et al.*, 2004). The label free method is usually applied for non-targeted analysis in a discovery phase research (Clough *et al.*, 2009). Label-free technology has better advantages with regards to sample preparations and experimental time frame when compared with the labelling method (Patel *et al.*, 2009).

3.4.2.1 Tissue Samples Preparation

Protein from the micro dissected OSCC and normal tissue was extracted using lysis buffer and followed by homogenization. The filtered peptides were then dissolved in 100 µl 0.1% Formic acid. Peptides were desalted using Ziptip® with 0.6 µL C18 resin (Millipore, Bedford, MA). The desalted peptides were dried down using SpeedVac concentrator (Savant, NY, USA). The peptides were re-dissolved in 2% Acetonitrile and 0.1% Formic acid. Finally, the proteins concentrations were determined using Bradford assay.

3.4.2.2 LC-MS

The re-dissolved were then loaded onto a PicoFrit C18 nanospray column (New Objective) using a Thermo Scientific Surveyor Autosampler operated in the no waste injection mode. The peptides were eluted from the column using a linear acetonitrile gradient from 5% to 45% acetonitrile for 230 min followed by high and low organic washes for another 20 min into an LTQ XL™ mass spectrometer (Thermo Scientific,

USA) via a nanospray source. The spray voltage was set to 1.8kV and the ion transfer capillary set was at 180 °C. A data-dependent Top 7 method was used for peptide identification with a full MS scan from m/z 400-1500, followed by MS/MS scans on the three most abundant ions.

3.4.2.3 Bioinformatics Analysis

The raw data obtained from mass spectrometry was analysed using Proteome Discoverer 1.3 (Thermo Scientific, Waltham, MA, USA) and the SEQUEST algorithm against the most recent species-specific database for human in NCBI (National Center for Biotechnology Information, USA) to identify proteins and to determine the number of missed cleavages. The search parameters were set under the trypsin digestion with up to two missed cleavage per peptide. Additionally, carbamidomethyl was used as a static modification while oxidation of Methionine was used as a variable modification. Proteins were then identified when two or more unique peptides had X-correlation scores greater than 1.5, 2.0, and 2.5 for respective charge of +1, +2, and +3 (El-Bayoumy *et al.*, 2012).

Label free quantitation was performed using Sieve 2.1 software (Thermo Scientific, USA) to execute component detection, peak alignment, background subtraction and differential analysis. The quantitation was performed using default parameters except for frames thresholds which were configured to 8000. SEQUEST parameters were used for peptide filtering. The loading amount was not normalized since the peptides were used as starting material. The ratios reported were corrected based on the sample concentration. The comparison was performed between OSCC sample and the adjacent normal. Using ProteoWizard, the raw data obtained from the analysis was converted to the mzXML file format (Kessner *et al.*, 2008) and further visualized using Insilicos Viewer Version 1.5.4 and further analysed through XCMS

Online (https://xcmsonline.scripps.edu/index.php) (Gowda *et al.*, 2014) and Mzmine 2.0.

3.5 Functional Annotation and Pathway Analysis of the Identified Biomarkers

Functional annotation analyses were performed on the candidate biomarkers identified through both genomic and proteomic approach. Approval for this study was obtained from the Medical Ethics Committee for the Faculty of Dentistry, UM (reference no: DF OC1703/0024(U)).

Functional annotation and protein interactions were performed using web-based bioinformatics tools. ConsensusPathDB (http://cpdb.molgen.mpg.de/CPDB), a web-based database developed by Max Planck Institute for Molecular Genetics was used as the major enrichment tool. ConsensusPathDB-human is a consensus database to integrate human molecular interaction network using 32 publically accessible database (Herwig *et al.*, 2016; Kamburov *et al.*, 2008). The publically accessible database integrated in this tool includes Reactome, KEGG, HumanCyc, PID, BioCarta, NetPath, IntAct, Dip, HPRD, BioGRID, SPIKE and more.

To further support and validate the functional annotation data from ConsensusPathDB, DAVID v6.8 (Database for Annotation, Visualization and Integrated Discovery) at http://david.abcc.ncifcrf.gov was used. DAVID v6.8 is a web-based integrated biological knowledgebase and analytical tool that systematically extracts biological meaning from large lists of genes and proteins (Huang *et al.*, 2008). The functional annotation was considered to be significant when a p-value of less than 0.05 ($p<0.05$) was obtained.

Additionally, the identified genes and proteins were evaluated using web-based resources; STRING v10.1 (Search Tool for the Retrieval of Interacting Genes) at http://string-db.org/ to examine protein–protein interaction networks (Franceschini *et al.*, 2013). This was done to explore known and predicted interactions between proteins in each identified pathways (Kuhn *et al.*, 2008).

CHAPTER 4: RESULTS

4.1 Demographic Profiling

4.1.1 Study Populations

A total of 208 OSCC patients (cases) and 134 non-OSCC patients (control) representing a mean age of 58.8±14.2 and 33.5±8.6 respectively was included in this investigation. Due to inadequate clinical and demographic information, samples from two patients were excluded from the study. The socio-demographic profiles of OSCC patient recruited in this study revealed that female (67.0%) Indians (49.5%) had the highest number of patients diagnosed with OSCC (Table 4.1). Among the risk habits recorded in the study populations, tobacco smoking, alcohol drinking and betel quid chewing were the most common risk habits in OSCC. Based on the statistics, betel quid chewing (43.7%) was ranked as the highest possible risk factor in the development of OSCC (Table 4.2).

Table 4.1: Socio-demographic profile of Patients with OSCC and non-OSCC patients.

	OSCC Samples (n=206)		Control Samples (n=134)	
	No. of Patients	%	No. of Patients	%
Gender				
Male	68	33.0	83	61.9
Female	138	67.0	51	38.1
Ethnicity				
Malay	47	22.8	91	67.9
Chinese	35	17.0	30	22.4
Indian	102	49.5	12	9.0
Others	22	10.7	1	0.7
Age (mean ± SD)	205	58.8± 14.2	133	33.5± 8.6

Table 4.2: Social habits (etiologic risk factors of OSCC) of patients (n=206) with OSCC.

Social Habits	No. of Patients	%
Tobacco Smoking		
Smoker	48	23.3
Non-Smoker	136	66.0
Not- Available	22	10.7
Alcohol Drinking		
Alcoholic	50	24.3
Non-Alcoholic	134	65.0
Not- Available	22	10.3
Betel Quid Chewing		
Chewing	90	43.7
Not-Chewing	94	45.6
Not- Available	22	10.7

4.1.2 Serological Analysis

To evaluate seropositivity of HPV and EBV in OSCC patient (n=206) and control (n=134), HPV16-specific ELISA and EBV VCA ELISA assays were used to detect both Immunoglobulin G (IgG) and Immunoglobulin M (IgM) of both infections. HPV16-specific ELISA analysis showed positivity in 95.6% of OSCC patients and 66.4% of control for HPV IgG, whereas for HPV16 IgM, only 20.4% of OSCC patients were seen positive and HPV16 IgM was not detected in the control samples. Based on the EBV VCA ELISA analysis, 96.6% of OSCC patients and 97.2% of control were seen positive for EBV VCA IgG, whereas EBV VCA IgM was not detected in both OSCC patients and control (Table 4.3).

Table 4.3: Percentage of distribution for HPV16 IgG/IgM and EBV VCA IgG/IgM antibodies among the OSCC patients and the control group.

	OSCC Samples (n=206)		Control Samples (n=134)	
	No. of Patients	%	No. of Patients	%
HPV16 IgG				
Positive	197	95.6	89	66.4
Negative	9	4.4	45	33.6
HPV16 IgM				
Positive	42	20.4	0	0
Negative	164	79.6	134	100
EBV VCA IgG				
Positive	199	96.6	130	97.2
Negative	7	3.4	4	2.8
EBV VCA IgM				
Positive	0	0	0	0
Negative	100	100	100	100

4.1.3 HPV Detection using Nested PCR

Using nested PCR assay, the presence of HPV was validated in a total of 84 OSCC genomic DNA. The samples used in this analysis were from the similar patients used in the HPV 16 serological assay. The quality of the genomic samples used in this assay was screened prior nested PCR, through PCR assay using primers PC04 and GH20 that targets the β-globin-specific gene. Positive amplification for the β-globin-specific gene was seen in all samples; therefore all the samples were further analysed through nested PCR (nPCR) to validate the presence of HPV in the OSCC samples.

Nested PCR was performed using two general consensus PCR primers (PGMY09/11 and $GP5^{+}/6^{+}$). Subsequently, by employing agarose gel electrophoresis, the resulting PCR product was examined, to further determine the presences of near identical bands size to the expected band (Figure 4.1). Based on the gel electrophoresis

analysis, amplification of PGMY09/11 was detected in 8/84 OSCC DNA samples, while amplification of nested $GP5^{+}/6^{+}$ was detected in 18/84 DNA samples (Table 4.4).

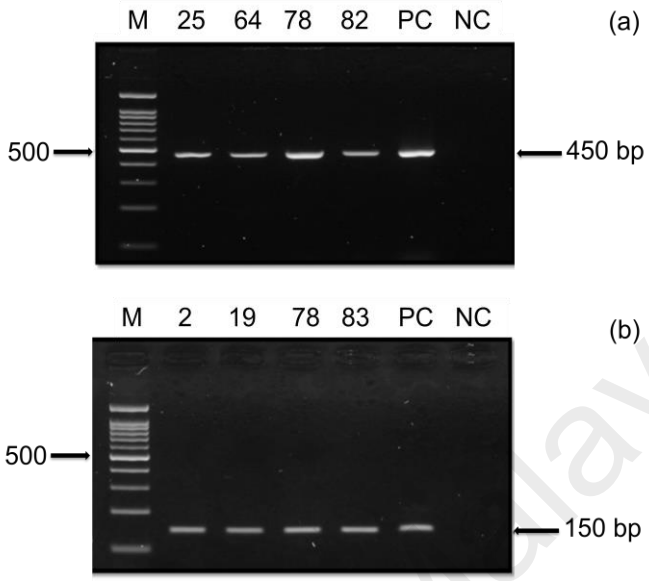


Figure 4.1: HPV nested PCR.
The presence of HPV in the OSCC samples were tested using (a) HPV primer PGMY09/11 and (b) nested HPV primer $GP5^{+}/6^{+}$. A representative gel is shown from 84 samples tested with the two consensus primers displaying positive PCR amplification for PGMY09/11 (450bp) and nested $GP5^{+}/6^{+}$ (150bp). Positive Control (PC) using DNA from HeLa cells and negative control (NC) were included in the experiments. Marker and sample ID are indicated.

Table 4.4: Frequency percentage of HPV detection using PGMY09/11 and GP5+/6+ nested PCR.

	PGMY09/11		PGMY09/11 / $nGP5^{+}/6^{+}$	
	No. of patients	%	No. of patients	%
Positive	8	9.5	18	21.4
Negative	76	90.5	66	78.6
Total	84	100	84	100

The demographic profiles obtained were further used to identify and set the criteria that were used to determine the characteristic and inclusion/exclusion criteria for the exome sequencing cohort. These demographic profiles were also used to further

analyse the exome sequencing data to understand the association of OSCC patients demographic with the mutations identified from next generation sequencing.

4.1.4 Incidence Rate and Prediction of OSCC

Using the data obtained from the demographic profiling, a logistic regression analysis was employed to evaluate whether the patients characteristic or variables could significantly predict the likelihood of OSCC. Four independent variables (i.e. HPV 16 IgG, gender, race and age) were identified from this analysis. These variables were recognized as a unique statistically significant contributor in predicting the likelihood of OSCC (Table 4.5). To further elaborate, this logistic regression analysis identified the strongest predictor as HPV 16 IgG with an odd ratio of 13.6, which was followed by female (gender) with the significant ratio of 4.01. Additionally, when compared with other races, Indian (race) was seen as the most significant predictor for OSCC. Lastly, the model predicted that with every additional year of age, the chances of OSCC were 1.15 times higher.

Table 4.5: Logistic Regression analyses in predicting the risk factors of OSCC.

	B	SE	Odd Ratio	P-value	95% C.I for Lower	95% C.I for Upper
HPV16 IgG	2.61	0.64	13.59**	0.00	3.89	47.51
EBV VCA IgG	-0.14	1.11	0.87	0.90	0.10	7.67
Gender						
(Female)	1.39	0.47	4.01**	0.00	1.59	10.07
Indian				0.00		
Race						
(Indian vs Malay)	-2.26	072	0.10**	0.00	0.03	0.43
Race						
(Indian vs Chinese)	-1.71	0.81	0.18*	0.03	0.04	0.88
Race						
(Indian vs Others)	1.11	1.71	3.04	0.52	0.11	86.71
Age	0.14	0.02	1.15**	0.00	1.10	1.19
Constant	-6.55	1.69	0.00	0.00		

Note: R^2 = 0.561 (Cox and Snell), 0.774 (Nagalkerke). Model χ^2 (7) = 254.3, p<0.001. *p<0.05, **p<0.01.

4.2 Exome Sequencing and Bioinformatics Analysis

4.2.1 Patients Characteristics

To identify somatic mutations in OSCC, exome sequencing was performed on tumour (n=12) and matched non-malignant (normal) adjacent tissues (n=12). All patients were not treated prior sample collection to obtain tissue with the natural occurring state. All patients and their immediate family members did not have any history of past OSCC or another form of cancers. Samples were collected from both female Indian population and other major races in Malaysia. Samples were also collected from patients without risk habits (smoking, alcohol consumption and betel quid chewing) and patients with statistically prominent risk habit (betel quid chewing).

4.2.2 Target Enrichment and Sequencing

The high throughput sequencing produces raw sequencing data which contains sequence adapter, reads with high content of unknown base and low quality reads (Appendix A.1) which was filtered subsequently. This low quality reads are reads with more than half of low-quality base (base quality ≤ 5). Reads with unknown bases (N) were considered low quality if more than 10% of the bases were unknown.

Through these filtering, a clean sequence reads was obtained and mapped or aligned to the reference genome. The alignment data includes sequences of the reads, position of reads, mapping quality and several other statistics on alignment (Appendix B.2). These quality controls of the alignments allow the identification of samples with low quality from the raw data. It also allows quality control evaluation, removal of unwanted reads or alignments and to evaluate the median depth (Guo *et al.*, 2013). This was done to obtain high confidence in the sequence mapping. Once the alignment was completed, only unique mapped reads were reported.

The distribution of per-base sequencing depth and cumulative depth distribution in target regions were also plotted (Figure 4.2). Distribution of per-base sequencing depth approximately followed a Poisson distribution, which showed the exome-capturing target region was evenly sampled. The x-axis denotes sequencing depth, while y-axis indicates the percentage of total target region under a given sequencing depth. In the plot of cumulative depth distribution in target regions, x-axis denotes sequencing depth, and the y-axis indicates the fraction of bases that achieves at or above a given sequencing depth.

a)

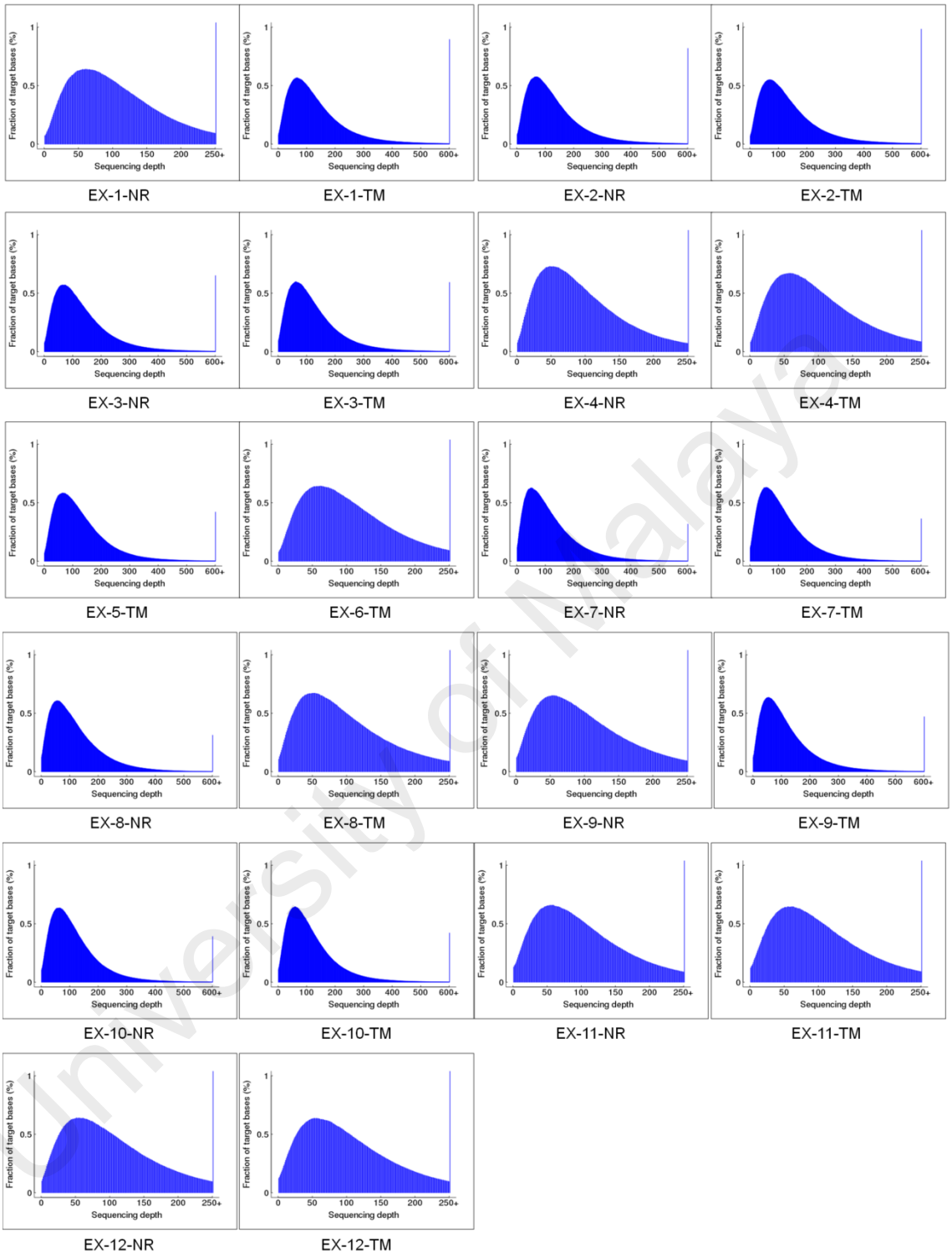


Figure 4.2: Depth distributions.
a) distribution of per-base sequencing depth b) cumulative depth distribution.

b)

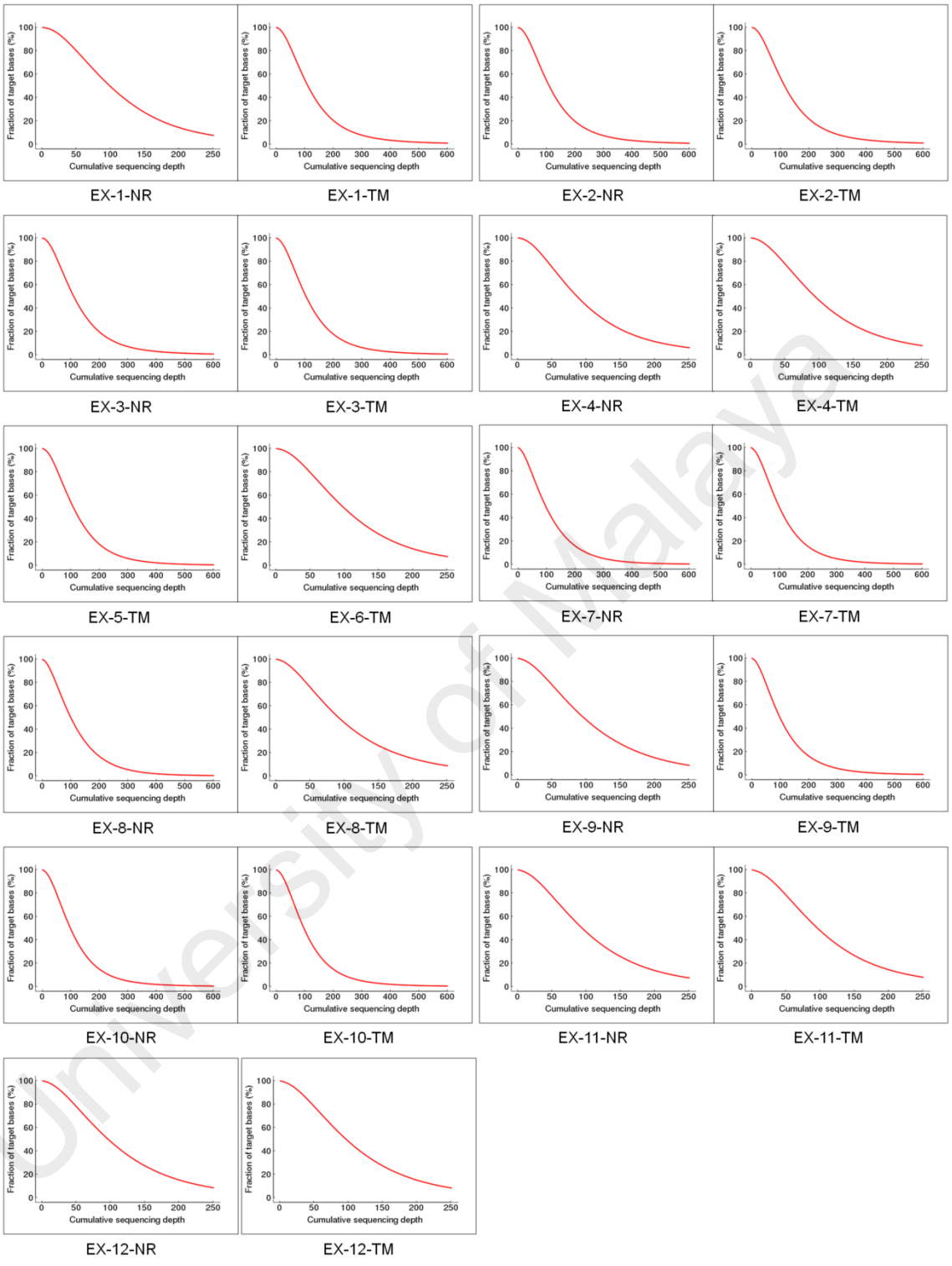


Figure 4.2, Continue

4.2.3 Variant Calling

SNPs in OSCC were identified using SOAPsnp followed by annotation and classification using ANNOVAR. The statistics on the mutations and distribution of SNPs in different gene regions were presented in Appendix A.3. To further identify somatic mutation or SNVs in this study, VarScan was used followed by annotation and classification using ANNOVAR. The statistics on the identification and distribution of SNVs in different gene regions were presented in Appendix A.4.

The data obtained through variant calling were filtered to remove non-relevant or germline mutations. A total of 114 mutations were removed from both 1000 genome database and dbsnp135 database, 9 mutations were removed from 1000 genome database and 139 mutations were removed from dbsnp135 database. Finally, through variant calling a total of 4,610 mutations were identified in 1,479 genes. 94% (4,348) of this total mutation were identified as novel mutation. 1,755 of this total mutation were located in the coding DNA sequence (CDS) which is also known as coding region (Figure 4.3).

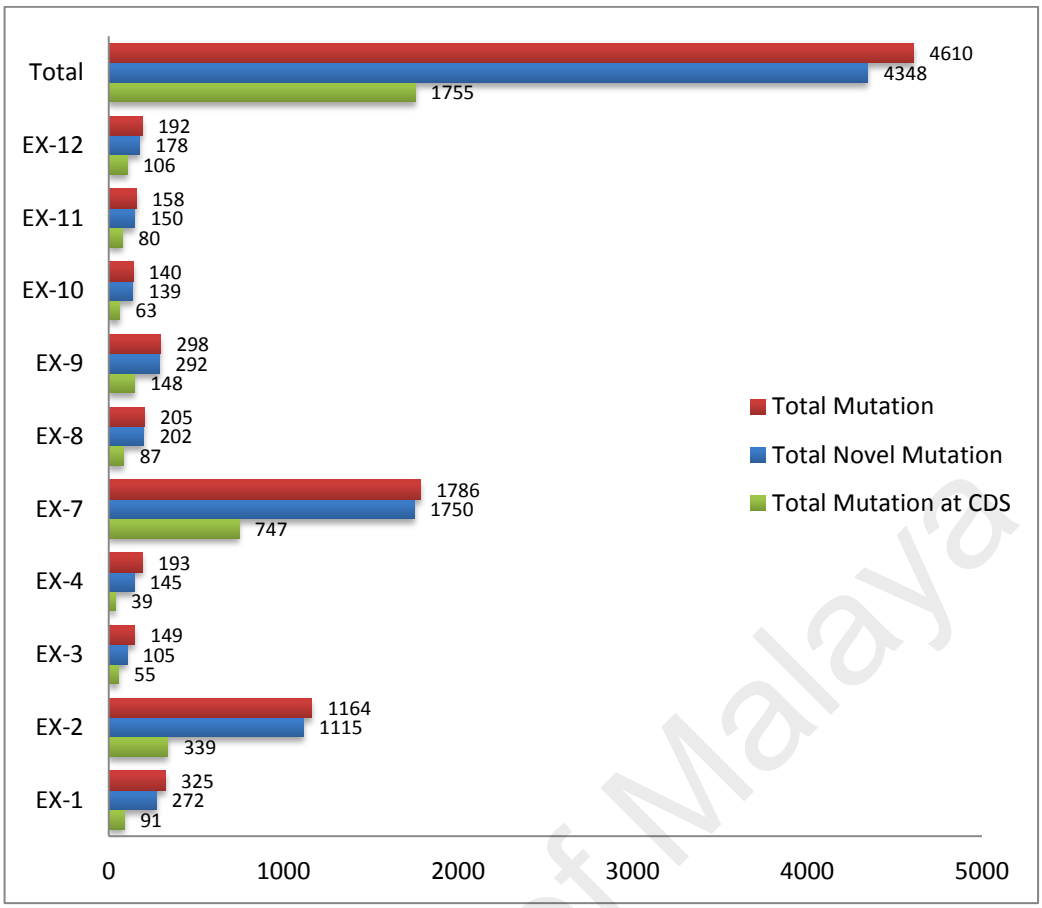


Figure 4.3: Mutation burden of 10 OSCC samples.
Total mutations identified in 10 OSCC samples using Exome sequencing. Mutations represent somatic mutations, novel mutations and mutations at the CDS.

A total of 4,603 mutations were identified as heterozygous and 7 mutations were identified as homozygous. Out of the total mutations identified, 1,755 mutations leads to amino acid changes which includes 1,071 (61%) missense mutations, 76 (4%) stopgain mutations (nonsense alteration) and 35 (2%) splice site mutations (splicing) (Figure 4.5a). In addition, a total of 573 (33%) synonymous (silent) mutations were also detected. The mean number of somatic SNVs per patients, including missense, stopgain, splicing and synonymous was 107 (range: 26-448), 8 (range: 0-26), 4 (range: 1-9) and 57 (range: 12-265) respectively. Missense mutation (non-synonymous mutation) in which a single nucleotide alteration results in amino acid changes, were seen as the frequent mutation in all of the OSCC samples (Figure 4.4).

a) b)

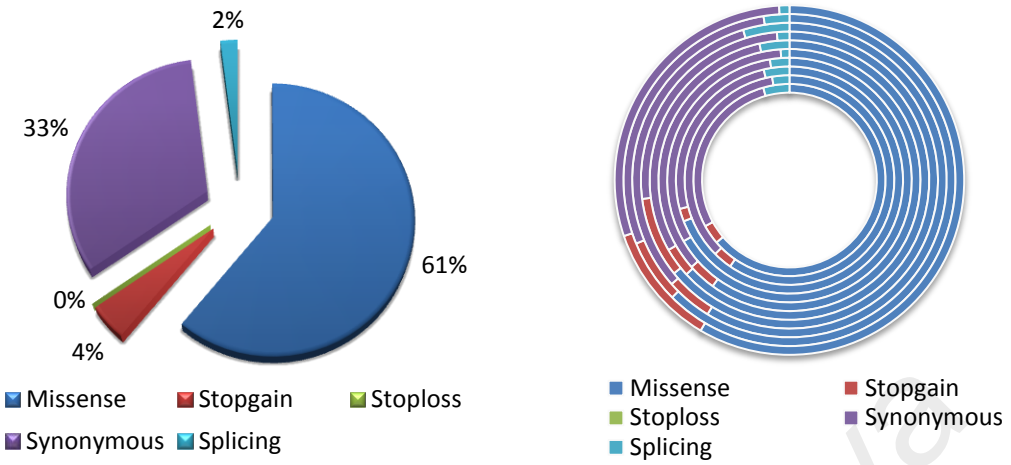


Figure 4.4: Mutation distribution of 10 OSCC samples.
(a) The proportion of somatic mutation (SNVs) in the exonic and splice site region with the breakdown of the mutation types, including Missense, Stopgain, Stoploss, Synonymous and Splicing. (b) The breakdown of different mutation type on each individual OSCC sample (The order of the samples begins with the innermost ring as EX- 1 followed by EX- 2 to EX-12 respectively).

4.3 Evaluation and Validation of Candidate Mutated Gene

4.3.1 Identifying Candidate Mutated Gene

To further discover potential genes associate to the development or progression of OSCC, candidate mutated genes were determined by identifying recurrently mutated genes. Mutation burden allows the identification of recurrently mutated genes and candidate somatic mutation. To elucidate the mutation burden, the clean data obtained through exome sequencing were calculated and plotted (Figure 4.5) with GenVisR waterfall function using R i386 version 3.2.2. The GenVisR waterfall function illustrates the mutational burden of variants and further differentiate the mutation types (Dessau *et al.*, 2008; Team, 2013). The genomic events at the variant level were emphasized in this plot by ranking the most recurrently mutated genes.

