ORAL CANCER GENOMICS DATA MINING AND INTEGRATION FOR PREDICTIVE THERAPEUTICS

BERNARD LEE KOK BANG

FACULTY OF DENTISTRY
UNIVERSITY OF MALAYA
KUALA LUMPUR

2019

University of Malaya

ORAL CANCER GENOMICS DATA MINING AND INTEGRATION FOR PREDICTIVE THERAPEUTICS

BERNARD LEE KOK BANG

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

FACULTY OF DENTISTRY
UNIVERSITY OF MALAYA
KUALA LUMPUR

2019

University of Malaya

UNIVERSITY OF MALAYA

ORIGINAL LITERARY WORK DECLARATION

Name of Candidate: Bernard Lee Kok Bang

Matric No: DHA150001

Name of Degree: Doctor of Philosophy (Bioinformatics)

Title of Project Paper/Research Report/Dissertation/Thesis ("this Work"): Oral cancer genomics data mining and integration for predictive therapeutics

Field of Study: Oral Cancer

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:

University of Malaya

ORAL CANCER GENOMICS DATA MINING AND INTEGRATION FOR PREDICTIVE THERAPEUTICS

ABSTRACT

Global oral cancer incidence and mortality rates are increasing rapidly, with more than 350 000 new cases and 170 000 deaths recorded in 2018. Depressingly, standard treatments for oral cancer such as surgery, chemotherapy, and radiotherapy are associated with significant morbidity and a relatively static 5-year survival rate of around 50 – 60%. To date, three drugs - cetuximab, pembrolizumab, and nivolumab, are available for treating oral cancer. However, only a small fraction of oral cancer patients respond to these drugs. Discovery of further efficacious drugs in a cost-effective way through drug repurposing can potentially uncover the best combinatorial drug therapy against oral cancer. In this thesis, I aimed to create, using computational and statistical approaches, an integrative digital resource that can be mined to identify drug candidates that could be repurposed for oral cancer treatment. To this end, two bioinformatics tools were developed. The first tool – GENIPAC (Genomic Information Portal on Cancer Cell Lines), is a web resource for exploring, visualising, and analysing genomics information from 44 head and neck cancer cell lines. The second tool – DeSigN (Differentially Expressed Gene Signatures - Inhibitors), links the gene expression of oral cancer cell lines to the publicly available gene expression databases that have drug sensitivity data. To validate the efficacy of drug candidate shortlisted by DeSigN on a panel of oral cancer cell lines, several *in vitro* experiments were performed. Using gene expression signatures retrieved from the ORL Series in GENIPAC, DeSigN predicted bosutinib, an Src/Abl kinase inhibitor used for treating leukemia, to have inhibitory effect on oral cancer cell lines. Subsequent *in vitro* drug sensitivity validation showed that these oral cancer cell lines were susceptible to bosutinib treatment at IC_{50} of 0.8 – 1.2 μM. Later, anti-proliferative experiments confirmed the efficacy of bosutinib in controlling tumour

University of Malaya

growth in oral cancer cell lines. Technical evaluation of performance reliability of six gene signature similarity scoring algorithms showed that the Weighted Connectivity Score or the statistically significant Connectivity Map, are prime candidates for upgrading the current core algorithm of DeSigN, which is based on the Kolmogorov-Smirnov statistic. In conclusion, the present work has demonstrated that cancer genomics data mining and integration through GENIPAC and DeSigN is a viable approach to accelerating the drug development process for oral cancer. Importantly, application of these two tools led to the discovery of bosutinib as a new, promising drug candidate to be repurposed for treating oral cancer in the future.

Keywords: Connectivity Map, oral cancer, gene expression, gene signature similarity scoring algorithms, drug sensitivity

University of Malaya

PERLOMBONGAN DAN PERSEPADUAN DATA GENOMIK KANSER MULUT UNTUK RAMALAN TERAPEUTIK

ABSTRAK

Kadar kejadian dan kematian kanser mulut global meningkat dengan pesat, mencatatkan lebih daripada 350 000 kes baru dan 170 000 kematian pada tahun 2018. Yang menyedihkan, rawatan piawai untuk kanser mulut seperti pembedahan, kemoterapi dan radioterapi adalah dikaitkan dengan kematian yang nyata dan secara relatifnya kadar hidup 5 tahun adalah kekal sekitar 50 – 60%. Sehingga kini, tiga dadah – cetuximab, pembrolizumab, dan nivolumab boleh didapati untuk merawat kanser mulut. Namun demikian, hanya sebahagian kecil pesakit-pesakit kanser mulut yang bertindak balas terhadap dadah-dadah tersebut. Penemuan dadah mujarab yang berterusan dengan cara yang kos efektif melalui penggunaan semula dadah berpotensi untuk menyerlahkan kombinasi terapi dadah yang terbaik terhadap kanser mulut. Dalam tesis ini, saya mempunyai matlamat untuk menciptakan satu sumber digital bersepadu yang boleh dilombong, dengan menggunakan pendekatan-pendekatan pengkomputeran dan statistik, bagi mengenal pasti calon-calon dadah yang berkemungkinan untuk diguna semula untuk rawatan kanser mulut. Sehingga kini, dua perkakasan bioinformatik telah dibangunkan. Perkakasan yang pertama – GENIPAC (*Genomic Information Portal on Cancer Cell Lines*) merupakan sumber web untuk meneroka, menggambarkan dan menganalisis maklumat genomik daripada 44 susuran sel kanser leher dan kepala. Perkakasan yang kedua – DeSigN (*Differentially Expressed Gene Signatures - Inhibitors*) menghubungkan ekspresi gen susuran sel kanser mulut terhadap pangkalan data umum ekspresi gen yang mengandungi data kepekaan dadah. Beberapa eksperimen *in vitro* telah dijalankan untuk mengesahkan kemujaraban calon dadah yang disenaraipendekkan oleh DeSigN terhadap satu panel susuran sel kanser mulut. Dengan menggunakan corak ekspresi gen yang diperoleh daripada *ORL Series* dalam GENIPAC, DeSigN telah meramalkan bahawa

University of Malaya

bosutinib, suatu perencat kinase Src/Abl yang digunakan untuk merawat leukemia, mempunyai kesan perencatan terhadap susuran sel kanser mulut. Pengesahan kepekaan dadah secara *in vitro* yang berikutnya menunjukkan bahawa susuran sel kanser mulut adalah peka terhadap rawatan bosutinib pada nilai IC_{50} 0.8 – 1.2 μM. Selanjutnya, eksperimen anti-proliferasi telah mengesahkan kemujaraban bosutinib dalam mengawal pertumbuhan tumor dalam susuran sel kanser mulut. Penilaian teknikal dari segi kebolehpercayaan prestasi enam algoritma pemarkahan corak gen seiras menunjukkan bahawa *Weighted Connectivity Score* atau *statistically significant Connectivity Map* merupakan calon-calon algoritma utama untuk menaik taraf algoritma teras DeSigN sedia ada yang berasaskan statistik Kolmogorov-Smirnov. Kesimpulannya, hasil kerja ini telah menunjukkan bahawa perlombongan dan integrasi data genomik kanser melalui GENIPAC dan DeSigN merupakan pendekatan yang berdaya maju dalam mempercepatkan proses pembangunan dadah untuk kanser mulut. Yang pentingnya, aplikasi kedua-dua perkakasan tersebut telah membawa kepada penemuan bosutinib sebagai satu calon dadah yang baru dan boleh diharapkan untuk diguna semula bagi merawat kanser mulut pada masa depan.

Kata kunci: *Connectivity Map*, kanser mulut, ekspresi gen, algoritma pemadanan corak gen yang seiras, kepekaan dadah

University of Malaya

ACKNOWLEDGEMENTS

I am very grateful to my supervisors, Prof. Dato' Dr. Zainal Ariff Abdul Rahman, Prof. Dr. Cheong Sok Ching, and Dr. Khang Tsung Fei for their invaluable vision, support, encouragement, and advice given throughout this project. I appreciate the excellent grounding that they have given me in oral cancer research, as well as the opportunity to learn how to be a versatile researcher.

Special thanks to the Department of Oral and Maxillofacial Clinical Sciences and the staff of Oral Cancer Research and Coordinating Centre (OCRCC), Faculty of Dentistry, University of Malaya for their efforts in hosting me to carry out this research. I would also like to thank the staff of Cancer Research Malaysia, especially members of the Head and Neck Cancer Research Team for their invaluable support and input on this project.

The success of the present work is the result of excellent collaborative work with the most dedicated local and international collaborators. Specifically, I would like to thank Assoc. Prof. Dr. Aik Choon Tan (formerly at the University of Colorado), who has given me much mentorship in my Ph.D. journey. In particular, I had the excellent opportunity to spend three months in Dr. Tan's lab, where much of the improvement of DeSigN as a drug repurposing tool through performance evaluation of different gene signature similarity scoring algorithms were done under his guidance. Thanks and appreciation also go to members of the Data Intensive Computing Centre (DICC), University Malaya, especially Dr. Liew Chee Sun and Chang Jit Kang. Both of them have been instrumental in setting up the user interface for DeSigN and GENIPAC and for hosting these resources in their high-performance computers so that researchers from around the world can access these tools freely. To Dr. Tan Joon Liang from Multimedia University, Melaka, I thank you for helping with the analysis of copy number variation data for our cell lines. A personal token of gratitude has to be given to Dr. Silvio Gutkind and his post-doctoral

University of Malaya

scientist, Dr. Daniel Martin for kindly sharing with me the genomics data of OPC-22 lines, which are now hosted in GENIPAC.

I wish to extend my gratitude to my beloved family members, especially to my ever-supporting father; my wife, Dr. Ong Hui San; my daughter, Erin Lee Ching Lin; as well as my late mother for their ever-lasting love and support in my pursuit of this study. To them, I dedicate this thesis.

Last but not least, I would like to acknowledge funding from the University of Malaya High Impact Research Grant from the Ministry of Higher Education (HIR-MOHE) (UM.C/625/1/HIR/MOHE/DENT-03) and sponsorship of my Ph.D. studies by the Ong Hin Tiang & Ong Sek Pek Foundation from 2016 to 2019.

University of Malaya

TABLE OF CONTENTS

University of Malaya

University of Malaya

University of Malaya

University of Malaya

LIST OF FIGURES

University of Malaya

University of Malaya

LIST OF TABLES

University of Malaya

University of Malaya

LIST OF SYMBOLS AND ABBREVIATIONS

AHR	:	Aryl hydrocarbon receptor
ASR	:	Age-standard rate
CCL	:	Cancer cell lines
CCLE	:	Cancer Cell Line Encyclopedia
CCND1	:	Cyclin D1
CMap	:	Connectivity Map
CS	:	Connectivity score
CTRP	:	Cancer Therapeutics Response Portal
DEG	:	Differentially expressed gene
ECDF	:	Empirical cumulative distribution function
EGFR	:	Epidermal growth factor receptor
ES	:	Enrichment score
FDA	:	US Food and Drug Administration
GDSC	:	Genomics of Drug Sensitivity in Cancer
GSEA	:	Gene Set Expression Analysis
HNSCC	:	Head and neck squamous cell carcinoma
KS statistic	:	Kolmogorov-Smirnov statistic
MoA	:	Mechanism of action
NCI	:	National Cancer Institute
OSCC	:	Oral squamous cell carcinoma
NOK	:	Normal oral keratinocytes
OPMD	:	Oral potentially malignant disorders
PPV	:	Positive predictive value
SEA	:	South East Asia

University of Malaya

sscMap	:	Statistically significant Connectivity Map
TCGA	:	The Cancer Genome Atlas
WTCS	:	Weighted Connectivity Score
XCos	:	eXtreme Cosine
XSum	:	eXtreme Sum

University of Malaya

LIST OF APPENDICES

University of Malaya

University of Malaya

CHAPTER 1: INTRODUCTION

1.1 Background

Oral cancer is among the most devastating head and neck squamous cell carcinoma (HNSCC) subtypes. The incidence and mortality rates are growing worldwide, recording more than 350 000 new cases and 170 000 deaths in 2018 based on the report from GLOBOCAN 2018 (Bray et al., 2018). While HNSCC that is detected early can be effectively treated with surgery and radiotherapy (Gilyoma et al., 2015; Joshi et al., 2014), about 75% of patients are diagnosed at a late stage where treatment options become limited. This is reflected in the overall 5-year survival rate of about 60% (Marur & Forastiere, 2016). In the Malaysian context, more than 70% of the oral cancer patients are diagnosed in their advanced stage with poor survival (Ghani et al., 2019).

Presently, three targeted therapies have so far been approved by the US Food and Drug Administration (FDA) to treat oral cancer. Cetuximab, a monoclonal antibody that inhibits epidermal growth factor receptor (EGFR) signaling, has been the only molecular-targeted therapy approved for the treatment of recurrent and metastatic HNSCC for the past ten years (Vermorken et al., 2008). Only very recently two inhibitors of the immune checkpoint molecule PD-1: pembrolizumab, and nivolumab have been approved for the treatment of platinum-refractory HNSCC (Bauml et al., 2017; Ferris et al., 2016). While this is an improvement in the repertoire of therapeutic options for recurrent and metastatic HNSCC, these treatments are only effective in less than 20% of HNSCC patients (Bauml et al., 2017; Ferris et al., 2016; Mehra et al., 2018), thus underscoring the urgent need to develop more effective therapies and those that are associated with less side effects.

One of the innovative approaches to identifying effective therapies is to match inherent gene expression signatures with potentially efficacious drug candidates. This concept was

University of Malaya

first demonstrated through the Connectivity Map (CMap) project by Lamb et al. in 2006. One key component of CMap concept is the 'gene expression changes', which is used to connect a disease-specific gene signature (up-regulated and down-regulated genes) to a reference database containing drug-specific gene expression profiles. Following the inception of CMap, more recently, a couple of large-scale public pharmacogenomic studies, such as the Genomics of Drug Sensitivity in Cancer (GDSC) (Garnett et al., 2012; Iorio et al., 2016), Cancer Cell Line Encyclopedia (CCLE) (Barretina et al., 2012), and Cancer Therapeutics Response Portal (CTRP) (Basu et al., 2013) have since been developed. While CMap focuses on drug-induced gene expression profiles, these newer pharmacogenomic studies instead emphasise on the drug sensitivity response and characterise the genomic profiles of more than a thousand cancer cell lines (CCL) at the baseline level. Notably, more than 700 drugs have since been tested on these CCL, representing one of the most substantial endeavour reported so far in trying to identify lists of drugs that could potentially be efficacious against certain cancers.

While these public pharmacogenomic databases are a valuable resource for association studies between genomic features and drug response, they cannot be readily integrated with experimental data generated by individual research laboratories. For example, Hsp90 inhibitor 17-AAG was shown to have a favourable response against the HNSCC cell lines (Garnett et al., 2012). However, predicting which new cell lines derived from HNSCC patients that are likely to respond to 17-AAG remains challenging.

Fortunately, the availability of these open-source pharmacogenomics studies offers an unprecedented opportunity for developing practical computational algorithms that could leverage on the availability of the comprehensive drug response as well as gene expression data. The use of computational algorithms to mine and integrate genomics data of cancers with public pharmacogenomics database will accelerate the identification

University of Malaya

of molecular features in cancers that are associated with sensitivity to specific drugs. Thus, the development of computational algorithm that could predict drug sensitivity in CCL is particularly crucial for cancers with limited therapeutic options, such as oral cancer.

1.2 Aims and Objectives

In this study, I aim to create an integrative resource for HNSCC that can be mined to repurpose existing drugs for effective treatment of oral cancer. To this end, four objectives will be met. They are listed as follows:

(i) To develop a user-friendly web resource for exploring, visualising, and analysing genomics information of commonly-used head and neck CCL;

(ii) To develop computational approaches that can associate the gene expression profile of oral CCL of interest to gene expression profiles that are augmented with drug sensitivity data in publicly available databases;

(iii) To experimentally validate the computational prediction of the approach in objective (ii) on oral CCL;

(iv) To evaluate different gene signature similarity scoring algorithms for optimal drug sensitivity prediction.

University of Malaya

CHAPTER 2: LITERATURE REVIEW

2.1 Oral Cancer

2.1.1 Epidemiology

Head and neck squamous cell carcinoma (HNSCC) (C00-C13) refers to a heterogeneous group of tumours that originate from various tissue types along the upper aerodigestive tract. It is the sixth most common cancer worldwide based on the GLOBOCAN 2018 report (Bray et al., 2018). Oral squamous cell carcinoma (OSCC) (C00-C06), meanwhile, is the most common subtype of HNSCC.

GLOBOCAN 2018 reported more than 350 000 new cases and 170 000 deaths due to oral cancer in 2018. Of these, approximately 65% (227 906 new cases) occurred in Asia (Figure 2.1) (Bray et al., 2018). Similarly, the Asian continent reported the highest number of deaths due to this disease, with 129 939 patients reported to have succumbed to oral cancer in 2018 (Figure 2.1). Notably, within countries in Asia, about 68% of the new cases were from the South Asian countries (India, Pakistan, Bangladesh, and Sri Lanka) where the incidence of oral cancer is also among the highest in the world (Figure 2.2). Similarly, about 73% of the death cases (n = 94 537) were from these four South Asian countries (Figure 2.2). These alarming incidence and mortality rates are in part attributed to risk factors such as smoking, tobacco chewing (with or without areca nut) and/or heavy alcohol drinking (Sankaranarayanan et al., 2013).

University of Malaya

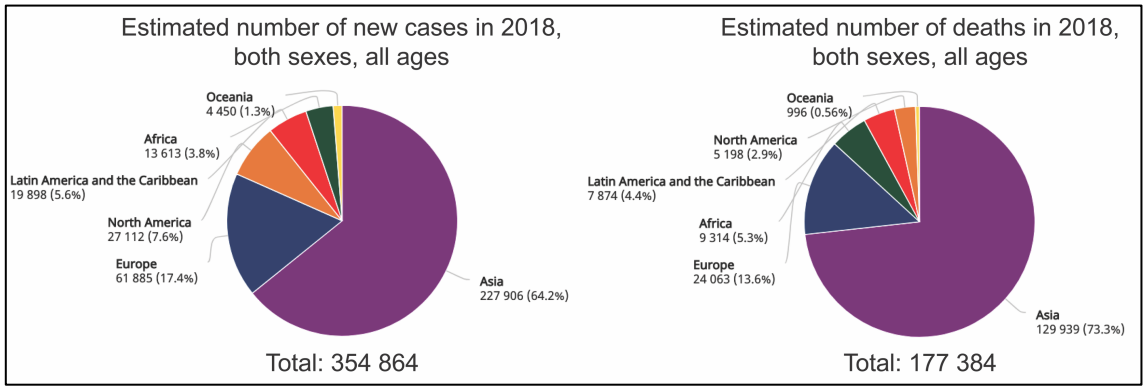

Figure 2.1: Incidence and mortality rates of oral cancer in 2018 for both sexes at all ages according to different continents. The Asian continent has the highest incidence and mortality rates, with 227 906 new cases and 129 939 deaths occurring in 2018. The data was retrieved and adapted from GLOBOCAN 2018 (URL: https://gco.iarc.fr/today/home).