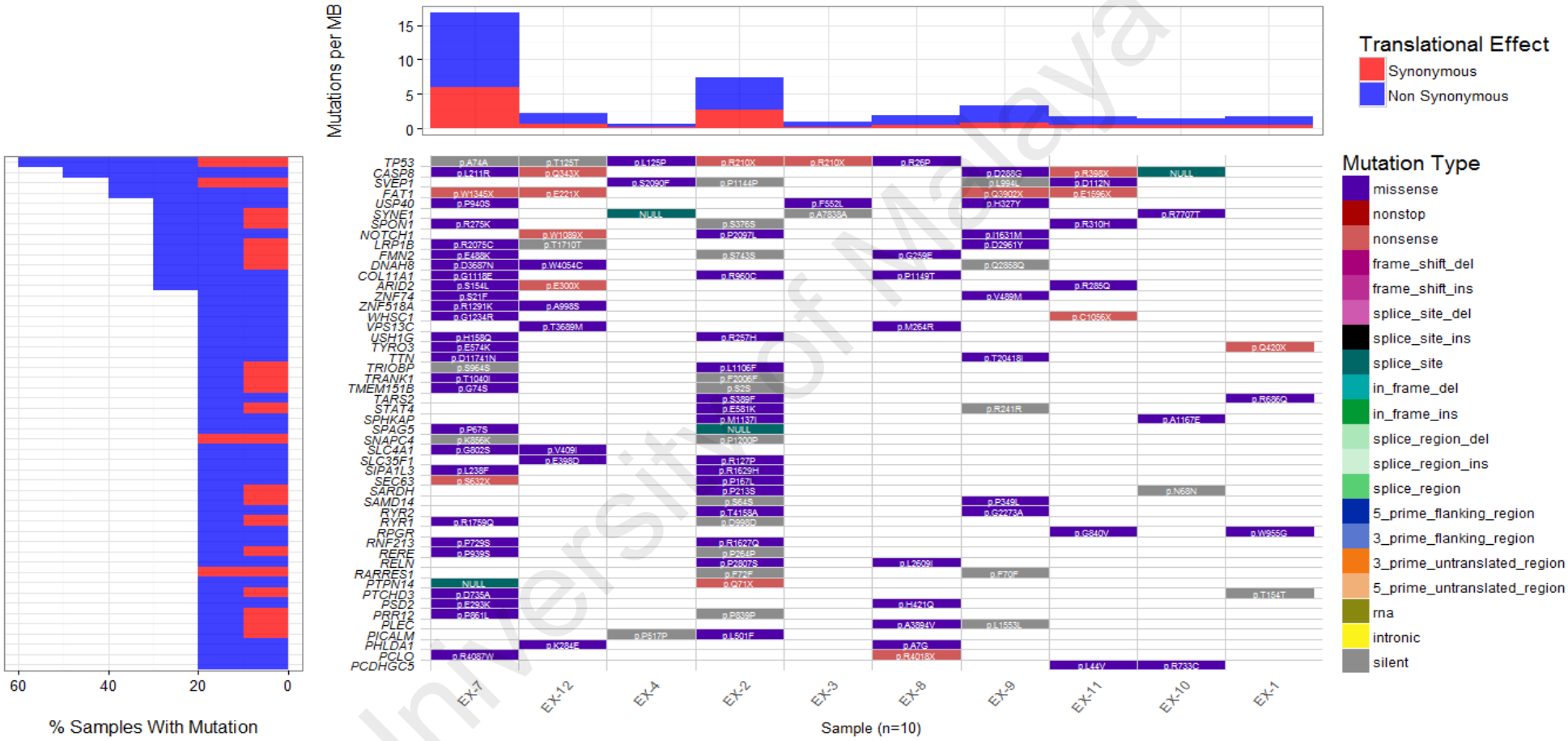


Figure 4.5: Mutation landscape of OSCC.

Mutation landscape illustrated with waterfall plot indicates the most recurrently mutated genes across 10 OSCC samples. The plot includes mutation burden, mutation types indicated by the color grid and amino acid changes. TP53 was ranked as the most mutated gene in OSCC, followed by CASP8. Missense mutation was also seen as the highest non-synonymous point mutation in OSCC.

Based on the mutation burden analysis, sample EX-7 was identified with the highest percentage of mutations; which was more than 15 mutations for every one millions base pairs. This was followed by EX-2 which had more than 5 mutations for every one million base pairs. The most recurrent and significant mutated genes in OSCC were ranked and displayed on the left panel of Figure 4.5. *TP53* the most important tumour suppressor gene in carcinogenesis was seen altered in 60% of all the samples, followed by *CASP8* (50%), *SVEP1* (40%), *FAT1* (40%); *USP40* (30%), *SYNE1* (30%), *SPON1* (30%), *NOTCH1* (30%), *LRP1B* (30%), *FMN2* (30%), *DNAH8* (30%), *COL11A1* (30%), *and ARID2* (30%). The remaining mutations were only identified in 20% of the samples. Missense mutation was also seen as the highest non-synonymous point mutation. In addition, the amino acid changes resulting from the mutations were also illustrated in the waterfall plot.

The mutation spectrum of the identified SNV was determined using the GenVisR TiTv (transition/transversion) graph (Figure 4.6). Based on the analysis, the transition G>A/C>T were observed higher than the other nucleotide substitution (Transition: G>A/C>T, A>G/T>C, Transversion: G>T/C>A, G>C/C>G, A>T/T>A, A>C/T>G). This was seen in concordance with the findings by Do *et al.* (2012), where the transition was expected to occur more frequent than transversion. Figure 4.7a and Figure 4.7b illustrates the mutation spectrum representing overall SNV and SNVs at the CDS region respectively for each 10 OSCC samples.

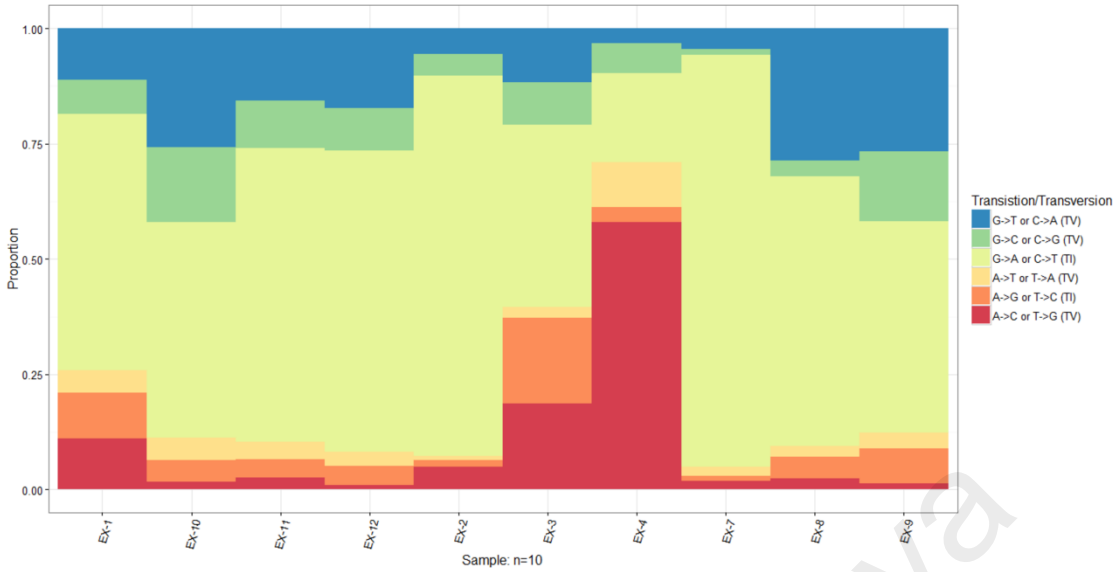


Figure 4.6: Somatic Mutation spectrum of transition and transversion SNV. GenVisR TiTv (transition/transversion) graph illustrates spectrum of mutation of OSCC samples.

a)

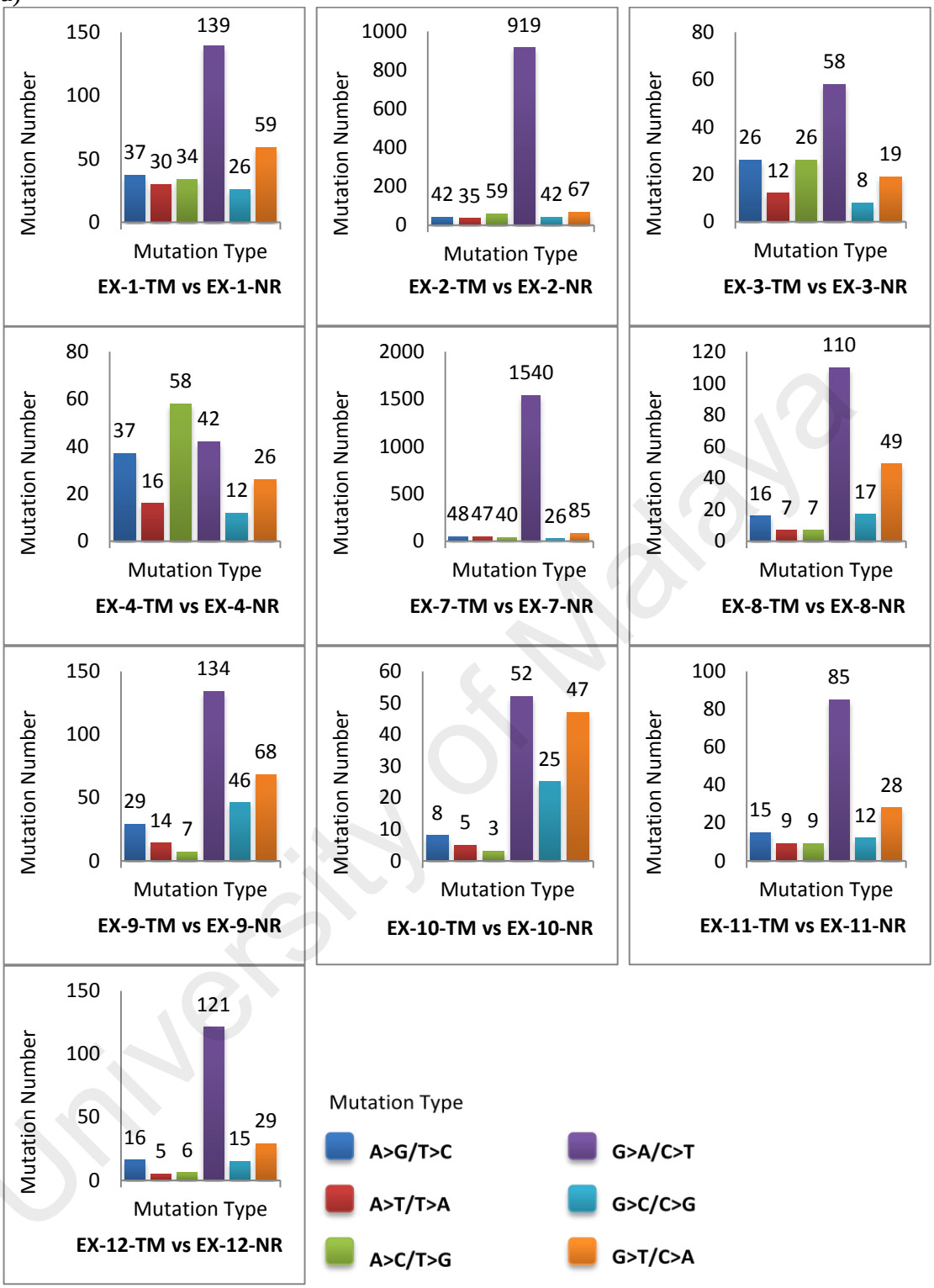


Figure 4.7: Transition and transversion SNV across 10 OSCC samples.
a) Somatic mutation spectrum of overall SNV in all 10 OSCC samples. b) Somatic mutation spectrum of SNV located in the CDS region in all 10 OSCC.

b)

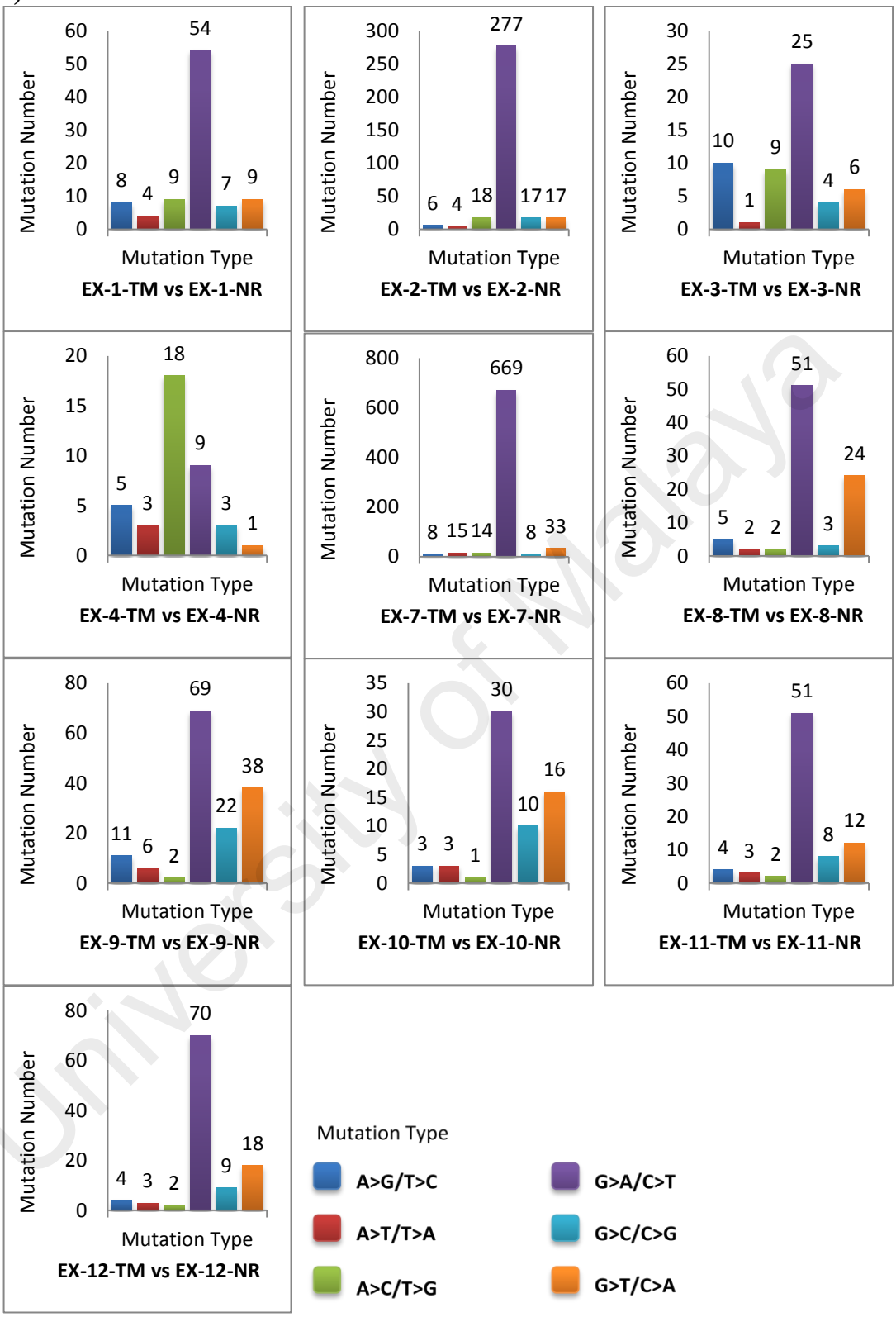


Figure 4.7, Continue

From the large numbers of somatically mutated genes identified through exome sequencing, 13 genes were found to be recurrently mutated in 3 or more OSCC samples (Figure 4.5). These genes were Tumour protein p53 (*TP53)*, Caspase 8 (*CASP8)*, Sushi, von Willebrand factor type A, EGF and pentraxin domain containing 1 (*SVEP1)*, FAT atypical cadherin (*FAT1)*, Ubiquitin specific peptidase 40 (*USP40)*, Spectrin repeat containing nuclear envelope protein 1 (*SYNE1)*, Spondin 1 (*SPON1)*, Notch 1 (*NOTCH1)*, LDL receptor related protein 1B *(LRP1B)*, Formin 2 *(FMN2)*, Dynein axonemal heavy chain 8 (*DNAH8)*, Collagen type XI alpha 1 chain *(COL11A1)*, and AT-rich interaction domain 2 (*ARID2*). These genes were then compared to the Sanger COSMIC and OrCGDB cancer database. This was done to select candidate genes to be further studied. The candidate genes were selected based on novelty and previously reported association with OSCC (Table 4.6). Candidate genes were also selected based on the highest number of missense mutations. From these 13 somatically muted genes, *CASP8*, *USP40*, *NOTCH1* and *COL11A1* were selected as the candidate genes. *CASP8* (2 missense) and *NOTCH1* (3 missense) were previously reported in OSCC based on the Sanger COSMIC database. However, the association of *COL11A1* (4 missense) and *USP40* (3 missense) with OSCC was not previously reported in Sanger COSMIC database.

Table 4.6: Comparison of the candidate somatic genes with Sanger COSMIC database and OrCGDB database.

No	Gene Symbols	Gene Name	OrCGDB	COSMIC
1	TP53	Tumor protein p53	√	√
2	CASP8	Caspase 8	X	√
3	SVEP1	Sushi, von Willebrand factor type A, EGF and pentraxin domain containing 1	X	X
4	FAT1	FAT atypical cadherin 1	√	√
5	USP40	Ubiquitin specific peptidase 40	X	X
6	SYNE1	Spectrin repeat containing nuclear envelope protein 1	X	X
7	SPON1	Spondin 1	X	X
8	NOTCH1	Notch 1	X	√
9	LRP1B	LDL receptor related protein 1B	X	X
10	FMN2	Formin 2	X	X
11	DNAH8	Dynein axonemal heavy chain 8	X	X
12	COL11A1	Collagen type XI alpha 1 chain	X	X
13	ARID2	AT-rich interaction domain 2	X	√*

* Not associated with Head & Neck Cancer

4.3.2 Evaluation of Candidate Mutated Gene

The candidate genes that were selected for this study consist of several SNVs from the 4,348 novel somatic mutations identified through exome sequencing. In total 12 SNVs from *CASP8, USP40, NOTCH1* and *COL11A1* (2, 3, 3, 4 SNVs respectively) were identified to be further studied. Figure 4.8 illustrates the association of the candidate SNVs (presented in green) with the total novel somatic mutations. The location, exon position, nucleotide position, SNVs variation of the genes were further identified and presented in Table 4.7.

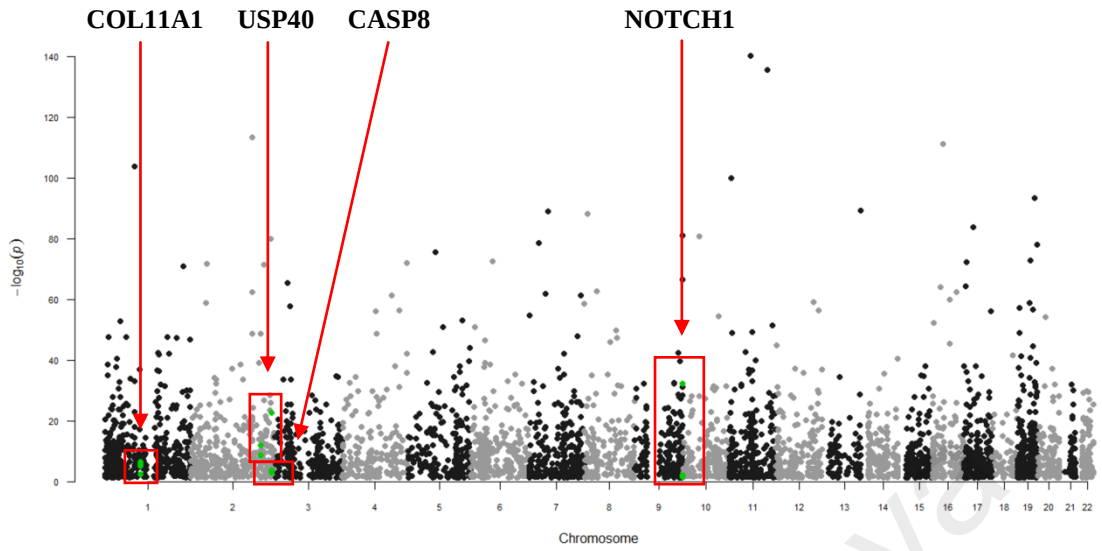


Figure 4.8: Manhattan plot for exome sequencing.
The association results of the SNVs analysis from the candidate genes were plotted against the genomic position. The identified SNVs are presented in green. Candidate genes *COL11A1* and *NOTCH1* were located at chromosome 1and 9 respectively. Candidate genes *USP40* and *CASP8* were located on chromosome 2.

Table 4.7: Novel candidate mutated gene associated with OSCC. All mutations were identified as missense mutation.

No	Sample	Gene	Location	Exon Count	Position	Reference	Variation	Somatic p-value
1	EX-7	CASP8	2q33-q34	16	202137404	T	G	2.43E-09
2	EX-9				202149644	A	G	1.35E-12
3	EX-3	USP40	2q37.1	35	234436155	G	C	1.79E-04
4	EX-7				234405409	G	A	2.23E-23
5	EX-9				234457770	G	A	9.45E-04
6	EX-2	NOTCH1	9q34.3	34	139391901	G	A	4.73E-33
7	EX-9				139399250	G	C	0.006924
8	EX-9				139412651	C	A	0.033698
9	EX-2	COL11A1	1p21	71	103412455	G	A	0.003948
10	EX-7				103387081	C	T	9.45E-07
11	EX-8				103381210	G	T	3.41E-07
12	EX-8				103483428	C	T	8.96E-06

To further understand the association of this mutation with its associated protein, the amino acid changes of each SNV were identified. Table 4.8 describes the identified SNVs with the resulting codon changes and amino acid changes. To contextualize this mutation within a structurally related family of proteins, lollipop plot was generated (Figure 4.9). Lollipop plot which was manually curated and generated using MutationMapper, identified the position and the frequency of the mutations in *CASP8*, *USP40*, *NOTCH1* and *COL11A1* in the context of Pfam protein domain (Vohra & Biggin, 2013). All information on the protein and the protein domain was retrieved from the protein database UniProt (www.uniprot.org).

Table 4.8: Amino acid changes in candidate mutated gene associated with OSCC.

No	Sample	Gene	Position	Reference	Variation	Codon Changes	Amino Acid Changes
1	EX-7	CASP8	202137404	T	G	c.T632G	p.L211R
2	EX-9		202149644	A	G	c.A863G	p.D288G
3	EX-3	USP40	234436155	G	C	c.C1656G	p.F552L
4	EX-7		234405409	G	A	c.C2818T	p.P940S
5	EX-9		234457770	G	A	c.C979T	p.H327Y
6	EX-2	NOTCH1	139391901	G	A	c.C6290T	p.P2097L
7	EX-9		139399250	G	C	c.C4893G	p.I1631M
8	EX-9		139412651	C	A	c.G1193T	p.C398F
9	EX-2	COL11A1	103412455	G	A	c.C2878T	p.R960C
10	EX-7		103387081	C	T	c.G3353A	p.G1118E
11	EX-8		103381210	G	T	c.C3445A	p.P1149T
12	EX-8		103483428	C	T	c.G1013A	p.G338D

The lollipop plot schematically represents the location of the somatic mutations and the protein structure of each candidate genes. The amino acid changes resulting from the non-synonymous (missense) mutation were plotted on the protein domain. Lollipop plot revealed 2 non-synonymous mutations (L211R, D288G) in *CASP8* gene which were located in the various sites of protein including in the Peptidase_C14 (caspase) domain. Caspases play an important role in inflammation and as apoptosis

initiator (Eckhart *et al.*, 2008). 3 non-synonymous mutations (H327Y, F552L, and P940S) in *USP40* genes were located in the several sites of the protein including in the Ubiquitin carboxyl-terminal hydrolase (UCH) domain. 3 non-synonymous mutations (C398F, I1631M, and P2097L) in *NOTCH1* genes were located in several sites in the protein and in various domains. These domains include ankyrin repeat domain and human growth factor-like EGF (EGF-like domain). Lastly, the lollipop plots revealed 4 non-synonymous mutations (G338D, R960C, G1118E, and P1149T) in *COL11A1* gene which was located in collagen domain (collagen triple helix repeat).

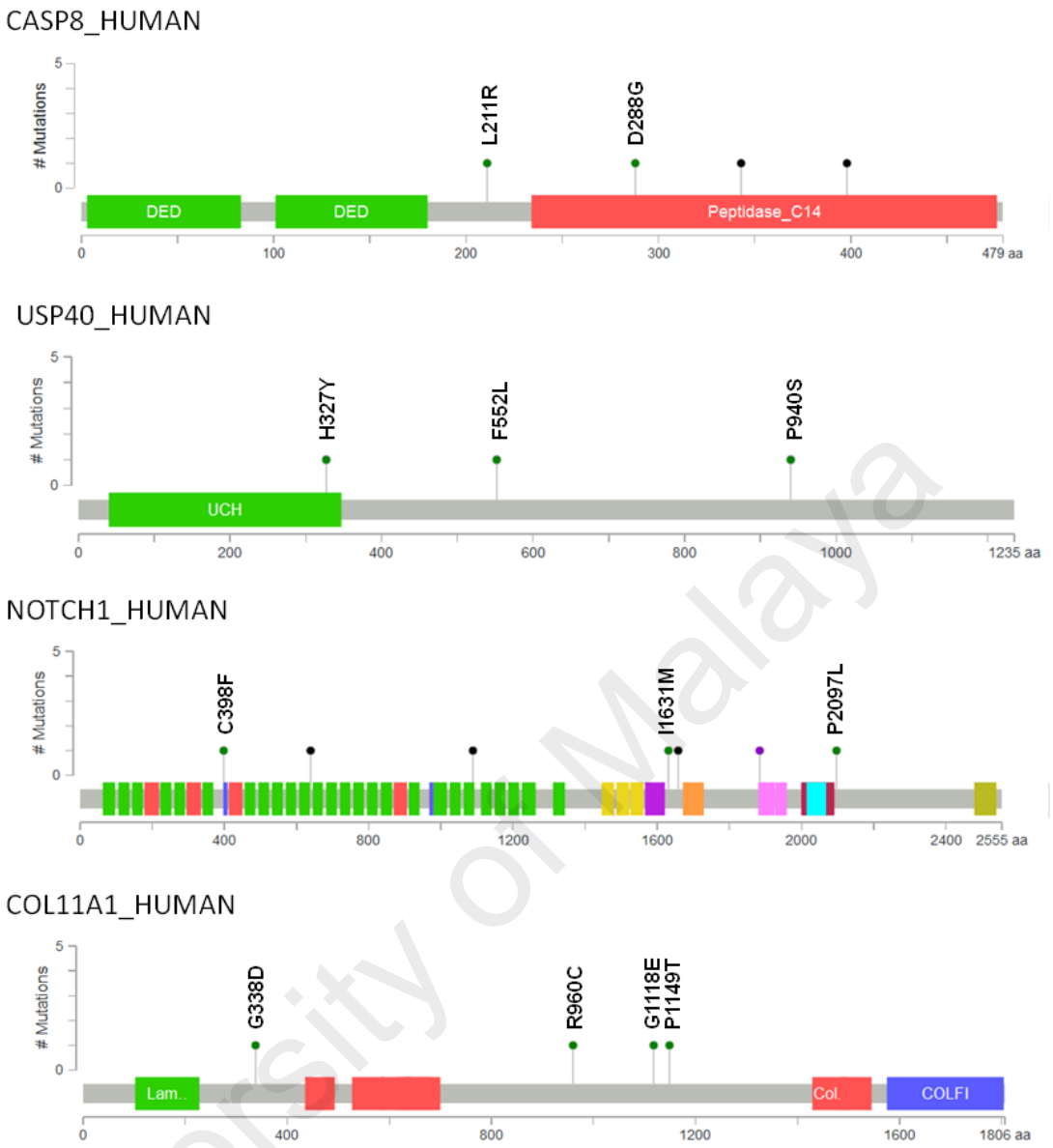


Figure 4.9: Lollipop plot of amino acid variants along the CASP8, USP40, NOTCH1 and COL11A1 protein schematic structure.

The mutations and the represented amino acid changes were shown on the x-axis, while the frequency of the mutations was shown on the y-axis. Missense mutations are depicted in green, while truncating mutations (including nonsense, nonstop, frameshift deletion, frameshift insertion, and splice site) was depicted in black. Other mutations such as silent mutation are depicted in purple.

The 12 targeted SNVs from candidate mutated gene *CASP8, USP40, NOTCH1* and *COL11A1* were further evaluated to understand the effects of these mutations on protein function. To predict whether this amino acid substitution is neutral or deleterious/dangerous, consensus classifier PredictSNP which is a combination of eight established prediction tools (MAPP, nsSNPAnalyzer, PANTHER, PhD-SNP, PolyPhen-1, PolyPhen-2, SIFT and SNAP) was used (Bendl *et al.*, 2014). PredictSNP provides predictions based on the evolutionary, structural characteristic, physico-chemical and supplementation of experimental annotation from UniProtKB databases and Protein Mutant (Bendl *et al.*, 2014; Tandale *et al.*, 2016).

Table 4.9 lists the predictions of the 12 amino acid substitutions found in the candidate mutated gene *CASP8, USP40, NOTCH1* and *COL11A1* in this study. The PredictSNP score reflects the decree of confidence in a percentage value. Based on individual prediction, the PredictSNP platform identified all targeted amino acid substitution as a deleterious variant with at least one of the 9 algorithms with a confidence interval between 40-91%. These amino acid substitutions were considered as disease-associated variants. However, this was exceptional for the amino acid substitution p.F552L of the gene *USP40* which was seen as neutral.

Table 4.9: Amino acid substitution pathogenecity prediction of the mutations found in OSCC using PredictSNP consensus classifier.

Gene	Wild	Position	Target	PredictSNP prediction	MAPP prediction	PhD-SNP prediction	Poly Phen-1 prediction	Poly Phen-2 prediction	SIFT prediction	SNAP prediction	nsSNP Analyzer prediction	PANTHER prediction
CASP8	L	211	R	87%*	NILL	86%*	74%*	68%*	79%*	85%*	NILL	74%*
CASP8	D	288	G	60%	66%*	61%*	67%	71%	75%	62%*	65%	69%*
USP40	H	327	Y	60%	63%	55%	67%	47%*	45%*	56%*	NILL	47%
USP40	F	552	L	83%	75%	72%	67%	68%	81%	61%	NILL	71%
USP40	P	940	S	74%	76%	55%	67%	40%*	71%	67%	NILL	NILL
NOTCH1	C	398	F	87%*	NILL	73%*	NILL	NILL	79%*	89%*	63%*	NILL
NOTCH1	I	1631	M	83%	75%	78%	NILL	NILL	76%	77%	63%*	NILL
NOTCH1	P	2097	L	87%*	NILL	86%*	NILL	NILL	79%*	56%*	63%*	47%
COL11A1	G	338	D	87%*	57%*	77%*	74%*	81%*	79%*	89%*	NILL	NILL
COL11A1	R	960	C	61%*	71%	82%*	NILL	NILL	53%*	62%*	NILL	NILL
COL11A1	G	1118	E	87%*	91%*	86%*	NILL	NILL	79%*	87%*	NILL	NILL
COL11A1	P	1149	T	75%	73%	72%	NILL	NILL	46%*	71%	NILL	65%

Percentages with * indicate the reliability of pathogenicity prediction in pathogenic mutations (deleterious). NILL represents unclassified result

4.3.3 Candidate Mutated Gene Screening

The identified SNV were further validated using the Fluidigm 192.24 Dynamic Array Integrated Fluidic Circuit (IFC). The Fluidigm array used in this study was capable of high sample throughput SNP genotyping. It allowed the genotyping of 192 samples against 24 SNP assays in a single run, of which it provided outstanding data quality, and an accelerated workflow.

In this study, a total of 167 OSCC gDNA sample (10 - 20 ng/µl) along with two no-template control (NTC) were analysed against 12 targeted SNVs (CASP8_E7a, CASP8_E9a, USP40_E3a, USP40_E7a, USP40_E9a, NOTCH1_E2a, NOTCH1_E9a, NOTCH1_E9b, COL11A1_E2a, COL11A1_E7a, COL11A1_E8a, COL11A1_E8b) identified through exome sequencing. The 12 OSCC samples (EX-1, EX-2, EX-3, EX-4, EX-5, EX-6, EX-7, EX-8, EX-9, EX-10, EX-11, and EX-12) that were sequenced through exome sequencing were included to validate the identification of the targeted SNVs in the respected samples. Several samples were duplicated to allow the maximum use of the IFC. All samples were amplified using Specific Targeted Amplification (STA) protocol on the 12 targeted SNVs. The amplified samples were diluted 1:100 with DNA Suspension Buffer after STA and genotyped on the 192.24 Dynamic Array Integrated Fluidic Circuit (IFC). Primer assays were designed prior genotyping. All design showed high rank assay design with higher confidence in SNVs detection. This was an exception for three assays (NOTCH1_E9a, NOTCH1_E9b, and COL11A1_E8a) that showed medium rank design due to GC content that was outside product specification.

The IFC chip was thermal cycled using the pre-set conditions. The end-point fluorescence values were measured using the BioMark^{TM} system and analysed using the Fluidigm SNP Genotyping Analysis software to obtain genotype calls. An average call

rate of 72.54 % (Appendix A.5) was achieved from the total genotyped samples against the 12 SNVs.

Raw image of IFC Chip in both FAM (excitation peak = 495 nm and emission peak = 520 nm) and HEX (excitation peak = 538 nm and emission peak = 544 nm) fluorescent channels along with the computer generated image of the genotype call for the sample reaction chambers are shown in Figure 4.10. Each row in the IFC chip represents data from one OSCC gDNA sample from each sample inlet.

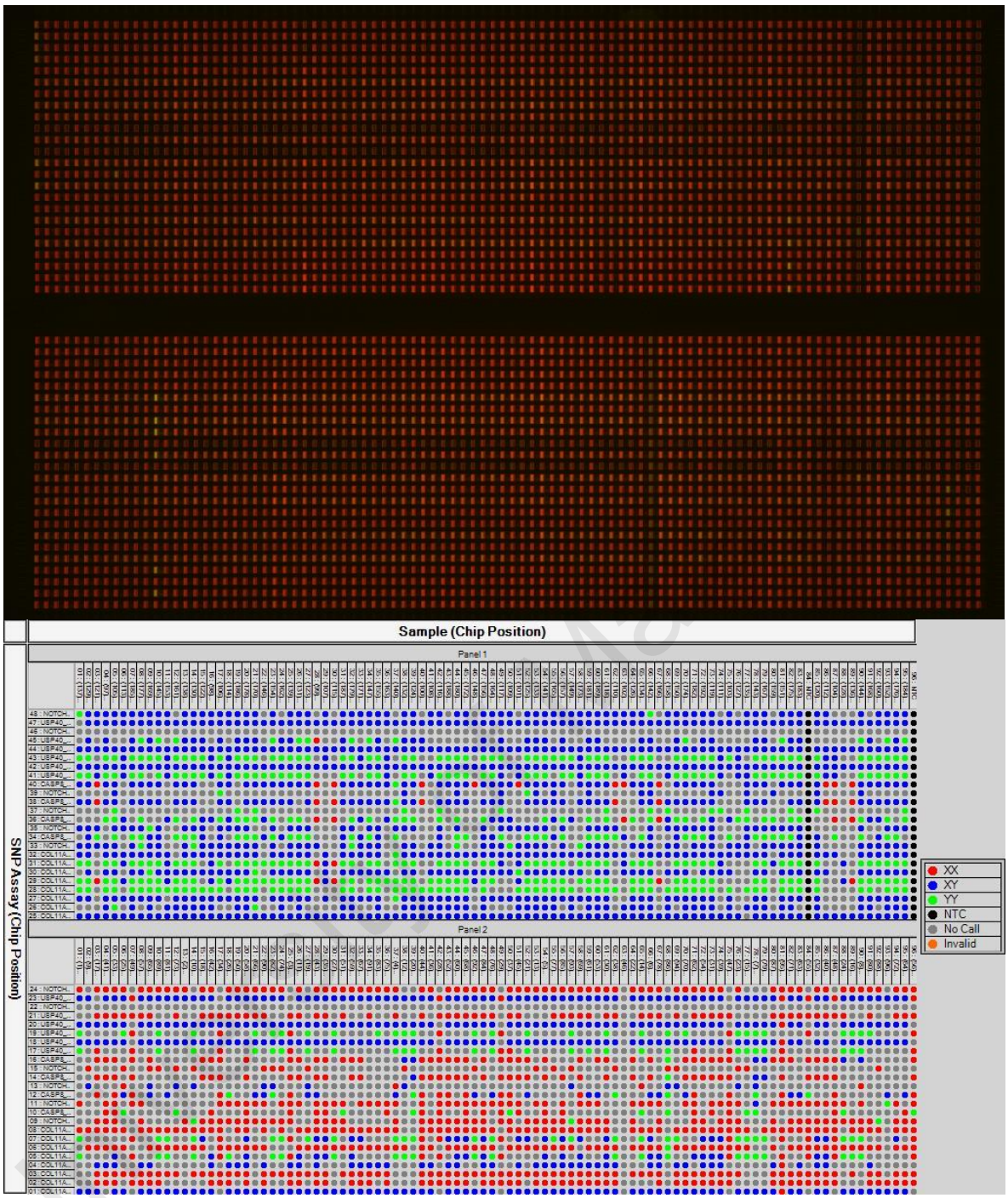


Figure 4.10: SNP Genotyping using Fluidigm 192.24 Dynamic Array IFC. The analysis on the end-point image of a genotyping chip run and the genotyping calls for each sample was performed using the Fluidigm SNP Genotyping Analysis software. (A) Call map view of the genotyping calls for each of reaction chambers. (B) Raw image of 192.24 Dynamic Array IFC chip run in both FAM and HEX fluorescent channels.

The software calculated the fluorescent signals from both FAM and HEX channels and plots each sample on a scatter plot (Figure 4.11). The fluorescent signal from each sample was represented as an independent data point on the scatter plot. The X-axis was used to plot the FAM (ROX) fluorescent value and the Y-axis was used to plot the HEX fluorescent value for each sample. K-means clustering algorithm based on the nearest-centroid sorting was used to automatically classify samples into their respective genotype groups including the NTC. This genotyping include for homozygous XX (red), homozygous YY (green), heterozygous XY (blue) and NTC (black). Samples with low confidence (No Call) were marked in grey.

Fluidigm assay confirmed the presences of 9 targeted SNV in the original OSCC samples. This was exceptional for the SNV NOTCH1_E9a, NOTCH1_E9b, and COL11A1_E8a due to poor assay design. As for USP40_3a, the targeted SNV fell in the cluster suggesting the SNV may represent the population of OSCC sample. Further study on USP40 is required to understand the function of the gene in OSCC.

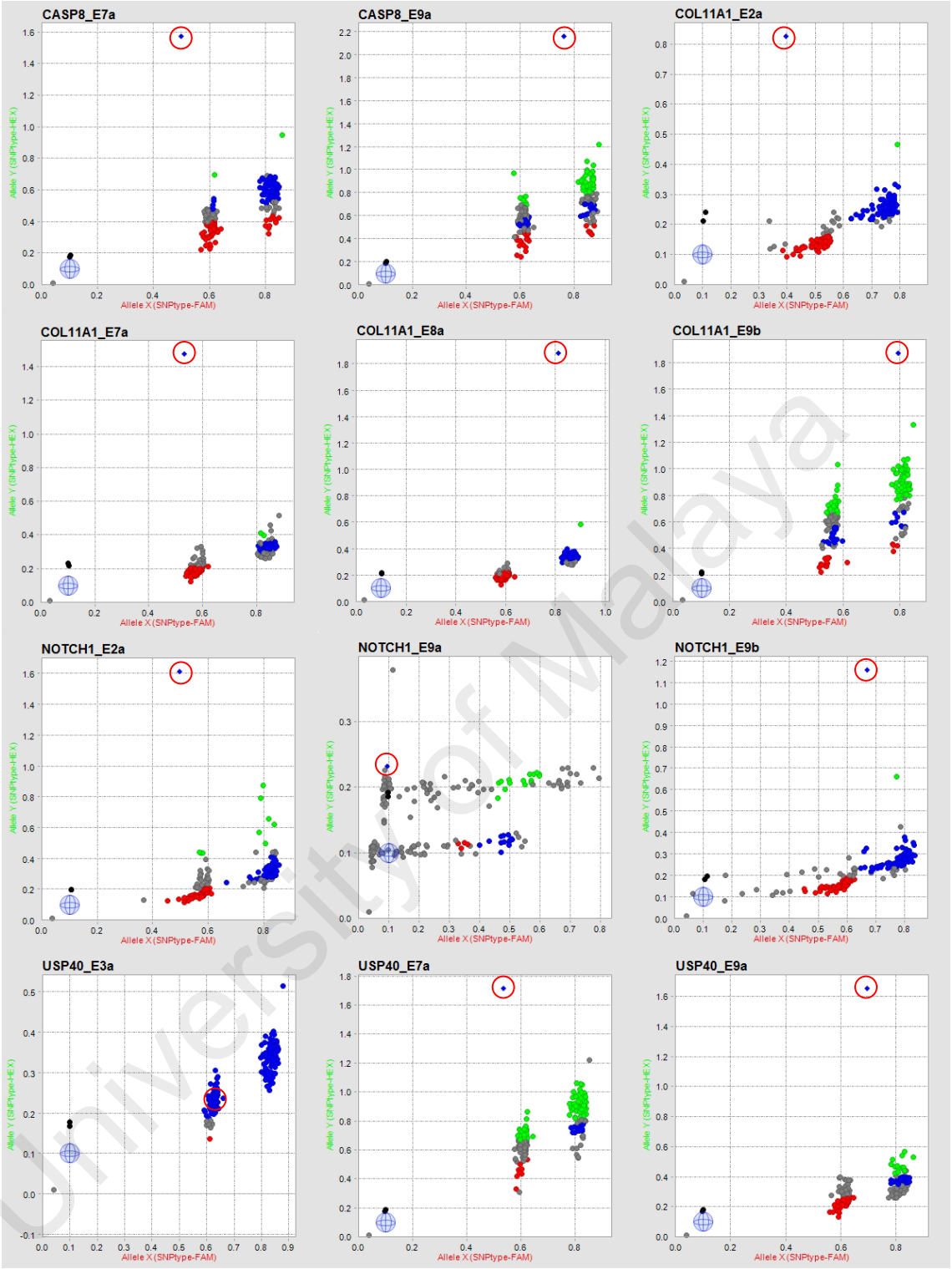


Figure 4.11: SNV Genotype calls.
Genotyping scatter plot of OSCC samples. Y-axis represents HEX fluorescence intensity and the x-axis represents FAM fluorescence intensity. Both intensity values were normalized by ROX fluorescence.

4.3.4 OSCC Mutation Association Study

To further understand the association of patients' demographic and risk factors with the identified mutation, patients' data were plotted against the mutation spectrum. The mutation spectrum covers 50 recurrent mutated genes identified through exome sequencing. The EBV and HPV status of three samples (EX8, and EX10) were un-available due to insufficient sample. These associations were depicted in Figure 4.12.

Based on this association study, higher mutation burden was observed in sample EX7 and EX2. In general non-synonymous mutations, predominantly missense mutation represented as the frequent mutation identified in this study. Higher mutation rate was observed in female Indian patients exposed to betel quid. Patients with EBV and HPV seropositivity showed an increase in mutation burden. However, irregularities of mutation burden were observed when compared to the stages of the cancer stages.

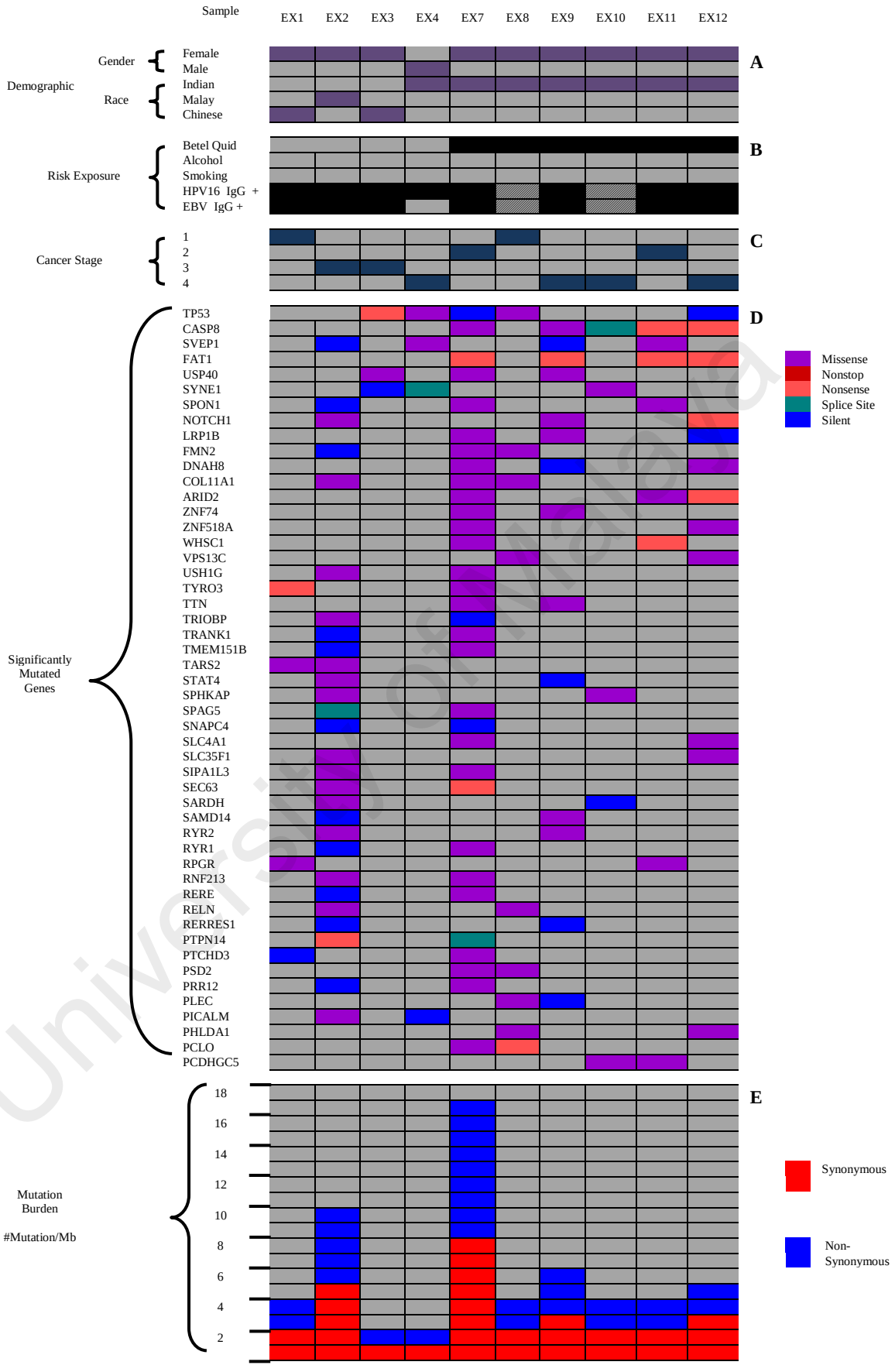


Figure 4.12: Association of demographics and mutation profile of OSCC patient.

4.4 Proteomic Analysis

Studies have shown that proteomics along with genomic study allow the identification of a reliable diagnostic molecular marker. In addition to the genomic study, a well-characterized proteomic analytical platform combining two-dimensional gel electrophoresis (2-DE), mass spectrometry (MS), immunoproteomic and label-free LCMS were used in this study to identify biological markers for OSCC and to further understand the molecular events underlying OSCC.

4.4.1 Oral Cancer Serum Proteomics

4.4.1.1 Identification of Possible Biomarkers Using 2-DE

To generate proteome profiles for OSCC patients and normal control, unfractionated serum samples were separated using 2-DE. Using PDQuest™ 2-D gel analysis software, all protein spots appeared in the 2-DE gels of OSCC patients (n=25) and normal controls (n=25) were analysed. Based on the analysis, several up- and down-regulated proteins in the sera of OSCC patients were identified. Using MASCOT search engine and NCBI database, the MS (MALDI-TOF/TOF) analysis of the proteins spot further recognized seven OSCC proteins: leucine-rich α2-glycoprotein (LRG), alpha-1B-glycoprotein (A1BG), clusterin (CLU), PRO2044, haptoglobin (HAP), proapolipoprotein A1 (proapo-A1) and retinol-binding protein 4 precursor (RBP4). In addition, the complement component 3 (C3) proteins which found to be immunoreactive was identified as additional proteins through MS analysis. Data regarding the protein identification, MASCOT accession number, isoelectric point (pI), theoretical mass, MASCOT protein details and fold change for each protein were presented in Table 4.10.

Table 4.10: MS identification of 8 OSCC proteins.

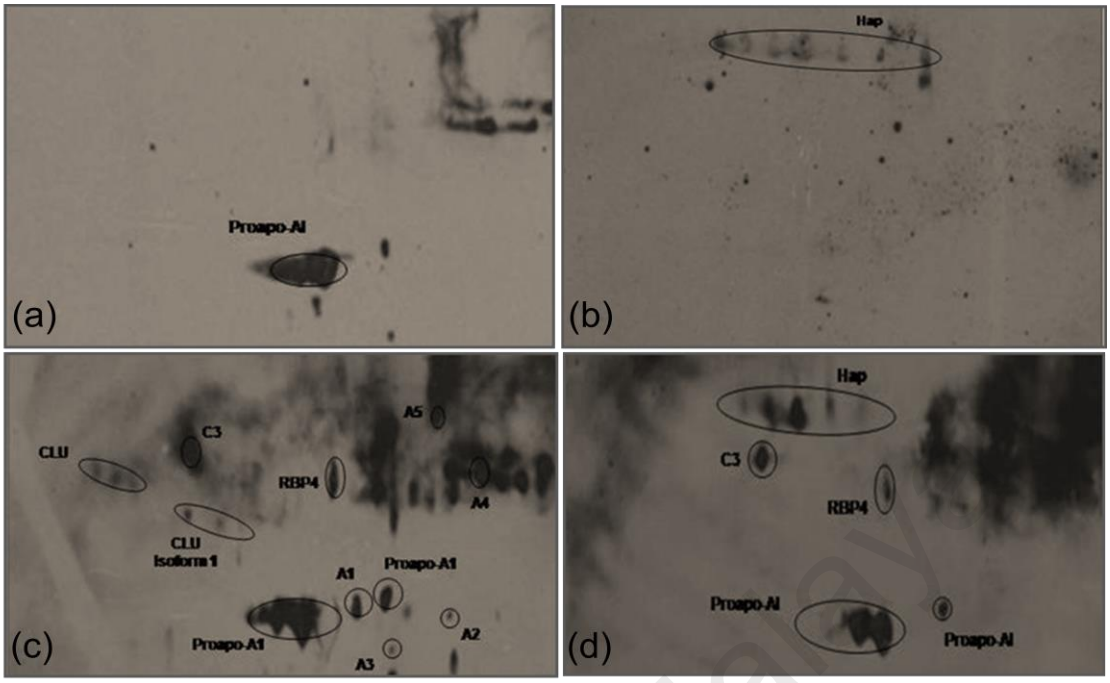
Protein Name	Protein Entry Name	MASCOT accession number	pI	Theorectical mass	Sequence coverage	Search score	Queries match	Expected value	Fold Change
Up-Regulated Protein									
Haptoglobin	HAP	gi|3337390	6.14	38722	32%	264	11	4.6e-020	1.47
Proapollipoprotein	Proapo-A1	gi|178775	5.45	28944	53%	503	20	1.1e-045	1.82
Retinol binding protein 4	RBP4	gi|1808832	5.76	23371	48%	276	14	2.9e-021	2.66
Down-Regulated Protein									
Leucine-rich alpha-2-glycoprotein precursor	LRG	gi|1641846	6.45	38382	40%	601	18	1.8e--055	0.21
Alpha-1B-glycoprotein – human	A1BG	gi|69990	5.65	52479	40%	824	24	9.1e-078	0.45
Clusterin	CLU	gi|2666585	5.60	16267	11%	39	4	29	0.60
PRO2044	PRO2044	gi|6650826	6.97	39984	45%	355	17	3.7e-029	0.63
Immunoreactive Protein									
Chain B, Human Complement Component 3	C3	gi|7810126	5.55	114238	20%	469	24	1.5e-040	

The fold change was obtained by dividing the average spot intensity of patients sample with the control. Therefore, the fold change measures the degree of changes in the protein of patient when compared to the control.

4.4.1.2 Immunogenic Protein Identification by Western Blotting

2-DE allows the identification of proteins in biological samples, however, numerous post-translational protein remain undetected. Therefore, to further improve the biomarker discovery in OSCC, the immunoproteomic approach was applied. The approach utilize 2-DE immunoblotting assay using OSCC patient and control sera. This was performed based on four conditions; (a) normal sera probed with normal sera; (s) normal sera probed with OSCC sera; (c) OSCC sera probed with normal sera; (d) OSCC sera probed with OSCC sera. Figure 4.13 illustrates the identification of OSCC protein spot using immunobloting technique based on the four categories.

Based on these immunobloting results, proapo-A1 was detected in condition (a) (Figure 4.13a) while HAP was seen immunoreactive in condition (b) (Figure 4.13b). As for condition (c) (Figure 4.13c), serum antigens CLU, C3, proapo-A1, and RBP4 were seen immunoreactive in healthy control sera. Lastly, C3, HAP, proapo-A1, and RBP4 were seen immunoreactive in condition (d) (Figure 4.13d).



Antigenic Proteins	Condition			
	(a)	(b)	(c)	(d)
1) Host specific proteins:				
CLU	-	-	/	-
HAP	-	/	-	/
C3	-	-	/	/
Proapo-A1	/	-	/	/
RBP4	-	-	/	/

/ Proteins of the patients or normal pooled sera recognized by the primary antibody
- Proteins of the patients or normal pooled sera not recognized by the primary antibody

Figure 4.13: 2-DE immunobloting.
Results from 2-DE immunoblots for (a) normal pooled sera probed with normal pooled sera, (b) normal pooled sera probed with OSCC pooled sera, (c) OSCC pooled sera probed with normal pooled sera, (d) OSCC pooled sera probed with OSCC pooled sera.

4.4.2 Label free LC-MS Relative Protein Quantitation

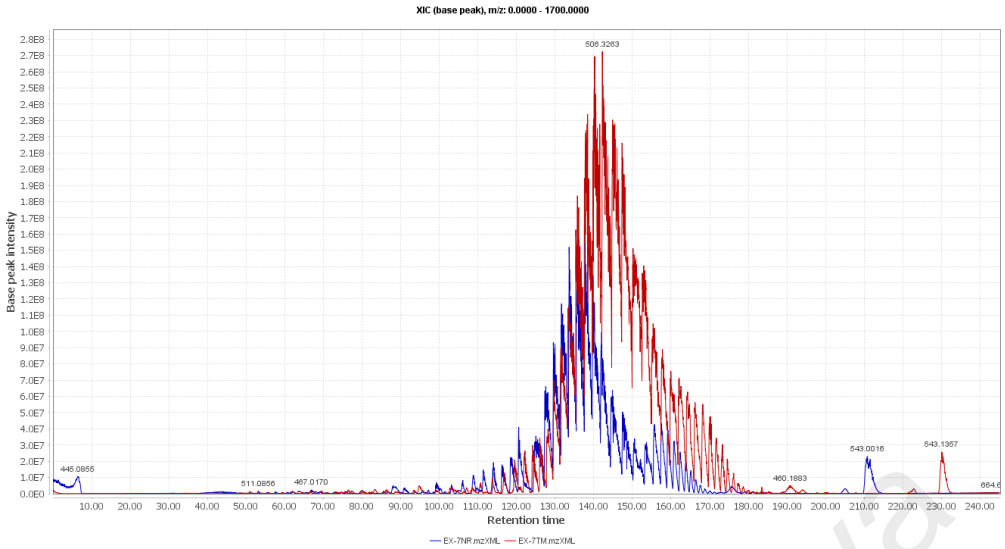
To identify additional potential biomarkers, label free LC-MS quantification was performed on OSCC (n=6) and matched non-malignant (normal) adjacent (n=6) tissues, that was previously used in exome sequencing. Proteins were extracted and measured using Bradford assay prior LC-MS. Protein concentration for both normal and OSCC samples were depicted in Appendix A.6.

This analysis was performed using LTQ XL mass spectrometer (Thermo Scientific, USA) on proteins extracted from OSCC and normal (adjacent) tissue. A data-dependent top 7 method was used, where a full MS scan from m/z 400 - 1500 was followed by MS/MS scan on the three most abundant ions. Raw data and MS peaks intensities were analysed using Proteome Discovery 1.3 (Thermo Scientific, USA) and SEQUEST algorithm against the most recent species-species database for human. This algorithm was downloaded from NCBI. Label-free quantification analysis was performed using Sieve 2.1 software with the frames thresholds set to 8000. A total of 6,339 peptides corresponding to 12,556 unique peptides were identified in these 6 pairs of OSCC samples (Table 4.11).

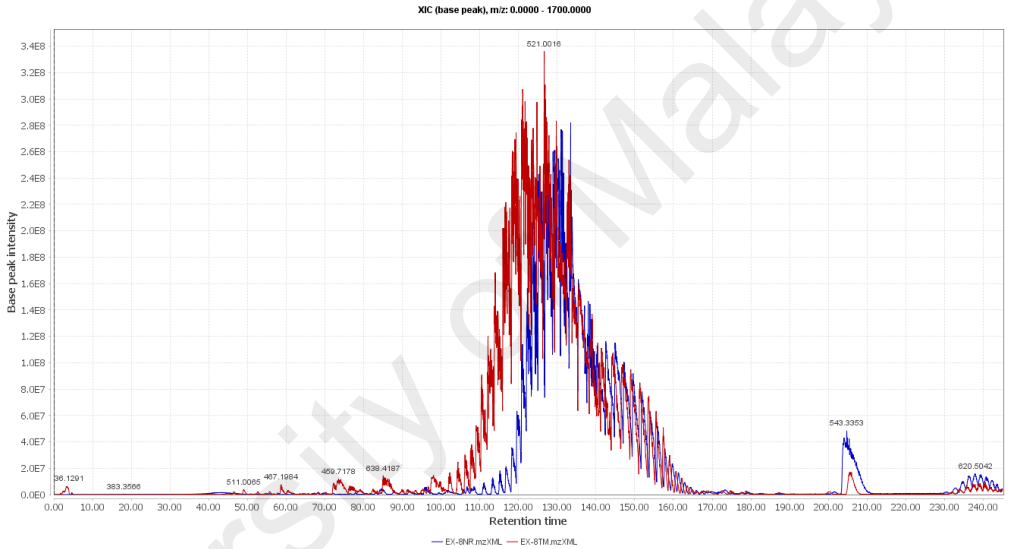
Table 4.11: Summary of protein identification in OSCC samples.

Sample	Protein	Total Unique Peptide	Total Peptide	PSM
EXp-7NR	303	833	425	495
EXp-7TM	4584	1379	700	811
EXp-8NR	2835	897	451	524
EXp-8TM	3058	969	482	575
EXp-9NR	2524	960	486	588
EXp-9TM	3029	892	448	529
EXp-10NR	3084	945	480	556
EXp-10TM	5063	1476	744	893
EXp-11NR	2713	868	431	560
EXp-11TM	4409	1454	740	939
EXp-12NR	1217	391	196	265
EXp-12TM	4169	1492	756	926
TOTAL	36988	12556	6339	7661

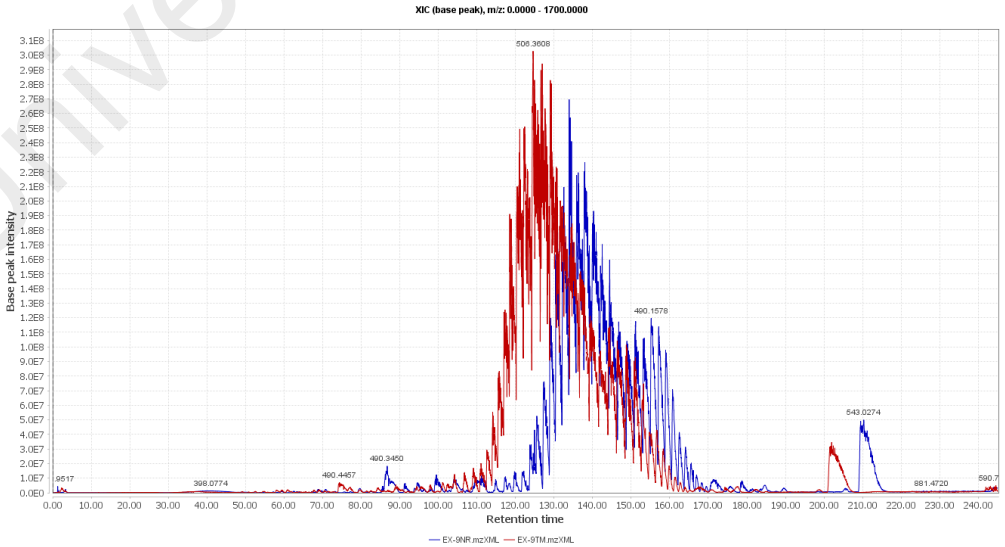
Protein changes between OSCC tumour and normal (adjacent) were compared in all 5 pairs of sample (EXp-7NR/TM, EXp-8NR/TM, EXp-9NR/TM, EXp-10NR/TM and EXp-11NR/TM). Samples EXp-12NR and EXp-12TM was excluded from the analysis due to defects in the ion chromatography (Figure 4.14).