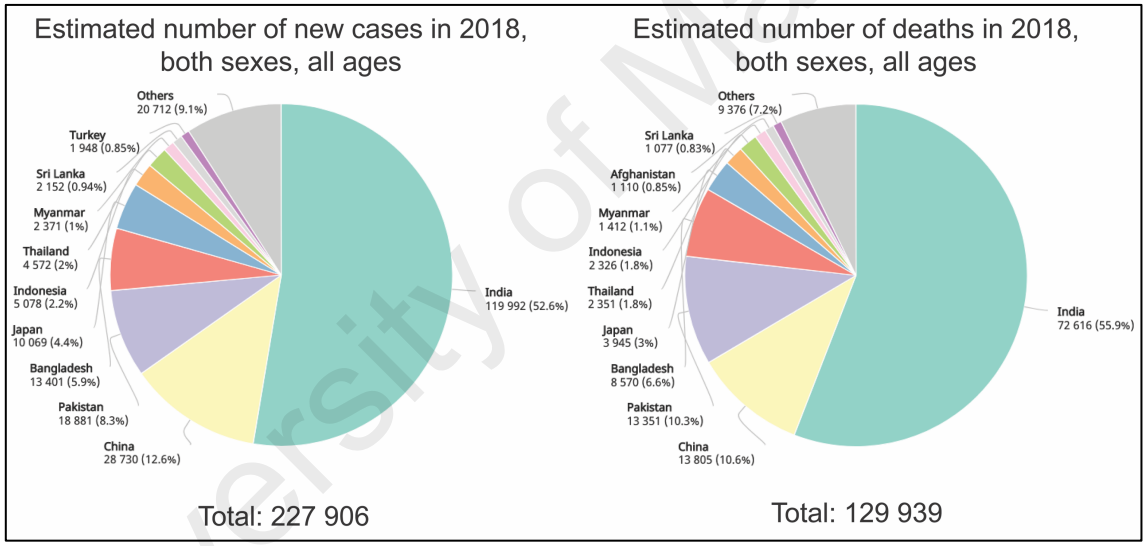

Figure 2.2: Top ten incidence and mortality rates of oral cancer for countries in Asia for both sexes at all ages. Oral cancer is frequently diagnosed in the South Asian countries such as India, Pakistan, Bangladesh, and Sri Lanka. South East Asian (SEA) countries such as Indonesia, Thailand, and Myanmar were also listed amongst the top ten countries in Asia for both incidence and mortality rates for oral cancer. The data was retrieved and adapted from GLOBOCAN 2018 (URL: https://gco.iarc.fr/today/).

Likewise, the incidence rate for oral cancer in South East Asian (SEA) (comprises of 11 countries) has been regarded as alarmingly high for many years (Warnakulasuriya, 2009). Although one report stated that oral cancer is more commonly diagnosed in females in Khon Kaen, Thailand (Vatanasapt et al., 2011), generally GLOBOCAN 2018

University of Malaya

reported that oral cancer is still a male-dominant disease, a trend that is shared across the globe as well as in SEA countries (Table 2.1). Focusing on the epidemiology data for SEA countries more closely, an estimation of 10 234 males and 6584 females were diagnosed with oral cancer in 2018, with a male-to-female ratio of 1.55:1. There is a marked variation in the age-standard rate (ASR) across the different SEA countries ranging from 0.81 to 6.9 per 100 000 for males, with Myanmar having the highest incidence; and 0.61 to 3.1 per 100 000 for females, with Thailand having the highest (Table 2.1). Meanwhile, the mortality to incidence ratio in SEA has been reported previously to be among the highest in Asia (Ng et al., 2015; Vatanasapt et al., 2011), and in 2018, the mortality due to oral cancer in SEA was estimated as 8542 cases, where 5327 and 3215 were men and women respectively, with a male-to-female ratio of 1.66:1 (Table 2.1). Based on the estimated deaths due to oral cancer, a wide range is observed across the SEA countries. Among males, the mortality rates were 1.2 to 4.4 per 100 000 persons, with Myanmar having the highest rates. Among female, the ASR was 0.42 to 2.1 per 100 000 persons, with Cambodia having the highest mortality rates (Table 2.1). Notably, looking at the ASR of incidence for oral cancer across the world as reported in GLOBOCAN 2018, only the females from Thailand (ranked 18; ASR: 3.1 per 100 000 persons) and Cambodia (ranked 20; ASR: 3.1 per 100 00 persons) were listed among the top 20 countries with the highest incidence of oral cancer. In terms of ASR of mortality, males from Myanmar (ranked 19; ASR: 4.4 per 100 000 persons) and females from Cambodia (ranked 14; ASR: 2.1 per 100 000 persons) and Laos (ranked 20; ASR: 1.8 per 100 000 persons) were among the top 20 countries in the world.

In the Malaysian setting, based on the GLOBOCAN 2018 report, the estimated number of cases of oral cancer for males was 335 (ASR = 2.2/100 000) and 332 for females (ASR = 2.1/100 000) (Table 2.1). As for the mortality rate, the number of cases of death was 179 (ASR = 1.2/100 000) for males and 148 (ASR = 0.94/100 000) for females (Table

University of Malaya

2.1). Concurrently, the latest Malaysian National Cancer Registry Report (2007-2011), which was published in 2016 (Azizah et al., 2016) stated that HNSCC, inclusive of oral cancer, is the fourth most common cancer amongst all ethnicity. Looking specifically at oral cancer (C00: lip, C01-C02: tongue, and C03-C06: mouth) and in accordance to the same clinical cataloguing system (International Statistical Classification of Diseases and Related Health Problems-10th Revision codes C00-C97) used by GLOBOCAN 2018, oral cancer was ranked as the 16th most common cancer across all ethnicities between 2007 and 2011 in Malaysia. Although it is not amongst the top ten cancers in Malaysia, it was ranked the sixth most common cancer amongst males (ASR = 4.8/100 000) and second for females (ASR = 10.0/100 000) of Indian origin.

Table 2.1: Estimated incidence and mortality rate of oral cancer in 2018 in SEA countries according to sex (GLOBOCAN 2018). Abbreviation: ASR = age-standard rate per 100 000 populations.

Population	Incidence		Mortality	
	Male (ASR)	Female (ASR)	Male (ASR)	Female (ASR)
Indonesia	3132 (2.5)	1946 (1.5)	1508 (1.2)	818 (0.63)
Thailand	2545 (5.1)	2027 (3.1)	1299 (2.6)	1052 (1.6)
Myanmar	1652 (6.9)	719 (2.5)	1012 (4.4)	400 (1.4)
Vietnam	1308 (2.6)	569 (0.92)	639 (1.3)	283 (0.42)
Philippines	813 (2.1)	614 (1.3)	430 (1.2)	297 (0.66)
Malaysia	335 (2.2)	332 (2.1)	179 (1.2)	148 (0.94)
Cambodia	213 (4.3)	211 (3.1)	141 (3)	136 (2.1)
Singapore	141 (2.8)	84 (1.4)	62 (1.2)	33 (0.56)
Laos	90 (3.8)	78 (3.1)	53 (2.3)	45 (1.8)
Timor-Leste	4 (1.2)	3 (0.87)	4 (1.2)	3 (0.87)
Brunei	1 (0.81)	1 (0.61)	-	-
Total	10 234	6 584	5 327	3 215

2.1.2 Risk Factors Associated with Oral Cancer

From the risk factors point of view, oral cancer is most commonly associated with the use of tobacco, both smoked and smokeless. This is most prevalent in South and SEA countries. For example, Indonesia and Timor-Leste are amongst the countries with the

University of Malaya

highest tobacco smoking rates in the world, where 72.3% and 96.5% respectively of the male population smoke (Sreeramareddy et al., 2014). In contrast, women from the SEA are among the highest users of smokeless tobacco globally (Sreeramareddy et al., 2014). In SEA, smokeless tobacco is often used as one of the ingredients of betel quid, a mixture of substances that contain areca nut, slaked lime, and other condiments (Boucher & Mannan, 2002). Notably, areca nut itself is a carcinogen (Secretan et al., 2009); the use of betel quid with or without smokeless tobacco is highly associated with oral potentially malignant disorders (OPMD) and oral cancer of the population in SEA (Kampangsri et al., 2013; Loyha et al., 2012). A recent report stated that 19.7% of women in Cambodia indulged in betel quid chewing and this was the most potent risk factor associated with OPMD with a relative risk of 6.7 (Chher et al., 2018).

2.1.3 Prognosis and Treatment of Oral Cancer

The prognosis for HNSCC is highly heterogeneous, with an average 5-year survival rate of around 60% (Marur & Forastiere, 2016). For patients who experience locoregionally recurrent or metastatic oral cancer, median survival is 8-10 months (Zandberg & Strome, 2014). In most cases, therapeutic options for HNSCC patients consist of either radical surgery, surgery plus neoadjuvant or postoperative radiation therapy, and/or chemotherapy and targeted therapies (Leemans et al., 2011). According to the National Comprehensive Cancer Network (NCCN) guidelines for oral cancer treatment, if a tumour is restricted to a limited region, surgery and radiation therapy would be the treatments of choice. In the event the cancer cells have spread into lymph nodes and distant parts of the body, a combination of therapies would be applied depending on the extent of the disease. This could include an addition of radiation and/or chemotherapy (cisplatin) following surgery. In the recurrent and metastatic setting, targeted therapy

University of Malaya

(cetuximab), and immunotherapy (pembrolizumab and nivolumab) are also indicated (Bauml et al., 2017; Ferris et al., 2016; Vermorken et al., 2007).

The chemotherapeutic agents currently approved by the US Food and Drug Administration (FDA) for the treatment of HNSCC include cisplatin, methotrexate, 5-fluorouracil (5-FU), bleomycin, and docetaxel. The treatment choice of either concomitant platinum-based chemoradiotherapy (CRT) or surgery followed by adjuvant radiation or chemoradiation is the current standard of care for patients with locally advanced (LA) HNSCC. For patients with recurrent and/or metastatic (R/M) HNSCC, platinum-based chemotherapy plus 5-FU has a response rate (RR) of 30-40% and median survival of 6-9 months (Cohen et al., 2004).

In contrast to standard cytotoxic chemotherapies, the research community is aiming to develop molecular-base targeted therapies that could offer more effective targeting of tumour cells based on the molecular mechanism driving the cancer. This was the basis for the development of the EGFR-targeted therapy cetuximab. In 2006, the US FDA approved cetuximab as a monoclonal antibody that inhibits epidermal growth factor receptor (EGFR) signaling. Cetuximab is approved to be used in combination with radiation for LA disease, in combination with platinum-based chemotherapy and 5-FU for first-line treatment of R/M HNSCC and as a monotherapy for R/M disease after patients fail platinum-based chemotherapy (Bonner et al., 2006; Vermorken et al., 2008; Vermorken et al., 2007). Cetuximab exerts anti-tumour activity by inhibiting cell proliferation, triggering antibody-dependent cell-mediated cytotoxicity and increasing the cytotoxic effects of chemotherapy and radiotherapy (Ang et al., 2002; Herbst & Hong, 2002; Needle, 2002; Schneider-Merck et al., 2010). However, HNSCC tumours display heterogeneity in drug response, with only 10% – 20% of patients reportedly having a favourable response to cetuximab as a monotherapy (Vermorken et al., 2007).

University of Malaya

Nonetheless, better clinical outcome was observed when cetuximab was used in combination with platinum-fluorouracil-based chemotherapy or radiotherapy (Bonner et al., 2006; Vermorken et al., 2008). For instance, the addition of cetuximab to platinum-fluorouracil chemotherapy improved overall survival (increased from 20% to 36%) when given as first-line treatment in patients with R/M HNSCC (Vermorken et al., 2008).

2.1.4 Immunotherapy of Oral Cancer

The better understanding of molecular targets of HNSCC, without doubt, has helped us to tailor better management strategies for HNSCC patients. Over the past years, one of the significant advancements in the field of cancer research is the success of immuno-oncology as a promising strategy for cancer therapy. The relevance of the PD-1: PD-L1 checkpoint in cancer immunity is highlighted by reports which demonstrate that blockade of PD-1 or PD-L1 by specific monoclonal antibodies can reverse the anergic state of tumour-specific T cells and thereby enhance the anti-tumour immunity (Dong et al., 2002; Strome et al., 2003). As a result, immune checkpoint inhibitors such as pembrolizumab or nivolumab, which target the interaction between programmed death receptor 1/programmed death ligand 1 (PD-1/PDL-1) and PDL-2, have been approved for the treatment of various malignancies (Bauml et al., 2017; Ferris et al., 2016; Mehra et al., 2018).

Following the failure of platinum-based chemotherapy, nivolumab, a monoclonal antibody that inhibits the interaction of the immune checkpoint receptor PD-1 with its ligands PD-L1 and PD-L2, has been approved as a single-agent in recurrent HNSCC patients. Ferris et al. (2016) in their phase III trial reported that an overall response rate of 13.3% (95% confidence interval (CI): [9.3%, 18.3%]) was observed in the nivolumab

University of Malaya

treatment group (n = 32 patients) versus 5.8% (95% CI: [2.4%, 11.6%]) in the standard-therapy group (n = 7) (CheckMate 141 ClinicalTrials.gov Identifier: NCT02105636).

Pembrolizumab, a monoclonal antibody with the same target as nivolumab, was also approved as a monotherapy in R/M HNSCC following the failure of platinum-based chemotherapy (Bauml et al., 2017; Seiwert et al., 2016; Sheth & Weiss, 2018). The evaluation of the efficacy of pembrolizumab on 171 HNSCC patients (phase II) by Bauml et al. (2017), reported an overall response rate of 16% (95% CI: [11%, 23%]). One patient achieved a complete response while 27 patients achieved partial response (KEYNOTE-055 ClinicalTrials.gov Identifier: NCT02255097).

2.2 Genomic Landscape of Cancer Cells

The few examples stated above show that cancer cells indeed display a broad spectrum of genetic alterations that include gene arrangements, point mutations, and gene amplification (Vargas & Harris, 2016). As defined by the National Cancer Institute (NCI), biomarkers are substances that are produced by cancer or by other cells of the body in response to cancer or certain benign (noncancerous) conditions. Most biomarkers are expressed at much higher levels in cancerous conditions as compared to the healthy cells.

Cancer biomarkers are used to help detect, diagnose, and manage some types of cancer. While it is true that targeted drugs work best when there is a biomarker, there are only a handful of cancer types such as breast, colorectal, leukemia, melanoma, and lung that have approved cancer biomarkers. Most cancers up to now do not have any approved and actionable biomarkers, and HNSCC is one of the cancers that has not received approved biomarkers by the US FDA. The current list of approved cancer biomarkers can be accessed through the NCI webpage: https://www.cancer.gov/about-cancer/diagnosis-staging/diagnosis/tumor-markers-fact-sheet#q1.

University of Malaya

More recently, with the advent of next generation sequencing, the genomics of cancers have been documented to unprecedented depth. For instance, the amplification of *CCND1* (cyclin D1) or the loss of *SMAD4* was shown to be associated with sensitivity to multiple EGFR-family inhibitors, including lapatinib and BIBW2992 (Garnett et al., 2012). Pharmacogenomics studies also identified elevated expression of the *AHR* gene (aryl hydrocarbon receptor) to be strongly correlated with sensitivity to the *MEK* inhibitor PD-0325901 in *NRAS*-mutant cancer cell lines (CCL), leading to the hypothesis that enhanced sensitivity of *NRAS*-mutant cell lines to *MEK* inhibitors might relate to a coexistent dependency on *AHR* function (Barretina et al., 2012). These data give rise to a slightly different context of identifying targeted therapies and their corresponding biomarkers where genetic patterns or gene expression signatures other than the genetic targets could be useful for predicting response to targeted therapies.

2.3 Gene Expression Patterns as an Alternative Drug Response Indicator

Besides examining the potential of using specific molecular targets as therapeutic targets, cancer researchers are turning attention to evaluate signatures of gene expression for their ability to help determine a patient's prognosis or response to therapy. For example, results of the NCI-sponsored Trial Assigning IndividuaLized Options for Treatment (Rx), or TAILORx (ClinicalTrials.gov Identifier: NCT00310180) showed that for women recently diagnosed with lymph node-negative, hormone receptor-positive, HER2-negative breast cancer who had undergone surgery, those with the lowest 21-gene (Oncotype Dx®) recurrence scores had low recurrence rates when given hormone therapy alone and thus can be spared chemotherapy (Sparano et al., 2015).

In fact, when examining the different types of molecular features including copy-number variation, gene expression, and whole exome sequencing, researchers reported

University of Malaya

that gene expression has the best predictive power for drug response (Costello et al., 2014). This conclusion was based on the 44 drug sensitivity prediction algorithms submitted by data scientists worldwide where mRNA gene expression microarrays were found to carry the most significant weight in their statistical models in predicting the sensitivity of 28 drugs on 53 breast CCL.

Indeed, large-scale human CCL pharmacogenomics studies such as GDSC reported the same observation. By using the *HSP90* inhibitor 17-AAG as an example, they found that the sets of genes overexpressed (up-regulated genes) in the sensitive CCL are down-regulated in the resistant CCL, and vice versa (Garnett et al., 2012) (Figure 2.3). These findings from large-scale pharmacogenomics exemplify the opportunity to predict drug response based on the gene expression signature.