EXp-7NR vs EXp-7TM

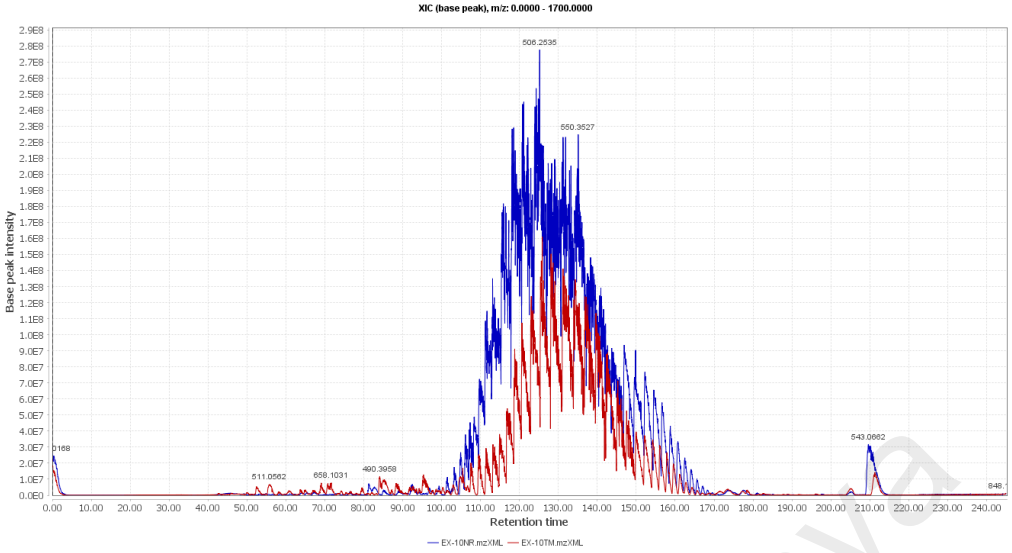


EXp-8NR vs EXp-8TM

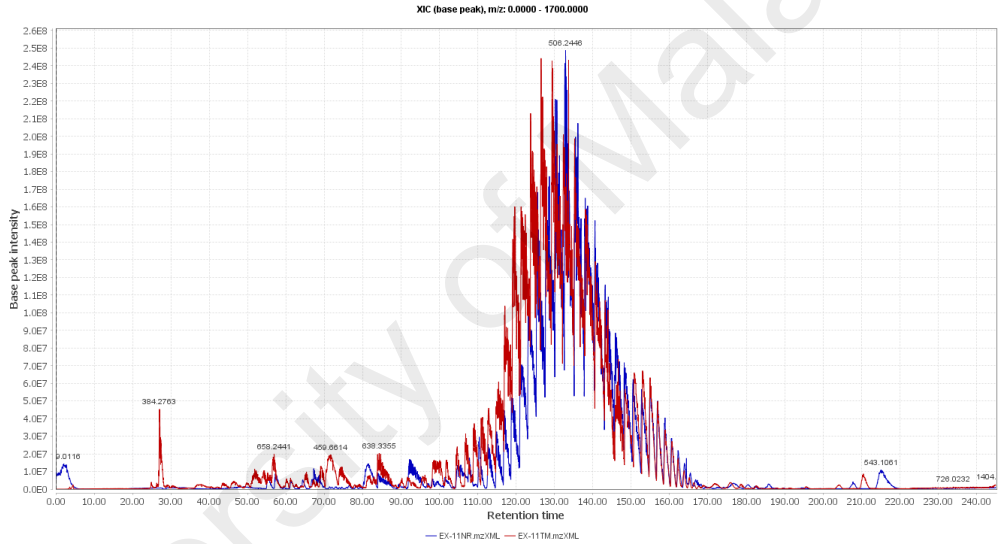


EXp-9NR vs EXp-9TM

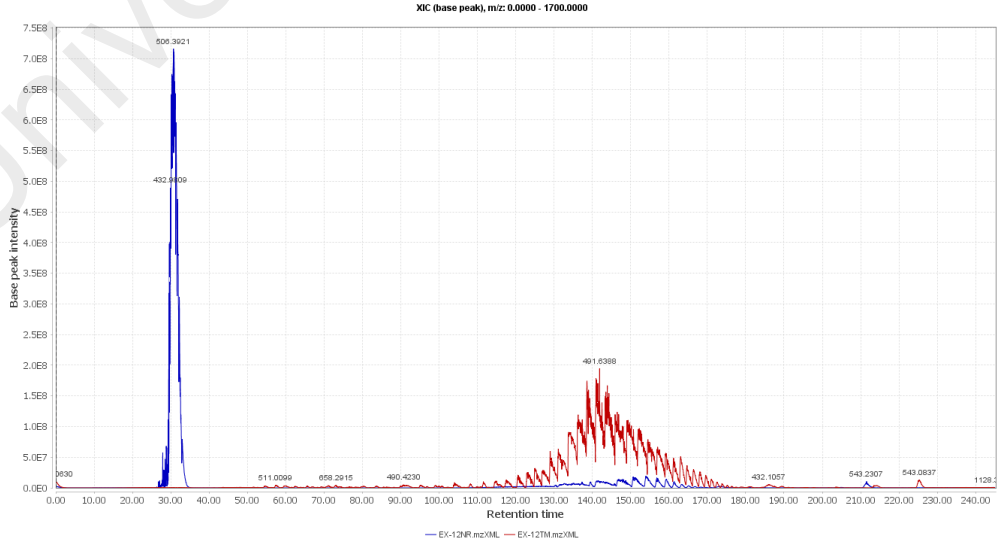
Figure 4.14: Ion chromatography of 6 pairs (tumour and adjacent normal) of OSCC samples generated analyzed using XCMS online and Mzmine 2.0.



EXp-10NR vs EXp-10TM



EXp-11NR vs EXp-11TM



EXp-12NR vs EXp-12TM

Figure 4.14, Continue

A total of 856 proteins were identified in OSCC samples when protein identification was compared between tumour and normal. Proteins that were not well described or unspecified in the NCBI and UniProt database were screened out and were not selected. The total proteins were then narrowed down to 271 well described and non-redundant proteins (≥95% confidence). Lastly, 19 significantly ($p \leq 0.05$) differentiated proteins were identified out of the 271 proteins as frequently observed proteins in OSCC samples and were further studied. The details of the proteins and the fold changes were described in Table 4.12.

Table 4.12: 19 proteins identified as differentially expressed between OSCC tumour and normal (adjacent) following label free LC-MS.

No	Protein Entry Name	Protein	Gene	Protein Name	AAs	MW [kDa]	Fold Change*
Up-regulated							
1	ACTB	P60709	ACTB	Actin, cytoplasmic 1/ Beta-Actin	375	41.7	3.21
2	ACTBM	Q9BYX7	POTEKP/ POTEI	Putative beta-actin-like protein 3	375	42.0	3.21
3	ACTC	P68032	ACTC1	Actin, alpha cardiac muscle 1	377	42.0	2.49
4	ACTG	P63261	ACTG1	Actin, cytoplasmic 2	375	41.8	3.21
5	ACTS	P68133	ACTA1	Actin, alpha skeletal muscle	377	42.1	2.49
6	HBB	P68871	HBB	Hemoglobin subunit beta	147	16.0	34.75
7	POTEE	Q6S8J3	POTEE	POTE ankyrin domain family member E	1075	121.4	3.21
8	POTEF	A5A3E0,	POTEF	POTE ankyrin domain family member F	1075	121.4	3.21
Down-regulated							
1	ALBU	P02768	ALB	Serum albumin	609	69.4	1.25
2	CRNS1	A5YM72	CARNS1	Carnosine synthase 1	827	88.5	0.98
3	E7ENN3	E7ENN3	SYNE1	Nesprin-1	8392	964.8	1.16
4	E7ERU0	E7ERU0	DST	Dystonin	5375	615.7	0.73
5	EF1DL	Q658K8	EEF1DP3	Putative elongation factor 1-delta-like protein	133	14.1	1.22
6	F8WAH6	F8WAH6	ELN	Elastin	786	68.4	1.22
7	FOCAD	Q5VW36	FOCAD	Focadhesin	1801	200.0	1.37
8	HBA	P69905	HBA1	Hemoglobin subunit alpha	142	15.3	0.45
9	SCND3	Q6R2W3	ZBED9/ SCAND3	SCAN domain-containing protein 3	1325	151.7	0.76
10	TITIN	Q8WZ42-8	TTN	Titin	34475	3829.8	0.68
11	TTL11	F8W6M1	TTLL11	Tubulin polyglutamylase TTLL11	800	87.6	0.69

*Fold change of ≥ 2 or < 2 in OSCC compared with adjacent normal was considered and reported as up- and down-regulation, respectively.

4.5 Functional Enrichment and Pathway Analysis of Potential Biomarkers

A total of 77 potential OSCC biomarkers (genes and proteins) were discovered through both genomic and proteomic platform applied in this study. From this total, 50 were identified using exome sequencing, 8 were identified using 2-DE and 19 were identified using label free LC-MS. In addition, 2 were identified by both exome sequencing and label free LC-MS platform and 1 was identified by both 2-DE and label free LC-MS (Figure 4.15).

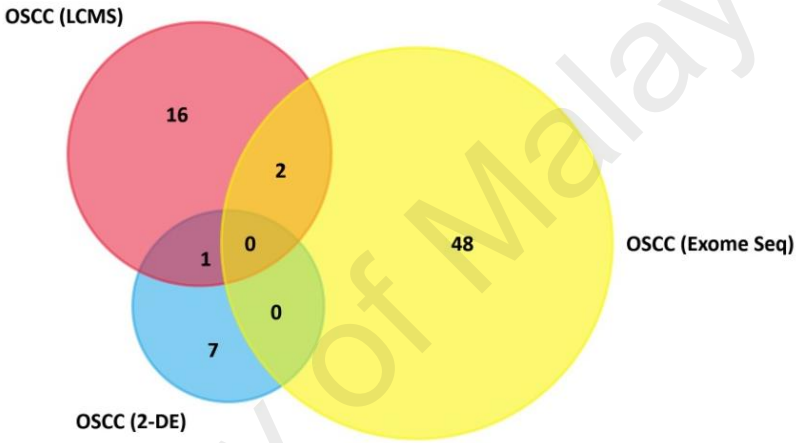


Figure 4.15: Total number of potential biomarkers identified through exome sequencing, 2-DE and label free LC-MS.

In addition to genomic-proteomic analysis, functional enrichment and pathway analysis was performed on the identified potential biomarkers. This was to further annotate and understand the biological function and interaction of these genes and proteins in OSCC. The 77 potential biomarkers were analysed using the web-based annotation tool; ConsensusPathDB-human (http://cpdb.molgen.mpg.de/CPDB) and DAVID v6.8 (http://david.abcc.ncifcrf.gov/). Based on the functional enrichment analysis, the biomarkers were classified based on the gene ontology term; a) biological process, b) cellular component, c) molecular function. Figure 4.16 depicts; a) the top gene ontology of the potential biomarkers and b) p-value of each gene ontology.

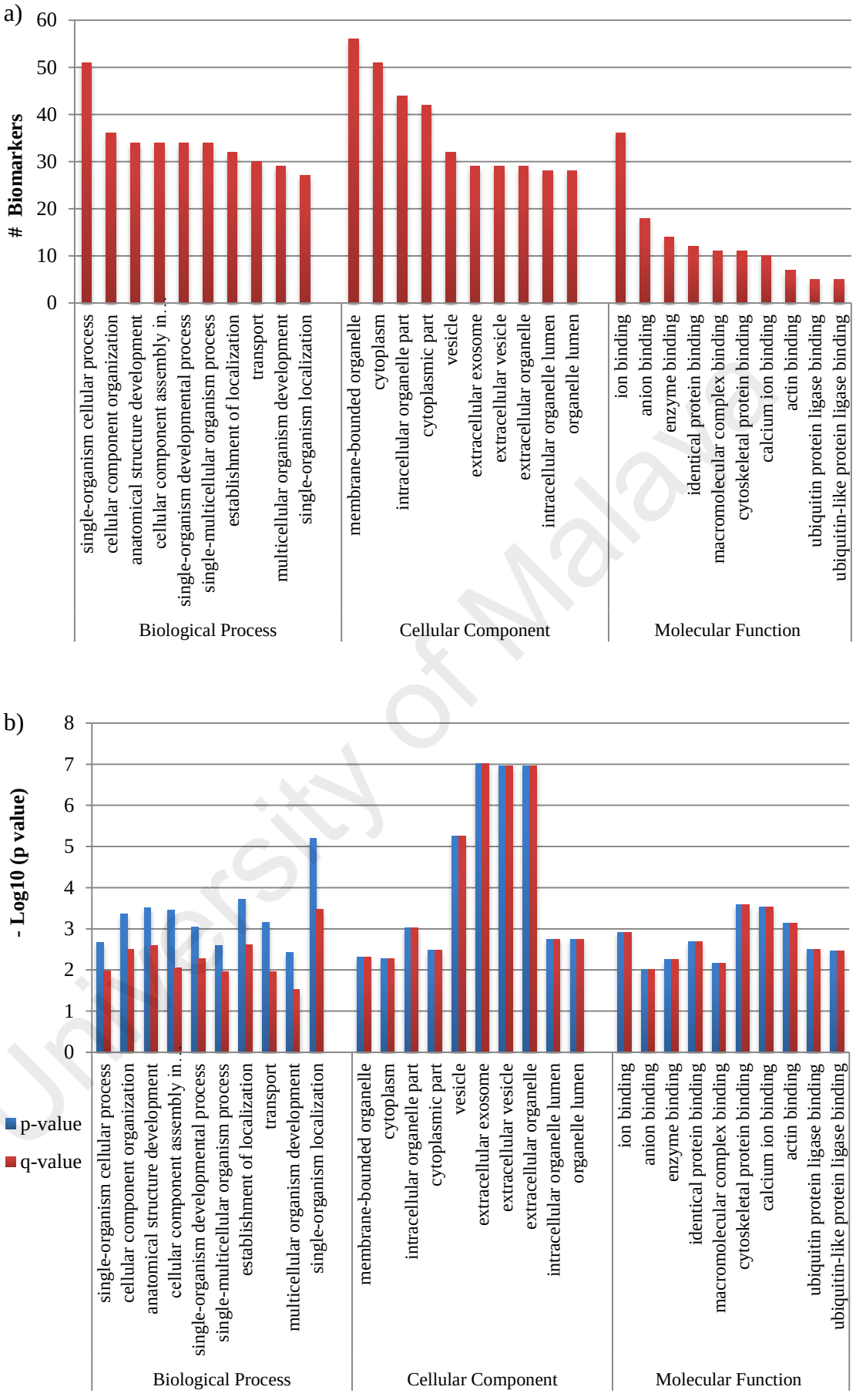


Figure 4.16: Gene ontology of the potential OSCC biomarkers.

Table 4.13 lists the significant over-represented gene ontology term based on the identified potential biomarker. In addition to the functional enrichment analysis, protein interaction network was generated using STRING v10.1 (http://string-db.org/) database to identify potential binding partners for the identified potential biomarkers. STRING database is a curated knowledge database of known and predicted protein-protein interactions. In addition, protein interaction networks representing the top identified pathways were also generated. Figure 4.17 depicts the interaction network of the identified biomarker in molecular action view which predicts the mode of action of the protein-protein interaction.

To further identify top networks associated with the potential biomarkers, pathway analysis was applied. ConsensusPathDB and DAVID v6.8 which applied several databases such as Kyoto Encyclopedia of Genes and Genomes (KEGG), BioCarta and Reactome database were used to conduct pathway analysis. Figure 4.18a depicts the top network associated with the potential biomarkers. These include Platelet activation, signalling and aggregation pathway, Focal adhesion pathway, Oxytocin signaling pathway, Apoptosis pathway, Thyroid hormone signalling pathway and Folate metabolism pathway. In addition, protein interaction networks of the top identified pathways were generated to identify interacting protein involved in the respective network Figure 4.18b.

Table 4.13: Functional annotation analysis of potential biomarkers using ConsensusPathDB and DAVID v6.8.

Gene Ontology Term	Genes
Biological Process	
single-organism cellular process	ACTA1; ACTG1; ACTC1; HP; HBB; SLC4A1; SYNE1; PHLDA1; APOA1; CARNS1; COL11A1; RERE; RYR1; RYR2; PCLO; RELN; RPGR; C3; TTLL11; SIPA1L3; FMN2; USH1G; VPS13C; A1BG; TTN; SPON1; TARS2; ELN; ARID2; DST; SPAG5; PLEC; PCDHGC5; HBA1; CLU; LRG1; SEC63; TYRO3; TP53; TRIOBP; DNAH8; ALB; NOTCH1; CASP8; FAT1; SARDH; RNF213; PTCHD3; RBP4; ACTB; PICALM
cellular component organization	ACTA1; ACTG1; ACTC1; HBB; SYNE1; APOA1; COL11A1; RERE; PCLO; TTN; RPGR; C3; SIPA1L3; FMN2; USH1G; VPS13C; RELN; TARS2; ELN; ARID2; DST; SPAG5; PLEC; PCDHGC5; HBA1; CLU; SEC63; TYRO3; TP53; TRIOBP; NOTCH1; CASP8; FAT1; RNF213; ACTB; PICALM
anatomical structure development	ACTA1; ACTG1; ACTC1; SEC63; SYNE1; PHLDA1; APOA1; ZNF74; COL11A1; RERE; RYR1; RYR2; PCLO; TTN; C3; SIPA1L3; USH1G; RELN; ELN; RBP4; CLU; LRG1; FMN2; POTEE; TYRO3; TP53; TRIOBP; NOTCH1; CASP8; PTPN14; FAT1; PTCHD3; ACTB; RNF213
cellular component assembly in morphogenesis	ACTC1; ACTA1; ACTG1; TTN
single-organism developmental process	ACTA1; ACTG1; ACTC1; SEC63; SYNE1; PHLDA1; APOA1; ZNF74; COL11A1; RERE; RYR1; RYR2; PCLO; TTN; C3; SIPA1L3; USH1G; RELN; ELN; RBP4; CLU; LRG1; FMN2; POTEE; TYRO3; TP53; TRIOBP; NOTCH1; CASP8; PTPN14; RNF213; PTCHD3; ACTB; PICALM
single-multicellular organism process	ACTA1; ACTG1; ACTC1; HBB; PHLDA1; APOA1; ZNF74; COL11A1; RERE; RYR1; RYR2; PCLO; RELN; POTEI; C3; POTEF; SIPA1L3; FMN2; USH1G; TTN; ELN; RBP4; CLU; LRG1; SEC63; POTEE; TYRO3; TP53; ALB; NOTCH1; CASP8; PTPN14; ACTB; RNF213
establishment of localization	ACTG1; RPGR; HP; SLC35F1; HBB; SLC4A1; SYNE1; APOA1; RERE; RYR1; RYR2; LRP1B; PCLO; TTN; C3; FMN2; VPS13C; RELN; SPAG5; RBP4; HBA1; CLU; SEC63; TYRO3; TP53; ALB; NOTCH1; CASP8; PTPN14; ACTB; A1BG; PICALM
transport	ACTG1; RPGR; HP; SLC35F1; HBB; SLC4A1; APOA1; RERE; RYR1; RYR2; LRP1B; PCLO; TTN; C3; FMN2; VPS13C; RELN; RBP4; HBA1; CLU; SEC63; TYRO3; TP53; ALB; NOTCH1; CASP8; PTPN14; ACTB; A1BG; PICALM
multicellular organism development	ACTA1; COL11A1; ACTC1; SEC63; PHLDA1; APOA1; ZNF74; RERE; RYR1; RYR2; PCLO; TTN; C3; SIPA1L3; USH1G; RELN; ELN; RBP4; CLU; LRG1; FMN2; POTEE; TYRO3; TP53; NOTCH1; CASP8; PTPN14; ACTB; RNF213
single-organism localization	ACTG1; RPGR; HBB; SYNE1; APOA1; RERE; RYR1; RYR2; PCLO; TTN; C3; FMN2; VPS13C; RELN; SPAG5; RBP4; HBA1; CLU; SEC63; TYRO3; TP53; ALB; NOTCH1; CASP8; ACTB; A1BG; PICALM
Molecular Function	
ion binding	ACTA1; ACTG1; CARNS1; ACTC1; HBB; APOA1; ZNF74; COL11A1; RERE; RYR1; RYR2; SVEP1; PSD2; LRP1B; PCLO; TTN; TTLL11; RELN; SPON1; TARS2; ARID2; DST; PCDHGC5; HBA1; ZNF518A; TYRO3; TP53; TRANK1; DNAH8; ALB; NOTCH1; FAT1; SARDH; RNF213; ACTB; PICALM

Table 4.13, Continued

Gene Ontology Term	Genes
anion binding	ACTA1; TP53; ACTG1; CARNS1; ACTC1; DNAH8; ALB; PSD2; SARDH; PCLO; TTN; TARS2; PICALM; ACTB; TTLL11; APOA1; TYRO3; TRANK1
enzyme binding	TP53; ACTG1; TRIOBP; RYR1; RYR2; ALB; NOTCH1; CASP8; PTPN14; SLC4A1; TTN; CLU; ACTB; APOA1
identical protein binding	TP53; ACTG1; DST; RYR2; ALB; CASP8; USH1G; SLC4A1; TTN; SYNE1; ACTB; APOA1
macromolecular complex binding	RERE; TP53; TRIOBP; DST; SVEP1; NOTCH1; CASP8; TTN; SYNE1; ACTB; APOA1
cytoskeletal protein binding	ACTA1; TRIOBP; ACTC1; DST; FMN2; PLEC; TTN; SYNE1; SLC4A1; ACTB; USH1G
calcium ion binding	LRP1B; RYR1; DST; RYR2; SVEP1; NOTCH1; FAT1; PCLO; TTN; PCDHGC5
actin binding	TRIOBP; DST; FMN2; SLC4A1; TTN; SYNE1; PLEC
ubiquitin protein ligase binding	TRIOBP; CLU; TP53; ACTG1; CASP8
ubiquitin-like protein ligase binding	TRIOBP; CLU; TP53; ACTG1; CASP8
Cellular Component	
membrane-bounded organelle	ACTA1; ACTG1; ACTC1; HP; HBB; LRG1; SLC4A1; SYNE1; RARRES1; PHLDA1; APOA1; ZNF74; COL11A1; RERE; STAT4; RYR1; ARID2; RYR2; USP40; SPHKAP; PCLO; TTN; POTEI; RPGR; C3; POTEF; POTEE; FMN2; VPS13C; A1BG; SPON1; TARS2; ELN; ZBED9; DST; SPAG5; PLEC; PCDHGC5; SNAPC4; HBA1; CLU; ZNF518A; SEC63; TYRO3; TP53; TRIOBP; ALB; NOTCH1; CASP8; PTPN14; FAT1; SARDH; RNF213; RBP4; ACTB; PICALM
cytoplasm	ACTA1; ACTG1; ACTC1; HP; HBB; SLC4A1; SYNE1; TTLL11; PHLDA1; APOA1; CARNS1; COL11A1; STAT4; RYR1; RYR2; SVEP1; USP40; SPHKAP; PCLO; RELN; RPGR; POTEF; FMN2; USH1G; VPS13C; A1BG; TTN; SPON1; TARS2; ELN; ZBED9; DST; SPAG5; PLEC; HBA1; CLU; SEC63; TYRO3; TP53; TRIOBP; DNAH8; ALB; NOTCH1; CASP8; PTPN14; FAT1; SARDH; RNF213; RBP4; ACTB; PICALM
intracellular organelle part	ACTA1; ACTG1; ACTC1; HP; HBB; SYNE1; PHLDA1; APOA1; ZNF74; COL11A1; RERE; STAT4; RYR1; ARID2; RYR2; TTLL11; TTN; RPGR; SIPA1L3; FMN2; USH1G; VPS13C; A1BG; SPON1; TARS2; ZBED9; DST; SPAG5; SNAPC4; HBA1; CLU; SEC63; TYRO3; TP53; TRIOBP; DNAH8; ALB; NOTCH1; CASP8; PTPN14; SARDH; RNF213; ACTB; PICALM
cytoplasmic part	ACTA1; ACTG1; ACTC1; HP; HBB; SLC4A1; SYNE1; PHLDA1; APOA1; CARNS1; COL11A1; RYR1; RYR2; SPHKAP; PCLO; TTN; RPGR; POTEF; FMN2; USH1G; VPS13C; A1BG; SPON1; TARS2; ELN; DST; SPAG5; PLEC; HBA1; CLU; SEC63; TYRO3; TP53; ALB; NOTCH1; CASP8; FAT1; SARDH; RNF213; RBP4; ACTB; PICALM
vesicle	ACTA1; ACTG1; ACTC1; HP; HBB; SLC4A1; APOA1; RYR1; RYR2; PCLO; TTN; POTEI; C3; POTEF; POTEE; VPS13C; A1BG; DST; PLEC; PCDHGC5; HBA1; CLU; LRG1; FMN2; ALB; NOTCH1; RARRES1; FAT1; SPHKAP; RBP4; ACTB; PICALM
extracellular exosome	ACTA1; ACTG1; ACTC1; HP; HBB; SLC4A1; APOA1; RYR1; RYR2; PCLO; TTN; POTEI; C3; POTEF; POTEE; VPS13C; A1BG; DST; PLEC; PCDHGC5; HBA1; CLU; LRG1; ALB; RARRES1; FAT1; SPHKAP; RBP4; ACTB

Table 4.13, Continued

Gene Ontology Term	Genes
extracellular vesicle	ACTA1; ACTG1; ACTC1; HP; HBB; SLC4A1; APOA1; RYR1; RYR2; PCLO; TTN; POTEI; C3; POTEF; POTEE; VPS13C; A1BG; DST; PLEC; PCDHGC5; HBA1; CLU; LRG1; ALB; RARRES1; FAT1; SPHKAP; RBP4; ACTB
extracellular organelle	ACTA1; ACTG1; ACTC1; HP; HBB; SLC4A1; APOA1; RYR1; RYR2; PCLO; TTN; POTEI; C3; POTEF; POTEE; VPS13C; A1BG; DST; PLEC; PCDHGC5; HBA1; CLU; LRG1; ALB; RARRES1; FAT1; SPHKAP; RBP4; ACTB
intracellular organelle lumen	COL11A1; HP; HBB; CLU; PHLDA1; APOA1; ZNF74; RERE; STAT4; TTN; ARID2; A1BG; SPON1; TARS2; ZBED9; SPAG5; SNAPC4; HBA1; SYNE1; FMN2; TP53; ALB; NOTCH1; CASP8; PTPN14; SARDH; ACTB; RNF213

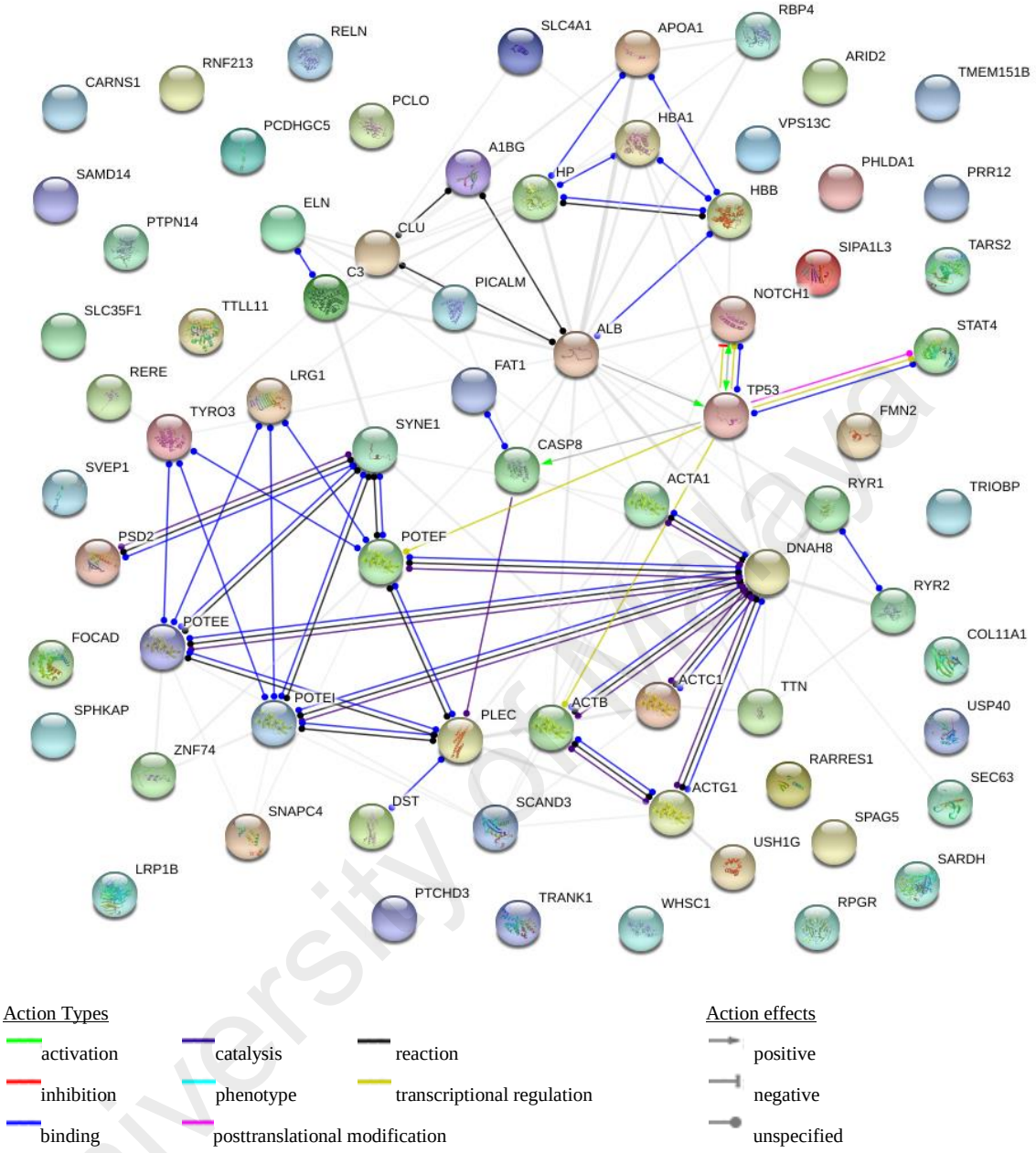
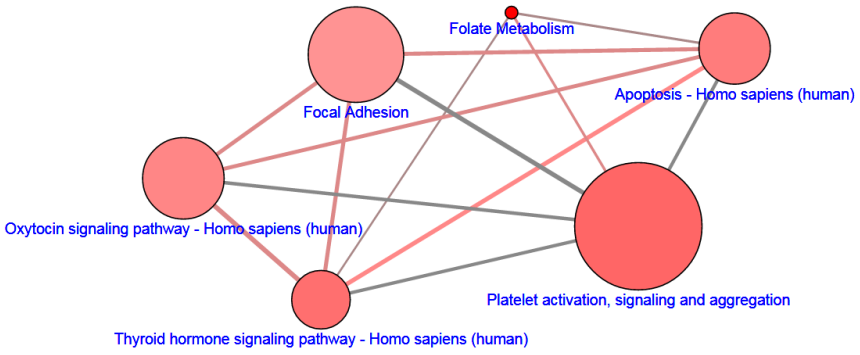
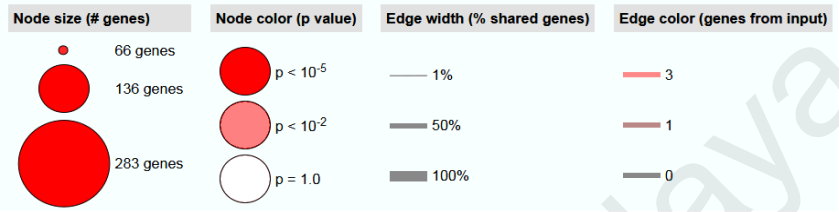


Figure 4.17: Interaction network of identified potential biomarkers and neighbouring protein using STRING v9.1.