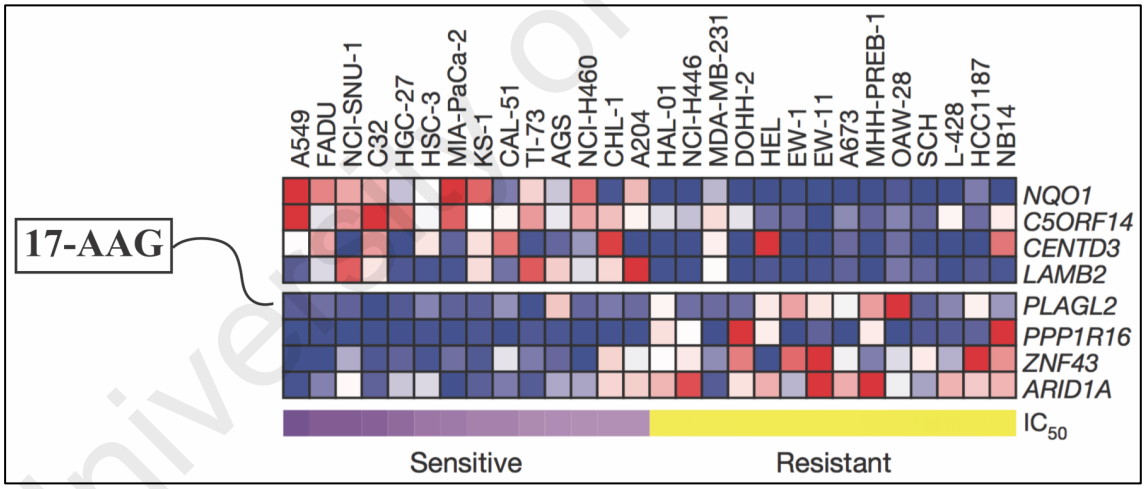

Figure 2.3: Heat map of highly significant genes associated with sensitivity and resistance to 17-AAG (HSP90 inhibitor). Cell line names are shown at the top of the heat map, followed by expression features (blue corresponds to lower expression, red for higher expression). To the right of the heat map is the list of genes that are associated with the response to 17-AAG. Bars in purple indicate expression features associated with sensitivity, and bars in yellow indicate features associated with resistance. In total, there are more than 250 drug sensitivity profiles currently hosted in GDSC web portal, with which each drug has its distinct gene expression signatures. Retrieved and adapted from (Garnett et al., 2012).

University of Malaya

One of the critical components to launch a clinical trial is to have an actionable molecular target to evaluate, such as oncogenic mutation of an essential gene. One such clinical trial success story concerns the patients with *BRAF* V600E mutation-positive metastatic melanoma that showed a response rate of approximately 50% to vemurafenib (*BRAF* inhibitor) (Chapman et al., 2011). In other words, the systematic collections of the patients' molecular profiles need to be in place before launching a clinical trial. Nonetheless, in the context of HNSCC, using gene expression signatures to predict drug response could be a more viable approach. This is mainly because thus far there are no apparent oncogenic mutations (except for *PIK3CA*) reported in HNSCC (Qiu et al., 2006) and the majority of the mutations reported in HNSCC are tumour suppressor mutations such as *TP53* and *NOTCH* (Cancer Genome Atlas Network, 2015) which are difficult to target therapeutically. This is because mutations will cause the inactivation or loss of normal cellular regulatory of tumour suppressor genes, and strategies to restore and maintain the functional copy of tumour suppressor genes to comparable level as in the normal cells have been proven to be technically challenging (Guo et al., 2014).

To facilitate genomics-driven drug response prediction, a key step is to set up a unified data repository that could host all available genomics data for HNSCC, in terms of the transcriptome, copy-number variation, and mutations data. These valuable genomic data could be shared amongst HNSCC researchers, thereby facilitating new biological discoveries as well as to promote quicker turnaround time for new treatment discoveries for HNSCC patients. In order to have an efficient way to share genomics information, one can take a lesson from the cBioPortal web portal set up by the Memorial Sloan-Kettering Cancer Center, USA. Five other multi-institutional teams, consisting of the Dana Farber Cancer Institute, Princess Margaret Cancer Centre in Toronto, Children's Hospital of Philadelphia, The Hyve in the Netherlands, and Bilkent University in Ankara, Turkey are also involved in setting up this comprehensive public cancer genomics

University of Malaya

database. The cBioPortal for Cancer Genomics (http://www.cbioportal.org/) is a web resource to explore, visualise, and analyse multidimensional cancer genomics data (Cerami et al., 2012; Gao et al., 2013). The cBioPortal currently provides access to genomic data from more than 10 000 tumour samples across 32 cancer types (as of April 8, 2019). By lowering the barriers of accessing complex genomics data, cBioPortal allows cancer researchers to translate large-scale cancer genomics datasets into biological insights and clinical applications.

To provide gene signatures as a means to predict drug response, a dedicated web resource for HNSCC cell line genomics data called GENIPAC will be set up as a research outcome of my Ph.D. study. The genomic information, particularly gene expression profiles, will be used to predict drugs that are efficacious against HNSCC. A detailed implementation of the GENIPAC database is given in Section 3.1.

2.4 The Connectivity Map Concept

One of the advancements in pharmacogenomics studies is the development of the Connectivity Map (CMap) (Lamb et al., 2006). The CMap concept is based on the observation that gene expression can be measured accurately and has shown promise as a "universal language" in disease characterisation and prognostication. Generally, the computational approach that utilised the CMap concept as the functional look-up table consists of three main components: a drug-sensitivity or drug perturbed gene expression database, a set of gene signatures given by users (a query), and a gene signature similarity scoring algorithm that correlates the user-defined gene signatures to the gene expression profiles in the reference database.

University of Malaya

The CMap database contains microarray-based gene expression profiles from cultured human cancer cell cells treated using a wide range of experimentally and clinically-used small molecules. Its goal is to create an extensive public database that collects as many genomics and drugs signatures as possible, where one then can query the CMap data using the web-based gene signature similarity scoring algorithm by inputting a gene expression profile of interest. The outcome of this similarity search is a list of ranked CMap drugs. A drug sensitivity prediction tool called DeSigN that leveraged on the concept of CMap will be built in this thesis. The DeSigN workflow will be described in detail in Section 3.2.

2.4.1 The CMap Datasets

The inception of first-generation CMap (Build 1) saw a total of 164 distinct small-molecule perturbagens profiled on five CCL, i.e., MCF7, ssMCF7 (breast), PC3 (prostate), HL60 (leukemia), and SKMEL5 (melanoma). To widen the coverage of the gene expression profiles, these cell lines were screened on 42 different concentrations (0.01 nM – 10 μM) at two time points: six, and 12 hours. A treatment "instance" was defined relative to three control treatments: DMSO, ethanol, or complete medium. These data were collected using Affymetrix GeneChip microarrays, HG-U133A (22 277 probe sets) and HT_HG-U133A (22 283 probe sets) and were preprocessed using the standard MAS 5.0 algorithm for microarrays. In total, 564 gene expression profiles were produced, representing 453 individual instances (i.e., one treatment-vehicle pair).

The updated version of CMap (Build 2) contains 6100 instances of unique treatment-control pairs, where treatment constitutes a selection of 1309 drugs, 156 different concentrations (0.01 nM – 10 μM), two time points (six hours and 12 hours) and five cell lines (HL60, MCF7, ssMCF7, PC3, and SKMEL5) against vehicle controls (either

University of Malaya

DMSO, ethanol or complete medium) for a parallel series of analysis. On top of the two Affymetrix GeneChip microarrays used previously in Build 1, one additional Affymetrix GeneChip microarray, HT_HG-U133A_EA (22 944 probe sets) was used to process the data in this updated CMap Build 2 version.

In CMap, a non-parametric, rank-based gene signature similarity scoring strategy based on the Kolmogorov-Smirnov (KS) statistic (Smirnov, 1939) was devised to detect similarities between the query signatures and the drug signatures of the reference gene expression profiles in the CMap dataset (Lamb et al., 2006). A query signature is any list of rank-ordered genes whose expression is correlated with a biological state of interest, carrying a sign that indicates whether it is up-regulated or down-regulated. Examples could be genes correlated with different time points of treatment (72 hours versus 24 hours) or enriched in specific biological pathways. The reference gene expression profiles in the CMap dataset are also represented in a non-parametric fashion. The genes on the array are sorted into decreasing order according to their differential expression values relative to the vehicle control, converted to a rank vector separately for each instance.

The query signature is then compared to each list of rank-ordered genes in the reference profile to determine whether up-regulated query genes tend to appear near the top of the list and down-regulated query genes near the bottom ("positive connectivity") or vice versa ("negative connectivity"), yielding a connectivity score (CS) ranging from -1 to +1. All instances in the database are then ranked according to their CS; those at the top are positively correlated to the query signatures, and those at the bottom are negatively correlated (Figure 2.4). The CMap Build 2 can be freely accessed at https://portals.broadinstitute.org/cmap/.

University of Malaya

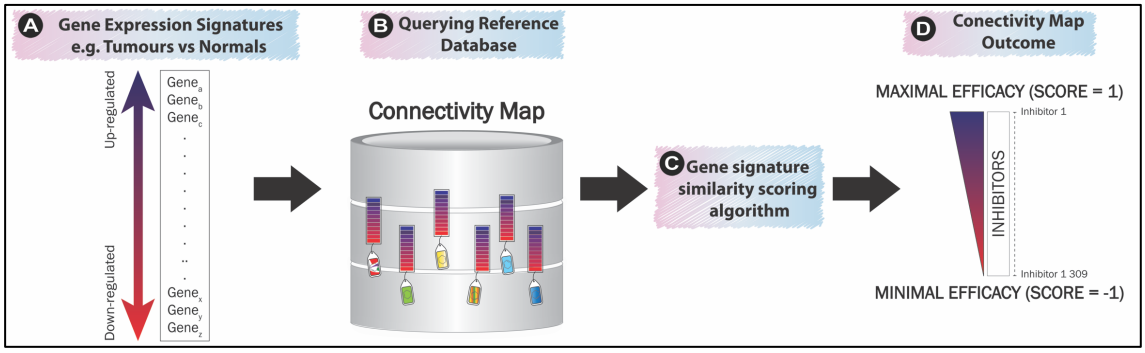

Figure 2.4: The CMap workflow. Users provide a pair of up-regulated and down-regulated genes (A) to query the CMap reference database (B). A gene signature similarity analysis would then be carried out using a gene signature similarity scoring algorithm to compute the gene expression similarity between the user-defined gene signatures and the reference profile (C). The outcome is a ranked list of inhibitors, with a CS ranging between 1 (maximal efficacy) and -1 (minimal efficacy) (D).

2.4.2 Application of CMap Datasets

Using CMap as the reference database, Jahchan et al. (2013) identified the tricyclic antidepressants (TCA) inhibitors as potent inducers of cell death in small cell lung cancer (SCLC) cells. They showed that treatment with two such TCA inhibitors: imipramine and promethazine disrupted the autocrine survival signals involving neurotransmitters and their G protein-coupled receptors. The potential of repurposing TCA inhibitors, as seen for treating SCLC, was also observed in other neuroendocrine tumours, such as Merkel cell carcinoma, and neuroblastoma tumour cells, thus highlighting the importance of autocrine mechanisms in promoting the growth of neuroendocrine tumour cells. Their findings led to the initiation of a phase IIA clinical trial, assessing the efficacy of the TCA inhibitor desipramine in SCLC and other high-grade neuroendocrine tumours (ClinicalTrials.gov Identifier: NCT01719861).

In combating epithelial ovarian cancer (EOC) through the identification of novel therapeutics, Raghavan et al. (2016) used the EOC gene expression signatures derived from The Cancer Genome Atlas (TCGA) ($n = 407$) and Mayo Clinic ($n = 326$) participants to query CMap. They identified 11 drugs to have potential efficacy on EOC. Notably,

University of Malaya

five of those drugs (mitoxantrone, podophyllotoxin, wortmannin, doxorubicin, and 17-AAG) were known *a priori* to be cytotoxic to the EOC cells. A significant reduction in cell viability was observed upon treatment of these five drugs on a set of 10 EOC cell lines following 72 hours of drug treatment. Therefore, it will be interesting to know how the remaining short-listed six drugs would fare when tested *in vitro*.

In the context of HNSCC, Wei et al. (2019) used 401 differentially expressed genes (201 up-regulated and 200 down-regulated genes) obtained from two public databases: TCGA and Genotype-Tissue Expression Project (GTEx) to query the CMap and discovered that most of these genes are highly dysregulated in cell cycle and p53 signaling pathway. A further protein-protein interactions (PPI) analysis found that these highly dysregulated genes form two hub genes: *PCNA* and *CCND1*. In total, 22 drugs corresponding to the two pathways were chosen as the candidate drugs for HNSCC, and seven of these drugs had no previous indication for cancer-combating properties. Subsequent molecule docking analysis revealed that two drugs: bepridil and MG-262, have a strong binding affinity with *PCNA*, suggesting their possible roles in perturbing the development of HNSCC through targeting the *PCNA* gene.

In addition to the CMap reference database, several public pharmacogenomic databases that incorporate high-throughput drug testing on several orders of magnitude more cell lines as compared to CMap have started to emerge more recently. In this thesis, the Genomics of Drug Sensitivity (GDSC) study will be the key pharmacogenomic database used to develop the drug repurposing tool meant for predicting potential drugs for effective treatment of oral cancers.

University of Malaya

2.5 The Pharmacogenomic Datasets

2.5.1 Genomics of Drug Sensitivity in Cancer

The Genomics of Drug Sensitivity in Cancer (GDSC) database (https://www.cancerrxgene.org/) is one of the most extensive public resources for information on drug sensitivity in cancer cells and molecular markers of drug response. In 2012, GDSC launched their first version of the datasets, containing drug sensitivity data for almost 75 000 experiments, describing the response of 138 anticancer drugs across almost 700 CCL (Garnett et al., 2012; Yang et al., 2013). GDSC provides unique resources incorporating enormous drug sensitivity and genomic datasets to facilitate the discovery of new therapeutic targets for cancer therapies. The collection of compounds available in GDSC include cytotoxic chemotherapeutics as well as targeted therapeutics from commercial sources, academic collaborations, and the biotechnology and pharmaceutical industries.

The updated version (2016) of the GDSC currently has more than a thousand CCL genomics datasets (Iorio et al., 2016). The genomic information available for each cell line includes somatic mutation of 75 cancer genes, genome-wide gene copy number for amplification and deletion, targeted screening for seven gene rearrangement, markers of microsatellite instability, tissue type and transcriptional data. Various statistical approaches, such as multivariate analysis of variance (MANOVA) and elastic net regression, are used to correlate drug sensitivity with genomic alterations in cancer.

The number of cell lines available in GDSC varies according to different tissue types (Table 2.2). For example, the lung has the highest number of cell lines (n = 215), while the thyroid has only 17 cell lines currently hosted in GDSC. Meanwhile, HNSCC, forming part of the aerodigestive tract, has 42 cell lines in GDSC. Due to its large number of cell lines as well as drug sensitivity data, GDSC datasets were used as the reference

University of Malaya

profile in this thesis for drug sensitivity prediction. The detailed implementation of GDSC datasets as the drug sensitivity reference database will be described in Section 3.2.1.

Table 2.2: Breakdown of the number of cell lines based on tissue types in GDSC (version 2016).

Tissue type	Number of cell lines
Lung	215
Blood	182
Urogenital system	114
Digestive system	107
Nervous system	92
Aerodigestive tract	82
Skin	67
Breast	53
Bone	44
Kidney	35
Pancreas	32
Soft tissue	21
Thyroid	17

There is, however, one pitfall with regards to GDSC drug sensitivity datasets that one must take note. In many cases, the IC_{50} values of the tested drugs could not be computed for all cell lines, as the drug concentration necessary to inhibit 50% of the cell's growth was not reached. As depicted in Figure 2.5, with the screening concentration of palbociclib between 0.0156 μM and 4 μM, only about 43% (n = 367) of the 852 cell lines have IC_{50} values that fall within this screening concentration. For the rest of the 485 cell lines, a Bayesian sigmoid model is used to extrapolate their IC_{50} values.

University of Malaya

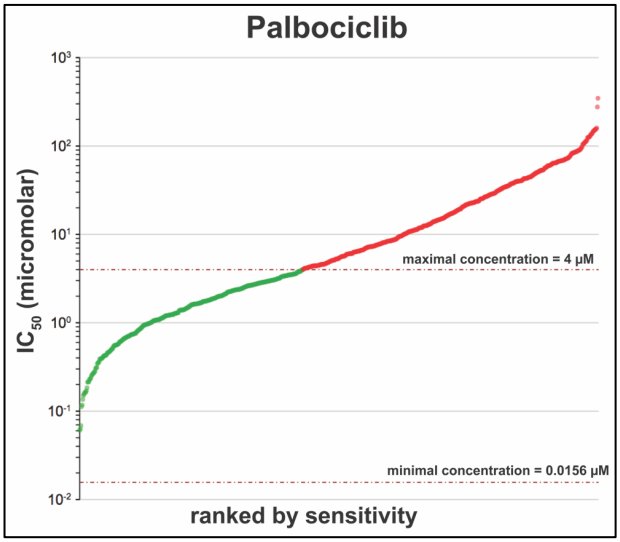

Figure 2.5: The IC_{50} values of the cytostatic drug palbociclib treated on 852 cell lines. The screening concentration ranges from 0.0156 µM (minimal) to 4 µM (maximal). The green dots represent cell lines with IC_{50} values that fall within the tested screening concentration, while red dots represent cell lines with extrapolated IC_{50} values estimated using the Bayesian sigmoid model. Retrieved and adapted from GDSC web portal: https://www.cancerrxgene.org/.

2.5.1.1 Application of GDSC Datasets

Bladder cancer remains one of the most deadly cancer diseases, with roughly 79 000 new cancer cases and 17 000 cancer-related deaths reported in the United States in 2017 (Siegel et al., 2017). Adopting the idea that biomarkers of therapeutic response developed in one cancer type can be effectively applied across multiple cancer types (Barretina et al., 2012; Garnett et al., 2012; Goodspeed et al., 2016), Goodspeed et al. (2018) first derived a novel 67-gene signature from 68 colorectal cancer patients that was associated with sensitivity response to several *EGFR* inhibitors. Using this 67-gene signature that is known for association with response to cetuximab (*EGFR* monotherapy) in colorectal cancer, they successfully identify a subset of bladder CCL (n = 5) that harbour the same gene expression signature. Indeed, these subset of bladder CCL were later found out to be sensitive to afatinib (*EGFR*/*HER2* tyrosine kinase inhibitor) according to published IC_{50} values provided in GDSC (Goodspeed et al., 2018). Additionally, using the GDSC datasets, they found that for those bladder cell lines that were resistant to *EGFR* inhibitors,

University of Malaya

they are sensitive to *PI3K* and *mTOR* inhibitors such as temsirolimus. Notably, the concept of leveraging on biomarkers of response from other cancer types was also adopted by the NCI-MATCH clinical trials, which use a panel of single genomic biomarkers to identify therapies for cancer patients independent of cancer type (ClinicalTrials.gov Identifier: NCT02465060).