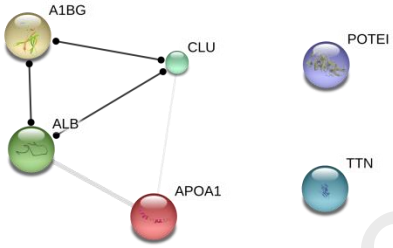
STRING database is a curated knowledge database of known and predicted protein-protein interaction. Most of the identified biomarkers have an established link with each other in the interaction network. Several

a)

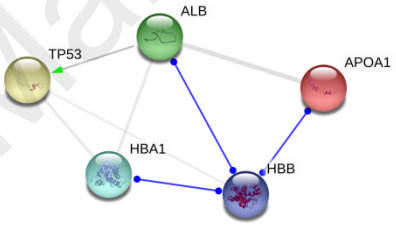




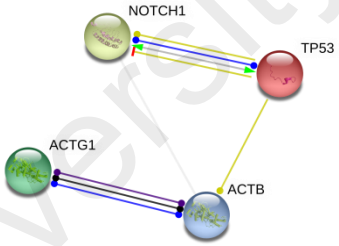
b)



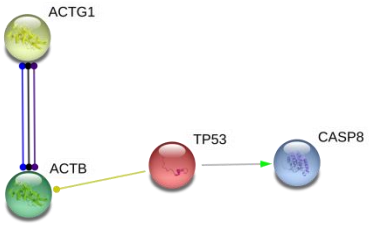
Platelet Activation, Signalling and Aggregation Pathway



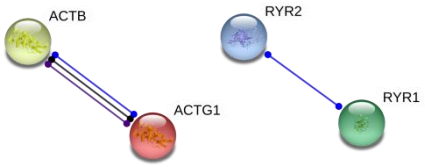
Folate Metabolism Pathway



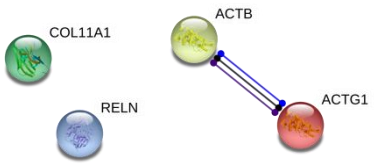
Thyroid Hormone Signalling Pathway



Apoptosis Pathway



Oxytocin Signalling Pathway



Focal Adhesion Pathway

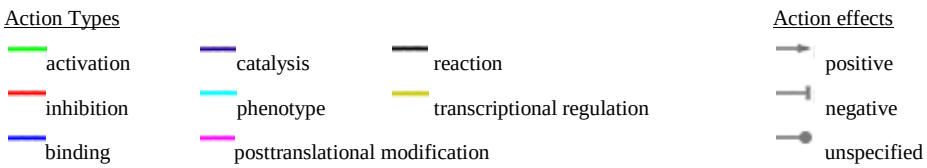


Figure 4.18: Top pathways associated with OSCC potential biomarkers.

CHAPTER 5: DISCUSSION

Oral squamous cell carcinoma (OSCC) remains to be one of the most common cancers worldwide. In 2012, there were about 300,000 new cases diagnosed with oral cancer, with almost two-thirds occurring in males according to the Global Burden of Cancer (GLOBACAN) (Ferlay *et al.*, 2015). Based on the predicted number of cases and deaths for 2020, oral cancer incidence and mortality rate is expected to increase especially in all South East Asian countries (Cheong, S. C. *et al.*, 2017). According to Omar, E. (2015), failure to detect OSCC in its early stage attributes as a barrier to improve the survival and quality of OSCC patients' life. It was noted that just by a few months of delay in diagnosis, the survival rate could decrease from 80% to 40% (Rao et al., 2013). However, cancer screening and early detection using biomarkers has been known to decrease the risk of morbidity and mortality in OSCC patients (Masthan *et al.*, 2012) and further improve patient care (Mäbert *et al.*, 2014). Therefore, identifying reliable biomarkers will improve the clinical management and care of these cancer patients. In this study, next generation technologies including exome sequencing (NGS) and 2D-electrophoresis coupled label free LC-MS were employed to identify reliable biomarkers that may assist in the early detection of OSCC. Before beginning the genomic and proteomic experimentation, demographic profiling was conducted to determine the inclusion and exclusion criteria of the study population.

5.1 Demographic Profile of Study Population

Oral cancer has been previously reported to have higher incidence and mortality rate in men than women (Petti *et al.*, 2010). This trend has been widely observed in several countries in the Asian region such as Japan, and Taiwan (Rao *et al.*, 2013). However, studies have shown that the predominance of OSCC was also observed in female populations compared to male in countries such as India and Mexico (Gaitán-

Cepeda *et al.*, 2011; Rao *et al.*, 2013). In India, oral cancer was recorded as one of the most prevalent cancers with OSCC attributing to one third of the total cancer burden (Rao *et al.*, 2013; Singh *et al.*, 2015). Similarly to this study, higher prevalence of OSCC was also observed among the female and the Indian community. Chang *et al.* (2011) and Zain, R. *et al.* (2001) had also reported that the incidence of oral cancer was higher in the Indian and Indigenous ethnics in Malaysia.

The habit of chewing betel quid has always been reported as the major risk factor of OSCC in the Asia region (Gupta, B. & Johnson, 2014; Neville, B.W *et al.*, 2002) . Betel quid has been classified by the International Agency for Research on Cancer (IARC) as a human carcinogen (Liao *et al.*, 2014). The predominant personal habits of betel quid chewing (Ghani *et al.*, 2011) among the Indian women (Lee, C. H. *et al.*, 2011) had led to the significant increase in number of female Indians diagnosed with OSCC when compared with other major populations in Malaysia (Guha *et al.*, 2014). Similarly, betel quid chewing was also seen in these findings as the most common risk factor of OSCC especially among the female Indian population. Alcohol drinking and tobacco smoking were ranked the second and the third most common risk factor for OSCC, respectively.

Although high incidence of oral cancer is generally associated with tobacco smoking, betel quid chewing and alcohol consumption, viruses are also known to play a role in OSCC development (Kumar *et al.*, 2016; Polz-Gruszka *et al.*, 2014). The general causative viruses for OSCC includes HPV and EBV (Jalouli *et al.*, 2010; Sand *et al.*, 2014). It had been suggested that past or long-term exposure to viruses may contribute to the onset of OSCC (Gupta, K. & Metgud, 2013). Therefore, the presences of both IgG and IgM antibodies against these viruses were determined in the study cohort using ELISA. Through IgG and IgM screening, the patients' past and current exposure status

against HPV and EBV were confirmed. The presence of IgM antibodies against virus directly shows acute or current exposure, whereas, IgG antibodies against virus indicate past exposure to the virus.

From the results, HPV16 IgG was seen as the strongest predictor for OSCC when compared to other risk factors. This result further suggests that long term exposure to HPV16 virus might develop the onset of OSCC. Past exposure to HPV16 has been previously reported to increase the risk of OSCC (Sand *et al.*, 2014). Based on the EBV results, both OSCC patients and normal control were found to be seropositive to EBV IgG antibody. Correspondingly, EBV was reported as the most common and widespread human virus with lifelong latent infection (Evans, 2013). Therefore, the association between EBV and OSCC was not well demonstrated in this study, even though this association has been widely reported previously (Sand *et al.*, 2014). Conversely, only a few samples were found to be positive for HPV16 IgM and EBV IgM respectively. The association between viral IgM and OSCC was not well-defined due to the fact that the onset of carcinogenesis by viral infection is influenced by the past infection or the latency of a virus (Sand *et al.*, 2014).

In general, the demographic results were in consistent with previous studies. These findings had confirmed that female Indian community that practices betel quid chewing are most likely to develop OSCC. The analysis also suggests that the past exposure to HPV16 infection could contribute to the onset of OSCC. These findings had improved the current understanding on the demographic profiles of this study cohort which was further used to characterize and improve the selection of the study cohort for the following experimentation in this study.

5.2 OSCC Biomarker Discovery

Biomarkers are biological molecules (i.e. DNA and proteins) which could be measured and evaluated to indicate the presence or progression of cancer or the response to treatment (Mäbert *et al.*, 2014). By understanding the molecular biology of gene and protein, researchers are able to identify unique biomarkers for the early detection of OSCC. Robust genomic and proteomic technologies along with the advanced bioinformatics tools enabled the simultaneous analysis of numerous biological molecules (Kulasingam *et al.*, 2008) for the discovery of reliable biomarkers. Technologies including NGS and LCMS had generated high throughput data that allowed the identification of distinguished signatures in cancer making them an ideal modality in biomarkers discovery.

5.2.1 Identification of Potential OSCC Biomarkers Using Exome Sequencing (NGS)

Exome sequencing was applied in this study to detect somatic mutations in the exonic region of the genome. The exonic region which contains exons plays an important function in the translation of protein (Ng *et al.*, 2010). Mutations in this region are responsible for 85% of diseases (Ku *et al.*, 2012). Therefore, the discovery of novel genetic region and pathways through exome sequencing (Chung *et al.*, 2009) in this study had further identified the potential molecular markers for early detection and diagnosis of OSCC.

Through exome sequencing on 10 pairs of gDNA samples from OSCC and their adjacent non-malignant tissue, a total of 1,755 novel mutations were identified in more than 50 genes. Of these 50 genes, 13 genes which had mutations in 3 or more pairs of samples were identified. These lists of genes were then filtered and searched through both Sanger COSMIC and OrCGDB cancer database. Finally, based on the number of

missense mutation and novel/known association with OSCC, candidate genes were selected. Ultimately, 4 somatically mutated genes; *CASP8, USP40, NOTCH1, and COL11A1* which had the highest missense mutation (SNVs) were selected as the candidate genes. This final list of genes represented genes that were previously associated with OSCC and novel genes with potential association with OSCC.

CASP8 is a member of the cysteine proteases. Over the years, several *CASP8* genetic variants have been reported to have influences in the risk of various cancers (Ma, X. *et al.*, 2011) and was defined as an important cancer susceptibility gene candidate (Easton *et al.*, 2008). In this study, 2 SNVs (missense mutation) were identified in the *CASP8* gene. These SNVs represent amino acid substitution p.L211R and p.D288G located at the Peptidase_C14 (caspase) domain. CASP8 which encodes Caspase-8 plays an important role in inflammation, cytokine processing and as apoptosis initiator (Eckhart *et al.*, 2008; Kruidering & Evan, 2000). Searches in the COSMIC database (http://cancer.sanger.ac.uk/cosmic/curation) indicate that both SNVs were found to be novel in OSCC. Both mutations were found to be deleterious based on the consensus classifier PredictSNP. Similarly to this study, missense mutation (non-synonymous) in *CASP8* has been previously reported in various cancers such as head and neck squamous cell carcinoma, hepatocellular carcinoma, gastric cancer and colorectal cancer (Ando *et al.*, 2013). *CASP8* has been also previously reported in OSCC (India Project Team of the International *et al.*, 2013; Perez-Sayans *et al.*, 2009). Alteration in *CASP8* has been suggested to reduce the capacity to initiate cell apoptosis thus promoting the tumour outgrowth (Olsson & Zhivotovsky, 2011). However, mutated *CASP8* has been defined as a tumour suppressor gene due to its ability to promote apoptosis with the activation of nuclear factor-kB (NF-kB) signalling (Ando *et al.*, 2013; Ghavami *et al.*, 2009; Olsson *et al.*, 2011).

NOTCH1 is a gene that plays an important role during the embryonic development and cellular patterning (Lee, S. H. *et al.*, 2007; Yoshida *et al.*, 2013). The activation of this NOTCH pathway has been commonly associated with many types of cancers. *NOTCH1* has been defined as potential oncogene in several cancers including T-cell acute lymphoblastic leukaemia, colon cancer, pancreatic cancer, ovarian cancer and non -small cell lung cancer (Hijioka *et al.*, 2010). However, newer studies have suggested that *NOCTH1* may act as a tumour suppressor in cancers such as pancreatic cancer, nonsmall-cell lung cancer, hepatocellular cancer and also in head and neck squamous cell carcinoma (Agrawal, N. *et al.*, 2011; Yoshida *et al.*, 2013). Similarly to this study, mutations in *NOTCH1* was also identified in gingivo-buccal oral squamous cell carcinoma by the India Project Team of the International Cancer Genome Consortium (2013) further supporting the findings in this study. In these findings, a total of 3 SNV (missense mutations) were identified in the *NOTCH1* gene. These SNVs represents amino acid substitution p.P2097L, p.I1631M and p.C398F located at the ankyrin repeat domain and human growth factor-like EGF (EGF-like domain). Akryrin repeats are involved in cellular functions such as cell-cycle regulations, ion transportation, transcriptional initiation and signal transduction (Chakrabarty & Parekh, 2014). EGF-like domain which are shed from the cell surface, mediate intercellular signalling (Wouters *et al.*, 2005). The mutation findings in both domains were in consistence with the previous studies (Izumchenko *et al.*, 2015). The SNV were also compared to the COSMIC database and both substitution p.P2097L and p.I1631M were found to be a novel alteration in OSCC. However, the substitution p.C398F (missense mutation) was previously reported in esophageal cancer by Agrawal, N *et al.* (2012) similarly to the findings of this study. All SNVs in the *NOTCH1* gene were also defined as deleterious to human.

Collagen 11A1 (*COL11A1*), also known as *COL11A1* is a gene that codes the α1 chain of procollagen and mature collagen type X1 (Galván *et al.*, 2014). It is a major component of the interstitial extracellular matrix (ECM) which is important in the biological processes such as proliferation, migration, differentiation, apoptosis, skeletal development and carcinogenesis (Kleinert *et al.*, 2015; Wu, Y. *et al.*, 2014). In cancer, *COL11A1* expression was associated with the fibroblast-like stromal phenotype; however, the exact mechanism of *COL11A1* carcinogenesis was not well understood (Galván *et al.*, 2014). Studies have suggested that *COL11A1* mRNA was elevated in cancers of oesophagus, breast, lung, stomach, pancreas, colon and ovary (Jia *et al.*, 2016; Raglow & Thomas, 2015). Although several studies have reported the identification of *COL11A1* mutation in head and neck cancer, the exact association of *COL11A1* with OSCC cancer has not been fully understood (Liu *et al.*, 2008). In this study, a total of 4 SNVs (missense mutations) were identified in the *COL11A1* gene. These SNVs represents amino acid substitution p.R960C, p.G1118E, p.P1149T and p.G338D located in the collagen domain (collagen triple helix repeat). Collagen is extracellular structural protein with 20 copies of G-X-Y repeat forming triple helix and involved in the formations of connective tissue structure (Kadler *et al.*, 2007; Orgel *et al.*, 2011). All 4 SNVs were not recorded in the COSMIC database suggesting that the identified SNVs can be characterized as novel findings. All mutations were also found to be deleterious based on the consensus classifier PredictSNP further associating the SNV with oral cancer. However, further studies are needed to fully understand the effect of the alteration to OSCC patients.

USP40 is a gene that encodes the Ubiquitin carboxyl-terminal hydrolase 40 enzymes. *USP40* is involved in both ubiquitin-proteasome system (UPS) and in the cellular clearance of abnormal proteins as causal factor pathway for Parkinson Disease (Wu, Y.-R. *et al.*, 2010). Previously, researchers have had identified mutation in the

USP40 gene (c.1921G>T) when sequencing tumour tissue of the head and neck squamous cell carcinoma patient (Stransky *et al.,* 2011). However, to date little is known on the association of *USP40* with OSCC and the pathway involved. A total of 3 SNVs (missense mutations) were identified in the *USP40* gene. These SNVs represents amino acid substitution p.F552L, p.P940S and p.H327Y located at the Ubiquitin carboxyl-terminal hydrolase (UCH) domain. UCH is a deubiquitylating enzyme which impacts chromatin remodelling, DNA repair, cell-cycle control and several signalling pathways associated with cancer (Fraile *et al.*, 2012). All of the *USP40* SNVs were not recorded in the COSMIC database further suggesting that the identified SNVs could be characterized as a novel finding. All mutations were also found to be deleterious further associating the SNV with oral cancer. However, further studies are needed to fully understand the association of the mutation with OSCC.

The identified candidate SNVs were further analysed in a larger cohort using SNPtype Assay for SNP Genotyping to verify and to determine the allele frequency (frequency in the population) of the mutations in the candidate gene. The array validated the presence of the 9 SNVs in the respective samples. The remaining SNVs were unable to be verified, due to the unavoidable design and assay errors. Among all the identified SNVs, mutation c.C1656G (USP40_3a) was the only alteration found in a cluster represented by the validation cohort. In the NCBI dbSNP database (https://www.ncbi.nlm.nih.gov/snp) under the GRCh37.p13 assembly, there were no records on the identified *USP40* SNV c.C1656G. However, another SNV (rs771349240) at the similar location of 234,436,155 was reported in the dbSNP database. This SNV was reported as a synonymous variant that does not affect the protein sequence. The absence of this SNV in the database and the prediction of it as deleterious, further suggest that this alteration may represent as a novel finding in

OSCC. However, studies using larger OSCC and normal cohorts are required to verify these findings and to rule out the alteration as possible common polymorphism.

The mutation spectrum of OSCC obtained through exome sequencing was further compared with the patients demographic and risk factors (Figure 4.12). When compared the risk factors with the mutation spectrum, higher mutation burden was observed in patient (EX-2) with no habits but has viral infection. This was in agreement with previous reports where, it was suggested that patients that do not have risk habits could have a higher risk of developing OSCC with the influence of past viral (HPV) infection (Vargas-Ferreira *et al.,* 2012). In addition, mutation in *CASP8* gene was seen in all patients with betel quid chewing habits further suggesting that this mutation is associated with betel quid chewing. This was further supported by the previous report from the India Project Team of the International *et al.* (2013). The association of the remaining genes with the demographic profile of the patients remain unclear.

5.2.2 Identification of Potential OSCC Biomarkers Using the Proteomic Approach

Apart from the genomic approach applied in this study, proteomic analytical platform with the combination of two-dimensional gel electrophoresis (2-DE), mass spectrometry (MS), immunoproteomic and label-free LCMS were used in this study to identify biomarkers in OSCC. This paralleled analysis allows us to identify multiple potential biomarkers associated with OSCC. Moreover, this approach had successfully characterized proteins that are involved in the cancer progression.

5.2.2.1 Differently Expressed Proteins in OSCC Patients Identified Using 2-DE Technique.

The 2-DE and MS analysis in this study identified seven host specific proteins: leucine-rich α2-glycoprotein (LRG), alpha-1-B-glycoprotein (A1BG), clusterin (CLU),

PRO2044, haptoglobin (HP), proapolipoprotein A1 (proapo-A1) and retinol-binding protein 4 precursor (RBP4). Based on the analysis, proteins LRG, A1BG, CLU and PRO2044 were down-regulated while proteins HP, proapo-A1 and RBP4 were up-regulated. All of these identified proteins have been previously studied irrespective of the disease type. However, in contrast to some reports in the literature, there were few differences observed in the protein expressions.

LRG is a protein that has been associated with several malignancies including pancreatic (Kakisaka *et al.*, 2007), liver (Kawakami *et al.*, 2005) and lung (Okano *et al.*, 2006; Pang *et al.*, 2010) cancer. Several studies have reported that LRG was up-regulated in both lung and pancreatic cancer (Pang *et al.*, 2010). LRG was also observed to be up-regulated in esophageal squamous cell carcinoma (ESCC) (Zhao, Jia *et al.*, 2015). However, this was in contrast with the findings in this study, of which LRG was seen down-regulated in OSCC. Thus, LRG has a distinctive pattern in OSCC making it a potential biomarker. Nevertheless, LRG has been reported to be down-regulated in patients with inflammatory arthritis although it is produced during the inflammatory response (Weivoda *et al.*, 2008).

A1BG is a plasma glycoprotein from the immunoglobulin superfamily with unknown function (Song *et al.*, 2013). Up-regulation of A1BG was previously reported in malignancies, such as bladder cancer (Kreunin *et al.*, 2007), uterine cervix squamous cell carcinoma and non-small cell lung cancer (Jeong *et al.*, 2008). Similarly, in pancreatic ductal adenocarcinoma, A1BG was seen over expressed in patient's pancreatic juice (Tian *et al.*, 2008). However, based on the 2-DE and MS analysis A1BG was observed down-regulated in OSCC making it a unique marker and a potential biomarker for OSCC.

CLU is a protein highly associated with apoptosis and clearance of cellular debris (Jones & Jomary, 2002). This protein has been found to be expressed in numerous malignancies. The types of malignancy associated with CLU includes cancer in bladder (Stejskal & Fiala, 2006), breast (Redondo *et al.*, 2010; Rizzi & Bettuzzi, 2010), colorectal (Rodríguez-Piñeiro *et al.*, 2006), ovarian (Chen, Y. *et al.*, 2008), pancreatic and prostate (Koltai, 2014; Zellweger *et al.*, 2002). In this study, CLU was found to be down-regulated in OSCC. Similarly, down-regulations of CLU has been previously reported in cancers such as prostate (Scaltriti *et al.*, 2004), esophageal squamous cell carcinoma (Zhang, L. Y. *et al.*, 2003) and neuroblastoma (Santilli *et al.*, 2003). However, CLU was also found to be up-regulated in cancers such as pancreatic cancer (Chen, Q. *et al.*, 2011). This discrepancy may have been attributed by the differently expressed CLU isoforms which could exhibit distinct functions in cancer (Chen, Y. *et al.*, 2008; Wei, L. *et al.*, 2009). Down-regulation of CLU has been reported to be associated with the disease progression (Wu, J. *et al.*, 2013) depending on the type of cancer (Lourda *et al.*, 2007; Redondo *et al.*, 2010). However, it has been reported that CLU function may have been isoform dependent (Scaltriti *et al.*, 2004). Therefore, studying the differential expression of CLU isoforms in OSCC may contribute to a better understanding of this protein in OSCC.

The last and the most down-regulated protein identified in OSCC through the 2-DE/MS analysis was PRO2044. This protein is a C-terminal fragment of the protein albumin (ALB). Similar to this study, Kawakami *et al.* (2005) reported that PRO2044 was found to be down-regulated in patients with hepatocellular carcinoma. However, this was observed following a curative radiofrequency ablation. Furthermore, PRO2044 was found to increase in the cerebrospinal fluid of patients with Gullain-Barrè syndrome. Therefore, PRO2044 has the potential to be used as a biomarker for early detection of OSCC.

Haptoglobin (HP) is a protein that binds free plasma haemoglobin post-haemolysis (Levy *et al.*, 2010). The correlation of HP expression and cancer has been previously reported (Ahmed *et al.*, 2004). Similarly to this study, HP was reported as an up-regulated protein in OSCC (Lai *et al.*, 2010) further supporting the possibility of this protein to function as a biomarker for OSCC. High level of HP may indicate acute phase response of OSCC (Chen, Y. *et al.*, 2014), since HP is primarily produced by hepatocyte (Ahmed *et al.*, 2004). A study by (Ye *et al.*, 2003) suggests that HP may act as an angiogenic agent that contributes to endothelial cell differentiation and growth. Moreover, the involvement of HP in cell migrations further supports the role of HP in cancer (Zhao *et al.*, 2010).

The remaining up-regulated proteins identified in this study were proapo-A1 and RBP4. Proapo-A1protein represents a chain from Apolipoprotein A1 protein. It has been reported that reduced activity of proapo-A1 cleaving enzyme or high turnover of Apo-A1 may increase the level of proapo-A1 (Harn *et al.*, 2010; Huang, H. L. *et al.*, 2006). Up-regulations of proapo-A1 in this study were seen consistent with previous findings involving various malignancies. This includes breast (Huang, H. L. *et al.*, 2006), pancreatic (Mikuriya *et al.*, 2007), colorectal (Yu *et al.*, 2004), non-small cell lung (Huang, L.-J. *et al.*, 2006) and hepatocellular (Wang, H. Y., 2007) cancer. Similarly to proapo-A1, the RBP4 protein was also up-regulated in this study. This was found to be in consistent with previous studies on other cancers such as esophageal squamous cell carcinoma (Tsunoda *et al.*, 2009), ovarian (Cheng *et al.*, 2014) and pancreatic (Abulaizi *et al.*, 2011) cancers. It is suggested that high expression of RBP4 in cancer cells may have been in response to phosphatidylinositol-3 kinase (PI3K) activity inhibition (Farias *et al.*, 2005; Kuppumbatti *et al.*, 2001). In addition, it has been reported that retinoid depletion, which is common in cancer patients could influence RBP4 expression.

5.2.2.2 Identification of Immunogenic Protein by Western Blotting

To extend the proteomic analysis, 2-DE immunoblotting using patient and control sera were performed. The immunoblotting technique was performed under the following four conditions; (a) normal sera probed with normal sera; (b) normal sera probed with OSCC sera; (c) OSCC sera probed with normal sera; (d) OSCC sera probed with OSCC sera. Immunobloting techniques were applied in this study, due to the nature of the antibody as a natural defence mechanism against foreign antigens including cancer cells (Brandlein *et al.*, 2003). It has been reported that autoantibodies against specific cancer antigens were previously identified in several malignancies including cancer of the head and neck (Lin, H.-S. *et al.*, 2007). The specificity and sensitivity of the antibody response to antigens makes immunoblotting technique an ideal method to identify potential cancer biomarkers.

Condition (a) involves probing of normal sera with normal sera while condition (b) involves probing normal sera with OSCC sera. Normal sera were probed against normal and OSCC sera to verify that reactions were restricted to OSCC. Based on this analysis, proapo-A1 and HP were detected on both conditions respectively. Apart from its role in inflammatory response and as a scavenger, HP is known to be associated with immune suppression in cancer, regulation of epidermal cell transformation and angiogenesis (Ye *et al.*, 2003). In condition (c), protein CLU, C3, proapo-A1 and RBP4 were seen as the most immunoreactive proteins in OSCC patients. Immunoreactivity of these 4 proteins was consistent with the previous studies on various cancers including renal cancer (Koltai, 2014), uterine cervical cancer (Jeong *et al.*, 2008), colon cancer (Kim *et al.*, 2017) and bladder cancer (Li, F. *et al.*, 2012). The final condition in the immunoblotting assay involves probing of OSCC sera with OSCC sera. Four proteins (C3, HP, proapo-A1 and RBP4) were observed in this condition. This suggests that the

detected proteins were uniquely produced by the innate immune response of cancer cells.

As described, using the immunoblotting technique, a total of 5 proteins that were immunoreactive in OSCC patients were identified. These proteins were CLU, C3, HP, proapo-A1, and RBP4. These findings further supports the rationale of the OSCC proteins as biomarkers for early cancer detection and diagnosis (Mou *et al.*, 2009).

5.2.2.3 Differently Expressed Proteins in OSCC Patients Identified Through Label Free LC-MS

To support the proteomic results obtained from the 2-DE platform and to further expand the OSCC biomarkers panel, label-free LC-MS was employed. In the present study, label-free LC-MS quantification was performed on proteins extracted from six pairs of OSCC with adjacent normal tissue. Nineteen significantly differentiated proteins were identified as frequently observed proteins in the OSCC samples. Out of this total, eight proteins (ACTB, ACTBM, ACTC, ACTG, ACTS, HBB, POTEE, and POTEF) were up-regulated and eleven proteins (ALBU, CRNS1, E7ENN3, E7ERU0, EF1DL, F8WAH6, FOCAD, HBA, SCND3, TITIN and TIL11) were down-regulated in the OSCC samples.

All of the up-regulated proteins in this study have been previously described in various cancer studies. From the eight up-regulated proteins identified, actin proteins ACTB, ACTC, ACTG with proteins POTEE and POTEF from the POTE ankryrin domain were previously described in gastric cancer to be up-regulated (Ma, Y. *et al.*, 2014). As for actin protein ACTS, it was previously reported as an up-regulated protein in colorectal cancer (Sethi *et al.*, 2015) while HBB protein levels have been described up-regulated in serum from colorectal cancer patients (Choi, J. W. *et al.*, 2013).

However, up-regulated protein ACTBM, was previously reported to be down-regulated in cervical squamous cell carcinoma (Qing *et al.*, 2017).

Compared to all of the identified up-regulated proteins in this study, HBB protein was found to have the highest fold change across the up-regulated proteins identified. HBB or haemoglobin subunit beta is a protein that makes up haemoglobin and is known to be differently expressed in cancers such as colorectal and ovarian cancer (Choi, J. W. *et al.*, 2013; Woong-Shick *et al.*, 2005). Based on the analysis, HBB protein was found to be up-regulated in this study. However, a study by Roesch-Ely *et al.* (2007), HBB was reported to be down-regulated in the head and neck squamous cell carcinoma (HNSCC). Although the biological function of HBB in OSCC is not well understood, the irregularities of this protein expression in both HNSCC and OSCC may allow HBB to be used as a unique marker for OSCC.

In this study, ACTBM was observed as up-regulated in OSCC. However, in a recent study ACTBM was reported as down-regulated in cervical squamous cell carcinoma (Qing *et al.*, 2017). To date, ACTBM were reported to be associated with other cancers types. This suggests that ACTBM may have the potential to be classified as OSCC biomarker. Nevertheless, the specificity of ACTBM in OSCC should be taken into consideration as this protein was seen up-regulated in other diseases such psoriasis vulgaris (PV). Therefore, further studies should be conducted to broaden the current understandings on the role of ACTBM in OSCC.

Based on the LC-MS/MS data, eleven proteins (ALBU, CRNS1, E7ENN3, E7ERU0, EF1DL, F8WAH6, FOCAD, HBA, SCND3, TITIN and TIL11) were down-regulated in the OSCC samples. These eleven proteins were previously described in various cancer studies. However, the expression of several proteins in their respective cancers was not in concordance to the finding in this study. Similarly to this study,

proteins E7ENN3, E7ERU0, FOCAD, TIL11, TITIN and EF1DL were seen down-regulated in cancers of hepatocellular (Mizuno *et al.*, 2012), oral (Chen, Y. J. *et al.*, 2011), colorectal (Weren *et al.*, 2015), lung (Okamoto *et al.*, 2006) and breast (Krizman *et al.*, 2015; Rappa *et al.*, 2017). However, proteins CRNS1, HBA, F8WAH6 and SCND3 were seen up-regulated in cancers of breast (Beltran *et al.*, 2011), ovarian (Woong-Shick *et al.*, 2005), OSCC (Bagordakis *et al.*, 2016) and non-small cell lung cancer (Zhang, X. *et al.*, 2017). As a result, variation in the expression level of CRNS1, HBA, F8WAH6 and SCND3 when compared to other cancers, suggest that these protein has the potential to be classified as OSCC biomarker. However, further studies should be warranted to conclude the use of these proteins as potential biomarkers. Lastly, albumin a high abundant transport protein has been widely associated with various cancers in the past. Furthermore, albumin has been identified as either up-regulated or down-regulated in many cancers in previous studies. Therefore, this protein could not be used as a specific biomarker for OSCC.

In general, both 2-DE technique and label free LC-MS had successfully identified a total of 27 differently expressed proteins in OSCC. From this total, proteins LRG, A1BG, PRO2044, ACTBM, HBB, CRNS1, HBA, F8WAH6 and SCND3 were found to be uniquely expressed in OSCC when compared with other cancers. Therefore, it is suggested that these proteins may have the potential to be used as a specific biomarker for OSCC. However, further studies are required to validate this protein in a larger cohort and to fully understand the role of these proteins in OSCC.

5.2.3 Similarity in the Integrated 'Omics' Data and Its Association with Oral Disease

Recent development of various genomic and proteomic technologies, many potential biomarkers in the form of DNA, RNA and proteins has been discovered.

Unique markers of cancers has prompt researchers to look into multiparametric analysis of genes and proteins compared to the previous method of single-biomarker analysis (Kulasingam *et al.*, 2008). These multiple variables are able to provide an accurate information than single markers (Manne *et al.*, 2005). In this study, both genomic and proteomic analyses were applied to identify potential biomarkers for OSCC.

Based on the data obtained through exome sequencing and label free LCMS, the gene *SYNE1* (Nesprin-1) was frequently discovered in the OSCC tissue sample, suggesting that this gene may play a significant role in OSCC. Nesprin-1 is an outer nuclear membrane protein (Zhang, J. *et al.*, 2009). It has been reported to harbour potential loss of function variants in cancers, such as colorectal cancer (Tanskanen *et al.*, 2015). Therefore, mutation in this gene leads to the onset of tumourigenesis (Sur *et al.*, 2014). In addition, mutations in this gene were also reported in head and neck cancers including OSCC further supporting the findings in this study (Hedberg *et al.*, 2016; Nakagaki *et al.*, 2017). The LCMS results also showed down-regulation of Nesprin-1 in OSCC. This observation was in concordance with the previous studies (Doherty *et al.*, 2010; Stransky *et al.*, 2011).