In the context of HNSCC, De Cecco et al. (2015) successfully clustered the 46 upper aerodigestive tract cell lines available in GDSC into six molecular subtypes information based on their study from a cohort of 527 HNSCC samples (Table 2.3). They further evaluated the drug sensitivity profiles of HNSCC cell lines belonging to different clusters towards the drugs available in GDSC. Indeed, they found that lines in different subtypes have a statistically significant difference in drug sensitivity profile: paclitaxel for a subset of cell lines enriched for HPV-like pathway, Z-LLNle-CHO for those enriched for mesenchymal pathway, afatinib for hypoxia-associated cell lines, nutlin3a for defense response and immunoreactive related cell lines, and rapamycin for the cell lines enriched in classical pathway, respectively.

Table 2.3: Characteristic of HNSCC subtypes (n = 527) identified by De Cecco et al. (2015).

HNSCC subtypes	Functional pathways
CL1	HPV-like
CL2	Mesenchymal
CL3	Hypoxia-associated
CL4	Defense response
CL5	Classical
CL6	Immunoreactive

University of Malaya

2.5.2 The Ushijima Database

Contrary to the CMap database that profiles a thousand small molecule inhibitors on human CCL, but not necessarily restricted to anticancer compounds, Ushijima et al. (2013) developed a public gene expression database that focuses on cancer drugs. The database contains 83 anticancer compounds including 25 clinically used anticancer agents tested on five human CCL: H2228 (lung), HT29 (colon), K562 (leukemia), PC-9 (lung), and SKOV3 (ovary).

Ushijima et al. (2013) obtained the gene expression data by treating these five human CCL with the anticancer compounds at 11 concentrations (10 nm – 10mM) for six or 16 hours, generating a total of 129 treated samples. Gene expression changes were collected using the Affymetrix GeneChip Human Genome U133 Plus 2.0 Array. Unlike CMap gene expression data which were derived mainly from MCF7 (breast) or PC3 (prostate) cells, the majority of Ushijima gene expression datasets (107 out of 129 treatment samples; 83%) were acquired explicitly using the human colon cancer HT-29 cells. One unique feature about HT-29 cell lines is that it is a *TP53* mutant line (COSMIC GRCh38 Cell Line v87). Thus, the Ushijima database expands the coverage of the drug-associated gene expression profiles already available in CMap.

There are currently 22 chemotherapy drug-treated gene expression profiles hosted in the Ushijima database, including some of the common chemotherapy drugs such as 5-FU, cisplatin, docetaxel, and mitomycin that are not in the CMap database, thus making this database unique and an essential resource for cancer research. This is because although chemotherapy drugs are the mainstay of cancer treatment, which chemotherapy drugs work the best for each cancer patient remains challenging to predict. The need to have a tool that can guide the correct choice of chemotherapy drugs highlights the rationale of using gene expression to predict the efficacy of chemotherapy drug candidates, as

University of Malaya

chemotherapy remains the primary treatment for a recurrent and metastatic disease such as oral cancer. The Ushijima drug-treatment gene expression database can be freely accessible at http://scads.jfcr.or.jp/db/cs/.

2.5.3 Other Pharmacogenomic Datasets

2.5.3.1 Library of Integrated Network-based Cellular Signatures

The LINCS project is a National Institutes of Health Common Fund program that catalogues how human cells globally respond to chemical, genetic, and disease perturbations. Compared to CMap that profiles gene expression of CCL using microarrays upon treatment of compounds, the LINCS project additionally collects the genetic manipulation data of knocking-down genes by shRNA or over-expressing genes by cDNA. One unique feature about the LINCS project is that it constitutes an ongoing endeavour, which means the data in the LINCS project will be updated as and when new data are available. New data are still being generated to date (Musa et al., 2018). As of 2018, the LINCS program has generated almost 1.3 million profiles for over 20 000 drugs and 7494 genetic perturbations (e.g., single gene knockdown or overexpression assays). In contrast to CMap which uses only five cell lines, the LINCS dataset currently has data for over 70 different human cell types, including meta information about the experimental conditions and cell lines (Duan et al., 2014; Vidović et al., 2014).

Under the Common Fund programme, the same group of researchers who initiated the original CMap at the Broad Institute developed the third generation of CMap using a new technology called the L1000 platform. The L1000 technology measures only the expression of 978 genes (hence termed the 'L1000 landmark genes') (Subramanian et al., 2017). The expression value of the remaining transcriptome, each containing approximately 22 000 genes, are estimated by a model built from the processing of

University of Malaya

thousands of gene expression datasets from the Gene Expression Omnibus (GEO) (Peck et al., 2006). The rationale for using this computational modelling approach is that gene expression data contain a high degree of statistical dependencies between measured variables (mRNAs). The selected 978 landmark genes could capture most of the information contained within the entire transcriptome (Peck et al., 2006). Users can query, browse, and interrogate this third generation of CMap at the CLUE website (https://clue.io/).

As a whole, the L1000 technology is relatively newer compared to CMap. In the context of HNSCC, no studies thus far have used the L1000 technology to find novel biological discoveries.

2.5.3.2 NCI-60 Panel

Drug screening approach using a panel of 60 human CCL pioneered by the National Cancer Institute (hence the NCI-60 panel) represents the earliest endeavour in using human CCL to screen compounds for novel compounds with tumour-killing properties (Monks et al., 1991; Shoemaker, 2006). These cell lines (consisting of brain, colon, leukemia, lung, melanoma, ovarian, renal, breast, and prostate cancer) were molecularly-characterised to identify biomarkers of response, thus providing the first resource for cancer pharmacogenomics.

Some concerns about the limited number of lines available for any given cancer (6-7 lines for each cancer type) were raised for the NCI-60 panel. Wilding and Bodmer (2016) pointed out in their review paper that due to the molecular heterogeneity with cancers, given such sample sizes, there will not be enough statistical power to detect correlations with even a single key relatively common difference. The successful use of cell lines for evaluating drug responses concerning tumour properties depends critically on the use of

University of Malaya

substantial cell lines for adequate statistical power. Because of the limited number of cell lines hampers the full potential of NCI-60 panel, more recent large-scale pharmacogenomics screens, such as the Genomics of Drug Sensitivity in Cancer (GDSC), Cancer Cell Line Encyclopedia (CCLE), and Cancer Therapeutics Response Portal (CTRP), that contain several orders of magnitude more cell lines have since gained much popularity.

2.5.3.3 Cancer Cell Line Encyclopedia and Cancer Therapeutics Response Portal

The Cancer Cell Line Encyclopedia (CCLE) (https://portals.broadinstitute.org/ccle) is a public pharmacogenomics database containing genomic data from 947 human CCL. These data cover microarray and RNA-seq gene expression, chromosome copy number variation, DNA methylation, and Achilles shRNA knockdown (Barretina et al., 2012). Using the pharmacological profiles of 504 of the cell lines over 24 anticancer drugs in the CCLE database, Barretina et al. (2012) successfully applied naive Bayes and elastic net regression to predict the drug sensitivity profiles of these 504 cell lines. As a proof of principle for their approach, they presented plausible correlations of drug activity with aberrations in particular genes such as *IGF1R*, *AHR*, *NRAS*, and *SLFN11*. All CCLE-associated data are archived in the Gene Expression Omnibus (GEO) (Accession number: GSE36139).

The Cancer Therapeutics Response Portal (CTRP) (https://portals.broadinstitute.org/ctrp.v2.1/) is another public pharmacogenomics database that enables users to correlate genetic features to sensitivity in specific tissue lineages (Basu et al., 2013). Notably, CTRP was set up to complement drug response data hosted in CCLE that has a relatively smaller amount of drugs (n = 24). Although CTRP

University of Malaya

does not provide gene expression profile data, it contains the drug response data for cell lines that have been characterised previously by CCLE.

In Version 1, CTRP measured the sensitivity of 242 genomically-characterised CCL to 185 compounds that target many protein molecules, uncovering (i) genetic dependencies of these CCL as a result of specific cancer-genomic alterations such as the somatic mutations or translocation; and (ii) small-molecules that target these genetic dependencies. For example, using CTRP datasets, Basu et al. (2013) observed that *EGFR* mutant lung CCL are highly sensitive to neratinib - a dual *ERBB2/EGFR* inhibitor. A phase II trial evaluating the safety of neratinib in advanced non-small cell lung cancer patients has just recently concluded in 2018 (ClinicalTrials.gov Identifier: NCT00266877).

Moving on to Version 2, CTRP generated a set of 481 small-molecule probes and drugs that collectively modulate a broad array of cellular processes. The sensitivity of 860 CCL to these compounds was measured, and the association studies to connect the sensitivity of these 860 CCL to cancer features such as mutations, gene expression, copy-number variation, and lineage were carried out. CTRP performs statistics-based enrichment analyses that combined rank-based and parametric tests to identify genetic alterations and cellular features that are significantly enriched among sensitive (AUC < 3.5) or unresponsive (AUC > 5.5) CCL.

With this wealth of large-scale drug sensitivity studies, we are now poised to utilise these pharmacogenomic datasets to accelerate the identification of molecular features in cancers that are associated with sensitivity to specific drugs. To do this, computational algorithms that could mine and integrate both gene expression changes, as well as associated drug sensitivity response, are highly sought after. One prerequisite component

University of Malaya

to successfully carrying out such computational drug discovery approaches is to have effective gene signature similarity scoring algorithms.

2.6 Gene Signature Similarity Scoring Algorithms

To date, several computational algorithms for detecting gene signature similarity have been developed to make use of the perturbagen-induced signatures information contained in CMap. Here, six commonly-used gene signature similarity scoring algorithms were briefly introduced. These are (i) Kolmogorov-Smirnov (KS) statistic (Lamb et al., 2006); (ii) Weighted Connectivity Score (Subramanian et al., 2017); (iii) statistically significant Connectivity Map unordered (Zhang & Gant, 2008); (iv) statistically significant Connectivity Map ordered (Zhang & Gant, 2008); (v) eXtreme Sum (Cheng et al., 2014); and (vi) eXtreme Cosine (Cheng et al., 2013). Since few systematic evaluations of the performance of these gene signature similarity scoring algorithms are available (Cheng et al., 2014; Musa et al., 2017), it is necessary to evaluate the strengths and weaknesses of these six methods in this thesis. Some of the commonly-used performance evaluation metrics such as the positive predictive value, and enrichment analysis of drugs with similar mechanism of action could be useful for evaluating the performance of these gene signature similarity scoring algorithms (Powers, 2011).

2.6.1 Kolmogorov-Smirnov Statistic

The KS statistic is used for testing whether the distribution of observed data come from a hypothesised distribution, or whether the distributions of two sets of samples are the same (Smirnov, 1939). For a one-sample case, it is the supremum of the distance between the empirical cumulative distribution function (ECDF) and a hypothesised distribution function. In the two-sample case, the KS statistic is the supremum of the distance between

University of Malaya

the ECDF of the two samples. Figure 2.6 shows an example of the KS statistic obtained from comparing the ECDF of two standard normal distributions. Here, the KS statistic is given by the largest absolute deviation between the two ECDF (0.26, as indicated by the red arrow).

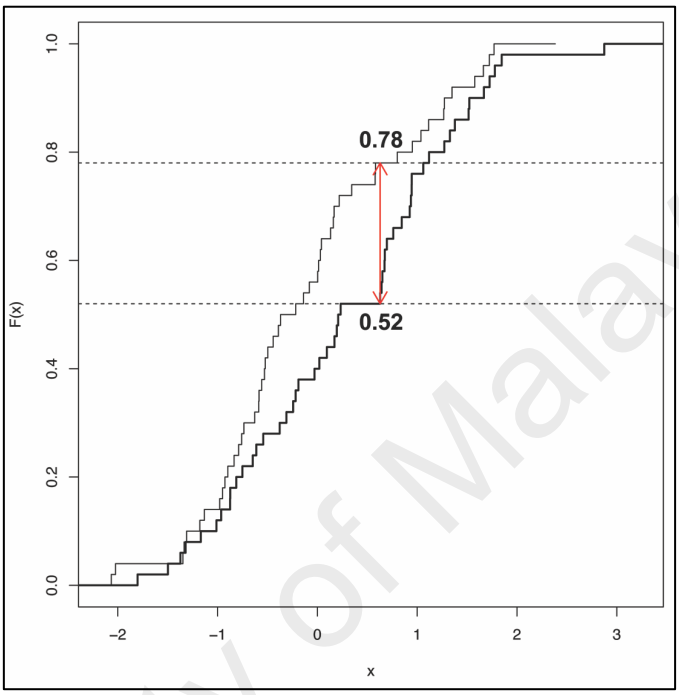

Figure 2.6: Empirical cumulative distribution function (ECDF) of two randomly generated standard normal distribution samples. The maximum height between the two ECDF is 0.26 (indicated by the distance between the red arrows).

The concept of the KS statistic is manifested in the use of the enrichment score (ES) in Gene Set Enrichment Analysis (GSEA; Subramanian et al., 2005). The ES is a KS-like statistic in the sense that it is calculated as the maximum deviation of the ES function from 0. Figure 2.7 gives an example of output from the GSEA. Note the changes to the ES as we walk down the ordered (according to some statistics such as the *t*-statistic) list of genes.

University of Malaya

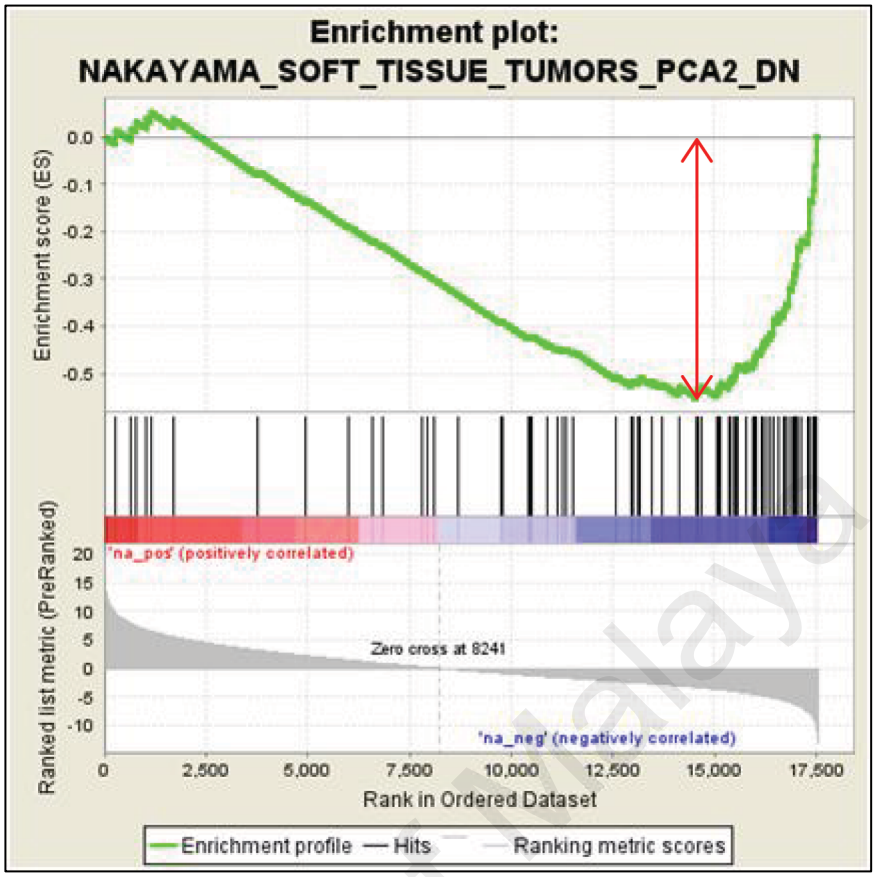

Figure 2.7: An example of ES output from GSEA. The red arrowed bar indicates the maximum deviation of the ES function from the flat line at 0.

To put things into the perspective of CMap, the KS statistic is used to perform gene signature similarity analysis of the up-regulated and down-regulated query genes to the compound-associated gene expression profile in CMap. Two KS values would be generated, one each for up-regulated (KS_{up}) and down-regulated (KS_{down}) genes. The connectivity score (CS) is calculated by computing the difference between KS_{up} and KS_{down}, with the exception that in the event both KS_{up} and KS_{down} have the same algebraic sign, then the CS takes on the zero value. See Section 3.2.3 for detailed implementation of KS statistic in terms of gene signature similarity analysis.

University of Malaya

2.6.2 Weighted Connectivity Score

The weighted connectivity score (WTCS; Subramanian et al., 2017) uses a weighted version of ES, a KS-like statistic described previously in Lamb et al. (2006). WTCS is a composite, bi-directional version of ES. The ES values for up-regulated and down-regulated genes would be computed separately for a given set of query gene set according to the ranked (in descending order) fold change values in the reference database. The different feature that WTCS has compared to KS statistics is that it does not take into consideration the ranking of the query gene signature. Instead, it calculates the ES of the instance in the CMap reference database that matches the query genes given by the users. By taking the normalised ES score, the WTCS values range between +1 and -1 (normalised using the absolute largest and smallest WTCS for individual positive and negative values, respectively). Signatures that are positively / negatively correlated will have positive / negative WTCS respectively. Those that are uncorrelated will have nearly zero WTCS. WTCS is zero when both ES values for up-regulated and down-regulated genes have the same algebraic sign. WTCS is implemented as the gene signature similarity scoring metric in the third generation of CMap using the L1000 platform (Subramanian et al., 2017). See Section 3.4.2.2 for details implementation of WTCS for gene signature similarity analysis.