Although biomarkers are a feasible tool for early detection of OSCC, the possibility of false positive results can be a major concern. Since biomarkers are produced from the interaction between a biological system and a potential hazard (Rifai *et al.*, 2006), it may result from a pathophysiological pathway represented by other diseases with a similar indicator. For example, oral cancer has been reported to mimic the clinical symptoms of advance periodontal disease (Fitzpatrick & Katz, 2010; Rezende, C. P. D. *et al.*, 2008). As a result, the diagnosis of OSCC which has similar clinical presentation of periodontal disease may be the result of a false positive detection. Therefore, it is important to exclude biological markers which have similar

expression. A more specific diagnostic biomarker may prevent difficulties and delays in early diagnosis of OSCC. To address this issue, the current findings were compared with a previous study on the periodontal biomarkers.

In the previous study, a total of 4 (HP, KNG1, AIAT, IGKC) differently expressed proteins were identified in the sera of patients with periodontitis using a similar proteomic approach (Kerishnan *et al.*, 2016). These immunogenic proteins could function as specific and sensitive markers for the early detection of periodontal disease. However, HP was found to be up-regulated in both OSCC and in mild chronic periodontitis. This suggests that HP should be excluded as a potential biomarker for OSCC. Apart from these 4 differentially expressed proteins, a previous study on periodontitis had also identified several other aberrantly expressed proteins similar to those found in this current study. Further studies are warranted to identify more possible proteins expression that may indicate false positive identification of OSCC.

5.2.4 Functional Enrichment and Pathway Analysis

Throughout the years, the application of next generation technologies allowed the identification of a vast number of genes and proteins associated with cancers. The identification of these gene and protein markers that either interacts directly or indirectly in the form of pathway or complex networks, allows researchers to evaluate its association to cancers (Wang, J. *et al.*, 2015). Moreover, giving a biological meaning to these lists of genes and proteins, allowed researchers to further understand the mechanism involved in the development and progression of cancer. Therefore, the key approach to elucidate the pathogenesis of a disease is to identify the biological function and pathways associated with the disease using these large number of distinguished genes and proteins (Zhao, Jinying *et al.*, 2015).

To elucidate the biological functions associated with OSCC in this study, functional enrichment analysis was applied on the large list of potential biomarkers obtained from both genomic and proteomic analysis in this study. Functional enrichment analysis was conducted using web-based bioinformatics annotation tools; ConsensusPathDB and DAVID v6.8 (Database for Annotation, Visualization and Integrated Discovery). This analysis classified the potential biomarkers based on the gene ontology term (biological process, cellular component and molecular function) and identified the most significant over-represented biological function (Figure 4.15). Based on the findings, the cellular components; extracellular exosome, extracellular vesicle and extracellular organelle, were found to be involved significantly in the biological function of OSCC. Extracellular exosome, extracellular vesicle and extracellular organelles are membranous particles released into the extracellular space by any type of cells (Mathivanan *et al.*, 2010; Minciacchi *et al.*, 2015; Raposo & Stoorvogel, 2013). These membranous particles are an organized structure outside the cell with distinctive morphology and functions (Mathivanan *et al.*, 2010).

Exosome is a type of membranous particle in the extracellular region (Mathivanan *et al.*, 2010; Minciacchi *et al.*, 2015; Raposo *et al.*, 2013). Exosome sometimes plays the role of a transporting agent that expels excess or non-functional cellular constituent (Kalluri, 2016). Exosomes from cancer cell are known to carry pathogenic elements such as DNA, protein, mRNA, transcriptional factors and lipids. Transportation of this pathogenic element allows long-distance crosstalk between distant organs and cancer cells resulting to pre-metastatic niche formation (Fujita *et al.*, 2016). Therefore, it may function as potential diagnostic and screening tool in cancer (Melo *et al.*, 2015). In some cases, exosome are suggested to be a transportation for drugs in cancer treatment or as a biomarker for treatment efficacy (Aubertin *et al.*,

2016). The involvement of the identified potential biomarkers in this cellular component further supports them as biomarkers for OSCC.

To further understand the functional working mechanism of OSCC, pathway analysis on the identified DNA/protein were also performed. The analysis was performed using ConsensusPathDB and DAVID v6.8 through the KEGG, BioCarta and Reactome database. In addition, protein-protein interaction (PPI) analysis using the identified biomarkers were performed using STRING v10.1 to provided valuable details for network-based identification and validation of this biomarkers (Chen, H. *et al.*, 2015). Based on the pathway analysis on the identified potential biomarkers, top 6 significant networks were identified. These networks are; a) platelet activation, signalling and aggregation pathway, b) focal adhesion pathway, c) oxytocin signalling pathway, d) apoptosis pathway, e) thyroid hormone signalling pathway and f) folate metabolism pathway. From these top 6 networks, the platelet activation, signalling and aggregation pathway was seen as the most over-represented pathway in OSCC. Protein AIBG, ALB, CLU, and APOA1 were identified as the binding or interacting partners in this pathway.

Platelet activation, signalling and aggregation pathway involves in the activation, shape changing, adhesion, and aggregation of platelet. This series of events are triggered by the exposure of platelets to sub endothelial tissue and leads to the formation of stable haemostatic plug. Studies have shown that platelet plays an active role in haemostasis, atherosclerosis, thrombosis, wound healing, immunity, inflammation and also in tumour metastasis (Li, Z. *et al.*, 2010). Platelets are also known to be associated with cancers. Platelets have been reported to be involved in cancer metastasis and some tumour cells are known to simulate platelets. Activated platelets release active molecules that regulate tumour growth and metastasis. Tumour

cells that aggregate with platelets avoid cytotoxicity mediated by the natural killer cell (Elaskalani *et al.*, 2017). Therefore, platelet activation, signalling and aggregation pathway plays an important role in the tumour cell survival further supporting the finding in this study.

Apart from platelet activation, signalling and aggregation pathway, other over-represented pathways identified in this study were folate metabolism pathway, thyroid hormone signaling pathway, apoptosis pathway, oxytocin signaling pathway and focal adhesion pathway. These pathways were also found to be associated with OSCC and other cancers.

Folate metabolism pathway involves in the synthesis of DNA and is influenced by polymorphisms in its associate genes such as methylenetetra-hydo folate reductase (*MTHFR*) gene (Nazki *et al.*, 2014). In OSCC, genetic variation in *MTHFR* gene had been reported to modulate the risk of OSCC by altering DNA synthesis/repair and the methylation process. For example, genetic variation in *MTHFR* gene had been reported to decrease the DNA mythylation efficiency further reducing the risk of cancer metastasis (Barbosa *et al.*, 2016).

Apoptosis pathway is important in the natural process of cell death and elimination of cell (Ashkenazi, 2008), whereas focal adhesion pathway is important not only in cell death but also in cell differentiation, cell proliferation and cell survival (Wolfenson *et al.*, 2013). In OSCC, protein caspase-8 which is a pro-apoptotic protein in the apoptosis pathway was reported to influence cell apoptosis during the cancer development (Ashkenazi, 2008; Coutinho-Camillo *et al.*, 2017). On the contrary, focal adhesion pathway was reported to be significantly regulated in OSCC, as a response to extracellular matrix (ECM)-receptor interaction pathway, prompting cancer cell

proliferation, movement and differentiation and further restricting cell death in OSCC (He *et al.*, 2016).

Lastly, both thyroid hormone signaling pathway and oxytocin signaling pathway are known to play a role in carcinogenesis. Thyroid hormone signaling pathways functions as regulators of cell proliferation, differentiation and apoptosis in thyroid cancer (Brent, 2012; Marini *et al.*, 2011). Whereas, oxytocin signaling pathway plays a crucial role as growth regulator and inhibit proliferation of neoplastic cell in breast cancer (Cassoni *et al.*, 2004). However, the exact mechanism of these 2 pathways in OSCC has not been reported previously.

CHAPTER 6: CONCLUSION

OSCC is a growing problem in many parts of the world due to late identification of the disease. Early detection of OSCC using reliable biomarkers may improve the survival rate of the patients. In the present study, potential biomarkers for OSCC were identified using next generation technologies including exome sequencing (NGS) and 2-DE coupled label free LCMS. Prior to the experimentation work, the demographic characteristic of the study cohort was determined. This study confirmed that local female Indian community that practices betel quid chewing are most likely to develop OSCC and past exposure to HPV16 infection could contribute to the onset of OSCC.

To identify somatic mutation underlying OSCC in the study cohort, exome sequencing technology was applied. Out of the total genes identified, 4 somatically mutated genes; *CASP8, USP40, NOTCH1, and COL11A1* which had the highest missense mutation (SNVs) were selected as candidate genes to be further validated and studied. *CASP8* and *NOTCH1* were previously reported as tumour suppressor genes in various cancers including OSCC. However, based on these findings, 2 SNVs from each gene were found to be novel in OSCC. Although mutations in *USP40* and *COL11A1* were previously reported in various cancers including the head and neck cancers, the exact association of these mutations were never described in OSCC. This study is the first to identify SNVs from both genes that are associated to OSCC. Deleterious mutations were confirmed for all of the identified SNVs in these 4 genes. Finally, based on the validation using Fluidigm SNP Genotyping, USP40 was found to be the most promising molecular biomarker for OSCC.

To further enhance the biomarker discovery in OSCC, proteomic approach was applied in this study. The proteomic platforms included was; 2-DE coupled MS, immunoproteomics and high throughput label-free LCMS. A total of 27 differently

expressed proteins in OSCC were identified. Among these proteins, LRG, A1BG, PRO2044, ACTBM, HBB, CRNS1, HBA, F8WAH6 and SCND3 were found to be uniquely expressed in OSCC when compared with other cancers. Therefore, it is suggested that these proteins may have the potential to be used as a specific biomarker for OSCC. In addition, *SYNE1* (Nesprin-1) was identified in both genomic and proteomic approaches in this study.

Finally, functional analysis was carried out to evaluate the association of the potential biomarkers with OSCC and to understand the intracellular signalling pathways that underlie the development of OSCC. The bioinformatics analysis tools used in achieving this objective were ConsensusPathDB, DAVID v6.8 and STRING v10.1. Based on the analysis, the most significant functional process in OSCC is the involvement of extracellular exosome, which is known to play an important role in cancer metastasis. Whereas, the most prominent pathway identified in the pathway analysis is the platelet activation, signalling and aggregation pathway which is also involved in cancer metastasis.

Taken together, the study had successfully identified a combination of 13 novel potential biomarkers and further improved the current understanding on the biological functions and pathways associated with OSCC. These biomarkers could potentially be used to identify risk of OSCC and as a screening tool for early cancer detection in a healthy asymptomatic individual. Early detection is a necessity and a key foundation to improve the survival rate of OSCC. However, further studies are required to validate these biomarkers in a larger cohort and to fully understand the role of these biomarkers in OSCC.

REFERENCES

Abulaizi, M., Tomonaga, T., Satoh, M., Sogawa, K., Matsushita, K., Kodera, Y., . . . Miyazaki, M. (2011). The application of a three-step proteome analysis for identification of new biomarkers of pancreatic cancer. *International Journal of Proteomics, 2011*.

Acharya, S., Ekalaksananan, T., Vatanasapt, P., Loyha, K., Phusingha, P., Promthet, S., . . . Pientong, C. (2015). Association of epstein-barr virus infection with oral squamous cell carcinoma in a case–control study. *Journal of Oral Pathology and Medicine, 44*(4), 252-257.

Aebersold, R., & Mann, M. (2003). Mass spectrometry-based proteomics. *Nature, 422*(6928), 198-207.

Agrawal, N., Frederick, M. J., Pickering, C. R., Bettegowda, C., Chang, K., Li, R. J., . . . Myers, J. N. (2011). Exome sequencing of head and neck squamous cell carcinoma reveals inactivating mutations in notch1. *Science, 333*(6046), 1154-1157.

Agrawal, N., Jiao, Y., Bettegowda, C., Hutfless, S. M., Wang, Y., David, S., . . . Shin, E. J. (2012). Comparative genomic analysis of esophageal adenocarcinoma and squamous cell carcinoma. *Cancer Discovery, 2*(10), 899-905.

Ah-See, K. W., Cooke, T. G., Pickford, I. R., Soutar, D., & Balmain, A. (1994). An allelotype of squamous carcinoma of the head and neck using microsatellite markers. *Cancer Research, 54*(7), 1617-1621.

Ahmed, N., Barker, G., Oliva, K., Hoffmann, P., Riley, C., Reeve, S., . . . Rice, G. (2004). Proteomic-based identification of haptoglobin-1 precursor as a novel circulating biomarker of ovarian cancer. *British Journal of Cancer, 91*(1), 129.

Al-hebshi, N. N., Li, S., Nasher, A. T., El-Setouhy, M., Alsanosi, R., Blancato, J., & Loffredo, C. (2016). Exome sequencing of oral squamous cell carcinoma in users of a rabian snuff reveals novel candidates for driver genes. *International Journal of Cancer, 139*(2), 363-372.

Altelaar, A. M., Munoz, J., & Heck, A. J. (2013). Next-generation proteomics: Towards an integrative view of proteome dynamics. *Nature Reviews Genetics, 14*(1), 35-48.

Ambatipudi, S., Gerstung, M., Gowda, R., Pai, P., Borges, A. M., Schäffer, A. A., . . . Mahimkar, M. B. (2011). Genomic profiling of advanced-stage oral cancers reveals chromosome 11q alterations as markers of poor clinical outcome. *PloS One, 6*(2), e17250.

Ando, M., Kawazu, M., Ueno, T., Fukumura, K., Yamato, A., Soda, M., . . . Mano, H. (2013). Cancer-associated missense mutations of caspase-8 activate nuclear factor-κb signaling. *Cancer Science, 104*(8), 1002-1008.

Arbyn, M., Tommasino, M., Depuydt, C., & Dillner, J. (2014). Are 20 human papillomavirus types causing cervical cancer? *The Journal of pathology, 234*(4), 431-435.

Arnott, D., & Emmert-Buck, M. R. (2010). Proteomic profiling of cancer—opportunities, challenges, and context. *The Journal of Pathology, 222*(1), 16-20.

Ashkenazi, A. (2008). Targeting the extrinsic apoptosis pathway in cancer. *Cytokine and Growth Factor Reviews, 19*(3-4), 325-331.

Aubertin, K., Silva, A. K., Luciani, N., Espinosa, A., Djemat, A., Charue, D., . . . Wilhelm, C. (2016). Massive release of extracellular vesicles from cancer cells after photodynamic treatment or chemotherapy. *Scientific Reports, 6*.

Baak, J., Path, F., Hermsen, M., Meijer, G., Schmidt, J., & Janssen, E. (2003). Genomics and proteomics in cancer. *European Journal of Cancer, 39*(9), 1199-1215.

Baer, B., & Millar, A. (2016). Proteomics in evolutionary ecology. *Journal of Proteomics, 135*, 4-11.

Bagan, J., Sarrion, G., & Jimenez, Y. (2010). Oral cancer: Clinical features. *Oral Oncology, 46*(6), 414-417.

Bagordakis, E., Sawazaki-Calone, I., Macedo, C. C. S., Carnielli, C. M., de Oliveira, C. E., Rodrigues, P. C., . . . Graner, E. (2016). Secretome profiling of oral squamous cell carcinoma-associated fibroblasts reveals organization and disassembly of extracellular matrix and collagen metabolic process signatures. *Tumor Biology, 37*(7), 9045-9057.

Bamshad, M. J., Ng, S. B., Bigham, A. W., Tabor, H. K., Emond, M. J., Nickerson, D. A., & Shendure, J. (2011). Exome sequencing as a tool for mendelian disease gene discovery. *Nature Reviews Genetics, 12*(11), 745.

Banerji, S., Cibulskis, K., Rangel-Escareno, C., Brown, K. K., Carter, S. L., Frederick, A. M., . . . Zou, L. (2012). Sequence analysis of mutations and translocations across breast cancer subtypes. *Nature, 486*(7403), 405-409.

Barbieri, C. E., Baca, S. C., Lawrence, M. S., Demichelis, F., Blattner, M., Theurillat, J.-P., . . . Stransky, N. (2012). Exome sequencing identifies recurrent spop, foxa1 and med12 mutations in prostate cancer. *Nature Genetics, 44*(6), 685.

Barbosa, A., dos Santos, M., de Podestá, J. R. V., Gouvêa, S. A., Von Zeidler, S. V., Louro, I. D., & de Freitas Cordeiro-Silva, M. (2016). Polymorphisms in methylenetetrahydrofolate reductase and cystathionine beta-synthase in oral cancer–a case–control study in southeastern brazilians. *Brazilian Journal of Otorhinolaryngology, 82*(5), 558-566.

Baxevanis, A. D., & Ouellette, B. F. (2004). *Bioinformatics: A practical guide to the analysis of genes and proteins* (Vol. 43): John Wiley & Sons.

Beltran, A. S., Russo, A., Lara, H., Fan, C., Lizardi, P. M., & Blancafort, P. (2011). Suppression of breast tumor growth and metastasis by an engineered transcription factor. *PloS One, 6*(9), e24595.

Bendl, J., Stourac, J., Salanda, O., Pavelka, A., Wieben, E. D., Zendulka, J., . . . Damborsky, J. (2014). Predictsnp: Robust and accurate consensus classifier for prediction of disease-related mutations. *PLoS Computational Biology, 10*(1), e1003440.

Benzian, H. (2013). *Ncds and oral cancer: Rationale for inclusion in sear ncd action plan and voluntary targets.* Paper presented at the Technical Working Group Meeting on Regional Action Plan and Targets for Prevention and Control of Noncommunicable Diseases.

Bhatnagar, R., Dabholkar, J., & Saranath, D. (2012). Genome-wide disease association study in chewing tobacco associated oral cancers. *Oral Oncology, 48*(9), 831-835.

Biesecker, L. G., Shianna, K. V., & Mullikin, J. C. (2011). Exome sequencing: The expert view. *Genome Biology, 12*(9), 1.

Bouvard, V., Baan, R., Straif, K., Grosse, Y., Secretan, B., El Ghissassi, F., . . . Galichet, L. (2009). A review of human carcinogens—part b: Biological agents. *The lancet oncology, 10*(4), 321-322.

Brandlein, S., Pohle, T., Ruoff, N., Wozniak, E., Muller-Hermelink, H. K., & Vollmers, H. P. (2003). Natural igm antibodies and immunosurveillance mechanisms against epithelial cancer cells in humans. *Cancer Research, 63*(22), 7995-8005.

Brent, G. A. (2012). Mechanisms of thyroid hormone action. *The Journal of Clinical Investigation, 122*(9), 3035-3043.

Buajeeb, W., Poomsawat, S., Punyasingh, J., & Sanguansin, S. (2009). Expression of p16 in oral cancer and premalignant lesions. *Journal of Oral Pathology and Medicine, 38*(1), 104-108.

Cassoni, P., Sapino, A., Marrocco, T., Chini, B., & Bussolati, G. (2004). Oxytocin and oxytocin receptors in cancer cells and proliferation. *Journal of Neuroendocrinology, 16*(4), 362-364.

Chai, R. C., Lim, Y., Frazer, I. H., Wan, Y., Perry, C., Jones, L., . . . Punyadeera, C. (2016). A pilot study to compare the detection of hpv-16 biomarkers in salivary oral rinses with tumour p16 ink4a expression in head and neck squamous cell carcinoma patients. *BMC Cancer, 16*(1), 178.

Chakrabarty, B., & Parekh, N. (2014). Identifying tandem ankyrin repeats in protein structures. *BMC Bioinformatics, 15*(1), 6599.

Challis, D., Yu, J., Evani, U. S., Jackson, A. R., Paithankar, S., Coarfa, C., . . . Yu, F. (2012). An integrative variant analysis suite for whole exome next-generation sequencing data. *BMC Bioinformatics, 13*(1), 8.

Chang, S.-W., Kareem, S. A., Kallarakkal, T. G., Merican, A., Abraham, M. T., & Zain, R. B. (2011). Feature selection methods for optimizing clinicopathologic input variables in oral cancer prognosis. *Asian Pacific Journal of Cancer Prevention, 12*, 2659-2664.

Chaturvedi, A. K., Anderson, W. F., Lortet-Tieulent, J., Curado, M. P., Ferlay, J., Franceschi, S., . . . Gillison, M. L. (2013). Worldwide trends in incidence rates for oral cavity and oropharyngeal cancers. *Journal of Clinical Oncology, 31*(36), 4550.

Chen, H., Zhu, Z., Zhu, Y., Wang, J., Mei, Y., & Cheng, Y. (2015). Pathway mapping and development of disease-specific biomarkers: Protein-based network biomarkers. *Journal of Cellular and Molecular Medicine, 19*(2), 297-314.

Chen, Q., Wang, Z., Zhang, K., Liu, X., Cao, W., Zhang, L., . . . Xia, C. (2011). Clusterin confers gmcitabine resistance in pancreatic cancer. *World Journal of Surgical Oncology, 9*(1), 59.

Chen, Y., Azman, S. N., Kerishnan, J. P., Zain, R. B., Chen, Y. N., Wong, Y.-L., & Gopinath, S. C. (2014). Identification of host-immune response protein candidates in the sera of human oral squamous cell carcinoma patients. *PloS One, 9*(10), e109012.

Chen, Y., Lim, B.-K., Peh, S.-C., Abdul-Rahman, P. S., & Hashim, O. H. (2008). Profiling of serum and tissue high abundance acute-phase proteins of patients with epithelial and germ line ovarian carcinoma. *Proteome science, 6*(1), 20.

Chen, Y. J., Liao, C. T., Chen, P. J., Lee, L. Y., Li, Y. C., Chen, I. H., . . . Yen, T. C. (2011). Downregulation of ches1 and other novel genes in oral cancer cells chronically exposed to areca nut extract. *Head and Neck, 33*(2), 257-266.

Cheng, Y., Liu, C., Zhang, N., Wang, S., & Zhang, Z. (2014). Proteomics analysis for finding serum markers of ovarian cancer. *BioMed Research Iinternational, 2014*.

Cheong, S., Chandramouli, G., Saleh, A., Zain, R. B., Lau, S., Sivakumaren, S., . . . Patel, V. (2009). Gene expression in human oral squamous cell carcinoma is influenced by risk factor exposure. *Oral Oncology, 45*(8), 712-719.

Cheong, S. C., Vatanasapt, P., Yi-Hsin, Y., Zain, R. B., Kerr, A. R., & Johnson, N. W. (2017). Oral cancer in south east asia: Current status and future directions. *Translational Research in Oral Oncology, 2*, 2057178X17702921.

Chi, A. C., Day, T. A., & Neville, B. W. (2015). Oral cavity and oropharyngeal squamous cell carcinoma—an update. *CA: A Cancer Journal for Clinicians, 65*(5), 401-421.

Choi, J. W., Liu, H., Shin, D. H., Yu, G. I., Hwang, J. S., Kim, E. S., & Yun, J. W. (2013). Proteomic and cytokine plasma biomarkers for predicting progression from colorectal adenoma to carcinoma in human patients. *Proteomics, 13*(15), 2361-2374.

Choi, M., Scholl, U. I., Ji, W., Liu, T., Tikhonova, I. R., Zumbo, P., . . . Sanjad, S. (2009). Genetic diagnosis by whole exome capture and massively parallel DNA sequencing. *Proceedings of the National Academy of Sciences, 106*(45), 19096-19101.

Choi, S., & Myers, J. N. (2008). Molecular pathogenesis of oral squamous cell carcinoma: Implications for therapy. *J Dent Res, 87*(1), 14-32.

Chong, I. Y., Cunningham, D., Barber, L. J., Campbell, J., Chen, L., Kozarewa, I., . . . Garcia-Murillas, I. (2013). The genomic landscape of oesophagogastric junctional adenocarcinoma. *The Journal of Pathology, 231*(3), 301-310.

Chung, C. C., Magalhaes, W. C., Gonzalez-Bosquet, J., & Chanock, S. J. (2009). Genome-wide association studies in cancer—current and future directions. *Carcinogenesis, 31*(1), 111-120.

Clark, M. J., Chen, R., Lam, H. Y., Karczewski, K. J., Chen, R., Euskirchen, G., . . . Snyder, M. (2011). Performance comparison of exome DNA sequencing technologies. *Nature Biotechnology, 29*(10), 908-914.

Clough, T., Key, M., Ott, I., Ragg, S., Schadow, G., & Vitek, O. (2009). Protein quantification in label-free lc-ms experiments. *Journal of Proteome Research, 8*(11), 5275-5284.

Connolly, J. L., Schnitt, S. J., Wang, H. H., Longtine, J. A., Dvorak, A., & Dvorak, H. F. (2003). Role of the surgical pathologist in the diagnosis and management of the cancer patient.

Coutinho-Camillo, C. M., Lourenço, S. V., Puga, R. D., Damascena, A. S., Teshima, T. H. N., Kowalski, L. P., & Soares, F. A. (2017). Profile of apoptotic proteins in oral squamous cell carcinoma: A cluster analysis of 171 cases. *Applied Cancer Research, 37*(1), 2.

Cox, J., & Mann, M. (2007). Is proteomics the new genomics? *Cell, 130*(3), 395-398.

Curtis, R. K., Orešič, M., & Vidal-Puig, A. (2005). Pathways to the analysis of microarray data. *Trends in Biotechnology, 23*(8), 429-435.

da Silva, S. D., Ferlito, A., Takes, R. P., Brakenhoff, R. H., Valentin, M. D., Woolgar, J. A., . . . Hier, M. P. (2011). Advances and applications of oral cancer basic research. *Oral Oncology, 47*(9), 783-791.

Dessau, R., & Pipper, C. (2008). ["r"--project for statistical computing]. *Ugeskrift for Laeger, 170*(5), 328-330.

Do, R., Kathiresan, S., & Abecasis, G. R. (2012). Exome sequencing and complex disease: Practical aspects of rare variant association studies. *Human Molecular Genetics, 21*(R1), R1-R9.

Doherty, J. A., Rossing, M. A., Cushing-Haugen, K. L., Chen, C., Van Den Berg, D. J., Wu, A. H., . . . Chenevix-Trench, G. (2010). Esr1/syne1 polymorphism and invasive epithelial ovarian cancer risk: An ovarian cancer association consortium study. *Cancer Epidemiology and Prevention Biomarkers, 19*(1), 245-250.

Dorschner, M. O. (2014). Next-generation sequencing. *Genomic Applications in Pathology*, 209.

Easton, D. F., & Eeles, R. A. (2008). Genome-wide association studies in cancer. *Human Molecular Genetics, 17*(R2), R109-R115.

Eckhart, L., Ballaun, C., Hermann, M., VandeBerg, J. L., Sipos, W., Uthman, A., . . . Tschachler, E. (2008). Identification of novel mammalian caspases reveals an important role of gene loss in shaping the human caspase repertoire. *Molecular Biology and Evolution, 25*(5), 831-841.

El-Bayoumy, K., Das, A., Russell, S., Wolfe, S., Jordan, R., Renganathan, K., . . . Somiari, R. (2012). The effect of selenium enrichment on baker's yeast proteome. *Journal of Proteomics, 75*(3), 1018-1030.

Elaskalani, O., Berndt, M. C., Falasca, M., & Metharom, P. (2017). Targeting platelets for the treatment of cancer. *Cancers, 9*(7), 94.

Ellis, M. J., Gillette, M., Carr, S. A., Paulovich, A. G., Smith, R. D., Rodland, K. K., . . . Rodriguez, H. (2013). Connecting genomic alterations to cancer biology with

proteomics: The nci clinical proteomic tumor analysis consortium. *Cancer Discovery, 3*(10), 1108-1112.

Erhart, S. M. M., Rivero, E. R. C., Bazzo, M. L., & Onofre, A. S. C. (2016). Comparative evaluation of the gp5+/6+, my09/11 and pgmy09/11 primer sets for hpv detection by pcr in oral squamous cell carcinomas. *Experimental and Molecular Pathology, 100*(1), 13-16.

Ernani, F., & LeProust, E. (2013). Agilent's sureselect target enrichment system: Bringing cost and process efficiency to next generation sequencing. *Available at:. Accessed on May, 21*.

Eskiizmir, G., Ermertcan, A. T., & Yapici, K. (2017). Chapter 17 - nanomaterials: Promising structures for the management of oral cancer. In E. Andronescu & A. M. Grumezescu (Eds.), *Nanostructures for oral medicine* (pp. 511-544): Elsevier.

Etzioni, R., Urban, N., Ramsey, S., McIntosh, M., Schwartz, S., Reid, B., . . . Hartwell, L. (2003). The case for early detection. *Nature Reviews Cancer, 3*(4), 243-252.

Evans, A. S. (2013). *Viral infections of humans: Epidemiology and control*: Springer Science & Business Media.

Farias, E. F., Marzan, C., & Mira-y-Lopez, R. (2005). Cellular retinol-binding protein-i inhibits pi3k/akt signaling through a retinoic acid receptor-dependent mechanism that regulates p85–p110 heterodimerization. *Oncogene, 24*(9), 1598.

Ferlay, J., Soerjomataram, I., Dikshit, R., Eser, S., Mathers, C., Rebelo, M., . . . Bray, F. (2015). Cancer incidence and mortality worldwide: Sources, methods and major patterns in globocan 2012. *International Journal of Cancer, 136*(5).

Fernández-Olavarría, A., Mosquera-Pérez, R., Díaz-Sánchez, R.-M., Serrera-Figallo, M.-A., Gutiérrez-Pérez, J.-L., & Torres-Lagares, D. (2016). The role of serum biomarkers in the diagnosis and prognosis of oral cancer: A systematic review. *Journal of Clinical and Experimental Dentistry, 8*(2), e184.

Fitzpatrick, S. G., & Katz, J. (2010). The association between periodontal disease and cancer: A review of the literature. *Journal of Dentistry, 38*(2), 83-95.

Forbes, S. A., Bindal, N., Bamford, S., Cole, C., Kok, C. Y., Beare, D., . . . Menzies, A. (2010). Cosmic: Mining complete cancer genomes in the catalogue of somatic mutations in cancer. *Nucleic Acids Research, 39*(suppl_1), D945-D950.

Fraile, J., Quesada, V., Rodriguez, D., Freije, J., & López-Otín, C. (2012). Deubiquitinases in cancer: New functions and therapeutic options. *Oncogene, 31*(19), 2373-2388.

Franceschini, A., Szklarczyk, D., Frankild, S., Kuhn, M., Simonovic, M., Roth, A., . . . Von Mering, C. (2013). String v9. 1: Protein-protein interaction networks, with increased coverage and integration. *Nucleic Acids Research, 41*(D1), D808-D815.

Fujita, Y., Yoshioka, Y., & Ochiya, T. (2016). Extracellular vesicle transfer of cancer pathogenic components. *Cancer Science, 107*(4), 385-390.

Futreal, P. A., Coin, L., Marshall, M., Down, T., Hubbard, T., Wooster, R., . . . Stratton, M. R. (2004). A census of human cancer genes. *Nature Reviews Cancer, 4*(3), 177-183.

Gaitán-Cepeda, L.-A., Peniche-Becerra, A.-G., & Quezada-Rivera, D. (2011). Trends in frequency and prevalence of oral cancer and oral squamous cell carcinoma in mexicans. A 20 years retrospective study. *Medicina Oral, Patología Oral y Cirugía Bucal, 16*(1), e1-5.

Galván, J. A., García-Martínez, J., Vázquez-Villa, F., García-Ocaña, M., García-Pravia, C., Menéndez-Rodríguez, P., . . . de los Toyos, J. R. (2014). Validation of col11a1/procollagen 11a1 expression in tgf-β1-activated immortalised human mesenchymal cells and in stromal cells of human colon adenocarcinoma. *BMC Cancer, 14*(1), 867.

Garnis, C., Baldwin, C., Zhang, L., Rosin, M. P., & Lam, W. L. (2003). Use of complete coverage array comparative genomic hybridization to define copy number alterations on chromosome 3p in oral squamous cell carcinomas. *Cancer Res, 63*(24), 8582-8585.

Ghani, W. M., Razak, I. A., Yang, Y.-H., Talib, N. A., Ikeda, N., Axell, T., . . . Zain, R. B. (2011). Factors affecting commencement and cessation of betel quid chewing behaviour in malaysian adults. *BMC Public Health, 11*(1), 82.

Ghavami, S., Hashemi, M., Ande, S. R., Yeganeh, B., Xiao, W., Eshraghi, M., . . . Halayko, A. J. (2009). Apoptosis and cancer: Mutations within caspase genes. *Journal of Medical Genetics, 46*(8), 497-510.

Ginsburg, G. S., & Haga, S. B. (2006). Translating genomic biomarkers into clinically useful diagnostics. *Expert Review of Molecular Diagnostics, 6*(2), 179-191.

Gnirke, A., Melnikov, A., Maguire, J., Rogov, P., LeProust, E. M., Brockman, W., . . . Russ, C. (2009). Solution hybrid selection with ultra-long oligonucleotides for massively parallel targeted sequencing. *Nature Biotechnology, 27*(2), 182-189.

Goldszmid, R. S., Dzutsev, A., & Trinchieri, G. (2014). Host immune response to infection and cancer: Unexpected commonalities. *Cell Host & Microbe, 15*(3), 295-305.

Gómez, I., Warnakulasuriya, S., Varela-Centelles, P., López-Jornet, P., Suárez-Cunqueiro, M., Diz-Dios, P., & Seoane, J. (2010). Is early diagnosis of oral cancer a feasible objective? Who is to blame for diagnostic delay? *Oral Diseases, 16*(4), 333-342.

Gowda, H., Ivanisevic, J., Johnson, C. H., Kurczy, M. E., Benton, H. P., Rinehart, D., . . . Arevalo, B. (2014). Interactive xcms online: Simplifying advanced metabolomic data processing and subsequent statistical analyses. *Analytical Chemistry, 86*(14), 6931-6939.

Gravitt, P. E., Peyton, C. L., Alessi, T. Q., Wheeler, C. M., Coutlee, F., Hildesheim, A., . . . Apple, R. J. (2000). Improved amplification of genital human papillomaviruses. *Journal of Clinical Microbiology, 38*(1), 357-361.

Greenblatt, M., Bennett, W. P., Hollstein, M., & Harris, C. (1994). Mutations in the p53 tumor suppressor gene: Clues to cancer etiology and molecular pathogenesis. *Cancer Research, 54*(18), 4855-4878.

Guha, N., Warnakulasuriya, S., Vlaanderen, J., & Straif, K. (2014). Betel quid chewing and the risk of oral and oropharyngeal cancers: A meta-analysis with implications for cancer control. *International Journal of Cancer, 135*(6), 1433-1443.

Guo, Y., Ye, F., Sheng, Q., Clark, T., & Samuels, D. C. (2013). Three-stage quality control strategies for DNA re-sequencing data. *Briefings in Bioinformatics, 15*(6), 879-889.

Gupta, B., & Johnson, N. W. (2014). Systematic review and meta-analysis of association of smokeless tobacco and of betel quid without tobacco with incidence of oral cancer in south asia and the pacific. *PloS One, 9*(11), e113385.

Gupta, K., & Metgud, R. (2013). Evidences suggesting involvement of viruses in oral squamous cell carcinoma. *Pathology Research International, 2013*.

Hanahan, D., & Weinberg, R. A. (2000). The hallmarks of cancer. *Cell, 100*(1), 57-70.

Hanash, S. M., Bobek, M. P., Rickman, D. S., Williams, T., Rouillard, J. M., Kuick, R., & Puravs, E. (2002). Integrating cancer genomics and proteomics in the post-genome era. *Proteomics, 2*(1), 69-75.

Harn, H. J., Chen, Y. L., Lin, P. C., Cheng, Y. L., Lee, S. C., Chiou, T. W., & Yang, H. H. (2010). Exploration of potential tumor markers for lung adenocarcinomas by two-dimensional gel electrophoresis coupled with nano-lc/ms/ms. *Journal of the Chinese Chemical Society, 57*(2), 180-188.

Hartwell, L., Mankoff, D., Paulovich, A., Ramsey, S., & Swisher, E. (2006). Cancer biomarkers: A systems approach. *Nature Biotechnology, 24*(8), 905-908.

Haws, A. L. F., He, Q., Rady, P. L., Zhang, L., Grady, J., Hughes, T. K., . . . Tyring, S. K. (2004). Nested pcr with the pgmy09/11 and gp5+/6+ primer sets improves detection of hpv DNA in cervical samples. *Journal of Virological Methods, 122*(1), 87-93.

Hayes, T. F., Benaich, N., Goldie, S. J., Sipilä, K., Ames-Draycott, A., Cai, W., . . . Watt, F. M. (2016). Integrative genomic and functional analysis of human oral squamous cell carcinoma cell lines reveals synergistic effects of fat1 and casp8 inactivation. *Cancer Letters, 383*(1), 106-114.

He, Y., Shao, F., Pi, W., Shi, C., Chen, Y., Gong, D., . . . Tang, K. (2016). Largescale transcriptomics analysis suggests over-expression of bgh3, mmp9 and pdia3 in oral squamous cell carcinoma. *PloS One, 11*(1), e0146530.

Hedberg, M. L., Goh, G., Chiosea, S. I., Bauman, J. E., Freilino, M. L., Zeng, Y., . . . Lui, V. W. (2016). Genetic landscape of metastatic and recurrent head and neck squamous cell carcinoma. *The Journal of Clinical Investigation, 126*(1), 169-180.

Herwig, R., Hardt, C., Lienhard, M., & Kamburov, A. (2016). Analyzing and interpreting genome data at the network level with consensuspathdb. *Nature Protocols, 11*(10), 1889-1907.

Heukeshoven, J., & Dernick, R. (1988). Improved silver staining procedure for fast staining in phastsystem development unit. I. Staining of sodium dodecyl sulfate gels. *Electrophoresis, 9*(1), 28-32.

Higa, M., Kinjo, T., Kamiyama, K., Chinen, K., Iwamasa, T., Arasaki, A., & Sunakawa, H. (2003). Epstein–barr virus (ebv)-related oral squamous cell carcinoma in okinawa, a subtropical island, in southern japan—simultaneously infected with human papillomavirus (hpv). *Oral Oncology, 39*(4), 405-414.

Hijioka, H., Setoguchi, T., Miyawaki, A., Gao, H., Ishida, T., Komiya, S., & Nakamura, N. (2010). Upregulation of notch pathway molecules in oral squamous cell carcinoma. *International Journal of Oncology, 36*(4), 817-822.

Hogg, R., Honorio, S., Martinez, A., Agathanggelou, A., Dallol, A., Fullwood, P., . . . Latif, F. (2002). Frequent 3p allele loss and epigenetic inactivation of the rassf1a tumour suppressor gene from region 3p21. 3 in head and neck squamous cell carcinoma. *European Journal of Cancer, 38*(12), 1585-1592.

Hood, B. L., Veenstra, T. D., & Conrads, T. P. (2004). *Mass spectrometry-based proteomics.* Paper presented at the International Congress Series.

Hu, J., Ge, W., & Xu, J. (2016). Hpv 16 e7 inhibits oscc cell proliferation, invasion, and metastasis by upregulating the expression of mir-20a. *Tumor Biology, 37*(7), 9433-9440.

Huang, D. W., Sherman, B. T., & Lempicki, R. A. (2008). Systematic and integrative analysis of large gene lists using david bioinformatics resources. *Nature protocols, 4*(1), 44-57.

Huang, D. W., Sherman, B. T., & Lempicki, R. A. (2009). Bioinformatics enrichment tools: Paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Research, 37*(1), 1-13.

Huang, H. L., Stasyk, T., Morandell, S., Dieplinger, H., Falkensammer, G., Griesmacher, A., . . . Huck, C. W. (2006). Biomarker discovery in breast cancer serum using 2-d differential gel electrophoresis/maldi-tof/tof and data validation by routine clinical assays. *Electrophoresis, 27*(8), 1641-1650.

Huang, L.-J., Chen, S.-X., Huang, Y., Luo, W.-J., Jiang, H.-H., Hu, Q.-H., . . . Yi, H. (2006). Proteomics-based identification of secreted protein dihydrodiol dehydrogenase as a novel serum markers of non-small cell lung cancer. *Lung Cancer, 54*(1), 87-94.

Huang, S. K., Darfler, M. M., Nicholl, M. B., You, J., Bemis, K. G., Tegeler, T. J., . . . Nguyen, L. (2009). Lc/ms-based quantitative proteomic analysis of paraffin-embedded archival melanomas reveals potential proteomic biomarkers associated with metastasis. *PloS One, 4*(2), e4430.

Hussein, A. A., Helder, M. N., de Visscher, J. G., Leemans, C. R., Braakhuis, B. J., de Vet, H. C., & Forouzanfar, T. (2017). Global incidence of oral and oropharynx cancer in patients younger than 45 years versus older patients: A systematic review. *European Journal of Cancer, 82*, 115-127.

India Project Team of the International, & Consortium, C. G. (2013). Mutational landscape of gingivo-buccal oral squamous cell carcinoma reveals new recurrently-mutated genes and molecular subgroups. *Nature Communications, 4*.

Izumchenko, E., Sun, K., Jones, S., Brait, M., Agrawal, N., Koch, W., . . . Velculescu, V. E. (2015). Notch1 mutations are drivers of oral tumorigenesis. *Cancer Prevention Research, 8*(4), 277-286.

Jalouli, J., Ibrahim, S. O., Mehrotra, R., Jalouli, M. M., Sapkota, D., Larsson, P.-A., & Hirsch, J.-M. (2010). Prevalence of viral (hpv, ebv, hsv) infections in oral submucous fibrosis and oral cancer from india. *Acta Oto-Laryngologica, 130*(11), 1306-1311.

Jemal, A., Siegel, R., Ward, E., Hao, Y., Xu, J., & Thun, M. J. (2009). Cancer statistics, 2009. *CA: A Cancer Journal for Clinicians, 59*(4), 225-249.

Jenson, H. B. (2011). Epstein-barr virus. *Pediatrics in Review-Elk Grove, 32*(9), 375.

Jeong, D. H., Kim, H. K., Prince, A.-E., Lee, D. S., Kim, Y. N., Han, J., & Kim, K. T. (2008). Plasma proteomic analysis of patients with squamous cell carcinoma of the uterine cervix. *Journal of Gynecologic Oncology, 19*(3), 173-180.

Jia, D., Liu, Z., Deng, N., Tan, T. Z., Huang, R. Y.-J., Taylor-Harding, B., . . . Walts, A. E. (2016). A col11a1-correlated pan-cancer gene signature of activated fibroblasts for the prioritization of therapeutic targets. *Cancer Letters, 382*(2), 203-214.

Johnson, N. W., Warnakulasuriya, S., Gupta, P., Dimba, E., Chindia, M., Otoh, E., . . . Kowalski, L. (2011). Global oral health inequalities in incidence and outcomes for oral cancer causes and solutions. *Advances in Dental Research, 23*(2), 237-246.

Jones, S. E., & Jomary, C. (2002). Clusterin. *The international journal of biochemistry & cell biology, 34*(5), 427-431.

Jurel, S. K., Gupta, D. S., Singh, R. D., Singh, M., & Srivastava, S. (2014). Genes and oral cancer. *Indian Journal of Human Genetics, 20*(1), 4.

Kadler, K. E., Baldock, C., Bella, J., & Boot-Handford, R. P. (2007). Collagens at a glance. *Journal of Cell Science, 120*(12), 1955-1958.

Kakisaka, T., Kondo, T., Okano, T., Fujii, K., Honda, K., Endo, M., . . . Moriyasu, F. (2007). Plasma proteomics of pancreatic cancer patients by multi-dimensional liquid chromatography and two-dimensional difference gel electrophoresis (2d-dige): Up-regulation of leucine-rich alpha-2-glycoprotein in pancreatic cancer. *Journal of Chromatography B, 852*(1-2), 257-267.

Kalluri, R. (2016). The biology and function of exosomes in cancer. *The Journal of Clinical Investigation, 126*(4), 1208-1215.

Kalyankrishna, S., & Grandis, J. R. (2006). Epidermal growth factor receptor biology in head and neck cancer. *Journal of Clinical Oncology, 24*(17), 2666-2672.

Kamburov, A., Wierling, C., Lehrach, H., & Herwig, R. (2008). Consensuspathdb—a database for integrating human functional interaction networks. *Nucleic Acids Research, 37*(suppl_1), D623-D628.

Karsani, S. A., Saihen, N. A., Zain, R. B., Cheong, S.-C., & Rahman, M. A. (2014). Comparative proteomics analysis of oral cancer cell lines: Identification of cancer associated proteins. *Proteome science, 12*(1), 1.

Kawahara, R., Bollinger, J. G., Rivera, C., Ribeiro, A. C. P., Brandão, T. B., Leme, A. F. P., & MacCoss, M. J. (2016). A targeted proteomic strategy for the

measurement of oral cancer candidate biomarkers in human saliva. *Proteomics, 16*(1), 159-173.

Kawakami, T., Hoshida, Y., Kanai, F., Tanaka, Y., Tateishi, K., Ikenoue, T., . . . Shiina, S. (2005). Proteomic analysis of sera from hepatocellular carcinoma patients after radiofrequency ablation treatment. *Proteomics, 5*(16), 4287-4295.

Kerishnan, J. P., Mohammad, S., Alias, M. S., Mu, A. K.-W., Vaithilingam, R. D., Baharuddin, N. A., . . . Chen, Y. (2016). Identification of biomarkers for periodontal disease using the immunoproteomics approach. *PeerJ, 4*.

Kessner, D., Chambers, M., Burke, R., Agus, D., & Mallick, P. (2008). Proteowizard: Open source software for rapid proteomics tools development. *Bioinformatics, 24*(21), 2534-2536.

Khatri, P., Sirota, M., & Butte, A. J. (2012). Ten years of pathway analysis: Current approaches and outstanding challenges. *PLoS Computational Biology, 8*(2), e1002375.

Kim, K., Kim, M. J., Kim, K. H., Ahn, S. A., Kim, J. H., Cho, J. Y., & Yeo, S. G. (2017). C1qbp is upregulated in colon cancer and binds to apolipoprotein ai. *Experimental and Therapeutic Medicine, 13*(5), 2493-2500.

Kleinert, R., Prenzel, K., Stoecklein, N., Alakus, H., Bollschweiler, E., Hölscher, A., & Warnecke-Eberz, U. (2015). Gene expression of col11a1 is a marker not only for pancreas carcinoma but also for adenocarcinoma of the papilla of vater, discriminating between carcinoma and chronic pancreatitis. *Anticancer Research, 35*(11), 6153-6158.

Koboldt, D. C., Zhang, Q., Larson, D. E., Shen, D., McLellan, M. D., Lin, L., . . . Wilson, R. K. (2012). Varscan 2: Somatic mutation and copy number alteration discovery in cancer by exome sequencing. *Genome Research, 22*(3), 568-576.

Koltai, T. (2014). Clusterin: A key player in cancer chemoresistance and its inhibition. *OncoTargets and Therapy, 7*, 447.

Kozaki, K. i., Imoto, I., Pimkhaokham, A., Hasegawa, S., Tsuda, H., Omura, K., & Inazawa, J. (2006). Pik3ca mutation is an oncogenic aberration at advanced stages of oral squamous cell carcinoma. *Cancer Science, 97*(12), 1351-1358.

Kreunin, P., Zhao, J., Rosser, C., Urquidi, V., Lubman, D. M., & Goodison, S. (2007). Bladder cancer associated glycoprotein signatures revealed by urinary proteomic profiling. *Journal of Proteome Research, 6*(7), 2631-2639.

Krizman, D., Darfler, M. M., Conrads, T. P., & Hood, B. L. (2015). Protein biomarkers of late stage breast cancer: Google Patents.

Kruidering, M., & Evan, G. I. (2000). Caspase-8 in apoptosis: The beginning of "the end"? *IUBMB life, 50*(2), 85-90.

Ku, C. S., Cooper, D. N., Polychronakos, C., Naidoo, N., Wu, M., & Soong, R. (2012). Exome sequencing: Dual role as a discovery and diagnostic tool. *Annals of Neurology, 71*(1), 5-14.

Kuhn, M., von Mering, C., Campillos, M., Jensen, L. J., & Bork, P. (2008). Stitch: Interaction networks of chemicals and proteins. *Nucleic acids research, 36*(suppl 1), D684-D688.

Kulasingam, V., & Diamandis, E. P. (2008). Strategies for discovering novel cancer biomarkers through utilization of emerging technologies. *Nature clinical practice Oncology, 5*(10), 588-599.

Kumar, M., Nanavati, R., Modi, T. G., & Dobariya, C. (2016). Oral cancer: Etiology and risk factors: A review. *Journal of Cancer Research and Therapeutics, 12*(2), 458.

Kuo, W. P., Whipple, M. E., Jenssen, T. K., Todd, R., Epstein, J. B., Ohno-Machado, L., . . . Park, P. J. (2003). Microarrays and clinical dentistry. *J Am Dent Assoc, 134*(4), 456-462.

Kuppumbatti, Y. S., Rexer, B., Nakajo, S., Nakaya, K., & Mira-y-Lopez, R. (2001). Crbp suppresses breast cancer cell survival and anchorage-independent growth. *Oncogene, 20*(50), 7413.

Lai, C.-H., Chang, N.-W., Lin, C.-F., Lin, C.-D., Lin, Y.-J., Wan, L., . . . Sing, Y.-T. (2010). Proteomics-based identification of haptoglobin as a novel plasma biomarker in oral squamous cell carcinoma. *Clinica Chimica Acta, 411*(13-14), 984-991.

Lambert, R., Sauvaget, C., de Camargo Cancela, M., & Sankaranarayanan, R. (2011). Epidemiology of cancer from the oral cavity and oropharynx. *European Journal of Gastroenterology and Hepatology, 23*(8), 633-641.

Lee, C. H., Ko, A. M. S., Warnakulasuriya, S., Yin, B. L., Zain, R. B., Ibrahim, S. O., . . . Utomo, B. (2011). Intercountry prevalences and practices of betel-quid use in south, southeast and eastern asia regions and associated oral preneoplastic disorders: An international collaborative study by asian betel-quid consortium of south and east asia. *International Journal of Cancer, 129*(7), 1741-1751.

Lee, S. H., Jeong, E. G., Yoo, N. J., & Lee, S. H. (2007). Mutational analysis of notch1, 2, 3 and 4 genes in common solid cancers and acute leukemias. *APMIS, 115*(12), 1357-1363.

Levy, A. P., Asleh, R., Blum, S., Levy, N. S., Miller-Lotan, R., Kalet-Litman, S., . . . Asaf, R. (2010). Haptoglobin: Basic and clinical aspects. *Antioxidants & redox signaling, 12*(2), 293-304.

Li, D., & Chan, D. W. (2014). Proteomic cancer biomarkers from discovery to approval: It's worth the effort: Taylor & Francis.

Li, F., Chen, D.-n., He, C.-w., Zhou, Y., Olkkonen, V. M., He, N., . . . Lan, K.-j. (2012). Identification of urinary gc-globulin as a novel biomarker for bladder cancer by two-dimensional fluorescent differential gel electrophoresis (2d-dige). *Journal of Proteomics, 77*, 225-236.

Li, T., Kung, H.-J., Mack, P. C., & Gandara, D. R. (2013). Genotyping and genomic profiling of non–small-cell lung cancer: Implications for current and future therapies. *Journal of Clinical Oncology, 31*(8), 1039-1049.

Li, Z., Delaney, M. K., O'brien, K. A., & Du, X. (2010). Signaling during platelet adhesion and activation. *Arteriosclerosis, Thrombosis, and Vascular Biology, 30*(12), 2341-2349.

Liao, C.-T., Wallace, C. G., Lee, L.-Y., Hsueh, C., Lin, C.-Y., Fan, K.-H., . . . Tsao, C.-K. (2014). Clinical evidence of field cancerization in patients with oral cavity cancer in a betel quid chewing area. *Oral Oncology, 50*(8), 721-731.

Lin, H.-S., Talwar, H. S., Tarca, A. L., Ionan, A., Chatterjee, M., Ye, B., . . . Yoo, G. H. (2007). Autoantibody approach for serum-based detection of head and neck cancer. *Cancer Epidemiology and Prevention Biomarkers, 16*(11), 2396-2405.

Lin, J., Gan, C. M., Zhang, X., Jones, S., Sjöblom, T., Wood, L. D., . . . Vogelstein, B. (2007). A multidimensional analysis of genes mutated in breast and colorectal cancers. *Genome Research, 17*(9), 000-000.

Liu, C. J., Liu, T. Y., Kuo, L. T., Cheng, H. W., Chu, T. H., Chang, K. W., & Lin, S. C. (2008). Differential gene expression signature between primary and metastatic head and neck squamous cell carcinoma. *The Journal of Pathology, 214*(4), 489-497.

Lourda, M., Trougakos, I. P., & Gonos, E. S. (2007). Development of resistance to chemotherapeutic drugs in human osteosarcoma cell lines largely depends on up-regulation of clusterin/apolipoprotein j. *International Journal of Cancer, 120*(3), 611-622.

Ludwig, J. A., & Weinstein, J. N. (2005). Biomarkers in cancer staging, prognosis and treatment selection. *Nature Reviews Cancer, 5*(11), 845-856.

Ma, X., Zhang, J., Liu, S., Huang, Y., Chen, B., & Wang, D. (2011). Polymorphisms in the casp8 gene and the risk of epithelial ovarian cancer. *Gynecologic Oncology, 122*(3), 554-559.

Ma, Y., Li, Y.-F., Wang, T., Pang, R., Xue, Y.-W., & Zhao, S.-P. (2014). Identification of proteins associated with lymph node metastasis of gastric cancer. *Journal of Cancer Research and Clinical Oncology, 140*(10), 1739-1749.

Mäbert, K., Cojoc, M., Peitzsch, C., Kurth, I., Souchelnytskyi, S., & Dubrovska, A. (2014). Cancer biomarker discovery: Current status and future perspectives. *International Journal of Radiation Biology, 90*(8), 659-677.

Macdonald, F., Ford, C., & Casson, A. (2004). *Molecular biology of cancer*: Taylor & Francis.

Macleod, K. (2000). Tumor suppressor genes. *Current Opinion in Genetics and Development, 10*(1), 81-93.

Mahmood, T., & Yang, P.-C. (2012). Western blot: Technique, theory, and trouble shooting. *North American Journal of Medical Sciences, 4*(9), 429.

Majewski, J., Schwartzentruber, J., Lalonde, E., Montpetit, A., & Jabado, N. (2011). What can exome sequencing do for you? *Journal of Medical Genetics, 48*(9), 580-589.

Malarkey, D. E., Hoenerhoff, M., & Maronpot, R. R. (2013). Carcinogenesis: Mechanisms and manifestations *Haschek and rousseaux's handbook of toxicologic pathology (third edition)* (pp. 107-146): Elsevier.

Mali, S. B. (2014). Proteomics for oral cancer. *Oral Oncology, 50*(11), e67.

Mamanova, L., Coffey, A. J., Scott, C. E., Kozarewa, I., Turner, E. H., Kumar, A., . . . Turner, D. J. (2010). Target-enrichment strategies for next-generation sequencing. *Nature methods, 7*(2), 111-118.

Manne, U., Srivastava, R.-G., & Srivastava, S. (2005). Keynote review: Recent advances in biomarkers for cancer diagnosis and treatment. *Drug discovery today, 10*(14), 965-976.

Marini, F., Luzi, E., & Brandi, M. L. (2011). Microrna role in thyroid cancer development. *Journal of Thyroid Research, 2011*.

Marouga, R., David, S., & Hawkins, E. (2005). The development of the dige system: 2d fluorescence difference gel analysis technology. *Analytical and Bioanalytical Chemistry, 382*(3), 669-678.

Martin, D. B., & Nelson, P. S. (2001). From genomics to proteomics: Techniques and applications in cancer research. *Trends in Cell Biology, 11*(11), S60-S65.

Mascolo, M., Siano, M., Ilardi, G., Russo, D., Merolla, F., Rosa, G. D., & Staibano, S. (2012). Epigenetic disregulation in oral cancer. *International Journal of Molecular Sciences, 13*(2), 2331-2353.

Mashberg, A. (2000). Diagnosis of early oral and oropharyngeal squamous carcinoma: Obstacles and their amelioration. *Oral Oncol, 36*(3), 253-255.

Masthan, K., Babu, N. A., Dash, K. C., & Elumalai, M. (2012). Advanced diagnostic aids in oral cancer. *Asian Pacific Journal of Cancer Prevention, 13*(8), 3573-3576.

Mathivanan, S., Ji, H., & Simpson, R. J. (2010). Exosomes: Extracellular organelles important in intercellular communication. *Journal of Proteomics, 73*(10), 1907-1920.

McDermott, J. E., Wang, J., Mitchell, H., Webb-Robertson, B.-J., Hafen, R., Ramey, J., & Rodland, K. D. (2013). Challenges in biomarker discovery: Combining expert insights with statistical analysis of complex omics data. *Expert Opinion on Medical Diagnostics, 7*(1), 37-51.

Meldrum, C., Doyle, M. A., & Tothill, R. W. (2011). Next-generation sequencing for cancer diagnostics: A practical perspective. *Clinical Biochemist, 32*(4), 177-195.

Melo, S. A., Luecke, L. B., Kahlert, C., Fernandez, A. F., Gammon, S. T., Kaye, J., . . . Rahbari, N. (2015). Glypican-1 identifies cancer exosomes and detects early pancreatic cancer. *Nature, 523*(7559), 177.

Mendes, R. A. (2013). Oncogenic pathways in the development of oral cancer. *Journal of Carcinogenesis & Mutagenesis, 2012*.

Mertes, F., ElSharawy, A., Sauer, S., van Helvoort, J. M., Van Der Zaag, P., Franke, A., . . . Brookes, A. J. (2011). Targeted enrichment of genomic DNA regions for next-generation sequencing. *Briefings in Functional Genomics, 10*(6), 374-386.

Mesri, M. (2014). Advances in proteomic technologies and its contribution to the field of cancer. *Advances in Medicine, 2014*.

Metzker, M. L. (2010). Sequencing technologies—the next generation. *Nature Reviews Genetics, 11*(1), 31-46.

Meurman, J. H. (2010). Infectious and dietary risk factors of oral cancer. *Oral Oncology, 46*(6), 411-413.

Mikuriya, K., Kuramitsu, Y., Ryozawa, S., Fujimoto, M., Mori, S., Oka, M., . . . Nakamura, K. (2007). Expression of glycolytic enzymes is increased in pancreatic cancerous tissues as evidenced by proteomic profiling by two-dimensional electrophoresis and liquid chromatography-mass spectrometry/mass spectrometry. *International Journal of Oncology, 30*(4), 849-855.

Milac, T. I., Randolph, T. W., & Wang, P. (2012). Analyzing lc-ms/ms data by spectral count and ion abundance: Two case studies. *Statistics and its interface, 5*(1), 75.

Minciacchi, V. R., Freeman, M. R., & Di Vizio, D. (2015). *Extracellular vesicles in cancer: Exosomes, microvesicles and the emerging role of large oncosomes.* Paper presented at the Seminars in Cell and Developmental Biology.

Mizuno, H., Honda, M., Shirasaki, T., Yamashita, T., Yamashita, T., Mizukoshi, E., & Kaneko, S. (2012). Heterogeneous nuclear ribonucleoprotein a2/b1 in association with htert is a potential biomarker for hepatocellular carcinoma. *Liver International, 32*(7), 1146-1155.

Mou, Z., He, Y., & Wu, Y. (2009). Immunoproteomics to identify tumor-associated antigens eliciting humoral response. *Cancer Letters, 278*(2), 123-129.

Mu, A. K.-W., Chan, Y. S., Kang, S. S., Azman, S. N., Zain, R. B., Chai, W. L., & Chen, Y. (2014). Detection of host-specific immunogenic proteins in the saliva of patients with oral squamous cell carcinoma. *Journal of Immunoassay and Immunochemistry, 35*(2), 183-193.

Muller, P. A., & Vousden, K. H. (2014). Mutant p53 in cancer: New functions and therapeutic opportunities. *Cancer Cell, 25*(3), 304-317.

Murugan, A. K., Munirajan, A. K., & Tsuchida, N. (2012). Ras oncogenes in oral cancer: The past 20 years. *Oral Oncology, 48*(5), 383-392.

Murugan, A. K., Munirajan, A. K., & Tsuchida, N. (2013). Genetic deregulation of the pik3ca oncogene in oral cancer. *Cancer Letters, 338*(2), 193-203.

Nagaraj, N. S. (2009). Evolving 'omics' technologies for diagnostics of head and neck cancer. *Briefings in Functional Genomics and Proteomics, 8*(1), 49-59.

Nagpal, J. K., & Das, B. R. (2003). Oral cancer: Reviewing the present understanding of its molecular mechanism and exploring the future directions for its effective management. *Oral Oncology, 39*(3), 213-221.

Nakagaki, T., Tamura, M., Kobashi, K., Koyama, R., Fukushima, H., Ohashi, T., . . . Tokino, T. (2017). Profiling cancer-related gene mutations in oral squamous cell carcinoma from japanese patients by targeted amplicon sequencing. *Oncotarget, 8*(35), 59113.

Nawroz, H., van der Riet, P., Hruban, R. H., Koch, W., Ruppert, J. M., & Sidransky, D. (1994). Allelotype of head and neck squamous cell carcinoma. *Cancer Research, 54*(5), 1152-1155.

Nazki, F. H., Sameer, A. S., & Ganaie, B. A. (2014). Folate: Metabolism, genes, polymorphisms and the associated diseases. *Gene, 533*(1), 11-20.

Neville, B. W., & Day, T. A. (2002). Oral cancer and precancerous lesions. *CA Cancer J Clin, 52*(4), 195-215.

Neville, B. W., & Day, T. A. (2002). Oral cancer and precancerous lesions. *CA: A Cancer Journal for Clinicians, 52*(4), 195-215.

Ng, S. B., Buckingham, K. J., Lee, C., Bigham, A. W., Tabor, H. K., Dent, K. M., . . . Nickerson, D. A. (2010). Exome sequencing identifies the cause of a mendelian disorder. *Nature Genetics, 42*(1), 30.

Ni, Y. h., Ding, L., Hu, Q. g., & Hua, Z. C. (2015). Potential biomarkers for oral squamous cell carcinoma: Proteomics discovery and clinical validation. *PROTEOMICS–Clinical Applications, 9*(1-2), 86-97.

Ogbureke, K. U., & Bingham, C. (2012). Overview of oral cancer *Oral cancer*: InTech.