2.6.3 eXtreme Sum and eXtreme Cosine

eXtreme Sum (XSum) (Cheng et al., 2014) and eXtreme Cosine (XCos) (Cheng et al., 2013) use differentially up-regulated and down-regulated genes to query the reference database. In contrast to KS and WTCS algorithms, which evaluate the queried genes that span across the entire reference profile, XSum and XCos calculation start by subsetting the reference database to top *N* up-regulated and bottom *N* down-regulated genes by fold change values between the compound treated and control samples for each drug instance

University of Malaya

in CMap. The default value of *N* is 500. The rationale for using this approach of focusing on the extreme ends is to ensure high specificity, as one can only pursue a limited number of drug-disease indication hypotheses, especially in a drug repositioning exercise. In other words, the aim is to sacrifice some true positives to keep the false positive low. XSum is computed by taking the difference between the sums of the fold change value for the genes that match the up-regulated and down-regulated genes, respectively. The XCos algorithm, meanwhile, requires the user to provide additional fold change values of their differentially expressed genes. By having both fold change values for the reference profile (vector A) and the query signatures (vector B), the XCos values can then be calculated as a dot product of vector A and vector B. See Section 3.4.2.3 for detailed implementation of XSum and XCos for gene signature similarity analysis.

2.6.4 sscMap

Zhang and Gant (2008) used three principles in developing the sscMap algorithm. First, treatment and control instances are treated similarly, making the effect of the treatment instances to be determined by differentially expressed genes. Second, the genes that are affected to greater extents by the treatment, that is, genes that are more highly differentially expressed, are given more weight in characterising the treatment. Finally, the up-regulated and down-regulated genes are treated equally in the sense that a two-fold up-regulation or two-fold down-regulation of a gene has the same relevance in constructing the reference profile. The genes are ordered based on their absolute value of the differential log of fold change, as the up-regulated and down-regulated genes are considered the same. A query gene signature can be an ordered gene list, or just a collection of genes without specific ordering, which will be referred to as ordered and unordered gene signature, respectively.

University of Malaya

Based on this ranking proposed by Zhang and Gant (2008), the importance of a gene is determined by the absolute value of its signed rank rather than its differential expression value. The signed rank (either + or – sign) would, therefore, dictate the direction of the regulation. The most significant gene will be at the top of the list, while most of the insignificant gene will be at the bottom. For the specific implementation of sscMap for gene signature similarity analysis, please see Section 3.4.2.4.

As a whole, an effective gene signature similarity scoring algorithm is a valuable tool for computational discovery of new indications for drugs. Given practical scoring algorithms and stronger validation datasets, we believe that computationally predicting efficacious drugs through CMap concept will prove to be an effective method to repurpose drugs across a broad range of diseases. Computational drug candidates prediction could be one viable approach to serve the unmet need of oral cancer patients for more therapeutic options.

University of Malaya

CHAPTER 3: MATERIALS AND METHODS

3.1 GENIPAC: Genomic Information Portal on Cancer Cell Lines

A key component of identifying potential efficacious drugs for HNSCC is to put together a database of query signatures. To do this, a cell line database containing genomics information on HNSCC cell lines including gene expression signatures was established. This interactive web portal called GENIPAC – Genomic Information Portal on Cancer Cell Lines, takes advantage of the recently developed cBioPortal (Cerami et al., 2012; Gao et al., 2013) which hosts and displays genomic data on head and neck cancer cell lines (CCL). By facilitating easy information access, GENIPAC provides gene expression data to mine for drug efficacy. In addition it enables users to explore complex datasets of head and neck cancer for the development of biological hypotheses based on commonly altered gene profiles and biological pathways.

GENIPAC (http://genipac.cancerresearch.my/) (Lee et al., 2018) uses the cBioPortal engine, which runs on Apache Tomcat and enables the visualisation, analysis, and downloading of large-scale genomic datasets (Cerami et al., 2012; Gao et al., 2013). The genomic datasets (mutations, copy number alterations, and mRNA expression) hosted in GENIPAC are freely available to the public as stated under the Affero GPLV3 license. Currently, GENIPAC contains datasets from three series of HNSCC cell lines: ORL Series (Fadlullah et al., 2016), OPC-22 (Martin et al., 2014), and H Series (Prime et al., 1990). The ORL Series established by Cancer Research Malaysia consists of 16 oral squamous cell carcinoma (OSCC) cell lines described in Fadlullah et al. (2016). The cell lines in this series were derived from Asian patients with diverse etiological factors, including those who consumed tobacco products (smokeless and smoked), those who chewed betel quid, and those who consumed alcohol. There are also cell lines which were

University of Malaya

derived from patients with no known risk habits. In terms of age at diagnosis, the age of the patients ranged from 36 to 79 years old, with the majority of them from the Indian ethnicity (n = 13). The demographic details of the patients from whom the ORL Series were derived have been described previously (Fadlullah et al., 2016). The second set of data was from the OPC-22 study (Martin et al., 2014), which consists of 22 HNSCC cell lines that are widely used in the HNSCC research field (Li et al., 2014; Lui et al., 2013; Zhao et al., 2011), with some of these lines dating back to the 1980s (Kimmel & Carey, 1986). This series of cell lines, which were recently described in detail, were derived from different anatomic sites, and some testing positive for human papillomavirus. Finally, the H Series of cell lines was obtained from the University of Bristol, United Kingdom. These cell lines were among the first OSCC cell lines to be established and used in oral cancer research (Prime et al., 1990).

3.1.1 Mutations and mRNA Expression

Transcriptomic analyses of these three head and neck cancer studies using RNA-seq were done in 2012. The transcripts from the ORL Series and H Series were mapped to the human reference genome (Ensembl GRCh37) with Tophat2 2.0.9 using default parameters (Trapnell et al., 2012). Variant calling was conducted with the use of GATK HaplotypeCaller 2.8 (DePristo et al., 2011). A series of variant calling and filtering criteria were applied as described previously (Fadlullah et al., 2016). Gene expression in raw read counts was extracted through the use of featureCounts software with default parameters (Liao et al., 2014). Genes with zero expression value across all samples were excluded. Raw read counts of the remaining genes were normalised and then log-transformed (base 2) in the R computing environment (R 3.4.0; R Core Team 2015) with the R package DESeq2 (Love et al., 2014).

University of Malaya

For the OPC-22 series, the RNA-Seq transcriptome analyses were carried out separately by Dr. Silvio Gutkind's lab in the National Institute of Health (Bethesda, USA) (Martin et al., 2014). Briefly, the splice junction aligner GSNAP mapped the RNA-seq transcriptome data to the human reference genome hg19 (Wu & Nacu, 2010). Reads were mapped according to the UCSC.hg19.KnownGene database and later counted and annotated with the GenomicFeatures (Lawrence et al., 2013), Rsamtools (Morgan et al., 2017), and org.Hs.eg.db (Carlson, 2017) packages in R. To run variant calling for OPC-22, whole exome sequencing reads were first mapped to hg19 genome with Novoalign aligner, processed according to recommended guidelines of GATK (DePristo et al., 2011), and lastly annotated using the ANNOVAR software (Wang et al., 2010). Variant effect analysis of the OPC-22 was performed using the PROVEAN tool (Choi et al., 2012).

3.1.2 Copy Number Alterations

The copy number (CP) changes of ORL Series cell lines (in Affymetrix .CEL format) derived from Genome Wide Human Cytoscan HD array (Affymetrix) were preprocessed in Chromosome Analysis Suite 3.2. The generated DNA CP of each sample was subjected to segmentation with the Circular Binary Segmentation algorithm (Olshen et al., 2004), implemented in R package as DNAcopy (Seshan & Olshen, 2017). Altered regions were tested whether they were the result of amplification or deletion using GISTIC 2.0 (Mermel et al., 2011), with the low-level amplification and deletion threshold of 0.1.

CP alterations for the OPC-22 cell lines were extracted from Martin et al. (2014). Briefly, Strand NGS software was applied to compute the CP variations of the OPC-22 lines (Strand Life Sciences, Bangalore, India). A pseudo-normal sample was computed from the average read depth of all the OPC-22 lines and then used to define the CP

University of Malaya

baseline against which all the OPC-22 cell lines were compared. The CP values reported in Martin et al. (2014) were defined as follows: CP < 0.8, homozygous deletion (–2); CP 0.8 to < 1.5, hemizygous deletion (–1); CP 1.5 to 4.0, neutral (0); CP > 4.0 to 8.0, gain (+1); and CP > 8.0, high-level amplification (+2).

3.1.3 Data Formatting

The cBioPortal platform supports several types of the data format as query input (see the official cBio Portal Documentation webpage at https://readthedocs.org/projects/cbioportal/). cBioPortal requires several files to be uploaded for it to work optimally, as well as information on the cancer study and mutation data. To speed up the data extraction and preparation processes, a script was developed to automate data extraction from the original unmodified datasets. The datasets were extracted into several files with the ASCII text files extension. The lists of files input into GENIPAC are as follows: (i) cancer study, which contains information on the type of cancer, description, and an identifier in GENIPAC; (ii) mutation data, which contain mutation data for each gene with the unique Entrez gene ID, chromosome number, variant classification, and protein position; (iii) discrete copy number data, which contain all copy number levels for each gene; (iv) expression data, which consist of expression values of the genes in each sample; and (v) anonymised clinical data from patients.

All the file formats mentioned above can be uploaded into GENIPAC through either the console or the SSH terminal. The data then undergo multiple sessions of data validation and correction with dataset validator tools, which are built-in for cBioPortal (see https://cbioportal.readthedocs.io/en/latest/Using-the-dataset-validator.html). The validated genomic data, particularly gene expression values will undergo differential analysis to generate up-regulated and down-regulated genes. These differentially

University of Malaya

expressed genes (DEG) would be used as the input query genes for DeSigN analysis in Section 3.2.

3.2 DeSigN: Differentially Expressed Gene Signatures – Inhibitors Platform

To identify drugs that could be repurposed for cancer, the current challenge is to develop discovery pipelines to prioritise testing of already approved drugs, particularly in cancers with limited chemotherapy options, such as oral cancer (Vermorken et al., 2008). DeSigN is a web-based bioinformatics tool for associating gene signatures with drug response phenotype, using IC_{50} data (Lee et al., 2017). The DeSigN platform (Figure 3.1) consists of three key components: (i) a reference database that contains a set of pre-defined gene expression profiles associated with drug response data to 140 drugs derived from Genomics of Drug Sensitivity in Cancer (GDSC) database (https://www.cancerrxgene.org/); (ii) a set of DEG signatures as query input; and (iii) KS statistic as gene signature similarity scoring algorithm for evaluating similarity between the query gene signature and drug-associated gene expression profiles in the reference database.

University of Malaya

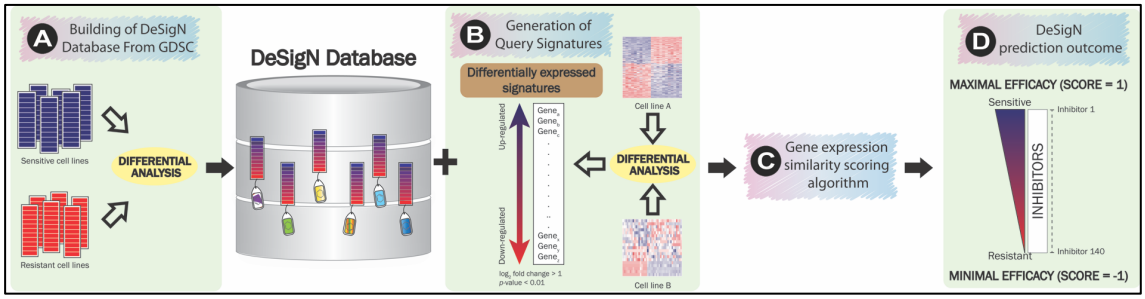

Figure 3.1: Principal workflow of DeSigN. (A) A reference database of cell lines that are sensitive and resistant to drugs available in the GDSC database was created. Version 1.0 contains 140 drugs with their unique ranked-based gene signatures. (B) Differentially expressed gene signatures are generated from differential expression analysis of cell lines from two distinct experimental conditions, e.g., cell line gene expression data from tumour samples versus normal samples. The up and down-regulated genes ($|\log_2$ fold change$| > 1$ and p-value < 0.01) thus selected will be used to query the DeSigN database. (C) Using a gene signature similarity scoring algorithm, the similarity analysis would then be carried out to compute the gene expression similarity between the query signatures and DeSigN database (D) A rank-based list of inhibitors is generated, with CS between 1 (maximal efficacy) and −1 (minimal efficacy). This allows users to prioritise the testing of these candidates.

3.2.1 Reference Database

The reference database was built using baseline microarray and drug sensitivity data obtained from the Genomics of Drug Sensitivity in Cancer (GDSC) database. The raw CEL microarray data files of solid tumour cell lines (normalised using the MAS 5.0 algorithm) were first downloaded from GDSC (Garnett et al., 2012). The probe sets were collapsed to gene symbols using the "Collapse Dataset" function provided in Gene Set Enrichment Analysis (GSEA) (Subramanian et al., 2005) with HT HG-U133A chip as the reference. This process produced 12 772 unique genes. For each drug, the CCL's drug response phenotype (resistant or sensitive) was classified in the following way: (i) the cell lines were first ranked by their IC_{50} values (lowest to highest); (ii) cell lines with IC_{50} that were U standard deviations larger than the median IC_{50} of all cell lines were considered to be resistant; those that were L standard deviations smaller were considered to be sensitive. The parameters U and L were chosen carefully on a case-by-case basis. These two cut-offs were generally values where sharp transitions in IC_{50} were observed in the

University of Malaya

scatter plot of $-\log_{10}(IC_{50})$ against rank. About 20 cell lines each from the sensitive and resistant phenotype were thus defined. The list of sensitive and resistant cell lines defined for the 140 drugs in DeSigN is provided in Appendix 1. Figure 3.2 shows an example of the scatter plot of $-\log_{10}(IC_{50})$ against rank for Mitomycin-C, a commonly-used drug for bladder and gastric cancer (Bosschieter et al., 2018; Murata et al., 2018). The scatter plot of $-\log_{10}(IC_{50})$ against rank for all the 140 drugs can be found in Appendix 2.

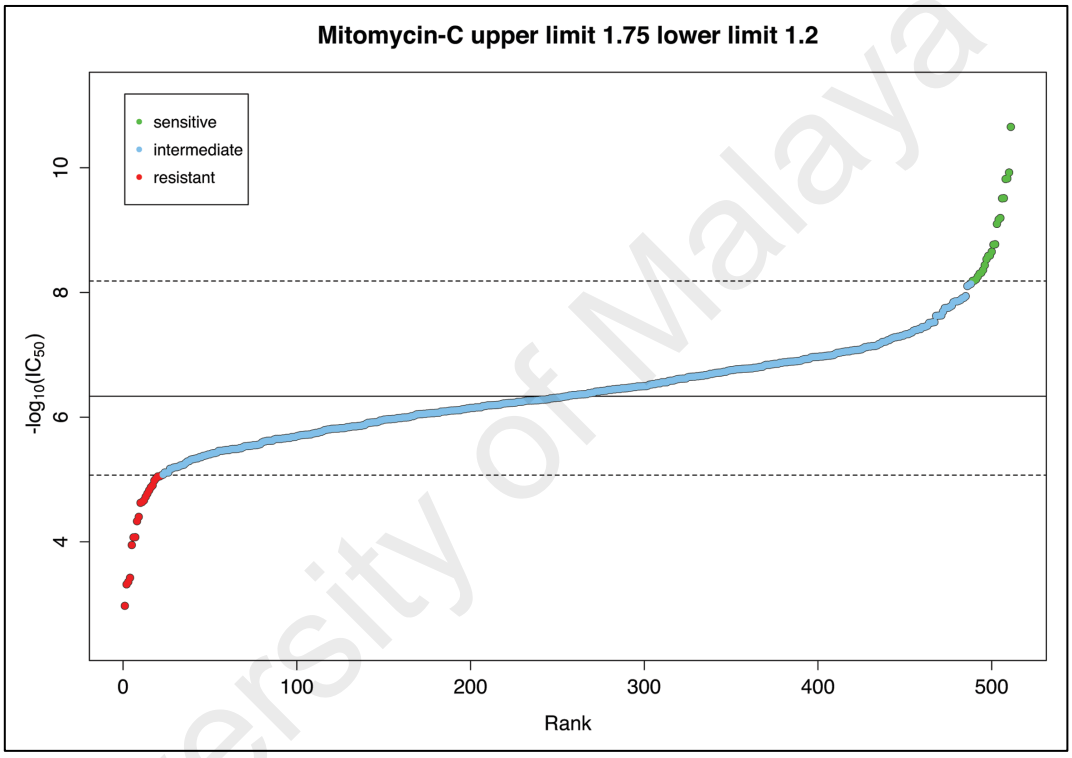

Figure 3.2: Example of $-\log_{10}(IC_{50})$ rank plot to define drug response phenotype. The solid line represents the median IC_{50} values of inhibitor mitomycin-C whereas the lower and upper dashed lines represent the cut-off for classifying cell lines into sensitive or resistant phenotypes, respectively.

Differential expression of microarray gene expression data between the sensitive and the resistant phenotype was done using the Linear Models for Microarray Data (limma) algorithm (Ritchie et al., 2015; Smyth, 2004). The result from limma for each drug was sorted and converted into ranked lists according to the gene's moderated *t*-statistic (rank 1 for the largest value). This reference database was used to connect the queries and return the rank-ordered list of drugs for a particular query (Figure 3.1A).

University of Malaya

Figure 3.3 shows an example of a limma output with the ranking of the genes ordered according to the moderated *t*-statistic in descending order. Upon treatment of a particular drug, the tested cell lines can be classified to sensitive (green cells) and resistant (red cells) lines. Limma would then be used to generate the *t*-statistic for the gene expression values comparison between the sensitive and resistant lines. The resulting moderated *t*-statistic is used to rank the genes in descending order. Figure 3.3 shows that *TP53* is ranked first among 15 000 genes because its moderated *t*-statistic value is the largest (15.558); conversely, *EGFR* is ranked lowest since its moderated *t*-statistic value is the smallest (-10.732).

Gene	Sensitive Cell Lines			Resistant Cell Lines			*t*-statistic	Ranking
	CAL-51	C32	IST-MEL1	LOXIMVI	RPMI-7951	CAL-29		
TP53	5.413	5.786	5.529	3.083	3.176	3.330	15.558	1
ERBB2	7.398	8.009	6.883	3.055	2.859	3.036	14.633	2
PIK3CA	5.090	5.110	5.288	4.082	4.219	4.231	8.025	3
FJX1	6.265	6.494	6.538	5.580	5.568	5.320	6.482	4
CDK4	5.922	7.533	5.532	3.114	3.896	3.505	4.927	5
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
NOTCH1	5.042	5.378	5.520	6.777	7.236	6.967	-8.552	14998
FAT1	4.501	5.073	4.744	6.441	6.631	6.529	-9.674	14999
EGFR	3.425	4.274	3.362	10.146	9.672	8.378	-10.732	15000

Figure 3.3: An example of limma output with the ranking of the genes ordered according to *t*-statistic in descending order. Cells highlighted green refers to sensitive cell lines while red indicates resistant cell lines in regards to the treatment of a particular drug. Limma is used to analyse the sensitive and resistant cell lines, with the moderated *t*-statistic used to rank the genes in descending order.

3.2.2 Query Signature

The query signatures were obtained from microarray or RNA-seq gene expression data of cell lines from two different phenotype classes. In this thesis, the query signature validation datasets were obtained from two sources. The first source, which is published drug sensitivity studies, were obtained from the NCBI GEO database. The second set of

University of Malaya

query signatures was obtained from OSCC lines, in which their gene expression data can be retrieved from GENIPAC under the ORL Series tab. Notably, these DEGs were selected using joint filtering of *p*-value and fold change (Xiao et al., 2014), with the threshold value set at $|\log_2$ fold change$| > 1$ and *p*-value < 0.01 (Figure 3.1B).

Figure 3.4 shows an example of the volcano plot for query signature generation using joint filtering method. Red dots indicate the down-regulation of genes and blue dots refer to genes that are up-regulated. The up-regulated and down-regulated genes selected in this way take both statistical significance (such as *p*-value) and biological relevance (such as fold change) into consideration (Xiao et al., 2014).

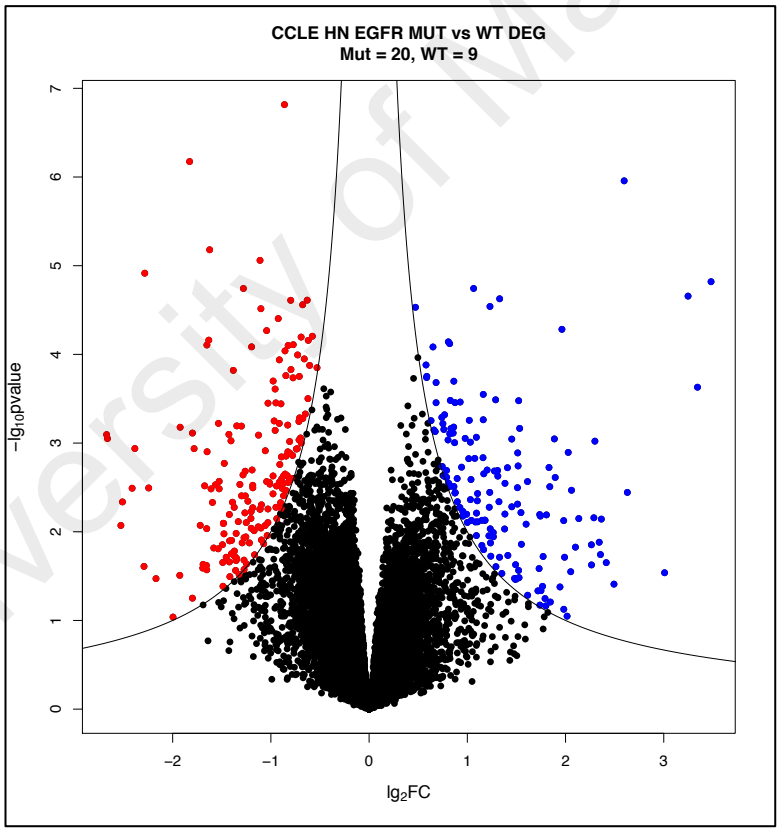

Figure 3.4: A volcano plot showing an example of the query signature generation using the joint filtering of *p*-value < 0.01 and $|\log_2$ fold change$| > 1$. Red dots indicate down-regulated genes while blue dots indicate up-regulated genes.

University of Malaya

3.2.3 Gene Signature Similarity Scoring Algorithm - Kolmogorov-Smirnov Statistic

A gene signature similarity scoring algorithm based on the non-parametric KS statistic (Lamb et al., 2006) was used to associate query signature to the drug-specific, rank-ordered gene expression profile database.

The CS is computed according to Lamb et al. (2006) as follows. Let *N* be the total number of genes in the reference database, and *T* the number of genes in the query signature for up-regulated or down-regulated genes. For every drug in the reference database, the rank-ordered (using moderated *t*-statistic) list *R* for all *N* genes was computed. Let *j* index the query genes in such a way that *R*(*j*), the rank of the *j*-th gene in the *N* total number of genes, is monotone increasing. For $j = 1, 2, \ldots, T$, the threshold a and b for up-regulated gene signature was computed as:

$$a = \max_{1 \le j \le T} \left\{ \frac{j}{T} - \frac{R(j)}{N} \right\}, \quad (1)$$

$$b = \max_{1 \le j \le T} \left\{ \frac{R(j)}{N} - \frac{(j-1)}{T} \right\}. \quad (2)$$

Similarly, the threshold a' and b' were computed for the down-regulated gene signature.

Using the ranking of the *ERBB2* and *CDK4* in Figure 3.3 as an example, the calculation of threshold *a* (Eq. 1) and *b* (Eq. 2) is shown in Table 3.1.

University of Malaya

Table 3.1: An example of threshold a and b calculations for an hypothetical up-regulated gene signature of size 2 derived from Figure 3.3. For this example, $a = 0.9999$ and $b = 0.00013$.

Gene	Threshold a	Threshold b
ERBB2	$\frac{1}{2} - \frac{2}{15000} = 0.4999$	$\frac{2}{15000} - \frac{(1-1)}{2} = 0.00013$
CDK4	$\frac{2}{2} - \frac{5}{15000} = 0.9999$	$\frac{5}{15000} - \frac{(2-1)}{2} = -0.4997$
Maximum	0.9999	0.00013

Subsequently, for each drug i, the KS-like statistics for up-regulated and down-regulated query gene signature, ks^i_{up} and ks^i_{down} are computed as (subscript omitted)

$$ks^i = \begin{cases} a, if\ a > b; & (3) \\ -b, if\ a < b. & (4) \end{cases}$$

Figure 3.5 below shows an example of KS statistic calculation considering the threshold a and threshold b using Eq. 1 and Eq. 2 respectively. In situation A, given an input of three differentially expressed genes by the user: *TP53*, *PIK3CA*, and *NOTCH1*, we have $a = 0.924 > 0.337 = b$, hence the KS value for Drug A in situation A is 0.924 (Eq. 3). In situation B, given the set of three differentially expressed genes by the user: *NOTCH1*, *FJX1*, and *CDK4*, we have $a = 0.493 < 0.767 = b$, hence the KS value for Drug B in situation B is -0.767 (Eq. 4).

A

Gene	Drug A	
	a	b
TP53	0.924	0.271
PIK3CA	0.024	-0.724
NOTCH1	0.633	0.337

KS for Drug A = 0.924

B

Gene	Drug B	
	a	b
NOTCH1	-0.567	0.767
FJX1	0.493	0.014
CDK4	0.074	0.498

KS for Drug B = -0.767

Figure 3.5: An example of KS value output considering the threshold of a and b respectively. (A) The KS value for Drug A is 0.924 (highlighted green) because a = $0.924 > 0.337 =$ b. (B) Drug B has the KS value of -0.767 (highlighted green) because a = $0.493 < 0.767 =$ b.

University of Malaya

The *ES* for the drug *i* (ES^i) in the reference database is set to zero if both ks^i_{up} and ks^i_{down} have the same algebraic sign; otherwise, $ES^i = ks^i_{up} - ks^i_{down}$. The CS ($S^i$) for non-zero instances is a normalised *ES* computed as:

$$S^i = \begin{cases} \frac{ES^i}{P}, if ES^i > 0; \\ -\left(\frac{ES^i}{Q}\right), if\ ES^i < 0, \end{cases}$$

where $P = \max_i ES^i$ and $Q = \min_i ES^i$ are the normalising constants. DeSigN returns a ranked list of drugs, with S^i ranging between 1 (maximal efficacy) and -1 (minimal efficacy) (Figure 3.1C).

Figure 3.6 below shows an example of KS statistic calculation implemented in DeSigN. The drugs are ranked according to their respective CS in descending order. Notably, in the event where both KS for up-regulated and down-regulated genes have the same algebraic sign, their CS is set to zero (e.g., 17-AAG and axitinib). In such situation, the drugs are ranked according to their KS value for up-regulated genes, by which 17-AAG is ranked higher than axitinib because it has higher KS value for up-regulated genes (KS_{up} = 0.614) compared to axitinib (KS_{up} = -0.789). The positive normalising constant (*P*) is 1.774 (cell highlighted green) whereas negative normalising constant (*Q*) is -1.852 (cell highlighted yellow).

Drug Name	KS_{up}	KS_{down}	ES	CS	Ranking
Afatinib	0.816	-0.958	1.774	1.000	1
Lenalidomide	0.787	-0.947	1.734	0.977	2
17-AAG	0.614	0.756	0.000	0.000	3
Axitinib	-0.789	-0.782	0.000	0.000	4
BMS-754807	-0.854	0.798	-1.625	-0.877	5
Nutlin	-0.869	0.983	-1.852	-1.000	6

Figure 3.6: An example of KS statistic calculation. The maximum value for *P* is 1.774 (cell highlighted green), and the minimum value of *Q* is -1.852 (cell highlighted yellow). The drugs are ranked according to their CS values.

University of Malaya

To evaluate the statistical significance of S^i, a permutation approach was used to simulate the null distribution of S^i. Thus, m random gene sets, each having the same size as the size of the input gene signature, were selected from the *N* total number of genes in the reference database. Each gene set then yields $S^i_{random}(k)$, where k indexes the random gene set. The p-value was computed as

$$p-value = \begin{cases} \frac{1}{m}\sum_{k=1}^{m} I_{\left(S^i_{random}(k) > S^i\right)}, & \text{if } S^i > 0; \\ \frac{1}{m}\sum_{k=1}^{m} I_{\left(S^i_{random}(k) < S^i\right)}, & \text{if } S^i < 0; \\ \max\left\{\frac{1}{m}\sum_{k=1}^{m} I_{\left(S^i_{random}(k) > S^i\right)}, \frac{1}{m}\sum_{k=1}^{m} I_{\left(S^i_{random}(k) < S^i\right)}\right\}, & \text{if } S^i = 0, \end{cases}$$

where I_A is the indicator function that takes the value 1 if event A occurs, and 0 otherwise. Here, $m = 1000$ was set.

3.2.4 The DeSigN Web Interface

The web interface of DeSigN uses PHP (v7.0) with the support of jQuery (version 1.4.2). It is hosted using the Apache Server. The reference database is generated and managed using the MySQL database (v5.5.49). DeSigN makes use of the AJAX feature to load the content quickly without reloading the pages. All queries are sent to the Java-based computing cluster to perform parallel computation. A help document providing a guide for users to query and navigate DeSigN is available on the website, with examples given. The DeSigN website is freely available at http://design.cancerresearch.my/.

University of Malaya

3.2.5 NCBI Gene Expression Omnibus Datasets

To demonstrate how DeSigN could be used to predict candidate drugs computationally, differentially expressed genes generated from studies published in the NCBI GEO database were used to validate DeSigN (Table 3.2). Several inclusion and exclusion criteria guided the inclusion of the studies: (i) the medians of the distribution of gene expression values of each sample were more or less equal; (ii) the subject of the drug sensitivity study was *Homo sapiens*; (iii) drug treatment was given for at least 24 hours; (iv) only one drug was used. Blood cancer-related studies were excluded. For each study, a list of DEG was identified and used to query DeSigN. Two studies - GSE9633 and GSE4342, were included.

The raw microarray gene expression CEL files from the two GEO studies were background-corrected, normalised, and summarised into probe sets values using the standard Robust Multichip Average (RMA) algorithm (Irizarry et al., 2003). The probe sets were then collapsed to gene symbols (maximum probe set value was chosen in the case where multiple probe sets mapped to the same gene) using the "Collapse Dataset" function provided in GSEA (Subramanian et al., 2005) with respective Affymetrix chip (HG-U133A_2: GSE9633; HG-U133A: GSE4342) as reference. Sensitive cell lines were defined as having $IC_{50} < 1$ μM and resistant cell lines as having $IC_{50} \geq 1$ μM. Subsequently, differential gene expression analyses between the sensitive and resistant cell lines of these two studies were processed as described in Section 3.2.2. The generated DEG were then subjected to DeSigN analysis to determine the rank of the intended drug for each study.

University of Malaya

Table 3.2: GEO studies to validate DeSigN prediction.

GEO reference	Drug	Response	Number of sensitive samples	Number of resistant samples	Reference
GSE9633	Dasatinib	Sensitive	11	5	Wang et al., 2007
GSE4342	Gefitinib	Sensitive	18	11	Coldren et al., 2006

3.3 Identifying Potential Drug Candidates for Oral Cancer

3.3.1 Computational Analyses of OSCC Cell Lines

To identify drugs that could be repurposed for oral cancer treatment, the RNA-seq data of five OSCC (ORL-48, ORL-150, ORL-156, ORL-196, and ORL-204) and three normal oral keratinocytes (NOK) cultures (Fadlullah et al., 2016) were subjected to differential expression analysis (OSCC versus NOK). The latter was carried out using DESeq2 (Love et al., 2014). Notably, the RNA-seq data of these lines can be obtained under the ORL Series tab in GENIPAC.

The list of DEG (149 up-regulated and 251 down-regulated genes) generated from DeSeq2 underwent further filtering parameters of $|\log_2 \text{fold change}| > 1$ and p-value < 0.01 as described in Section 3.2.2. These DEGs were then used to query DeSigN to shortlist candidate drugs for subsequent experimental validation. Following this, one of the candidate drugs, bosutinib, was selected for *in vitro* validation to evaluate its efficacy against the OSCC cell lines.

University of Malaya

3.3.2 Experimental Validation of Drugs Selected using DeSigN

3.3.2.1 Cell Culture

ORL cell lines and HSC-4 (sensitive control for response to bosutinib) were cultured in Dulbecco's Modified Eagle Medium (DMEM/F12) (1:1) supplemented with 10% (v/v/) heat-inactivated fetal calf serum (FBS), 100 IU penicillin/streptomycin and 0.5 µg/ml hydrocortisone as described previously (Fadlullah et al., 2016). NOK were cultured in keratinocyte serum-free media (KSFM; GIBCO, Carlsbad, CA, USA) supplemented with 25 µg/ml bovine pituitary extract, 0.2 ng/ml epidermal growth factor; 0.031 mM calcium chloride and 100 IU penicillin/streptomycin (GIBCO, Carlsbad, CA, USA) as described previously (Fadlullah et al., 2016). The breast CCL MCF7 (resistant control for response to bosutinib) was cultured in RPMI 1640 medium (GIBCO, Carlsbad, CA, USA) supplemented with 10% (v/v) heat-inactivated FBS and 100 IU penicillin/streptomycin. All cultures were incubated in a humidified atmosphere 5% CO_2 at 37°C.

3.3.2.2 Viability Assay using 3-(4,5-dimethylthiazol-2-yl)-2,5-diphenyltetrazolium bromide (MTT)

The effect of bosutinib on the selected OSCC cell lines was determined using MTT assay with $1.5 - 8 \times 10^3$ cells per well as described previously (Fadlullah et al., 2016). Cells were treated with $0.04 - 5$ µM of bosutinib, and cell viability was measured after 72 hours of treatment. DMSO (0.5%) served as vehicle control. The two-sample *t*-test was used to assess whether the difference in the sample mean of IC_{50} between the tested cell lines was statistically significant (p-value < 0.05). Experiments were repeated three times for ORL-204, HSC-4, and MCF7, four times for ORL-196, and five times for ORL-48.

University of Malaya

3.3.2.3 Apoptosis Assay

Apoptosis was quantified using a FITC Annexin V Apoptosis Detection Kit (BD Biosciences, San Jose, CA, USA) according to the manufacturer's instructions. Briefly, floating and attached cells were collected at 24, 48 and 72 hours after bosutinib treatment at 1 μM, and then stained using FITC Annexin V/Propidium iodide (PI). Apoptosis detection was performed using BD FACSCANTO™ II flow cytometer and data was analysed using the BD FACSDiva™ software (BD Biosciences, San Jose, CA, USA). For each of the three time points, the two-sample *t*-test was used to test whether the mean of the total number of apoptotic events differed significantly (p-value < 0.05) between bosutinib-treated cells and the vehicle control (0.01% DMSO) cells. Experiments were repeated two times for ORL-48, and ORL-204, and three times for ORL-196.

3.3.2.4 Proliferation Assay

The anti-proliferative effect of bosutinib on the OSCC cell lines was examined using Click-iT EdU Cell Proliferation Assay Kit (Invitrogen, Carlsbad, CA, USA) as previously described (Fadlullah et al., 2016). The cell lines ORL-48, ORL-196, and ORL-204 were treated with $0.3 - 3$ μM bosutinib, for 24 hours and cell proliferation evaluation was based on 5-ethynyl-2'-deoxyuridine (EdU) incorporation according to the manufacturer's protocol. Images were captured from 4 to 11 different fields of each treatment concentration and further analysed using EBImage (Pau et al., 2010). The percentage of EdU-labelled cells was expressed as the percentage of red fluorescent nuclei over the total number cells reflected by DAPI-stained nuclei, and the data was presented as the relative percentage compared to vehicle control cells (0 μM). The two-sample *t*-test was used to test whether the difference in the relative percentage of EdU^+ cells differed significantly (p-value < 0.05) between treatment and vehicle control for the three cell lines.

University of Malaya

Experiments were repeated two times for ORL-48 and ORL-204, and three times for ORL-196.

3.4 Evaluation of Gene Signature Similarity Scoring Algorithms

The KS statistic has been the most widely used gene signature similarity scoring algorithm to associate query signature to the drug-associated gene expression profile. More recently, several scoring algorithms have emerged since the inception of CMap in 2006, however, the systematic evaluation of the strengths and weaknesses of these algorithms remain limited (Cheng et al., 2014; Musa et al., 2017).