Okamoto, J., Onda, M., Hirata, T., Miyamoto, S., Akaishi, J., Mikami, I., . . . Shimizu, K. (2006). Dissimilarity in gene expression profiles of lung adenocarcinoma in japanese men and women. *Gender Medicine, 3*(3), 223-235.

Okano, T., Kondo, T., Kakisaka, T., Fujii, K., Yamada, M., Kato, H., . . . Hirohashi, S. (2006). Plasma proteomics of lung cancer by a linkage of multi-dimensional liquid chromatography and two-dimensional difference gel electrophoresis. *Proteomics, 6*(13), 3938-3948.

Olsson, M., & Zhivotovsky, B. (2011). Caspases and cancer. *Cell Death and Differentiation, 18*(9), 1441.

Omar, E. (2015). Current concepts and future of noninvasive procedures for diagnosing oral squamous cell carcinoma-a systematic review. *Head & Face Medicine, 11*(1), 6.

Omar, Z. A., Ali, Z. M., & Tamin, N. S. I. (2006). Malaysian cancer statistics-data and figure, peninsular malaysia 2006. *National cancer Registry, Ministry of Health Malaysia*.

Ono, K., Sugahara, K., Nomura, T., Takano, N., Shibahara, T., & Katakura, A. (2014). Multiple hpv subtypes infection in japanese oral squamous cell carcinoma. *Journal of Oral and Maxillofacial Surgery, Medicine, and Pathology, 26*(2), 128-132.

Orgel, J., San Antonio, J., & Antipova, O. (2011). Molecular and structural mapping of collagen fibril interactions. *Connective Tissue Research, 52*(1), 2-17.

Pang, W. W., Abdul-Rahman, P. S., Izlina Wan-Ibrahim, W., & Haji Hashim, O. (2010). Can the acute-phase reactant proteins be used as cancer biomarkers? *International Journal of Biological Markers, 25*(1), 1.

Patel, V. J., Thalassinos, K., Slade, S. E., Connolly, J. B., Crombie, A., Murrell, J. C., & Scrivens, J. H. (2009). A comparison of labeling and label-free mass spectrometry-based proteomics approaches. *Journal of Proteome Research, 8*(7), 3752-3759.

Perez-Sayans, M., Somoza-Martin, J. M., Barros-Angueira, F., Reboiras-Lopez, M. D., Gándara Rey, J. M., & García-García, A. (2009). Genetic and molecular alterations associated with oral squamous cell cancer (review). *Oncology Reports, 22*(6), 1277.

Petersen, P. E., Bourgeois, D., Ogawa, H., Estupinan-Day, S., & Ndiaye, C. (2005). The global burden of oral diseases and risks to oral health. *Bulletin of the World Health Organization, 83*, 661-669.

Petti, S., & Scully, C. (2010). Determinants of oral cancer at the national level: Just a question of smoking and alcohol drinking prevalence? *Odontology, 98*(2), 144-152.

Pickering, C. R., Zhang, J., Yoo, S. Y., Bengtsson, L., Moorthy, S., Neskey, D. M., . . . Drummond, J. (2013). Integrative genomic characterization of oral squamous cell carcinoma identifies frequent somatic drivers. *Cancer Discovery, 3*(7), 770-781.

Polz-Gruszka, D., Macielag, P., Foltyn, S., & Polz-Dacewicz, M. (2014). Oral squamous cell carcinoma (oscc)-molecular, viral and bacterial concepts. *Journal of Pre-Clinical and Clinical Research, 8*(2).

Qing, S., Tulake, W., Ru, M., Li, X., Yuemaier, R., Lidifu, D., . . . Rouziahong, R. (2017). Proteomic identification of potential biomarkers for cervical squamous cell carcinoma and human papillomavirus infection. *Tumor Biology, 39*(4), 1010428317697547.

Rabbani, B., Tekin, M., & Mahdieh, N. (2014). The promise of whole-exome sequencing in medical genetics. *Journal of Human Genetics, 59*(1), 5.

Raglow, Z., & Thomas, S. M. (2015). Tumor matrix protein collagen xiα1 in cancer. *Cancer Letters, 357*(2), 448-453.

Rahmani, A., Alzohairy, M., Babiker, A. Y., Rizvi, M. A., & Elkarimahmad, H. G. (2012). Clinicopathological significance of pten and bcl2 expressions in oral squamous cell carcinoma. *International Journal of Clinical and Experimental Pathology, 5*(9), 965-971.

Ralhan, R. (2007). Diagnostic potential of genomic and proteomic signatures in oral cancer. *International Journal of Human Genetics, 7*(1), 57.

Rao, S. V. K., Mejia, G., Roberts-Thomson, K., & Logan, R. (2013). Epidemiology of oral cancer in asia in the past decade-an update (2000-2012). *Asian Pacific Journal of Cancer Prevention, 14*(10), 5567-5577.

Raposo, G., & Stoorvogel, W. (2013). Extracellular vesicles: Exosomes, microvesicles, and friends. *Journal of Cell Biology, 200*(4), 373-383.

Rappa, G., Santos, M. F., Green, T. M., Karbanová, J., Hassler, J., Bai, Y., . . . Lorico, A. (2017). Nuclear transport of cancer extracellular vesicle-derived biomaterials through nuclear envelope invagination-associated late endosomes. *Oncotarget, 8*(9), 14443.

Redondo, M., Rodrigo, I., Alcaide, J., Tellez, T., Roldan, M. J., Funez, R., . . . Jiménez, E. (2010). Clusterin expression is associated with decreased disease-free survival of patients with colorectal carcinomas. *Histopathology, 56*(7), 932-936.

Rezende, C. P. d., Ramos, M. B., Daguíla, C. H., Dedivitis, R. A., & Rapoport, A. (2008). Oral health changes in with oral and oropharyngeal cancer. *Revista Brasileira de Otorrinolaringologia, 74*(4), 596-600.

Rezende, T. M. B., Freire, M. d. S., & Franco, O. L. (2010). Head and neck cancer: Proteomic advances and biomarker achievements. *Cancer, 116*(21), 4914-4925.

Ribeiro, I. P., Barroso, L., Marques, F., Melo, J. B., & Carreira, I. M. (2016). Early detection and personalized treatment in oral cancer: The impact of omics approaches. *Molecular Cytogenetics, 9*(1), 85.

Rifai, N., Gillette, M. A., & Carr, S. A. (2006). Protein biomarker discovery and validation: The long and uncertain path to clinical utility. *Nature Biotechnology, 24*(8), 971.

Riley, L. B., & Desai, D. C. (2009). The molecular basis of cancer and the development of targeted therapy. *Surgical Clinics of North America, 89*(1), 1-15.

Rizzi, F., & Bettuzzi, S. (2010). The clusterin paradigm in prostate and breast carcinogenesis. *Endocrine-Related Cancer, 17*(1), R1-R17.

Rizzo, G., Black, M., Mymryk, J., Barrett, J., & Nichols, A. (2015). Defining the genomic landscape of head and neck cancers through next-generation sequencing. *Oral Diseases, 21*(1), e11-e24.

Rocha-Zavaleta, L., Ambrosio, J. P., de Lourdes Mora-Garcia, M., Cruz-Talonia, F., Hernandez-Montes, J., Weiss-Steider, B., . . . Monroy-Garcia, A. (2004). Detection of antibodies against a human papillomavirus (hpv) type 16 peptide that differentiate high-risk from low-risk hpv-associated low-grade squamous intraepithelial lesions. *Journal of General Virology, 85*(9), 2643-2650.

Rodríguez-Piñeiro, A. M., de la Cadena, M. P., López-Saco, Á., & Rodríguez-Berrocal, F. J. (2006). Differential expression of serum clusterin isoforms in colorectal cancer. *Molecular & Cellular Proteomics, 5*(9), 1647-1657.

Roesch-Ely, M., Nees, M., Karsai, S., Ruess, A., Bogumil, R., Warnken, U., . . . Hofele, C. (2007). Proteomic analysis reveals successive aberrations in protein expression from healthy mucosa to invasive head and neck cancer. *Oncogene, 26*(1), 54.

Rossiter, D. (2012). Introduction to the r project for statistical computing for use at itc. *International Institute for Geo-information Science & Earth Observation (ITC), Enschede (NL), 3*, 3-6.

Ruddon, R. W. (2007). *Cancer biology*: Oxford University Press.

Saberkari, H., Shamsi, M., Heravi, H., & Sedaaghi, M. H. (2013). A novel fast algorithm for exon prediction in eukaryotic genes using linear predictive coding model and goertzel algorithm based on the z-curve. *International Journal of Computer Applications, 67*(17).

Sailasree, R., Abhilash, A., Sathyan, K., Nalinakumari, K., Thomas, S., & Kannan, S. (2008). Differential roles of p16ink4a and p14arf genes in prognosis of oral carcinoma. *Cancer Epidemiology and Prevention Biomarkers, 17*(2), 414-420.

Saini, R., Tang, T.-H., Zain, R. B., Cheong, S. C., Musa, K. I., Saini, D., . . . Santhanam, J. (2011). Significant association of high-risk human papillomavirus (hpv) but not of p53 polymorphisms with oral squamous cell carcinomas in malaysia. *Journal of Cancer Research and Clinical Oncology, 137*(2), 311-320.

Sakai, E., Rikimaru, K., Ueda, M., Matsumoto, Y., Ishii, N., Enomoto, S., . . . Tsuchida, N. (1992). The p53 tumor-suppressor gene and ras oncogene mutations in oral squamous-cell carcinoma. *International Journal of Cancer, 52*(6), 867-872.

Samuels, Y., Wang, Z., Bardelli, A., Silliman, N., Ptak, J., Szabo, S., . . . Riggins, G. J. (2004). High frequency of mutations of the pik3ca gene in human cancers. *Science, 304*(5670), 554-554.

Sand, L., & Jalouli, J. (2014). Viruses and oral cancer. Is there a link? *Microbes and infection, 16*(5), 371-378.

Santilli, G., Aronow, B. J., & Sala, A. (2003). Essential requirement of apolipoprotein j (clusterin) signaling for iκb expression and regulation of nf-κb activity. *Journal of Biological Chemistry, 278*(40), 38214-38219.

Santosh, A. B. R., Jones, T., & Harvey, J. (2016). A review on oral cancer biomarkers: Understanding the past and learning from the present. *Journal of Cancer Research and Therapeutics, 12*(2), 486.

Scaltriti, M., Brausi, M., Amorosi, A., Caporali, A., D'Arca, D., Astancolle, S., . . . Bettuzzi, S. (2004). Clusterin (sgp-2, apoj) expression is downregulated in low-and high-grade human prostate cancer. *International Journal of Cancer, 108*(1), 23-30.

Schottenfeld, D. (2006). *Cancer epidemiology and prevention*: Oxford University Press.

Schulz, W. (2005). *Molecular biology of human cancers: An advanced student's textbook*: Springer Science & Business Media.

Scully, C., & Bagan, J. (2009). Oral squamous cell carcinoma: Overview of current understanding of aetiopathogenesis and clinical implications. *Oral Diseases, 15*(6), 388-399.

Scully, C., & Bedi, R. (2000). Ethnicity and oral cancer. *The Lancet Oncology, 1*(1), 37-42.

Scully, C., Field, J., & Tanzawa, H. (2000). Genetic aberrations in oral or head and neck squamous cell carcinoma (scchn): 1. Carcinogen metabolism, DNA repair and cell cycle control. *Oral Oncology, 36*(3), 256-263.

Sethi, M. K., Thaysen-Andersen, M., Kim, H., Park, C. K., Baker, M. S., Packer, N. H., . . . Fanayan, S. (2015). Quantitative proteomic analysis of paired colorectal

cancer and non-tumorigenic tissues reveals signature proteins and perturbed pathways involved in crc progression and metastasis. *Journal of Proteomics, 126*, 54-67.

Shah, J. P., & Singh, B. (2006). Keynote comment: Why the lack of progress for oral cancer? *The Lancet Oncology, 7*(5), 356-357.

She, Y., Nong, X., Zhang, M., & Wang, M. (2017). Epstein-barr virus infection and oral squamous cell carcinoma risk: A meta-analysis. *PloS One, 12*(10), e0186860.

Shen, T., Yeat, N. C., & Lin, J. C.-H. (2015). Clinical applications of next generation sequencing in cancer: From panels, to exomes, to genomes. *Frontiers in genetics, 6*, 215.

Sherr, C. J. (2004). Principles of tumor suppression. *Cell, 116*(2), 235-246.

Shevchenko, A., Wilm, M., Vorm, O., & Mann, M. (1996). Mass spectrometric sequencing of proteins from silver-stained polyacrylamide gels. *Analytical Chemistry, 68*(5), 850-858.

Shyr, D., & Liu, Q. (2013). Next generation sequencing in cancer research and clinical application. *Biological Procedures Online, 15*(1), 1.

Singh, V., Husain, N., Akhtar, N., Kumar, V., Tewari, S., Mishra, S., . . . Khan, M. (2015). Do human papilloma viruses play any role in oral squamous cell carcinoma in north indians. *Asian Pacific Journal of Cancer Prevention, 16*, 7077-7084.

Song, H.-J., Xue, Y.-L., Qiu, Z.-L., & Luo, Q.-Y. (2013). Comparative serum proteomic analysis identified afamin as a downregulated protein in papillary thyroid carcinoma patients with non-131i-avid lung metastases. *Nuclear Medicine Communications, 34*(12), 1196.

Srinivas, P. R., Verma, M., Zhao, Y., & Srivastava, S. (2002). Proteomics for cancer biomarker discovery. *Clinical Chemistry, 48*(8), 1160-1169.

Stejskal, D., & Fiala, R. R. (2006). Evaluation of serum and urine clusterin as a potential tumor marker for urinary bladder cancer. *Neoplasma, 53*(4), 343-346.

Stransky, N., Egloff, A. M., Tward, A. D., Kostic, A. D., Cibulskis, K., Sivachenko, A., . . . McKenna, A. (2011). The mutational landscape of head and neck squamous cell carcinoma. *Science, 333*(6046), 1157-1160.

Strimbu, K., & Tavel, J. A. (2010). What are biomarkers? *Current Opinion in HIV and AIDS, 5*(6), 463.

Su, S.-C., Lin, C.-W., Liu, Y.-F., Fan, W.-L., Chen, M.-K., Yu, C.-P., . . . Li, W.-H. (2017). Exome sequencing of oral squamous cell carcinoma reveals molecular subgroups and novel therapeutic opportunities. *Theranostics, 7*(5), 1088.

Summerer, D. (2009). Enabling technologies of genomic-scale sequence enrichment for targeted high-throughput sequencing. *Genomics, 94*(6), 363-368.

Sur, I., Neumann, S., & Noegel, A. A. (2014). Nesprin-1 role in DNA damage response. *Nucleus, 5*(2), 173-191.

Tainsky, M. A. (2009). Genomic and proteomic biomarkers for cancer: A multitude of opportunities. *Biochimica et Biophysica Acta (BBA)-Reviews on Cancer, 1796*(2), 176-193.

Tandale, A., Joshi, M., & Sengupta, D. (2016). Structural insights and functional implications of inter-individual variability in β 2-adrenergic receptor. *Scientific Reports, 6*, 24379.

Tanskanen, T., Gylfe, A. E., Katainen, R., Taipale, M., Renkonen-Sinisalo, L., Järvinen, H., . . . Pitkänen, E. (2015). Systematic search for rare variants in finnish early-onset colorectal cancer patients. *Cancer Genetics, 208*(1), 35-40.

Team, R. C. (2013). R: A language and environment for statistical computing.

Termine, N., Giovannelli, L., Rodolico, V., Matranga, D., Pannone, G., & Campisi, G. (2012). Biopsy vs. Brushing: Comparison of two sampling methods for the detection of hpv-DNA in squamous cell carcinoma of the oral cavity. *Oral Oncology, 48*(9), 870-875.

Thompson, M. P., & Kurzrock, R. (2004). Epstein-barr virus and cancer. *Clinical Cancer Research, 10*(3), 803-821.

Tian, M., Cui, Y.-Z., Song, G.-H., Zong, M.-J., Zhou, X.-Y., Chen, Y., & Han, J.-X. (2008). Proteomic analysis identifies mmp-9, dj-1 and a1bg as overexpressed proteins in pancreatic juice from pancreatic ductal adenocarcinoma patients. *BMC Cancer, 8*(1), 241.

Torre, L. A., Bray, F., Siegel, R. L., Ferlay, J., Lortet-Tieulent, J., & Jemal, A. (2015). Global cancer statistics, 2012. *CA: A Cancer Journal for Clinicians, 65*(2), 87-108.

Tran, B., Dancey, J. E., Kamel-Reid, S., McPherson, J. D., Bedard, P. L., Brown, A. M., . . . Stein, L. (2012). Cancer genomics: Technology, discovery, and translation. *Journal of Clinical Oncology, 30*(6), 647-660.

Tsantoulis, P., Kastrinakis, N., Tourvas, A., Laskaris, G., & Gorgoulis, V. (2007). Advances in the biology of oral cancer. *Oral Oncology, 43*(6), 523-534.

Tsunoda, S., Smith, E., De Young, N. J., Wang, X., Tian, Z.-Q., Liu, J.-F., . . . Drew, P. A. (2009). Methylation of cldn6, fbn2, rbp1, rbp4, tfpi2, and tmeff2 in esophageal squamous cell carcinoma. *Oncology Reports, 21*(4), 1067-1073.

Tu, H.-F., Chang, K.-W., Chiang, W.-F., Liu, C.-J., Yu, E.-H., Liu, S.-T., & Lin, S.-C. (2011). The frequent co-expression of the oncogenes pik3ca and pak1 in oral carcinomas. *Oral Oncology, 47*(3), 211-216.

Turner, S. D. (2014). Qqman: An r package for visualizing gwas results using qq and manhattan plots. *bioRxiv*, 005165.

van den Brule, A. J., Pol, R., Fransen-Daalmeijer, N., Schouls, L. M., Meijer, C. J., & Snijders, P. J. (2002). Gp5+/6+ pcr followed by reverse line blot analysis enables rapid and high-throughput identification of human papillomavirus genotypes. *Journal of Clinical Microbiology, 40*(3), 779-787.

Varela, I., Tarpey, P., Raine, K., Huang, D., Ong, C. K., Stephens, P., . . . Teague, J. (2011). Exome sequencing identifies frequent mutation of the swi/snf complex gene pbrm1 in renal carcinoma. *Nature, 469*(7331), 539-542.

Vargas-Ferreira, F., Nedel, F., Etges, A., Gomes, A. P. N., Furuse, C., & Tarquinio, S. B. C. (2012). Etiologic factors associated with oral squamous cell carcinoma in non-smokers and non-alcoholic drinkers: A brief approach. *Brazilian Dental Journal, 23*(5), 586-590.

Voelkerding, K. V., Dames, S., & Durtschi, J. D. (2010). Next generation sequencing for clinical diagnostics-principles and application to targeted resequencing for hypertrophic cardiomyopathy: A paper from the 2009 william beaumont hospital symposium on molecular pathology. *The Journal of molecular diagnostics, 12*(5), 539-551.

Vogelstein, B., & Kinzler, K. W. (1993). The multistep nature of cancer. *Trends in Genetics, 9*(4), 138-141.

Vogelstein, B., & Kinzler, K. W. (2004). Cancer genes and the pathways they control. *Nature Medicine, 10*(8), 789-799.

Vohra, S., & Biggin, P. C. (2013). Mutationmapper: A tool to aid the mapping of protein mutation data. *PloS One, 8*(8), e71711.

Vokes, E. E., Weichselbaum, R. R., Lippman, S. M., & Hong, W. K. (1993). Head and neck cancer. *New England Journal of Medicine, 328*(3), 184-194.

Walden, M. J., & Aygun, N. (2013). *Head and neck cancer.* Paper presented at the Seminars in Roentgenology.

Wang, H. Y. (2007). Laser capture microdissection in comparative proteomic analysis of hepatocellular carcinoma. *Methods in Cell Biology, 82*, 689-707.

Wang, J., Zuo, Y., Man, Y.-g., Avital, I., Stojadinovic, A., Liu, M., . . . Ressom, H. W. (2015). Pathway and network approaches for identification of cancer signature markers from omics data. *Journal of Cancer, 6*(1), 54.

Warnakulasuriya, S. (2009). Global epidemiology of oral and oropharyngeal cancer. *Oral Oncology, 45*(4), 309-316.

Warnakulasuriya, S. (2010). Living with oral cancer: Epidemiology with particular reference to prevalence and life-style changes that influence survival. *Oral Oncology, 46*(6), 407-410.

Warr, A., Robert, C., Hume, D., Archibald, A., Deeb, N., & Watson, M. (2015). Exome sequencing: Current and future perspectives. *G3: Genes| Genomes| Genetics, 5*(8), 1543-1550.

Wei, L., Xue, T., Wang, J., Chen, B., Lei, Y., Huang, Y., . . . Xin, X. (2009). Roles of clusterin in progression, chemoresistance and metastasis of human ovarian cancer. *International Journal of Cancer, 125*(4), 791-806.

Wei, X., Walia, V., Lin, J. C., Teer, J. K., Prickett, T. D., Gartner, J., . . . Gershenwald, J. E. (2011). Exome sequencing identifies grin2a as frequently mutated in melanoma. *Nature Genetics, 43*(5), 442.

Weinberg, R. A. (1991). Tumor suppressor genes. *Science, 254*(5035), 1138-1146.

Weivoda, S., Andersen, J. D., Skogen, A., Schlievert, P. M., Fontana, D., Schacker, T., . . . Jemmerson, R. (2008). Elisa for human serum leucine-rich alpha-2-glycoprotein-1 employing cytochrome c as the capturing ligand. *Journal of Immunological Methods, 336*(1), 22-29.

Weren, R. D., Venkatachalam, R., Cazier, J. B., Farin, H. F., Kets, C. M., De Voer, R. M., . . . Kamping, E. J. (2015). Germline deletions in the tumour suppressor gene focad are associated with polyposis and colorectal cancer development. *The Journal of Pathology, 236*(2), 155-164.

Whibley, C., Pharoah, P. D., & Hollstein, M. (2009). P53 polymorphisms: Cancer implications. *Nature Reviews Cancer, 9*(2), 95-107.

Wiener, M. C., Sachs, J. R., Deyanova, E. G., & Yates, N. A. (2004). Differential mass spectrometry: A label-free lc− ms method for finding significant differences in complex peptide and protein mixtures. *Analytical Chemistry, 76*(20), 6085-6096.

Wilkins, M. R., Appel, R. D., Van Eyk, J. E., Chung, M., Görg, A., Hecker, M., . . . Paik, Y. K. (2006). Guidelines for the next 10 years of proteomics. *Proteomics, 6*(1), 4-8.

Winder, D. M., Ball, S. L., Vaughan, K., Hanna, N., Woo, Y. L., Fränzer, J.-T., . . . Goon, P. K. (2009). Sensitive hpv detection in oropharyngeal cancers. *BMC Cancer, 9*(1), 440.

Wolfenson, H., Lavelin, I., & Geiger, B. (2013). Dynamic regulation of the structure and functions of integrin adhesions. *Developmental Cell, 24*(5), 447-458.

Wong, D., Todd, R., Tsuji, T., & Donoff, R. (1996). Molecular biology of human oral cancer. *Critical Reviews in Oral Biology and Medicine, 7*(4), 319-328.

Woong-Shick, A., Sung-Pil, P., Su-Mi, B., Joon-Mo, L., Sung-Eun, N., Gye-Hyun, N., . . . Chong-Kook, K. (2005). Identification of hemoglobin-α and-β subunits as potential serum biomarkers for the diagnosis and prognosis of ovarian cancer. *Cancer Science, 96*(3), 197-201.

Worthey, E. A., Mayer, A. N., Syverson, G. D., Helbling, D., Bonacci, B. B., Decker, B., . . . Veith, R. L. (2011). Making a definitive diagnosis: Successful clinical application of whole exome sequencing in a child with intractable inflammatory bowel disease. *Genetics in Medicine, 13*(3), 255-262.

Wouters, M. A., Rigoutsos, I., Chu, C. K., Feng, L. L., Sparrow, D. B., & Dunwoodie, S. L. (2005). Evolution of distinct egf domains with specific functions. *Protein Science, 14*(4), 1091-1103.

Wu, J.-Y., Yi, C., Chung, H.-R., Wang, D.-J., Chang, W.-C., Lee, S.-Y., . . . Yang, W.-C. V. (2010). Potential biomarkers in saliva for oral squamous cell carcinoma. *Oral Oncology, 46*(4), 226-231.

Wu, J., Xie, X., Nie, S., Buckanovich, R. J., & Lubman, D. M. (2013). Altered expression of sialylated glycoproteins in ovarian cancer sera using lectin-based elisa assay and quantitative glycoproteomics analysis. *Journal of Proteome Research, 12*(7), 3342-3352.

Wu, Y.-R., Chen, C.-M., Chen, Y.-C., Chao, C.-Y., Ro, L. S., Fung, H.-C., . . . Lee-Chen, G.-J. (2010). Ubiquitin specific proteases usp24 and usp40 and ubiquitin thiolesterase uchl1 polymorphisms have synergic effect on the risk of parkinson's disease among taiwanese. *Clinica Chimica Acta, 411*(13), 955-958.

Wu, Y., Chang, T., Huang, Y., Huang, H., & Chou, C. (2014). Col11a1 promotes tumor progression and predicts poor clinical outcome in ovarian cancer. *Oncogene, 33*(26), 3432-3440.

Wulfkuhle, J. D., Liotta, L. A., & Petricoin, E. F. (2003). Proteomic applications for the early detection of cancer. *Nature Reviews Cancer, 3*(4), 267-275.

Ye, B., Cramer, D. W., Skates, S. J., Gygi, S. P., Pratomo, V., Fu, L. F., . . . Mok, S. C. (2003). Haptoglobin-alpha subunit as potential serum biomarker in ovarian cancer: Identification and characterization using proteomic profiling and mass spectrometry. *Clinical Cancer Research, 9*(8), 2904-2911.

Yoshida, R., Nagata, M., Nakayama, H., Niimori-Kita, K., Hassan, W., Tanaka, T., . . . Ito, T. (2013). The pathological significance of notch1 in oral squamous cell carcinoma. *Laboratory Investigation, 93*(10), 1068-1081.

Yu, B., Li, S. Y., An, P., Zhang, Y. N., Liang, Z. J., Yuan, S. J., & Cai, H. Y. (2004). Comparative study of proteome between primary cancer and hepatic metastatic tumor in colorectal cancer. *World Journal of Gastroenterology, 10*(18), 2652-2656.

Zain, R., & Ghazali, N. (2001). A review of epidemiological studies of oral cancer and precancer in malaysia. *Annals of Dentistry University of Malaya, 8*, 50-56.

Zain, R. B., Athirajan, V., Ghani, W. M. N., Razak, I. A., Latifah, R. J. R., Ismail, S. M., . . . Hussien, A. (2013). An oral cancer biobank initiative: A platform for multidisciplinary research in a developing country. *Cell and tissue banking, 14*(1), 45-52.

Zehbe, I., & Wilander, E. (1996). Two consensus primer systems and nested polymerase chain reaction for human papillomavirus detection in cervical biopsies: A study of sensitivity. *Human Pathology, 27*(8), 812-815.

Zellweger, T., Chi, K., Miyake, H., Adomat, H., Kiyama, S., Skov, K., & Gleave, M. E. (2002). Enhanced radiation sensitivity in prostate cancer by inhibition of the cell survival protein clusterin. *Clinical Cancer Research, 8*(10), 3276-3284.

Zhang, J., Felder, A., Liu, Y., Guo, L. T., Lange, S., Dalton, N. D., . . . Shelton, G. D. (2009). Nesprin 1 is critical for nuclear positioning and anchorage. *Human Molecular Genetics, 19*(2), 329-341.

Zhang, L. Y., Ying, W. T., Mao, Y. S., He, H. Z., Liu, Y., Wang, H. X., . . . Zhao, X. H. (2003). Loss of clusterin both in serum and tissue correlates with the tumorigenesis of esophageal squamous cell carcinoma via proteomics approaches. *World Journal of Gastroenterology, 9*(4), 650-654.

Zhang, X., Zhou, H., Zhang, Y., Cai, L., Jiang, G., Li, A., . . . Wang, E. (2017). Znf452 facilitates tumor proliferation and invasion via activating akt-gsk3β signaling pathway and predicts poor prognosis of non-small cell lung cancer patients. *Oncotarget, 8*(24), 38863.

Zhao, J., Fan, Y.-X., Yang, Y., Liu, D.-L., Wu, K., Wen, F.-B., . . . Zhao, S. (2015). Identification of potential plasma biomarkers for esophageal squamous cell carcinoma by a proteomic method. *International Journal of Clinical and Experimental Pathology*, *8*(2), 1535.

Zhao, J., Zhu, Y., Boerwinkle, E., & Xiong, M. (2015). Pathway analysis with next-generation sequencing data. *European Journal of Human Genetics*, *23*(4), 507-515.

Zhao, Q., Kirkness, E. F., Caballero, O. L., Galante, P. A., Parmigiani, R. B., Edsall, L., . . . Vasconcelos, A. T. R. (2010). Systematic detection of putative tumor suppressor genes through the combined use of exome and transcriptome sequencing. *Genome Biology, 11*(11), R114.

Zhu, W., Smith, J. W., & Huang, C.-M. (2009). Mass spectrometry-based label-free quantitative proteomics. *BioMed Research International, 2010*.

LIST OF PUBLICATIONS AND PAPERS PRESENTED

List of Publications:

1. Kerishnan JP, Gopinath SC, Kai SB, Tang TH, Ng HL, Rahman ZA, Hashim U, Chen Y. (2016). Detection of Human Papillomavirus 16-Specific IgG and IgM Antibodies in Patient Sera: A Potential Indicator of Oral Squamous Cell Carcinoma Risk Factor. *International Journal of Medical Sciences*, 2016;13(6):424.
2. Chen Y, Azman SN, Kerishnan JP, Zain RB, Chen YN, Wong YL, Gopinath SC (2014). Identification of host-immune response protein candidates in the sera of human oral squamous cell carcinoma patients. *Plos One*, 1;9(10):e109012

List of Papers Presented:

1. Kerishnan, JP, Chen Y. (2018). Discovery of Potential Biomarkers in Oral Squamous Cell Carcinoma Using Next Generation Sequencing Technology. International Conference on Oral Immunology & Oral Microbiology (ICOIOM), Balai Ungku Aziz, Faculty of Dentistry, University of Malaya, Kuala Lumpur, 14 – 15 August 2018.
2. Kerishnan, JP, Chen Y. (2016). Molecular Landscape of Oral Squamous Cell Carcinoma through Next Generation Technology. Dental Congregation, The Royal Chulan Damansara, Petaling Jaya, Malaysia, 13 – 14 August 2016.
3. Kerishnan, JP. Mutation Landscape of Oral Squamous Cell Carcinoma Through the Use of Next Generation Technology. Three Minute Thesis, Faculty of Dentistry, University of Malaya (2015 and 2016).
4. Kerishnan, JP, Mah MK, Mohd Fawzi NA, Tang TH, Chen Y. (2014). Evaluation of Human Papilomavirus Antibodies as Oral Squamous Cell Carcinoma Risk Factors Indicator. International Association of Dental Research (IADR) and South East Asia Association of Dental Education (SEAADE) Annual Scientific Meeting Annual Scientific Meeting, Kuching, Sarawak, Malaysia, 11 – 14 August 2014.