3.4.1 The Drug-associated Gene Expression Database for Algorithms Evaluation

To carry out the systematic evaluation of the various algorithms, the treatment-control pairs of CMap (Build 2) microarray gene expression profiles (Lamb et al., 2006) was used. A total of 7056 raw CEL microarray data files were downloaded from https://portals.broadinstitute.org/cmap/cel_file_chunks.jsp. In total, the CMap Build 2 contains 6100 unique treatment-control instances, where treatment constitutes a selection of 1309 drugs, administered at 156 different concentrations (0.01 nM – 10 μM), at two time points (six hours and 12 hours) on five cell lines (HL60, MCF7, PC3, SKMEL5 and ssMCF7) and drug responses were compared against vehicle controls (either DMSO, ethanol or complete medium). These raw CEL files were first background-corrected, normalised, and summarised into probe sets values using the standard Robust Multichip Average (RMA) algorithm (Irizarry et al., 2003). The probe sets for each of these instance were collapsed to gene symbols (maximum probe set value was chosen in the case where multiple probe sets mapped to the same gene) using the "Collapse Dataset" function provided in GSEA (Subramanian et al., 2005) with respective chip (either HG-U133A,

University of Malaya

HT-HG-U133A or HT_HG-U133A_EA) as reference. In total, 13 321 unique genes were generated.

After collapsing to the gene symbols, the resulting gene-level expressions were subjected to further analysis in order to create two types of CMap reference profile. Firstly, fold change of treatment to control values for each instance was calculated for each gene and sorted in decreasing order, thereby creating the CMap Build 2 reference profiles of ordered fold change values. Secondly, the ordered fold change values were ranked separately for each instance. Thus, the gene that was most up-regulated received rank 1 and most down-regulated received rank 13 321. The CMap reference database is a matrix of 13 321 genes x 6100 instances.

3.4.2 Gene Signature Similarity Scoring Algorithms

Following the development of the CMap reference profile, six gene signature similarity scoring algorithms were evaluated with respect to ranking analysis, positive predictive value, enrichment analysis of similar mechanism of action, and stability analysis. These six algorithms were chosen primarily because of their widespread use in drug repurposing research. To ease discussion, the following algebraic notations are introduced here: $\mathcal{C}$, the CMap reference database; i, the drug instance in the CMap reference database; j, the index of the query genes; L, the ranked list of genes; and N, the total number of genes in the CMap reference database.

3.4.2.1 Algorithm 1: Kolmogorov-Smirnov Statistic

The KS statistic algorithm described in Section 3.2.3 is the current gene signature similarity scoring algorithm implemented in DeSigN.

University of Malaya

3.4.2.2 Algorithm 2: Weighted Connectivity Score

The weighted connectivity score (WTCS; Subramanian et al., 2017) uses a composite, bi-directional version of *ES*, a KS-like statistic described previously in Subramanian et al. (2005). Let q_{up} and q_{down} be the query gene set of up-regulated and down-regulated genes respectively. The *ES* for q_{up} and q_{down} for every instance i in $\mathcal{C}$ (using the rank-ordered CMap reference profile) reflects the degree to which the query genes are overrepresented at the extremes (top or bottom) of the entire ranked list L in $\mathcal{C}$. The score is calculated by walking down the gene list L for every instance i, increasing a running-sum statistic when encountering a gene in the query and decreasing it when encountering genes not in the query. Let q_{member} be the total number of query genes (either q_{up} or q_{down}) encountered when going down the list L. If a gene is encountered in the query, then that particular gene receives the value of $1/q_{member}$; if not, it receives the value of $-1/(N - q_{member})$.

The CMap datasets used for WTCS analysis is the rank-ordered reference database, which means the rankings of the genes in the CMAp reference database for each instance *i* have been converted to the rank-ordered fold change value in descending order. Table 3.3 shows an example of how a running sum analysis of a rank-ordered dataset is computed. Genes in the list that are elements of the query gene set receive 0.25 (1/4), otherwise, they receive -0.167 (-1/(10-4)). The total number of genes in this example is $N = 10$, and the total number of encountered query genes, $q_{member} = 4$. The ES for WTCS is defined as the maximum deviation from zero value encountered in the running sum of the random walk down the list *L*, and in this particular example, the *ES* is -0.334. Figure 3.7 shows the corresponding running sum plot for Table 3.3.

University of Malaya

Table 3.3: An example of running sum analysis for a set of four encountered query genes (denoted by *). The maximum deviation from the zero value (*ES*), in this example, is -0.334 occurred at rank number two (highlighted in bold).

Gene	Rank	Running sum
PTEN	1	-0.167
CUL3	**2**	**-0.167 + (-0.167) = - 0.334**
*TSC1	3	-0.334 + 0.25 = -0.084
*REDD1	4	-0.084 + 0.25 = 0.166
TRAF3	5	0.166 + (-0.167) = -0.001
NFE2L2	6	-0.001 + (-0.167) = -0.168
*NSD1	7	-0.168 + 0.25 = 0.082
HLA-A	8	0.082 + (-0.167) = -0.085
*RB1	9	-0.085 + 0.25 = 0.165
AJUBA	10	0.165 + (-0.167) = -0.002

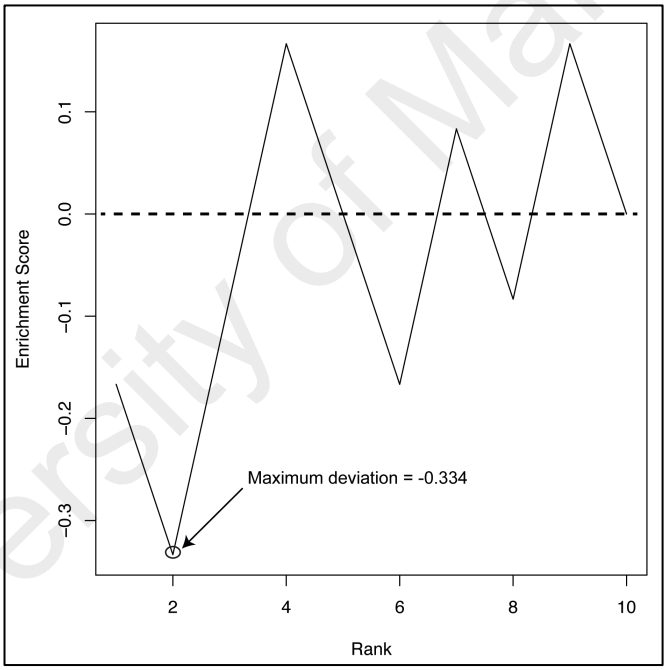

Figure 3.7: An example of running sum plot for a query set of four encountered genes. The *ES*, which is the maximum deviation from the zero dash line occurred at rank number two, -0.334.

For every query gene set, ES_{up} and ES_{down}, which denote the enrichment scores for the set of up-regulated and down-regulated genes respectively, are computed. The *CS* for the instance i in C is set to zero if both ES_{up}^{i} and ES_{down}^{i} have the same algebraic sign;

University of Malaya

otherwise, $CS^i = (ES^i_{up} - ES^i_{down})/2$. Subsequently, WTCS for non-zero instance *i* is a normalised *CS* computed as:

$$WTCS^i = \begin{cases} \frac{CS^i}{P}, & \text{if } CS^i > 0; \\ -\left(\frac{CS^i}{Q}\right), & \text{if } CS^i < 0, \end{cases}$$

where $P = \max_i CS^i$ and $Q = \min_i CS^i$ are the normalising constants. WTCS returns a ranked list of drug instances, with values range between -1 and 1. In the event where WTCS is zero, the ranking of the drug instances is ordered according to the latters'.

Table 3.4 shows an example of the WTCS analysis for six drugs. The drugs are ranked in descending order based on the sign and magnitude of WTCS. Notably, the ranking of docetaxel and AZD-0530 are ordered based on their ES_{up} values in descending order since they have zero values for WTCS.

Table 3.4: An example of WTCS calculation.

Drug	ES_{up}	ES_{down}	*CS*	*WTCS*	Ranking
BIBW2992	0.684	-0.382	0.533	1.000	1
AICAR	0.545	-0.380	0.463	0.869	2
Docetaxel	0.397	0.062	0.000	0.000	3
AZD-0530	0.393	0.191	0.000	0.000	4
Pazopanib	-0.453	0.270	-0.362	-0.995	5
GSK269962A	-0.353	0.374	-0.364	-1.000	6

3.4.2.3 Algorithm 3 and 4: eXtreme Sum and eXtreme Cosine

eXtreme Sum (XSum) (Cheng et al., 2014) uses two gene sets (Q_{up} and Q_{down}) to query the CMap reference database ($\mathcal{C}$), with $Q_{up}, Q_{down} \subseteq \mathcal{C}$. By default, *N* was set to 500. Let C_{500} and C_{-500} be the set of top and bottom 500 up-regulated and down-regulated genes by $\log_2$FC, respectively. Then, let $X_{up} = Q_{up} \cap C_{500}$ and $X_{down} =$

University of Malaya

$Q_{down} \cap C_{-500}$. Let $\text{LFC}_{\text{up}} = \{\log_2 FC_i : i \in X_{up}\}$ and $\text{LFC}_{\text{down}} = \{\log_2 FC_j : j \in X_{down}\}$. The XSum score for each instance of i in $\mathcal{C}$ was computed as

$$XSum = \sum LFC_{up} - \sum LFC_{down} = \sum_{i \in X_{up}} \log_2 FC_i - \sum_{j \in X_{down}} \log_2 FC_j.$$

Figure 3.8 shows an example of the reference database derived from a particular drug treatment. First, the top N number of genes to focus on need to be set, in this case, $N = 5$. Assuming that the user supplied $Q_{up} = \{CDKN2A, FAT1, TP53, CASP8, NOTCH1\}$ and $Q_{down} = \{HRAS, FBXW7\}$. Then $C_5 = \{CDKN2A, FAT1, TP53, CASP8, AJUBA\}$ and $C_{-5} = \{FBXW7, HRAS, TGFBR2, HLA\text{-}A, NSD1\}$, hence $X_{up} = \{CDKN2A, FAT1\}$ and $X_{down} = \{FBXW7, HRAS\}$. Thus, $XSum = (5.432 + 5.145) - (-1.117 + (-5.208)) = 16.902$ (Table 3.5). The fold change value for *NOTCH1* is not taken into consideration because XSum focuses on the extreme ends of the fold change value (top and bottom $N = 5$), ignoring those that fall in between the extreme ends. The ranking of the drugs would then be arranged in descending order of the XSum values.

Similar to the XSum algorithm, eXtreme cosine (XCos) similarity score (Cheng et al., 2013) retains the set of N top and N bottom up-regulated and down-regulated genes ranked according to $\log_2$FC. By default, N is set to 500. XCos differs from XSum in that XCos requires additional input about estimated $\log_2$FC for the query genes. The cosine similarity between the sets X_{up} and X_{down} is defined as a ratio, with the numerator being sum of the product of $LFC_{k,database}$ and $LFC_{k,user}$, over $k \in \{\text{up,down}\}$; and the denominator being the product of the Euclidean norms of $LFC_{database}$ and LFC_{user}. Thus,

$$XCos = \frac{\sum_{k \in \{up,down\}} LFC_{k,database} \cdot LFC_{k,user}}{\|LFC_{database}\| \times \|LFC_{user}\|}.$$

University of Malaya

Table 3.5 shows an illustration of the calculation for XCos using the same genes listed in Figure 3.8. Given the same condition for the case in XSum, thus,

$$\text{XCos} = (13.531 + 7.1 + 1.262 + 5.755)/(\sqrt{(29.507 + 26.471 + 1.248 + 27.123)} \times \sqrt{(6.205 + 1.904 + 1.277 + 1.221)}) = 0.924.$$

The drugs are then ranked in descending order of the XCos values.

	Gene	LogFC
Top *N* = 5	CDKN2A	5.432
	FAT1	5.145
	TP53	4.957
	CASP8	4.441
	AJUBA	3.978
	PIK3CA	3.478
	NOTCH1	3.122
	KMT2D	2.781
Bottom *N* = 5	NSD1	1.589
	HLA-A	0.498
	TGFBR2	0.389
	HRAS	-1.117
	FBXW7	-5.208

Figure 3.8: An example of the reference database used for XSum and XCos analysis. Cells highlighted green and red denote overlapping of up-regulated and down-regulated genes respectively with the reference database.

University of Malaya

Table 3.5: An example of XSum and XCos calculation. Cells highlighted green and red denote overlapping of up-regulated and down-regulated genes respectively with the reference database.

Gene	Log_2FC (database)	Log_2FC (user)	Log_2FC (database) x Log_2FC (user)	$[Log_2FC$ (database)$]^2$	$[Log_2FC(user)]^2$
CDKN2A	5.432	2.491	13.531	29.507	6.205
FAT1	5.145	1.380	7.100	26.471	1.904
TP53	4.957	3.746	18.569	24.572	14.033
CASP8	4.441	-1.775	-7.883	19.722	3.151
AJUBA	3.978	0.456	1.814	15.824	0.208
PIK3CA	3.478	3.142	10.928	12.096	9.872
NOTCH1	3.122	-1.113	-3.475	9.747	1.239
KMT2D	2.781	-0.488	-1.357	7.734	0.238
NSD1	1.589	2.945	4.680	2.525	8.673
HLA-A	0.498	1.389	0.692	0.248	1.929
TGFBR2	0.389	2.212	0.860	0.151	4.893
HRAS	-1.117	-1.130	1.262	1.248	1.277
FBXW7	-5.208	-1.105	5.755	27.123	1.221

3.4.2.4 Algorithm 5 and 6: sscMap unOrdered and sscMap Ordered

CMap reference profiles of ordered fold change values were used in evaluating the statistically significant Connectivity Map (sscMap) algorithm (Zhang & Gant, 2008). The main difference between sscMap and the other four algorithms is that all the genes for every instance i in CMap reference profile $\mathcal{C}$ were first ranked in ascending order using the absolute value of log_2FC. The ranks are then multiplied with the sign of the log_2FC value, so that they become signed ranks. Subsequently, the final ranks of the genes for each instance i in CMap reference profile $\mathcal{C}$ were re-ordered based on the magnitude of the signed rank in descending order. In this way, the importance of a gene is indicated by the magnitude of the rank, with the sign indicating its regulation status.

Table 3.6 shows an example of the sscMap reference database for a particular drug instance. After ordering based on the magnitude of the signed ranked in descending order,

University of Malaya

both *FBXW7* and *HRAS*, for example, get the rank of two and nine respectively, with the negative sign indicating their down-regulation status.

Table 3.6: An example of the sscMap reference database for one particular drug instance. The genes are first ordered in ascending order using the absolute value of log_2FC. The sign of the log_2FC value was then used to multiply the rank to generate the signed rank. The final rank of the genes, meanwhile, takes on the magnitude of the signed rank in descending order.

Gene	Log_2FC	Signed rank
CDKN2A	5.432	10
FBXW7	-5.208	-9
FAT1	5.145	8
TP53	4.957	7
PIK3CA	3.478	6
NOTCH1	3.122	5
KMT2D	2.781	4
NSD1	1.589	3
HRAS	-1.117	-2
HLA-A	0.498	1

A query gene set, meanwhile, can be an ordered gene list, or just a collection of genes without specific ordering, which will be referred to as *ordered* and *unordered* gene list respectively. For an ordered gene set, the genes in the list are ranked the same way as the CMap reference profile $\mathcal{C}$ (see Table 3.6). Thus, the gene with the strongest degree of differential expression in the set of *m* query genes will receive the signed rank of *m* or –*m*; whereas the one with the weakest degree of differential expression will receive the signed rank of 1 or -1. The connection strength between the set of query genes of size *m* (Q_m) and the set of genes in the database of size *N* ($\mathcal{C}_N$) is defined as

$$C(Q_m, \mathcal{C}_N) = \sum_{i \in Q_m} \text{signed rank}_{i,database} \times \text{signed rank}_{i,query}.$$

University of Malaya

Table 3.7 shows an example of the connection strength calculation for one particular drug instance.

Table 3.7: An example of sscMap connection strength calculation for one particular drug instance.

Gene	Query signed rank	Database signed rank	Connection strength
TNFSF9	5	10491	52455
RAB40B	-4	-8064	32256
PMM1	3	881	2643
TACC1	-2	12289	-24578
LGALS1	1	-12643	-12643
Sum			50133

For an ordered gene list, after calculating the connection strength for every instance *i* in CMap reference profile $\mathcal{C}$, the maximum positive connection strength is given by:

$$C^{o}_{max}(Q_m, C_N) = \sum_{j=1}^{m}(N - j + 1)(m - j + 1).$$

Similarly, maximum negative connection strength = $-C^{o}_{max}(Q_m, C_N)$.

For an unordered query gene list, the signed rank of $Q_{i,query}$ is replaced with the sign function for up-regulation (+1) and down-regulation (-1). Thus, the maximum magnitude of connection strength for an unordered query list is given by

$$C^{u}_{max}(Q_m, C_N) = \sum_{j=1}^{m}(N - j + 1).$$

Given a query gene set and a reference gene expression profile, the connection score is computed as the normalised connection strength, i.e., $c^{o} = C(Q_m, \mathcal{C}_N)/C^{o}_{max}(Q_m, \mathcal{C}_N)$ for ordered query gene list, and $c^{u} = C(Q_m, \mathcal{C}_N)/C^{u}_{max}(Q_m, \mathcal{C}_N)$ for unordered query gene list.

University of Malaya

Table 3.8 below shows an example of sscMap calculation. The maximum value of the connection strength, which is 53893586 (highlighted bold in Table 3.8) is used to normalise the sscMap score, c. The value $c = 1$ means that the query gene set has the maximum positive connection strength with the instance i in the CMap reference profile $\mathcal{C}$; and $c = -1$ indicates that the query gene set and the instance i are most inversely correlated. The range of c is from -1 to 1.

Table 3.8: An output example of sscMap calculation.

Drug	Connection strength	sscMap score, c	Rank
Afatinib	53893586	1.000	1
Bosutinib	53116288	0.986	2
Cisplatin	11813766	0.219	3
Dasatinib	-20280839	-0.376	4
Elesclomol	-26889776	-0.499	5
Maximum	53893586		

3.4.3 Query Signatures

The 39 query signatures extracted from the Ushijima dataset was used to test the prediction performance of these six gene signature similarity scoring algorithms (http://scads.jfcr.or.jp/db/cs/download_csv2.html). Notably, these 39 query signatures were derived from 19 compounds relative to untreated cells as negative controls. These 19 compounds mainly consist of clinically-used standard anticancer agents and related drugs (Ushijima et al., 2013). The details of these 39 signatures are listed in Table 3.9 and the list of up-regulated and down-regulated genes for each signature can be found in Appendix 3. These 39 query signatures were subjected to gene signature similarity analysis using CMap as the reference profile database, by which performance of these algorithms was evaluated. Some of the signatures (indicated by * in Table 3.9) were excluded for positive predictive value evaluation. The reasons for these are explained in the Results section in Chapter 4.

University of Malaya

Table 3.9: Details of 39 Ushijima signatures. Signatures denoted by * were excluded from positive predictive value analysis. Abbreviation: MoA = mechanism of actions.

Signature	Compound	# up-regulated genes	# down-regulated genes	Target/ MoA
C001	2-deoxy-D-glucose	279	212	Glycolysis
C003	Thapsigargin	148	75	SERCA
C006	Trichostatin A	716	580	HDAC
C007	Vorinostat	698	588	HDAC
C009	MG-132	219	274	Proteasome
C010*	Geldanamycin	16	20	Hsp90
C011*	Tanespimycin	40	21	Hsp90
C013*	Paclitaxel	16	4	Tubulin
C023*	Methotrexate	21	31	DHFR
C024*	Mercaptopurine	19	31	Purine
C029	Irinotecan	280	1 151	Topoisomerase I
C030	Camptothecin	342	1 434	Topoisomerase I
C032	Doxorubicin	227	601	DNA intercalator / Topoisomerase II
C033*	Etoposide	23	77	Topoisomerase II
C034	Mitoxantrone	162	642	DNA intercalator / Topoisomerase II
C044*	Gefitinib	20	17	EGFR
C045*	Gefitinib	104	54	EGFR
C047	Vorinostat	950	934	HDAC
C049*	Geldanamycin	5	934	Hsp90
C050*	Tanespimycin	6	4	Hsp90
C052	Paclitaxel	95	147	Tubulin
C056	Vorinostat	765	652	HDAC
C058	Vorinostat	1 021	953	HDAC
C064	Methotrexate	356	249	DHFR
C065*	Mercaptopurine	173	139	Purine
C069	Etoposide	271	139	Topoisomerase II
C090*	Decitabine	3	16	DNA methyltransferase
C101	Irinotecan	351	888	Topoisomerase I
C102	Doxorubicin	187	1 104	DNA intercalator / Topoisomerase II
C104	Irinotecan	458	941	Topoisomerase I
C105*	Doxorubicin	95	104	DNA intercalator / Topoisomerase II
C106*	Imatinib	137	134	Bcr-Abl/KIT
C111	Irinotecan	234	843	Topoisomerase I
C112	Doxorubicin	57	88	DNA intercalator/ Topoisomerase II
C115*	Gefitinib	90	100	EGFR

University of Malaya

C116*	Gefitinib	130	107	EGFR
C118*	Celecoxib	666	640	COX2
C128	Irinotecan	129	695	Topoisomerase I
C129	Doxorubicin	11	857	DNA intercalator ⁄ Topoisomerase II

3.4.4 Algorithm Performance Evaluation

To systematically evaluate the performance of these gene signature similarity scoring algorithms, four performance evaluation metrics, i.e., ranking analysis, positive predictive value (PPV), enrichment analysis of similar mechanism of action (MoA), and ranking analysis were employed. In particular, PPV was chosen because it focuses on the reliability of positive results. Table 3.10 shows a 2 x 2 contingency table for summarising the four possible types of truth-prediction outcomes.

Table 3.10: A 2 x 2 contingency table for algorithm performance metric evaluation.

		True condition		
	Total population	Positive instance (*P*)	Negative instance (*N*)	
Predicted condition	Predicted positive instance (*T*)	**True positive (*TP*)**	**False positive (*FP*)**	$PPV = \frac{TP}{TP + FP}$
	Predicted negative instance (*N*)	**False negative (*FN*)**	**True negative (*TN*)**	

In total, there are four possible output in a 2 x 2 contingency table. The green diagonal cells in Table 3.10 represent correct predictions, and the pink diagonal cells indicate incorrect predictions. For a general understanding, if the sample is positive and it is predicted as positive, i.e., correctly predicted positive sample, it is counted as a *true positive* (*TP*); if it is predicted as negative, it is considered as a *false negative* (*FN*). If the sample is negative and it is predicted as negative it is considered as *true negative* (*TN*); if it is predicted as positive, it is counted as *false positive* (*FP*). In this particular algorithm

University of Malaya

performance evaluation analysis, the "sample" is hereinafter referred to the intended compound in CMap reference database.

3.4.4.1 Ranking Analysis

The goal of the ranking analysis is to gauge which scoring algorithm performs better in returning the highest ranking of the drug instance associated with the respective Ushijima signature. Of note, each signature could be associated with more than one drug instance in CMap reference database. For example, out of the total number of 1309 drug instances in CMap reference database, 182 of those drug instances are associated with vorinostat (HDAC inhibitor) derived from different treatment concentrations and time points. Therefore, to evaluate the ranking performance of the scoring algorithm using, for example, the Ushijima Signature C007 (vorinostat), the highest ranking returned by any of these 182 drug instances would be taken into consideration. The same analogy applies to the other 38 Ushijima signatures.

3.4.4.2 Positive Predictive Value

The second performance metric is the PPV. It represents the proportion of positive samples that were correctly classified among total number of positive predicted samples (Sokolova et al., 2006) as indicated in the following equation (see Table 3.10):

$$PPV = \frac{TP}{TP + FP}.$$

Given a gene signature, the PPV analysis is carried out by evaluating the performance of the six algorithms at different interval of *K*. Here, the interval of *K* is set at {1, 5, 10, 15, 20, 25, 30, 35, 40, 45, 50}. To ease understanding, *K* is set at 10 (i.e., *K* = 10) when PPV is used to analyse the proportion of correctly classified positive samples

University of Malaya

within the top-10 of the prediction result returned by DeSigN. The same analogy applies to other interval of *K*.

3.4.4.3 Mechanism of Action Enrichment Analysis

The goal to carry out the MoA enrichment analysis is to reflect the degree to which the drug instances of a similar MoA is overrepresented at the extremes (top or bottom) ends of the entire ranked list of the 1309 drug instances in the CMap reference database. For example, if a signature is associated with *EGFR* inhibition, inhibitors such as gefitinib, erlotinib, afatinib, and lapatinib that target the *EGFR* signaling should theoretically appear at the top of the returned ranked list of inhibitors upon querying the CMap reference database. In such a case, the MoA enrichment analysis would take on a high positive value, and vice versa. The idea of enrichment analysis for similar MoA leverages on the seminal work of GSEA by Subramanian et al. (2005). To translate the concept of GSEA into the context of similar MoA, the enrichment analysis is calculated by walking down the ranked list of instances, increasing a running-sum statistic when encountering an instance with the same MoA and decreasing it when encountering instances with different MoA.

3.4.4.4 Stability Analysis

The stability analysis of the gene signature similarity scoring algorithms was carried out to evaluate the sensitivity of each algorithm towards varying query size. An algorithm is considered good and practical if it consistently returns the intended compound upon querying the CMap reference database under varying sample sizes of the original query. Xu et al. (2005) first devised the stability analysis on their new algorithm - the top-scoring pair (TSP) classifier, in identifying prostate cancer biomarkers from various microarray

University of Malaya

training datasets. To perform stability analysis on gene signature similarity scoring algorithms using varying query size, a subset of the original Ushijima query signature (C006 and C058) was randomly sampled, and simulated rankings from the reduced rank-ordered query datasets were generated. After repeating for ten times (n = 10) with the different values of simulated sample size B (B = 50, 100, 200, 400 and 800), the mean ranks of each sample size B was calculated for each algorithm. The list of up-regulated and down-regulated genes of respective simulated sample sizes for signature C006 and C058 can be found in Appendix 4.

3.5 Computational Work

All methods, unless stated otherwise, are implemented in R computing environment (R 3.5.1; R Core Team 2018). All the R codes developed for the computational analyses are available at https://gitlab.com/blkb0427/phd-thesis.

University of Malaya

CHAPTER 4: RESULTS

4.1 GENIPAC: A Platform to Visualise Genomic Data from HNSCC Cell Lines

The compilation of all available genomic data from HNSCC cell lines into a single web resource is a critical first step towards enabling routine querying of this resource for prediction of drug efficacy against HNSCC cell lines. To this end, an interactive web resource – GENIPAC (http://genipac.cancerresearch.my/) was built to enable exploration of the compiled genomic data from HNSCC cell lines. GENIPAC runs on the cBioPortal engines; it hosts mutations, mRNA gene expression, and copy number alterations data for 44 HNSCC cell lines taken from three individual studies (Figure 4.1A) (Fadlullah et al., 2016; Martin et al., 2014; Prime et al., 1990). The summary view page, assessed through the "Summary" icon (Figure 4.1B) presents overview information, such as the names of the HNSCC cell lines and clinical data associated with the selected dataset. For example, the ORL Series dataset consists of 16 HNSCC cell lines derived from an Asian population. The clinical information (e.g. patient demographics, risk factors, and primary sites of the tumour) associated with these cell lines is summarised in Figure 4.1C. Appendix 5 contains details of demographics of patients who contributed the cell lines in all three HNSCC studies.

To explore, visualise, and analyse the cancer genomics data, we can query the datasets in GENIPAC using gene sets of interest from a single study, or multiple studies (Figure 4.1D). Besides querying the three datasets directly on the web interface, the genomic data can be downloaded through the "Download Data" tab for offline analysis (Figure 4.1E). GENIPAC enables researchers to examine useful summary statistics, such as the frequency of unique genetic alterations. Additionally, it makes it easy to identify HNSCC

University of Malaya

cell lines with specific genomic profiles – a useful step toward crafting hypothesis-driven research projects in oral cancer biology.

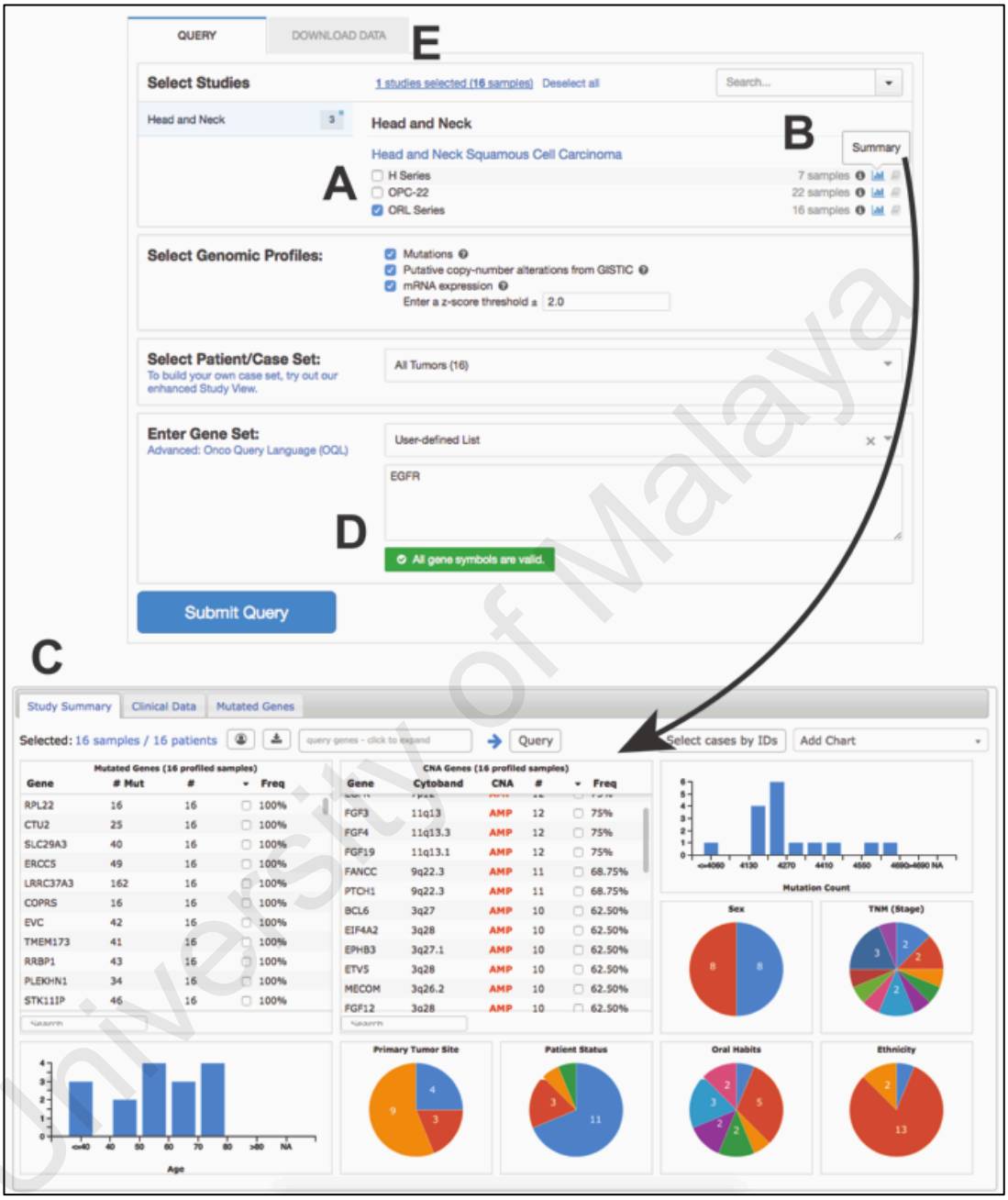

Figure 4.1: Query page of the GENIPAC. (A) Users have the option of choosing which head and neck cancer studies and genomics datasets to query from this panel. (B, C) Additional information of each study can be obtained under the "Summary" tab. (D) Users can key in their gene sets of interest here in order to query the three datasets. (E) Option to download the datasets for offline analysis is also available.

University of Malaya

Many of the gene alterations reported for HNSCC are found in the cell lines in GENIPAC. For example, the mutational status of the top five most significant mutated genes (*TP53*, *FAT1*, *CDKN2A*, *PIK3CA*, and *NOTCH1*) were assessed in 279 HNSCC tumours from the TCGA database (Cancer Genome Atlas Network, 2015) and the cell lines within GENIPAC. Mutations observed in HNSCC tumours were consistent with those in HNSCC cell lines in GENIPAC (Figure 4.2 A, B), albeit with some variation in the frequencies across the different datasets. An overview of the top five mutated genes between cell lines and tumours showed good representation of the unique mutational patterns within HNSCC. A closer examination of these top five mutated genes revealed a distinct HNSCC tissue-specific mutational pattern that is different from other solid tumours available in TCGA, such as melanoma, leukemia, pancreatic, and breast cancers (Appendix 6). Besides reporting the percentage of a mutation within a HNSCC study, additional details about the mutation can be viewed through a mouse-over operation (Figure 4.2C).

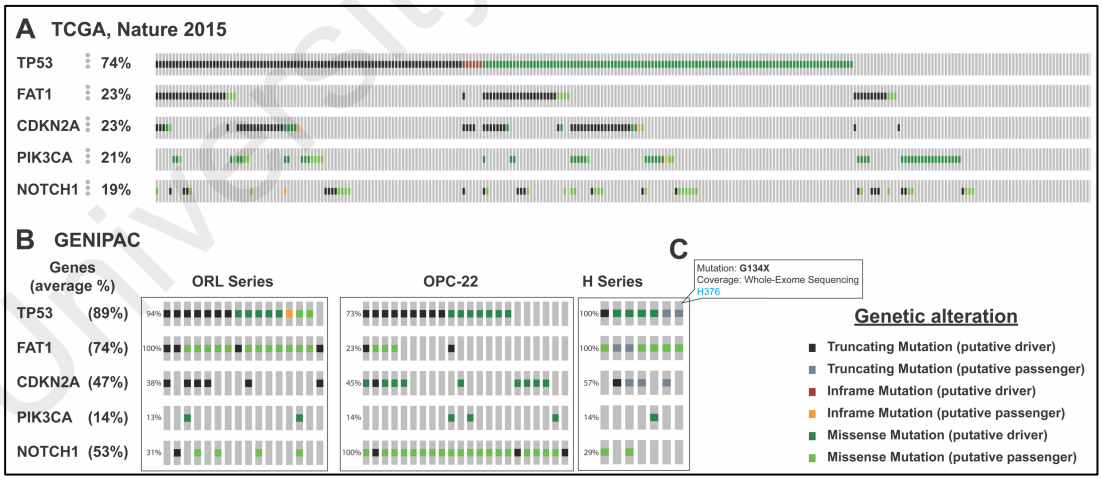

Figure 4.2: Overview of the mutational distribution pattern of the top five most mutated genes in HNSCC. (A) TCGA and (B) GENIPAC. *TP53* is the most mutated gene in TCGA (74%) and GENIPAC (89%). The percentage shown is the number of specimens or cell lines that contain any genetic change within the gene. (C) Details of the mutations can be viewed through mouse-over operations on the cell representing each cell line.

University of Malaya

To illustrate the functionality of GENIPAC, *TP53* was used as a case study. By querying GENIPAC with *TP53*, most of the *TP53* mutations (54/92; 59%) were observed to be missense mutations that frequently occurred within the DNA-binding domain (amino acid positions: 95 to 289; Figure 4.3A), in concordance with previous reports (Chang et al., 2016). Notably, some of the most common mutational hotspots, such as at the amino acid position of R175, H179, R196, R273, and R282 were also reported in Catalogue of Somatic Mutations in Cancer (COSMIC) database (Figure 4.3A, Appendix 7). The HNSCC lines that harbour these *TP53* mutations, such as ORL-166, ORL-204, ORL-215, CAL33, DETROIT-562, WSU-HN6, and WSU-HN8 can be easily identified using GENIPAC (Figure 4.3B). Additionally, different features associated with these *TP53* mutations, including their frequency in COSMIC as well as clinical implication, and biological effects are well-documented in tabular form as shown in Figure 4.3B.

University of Malaya

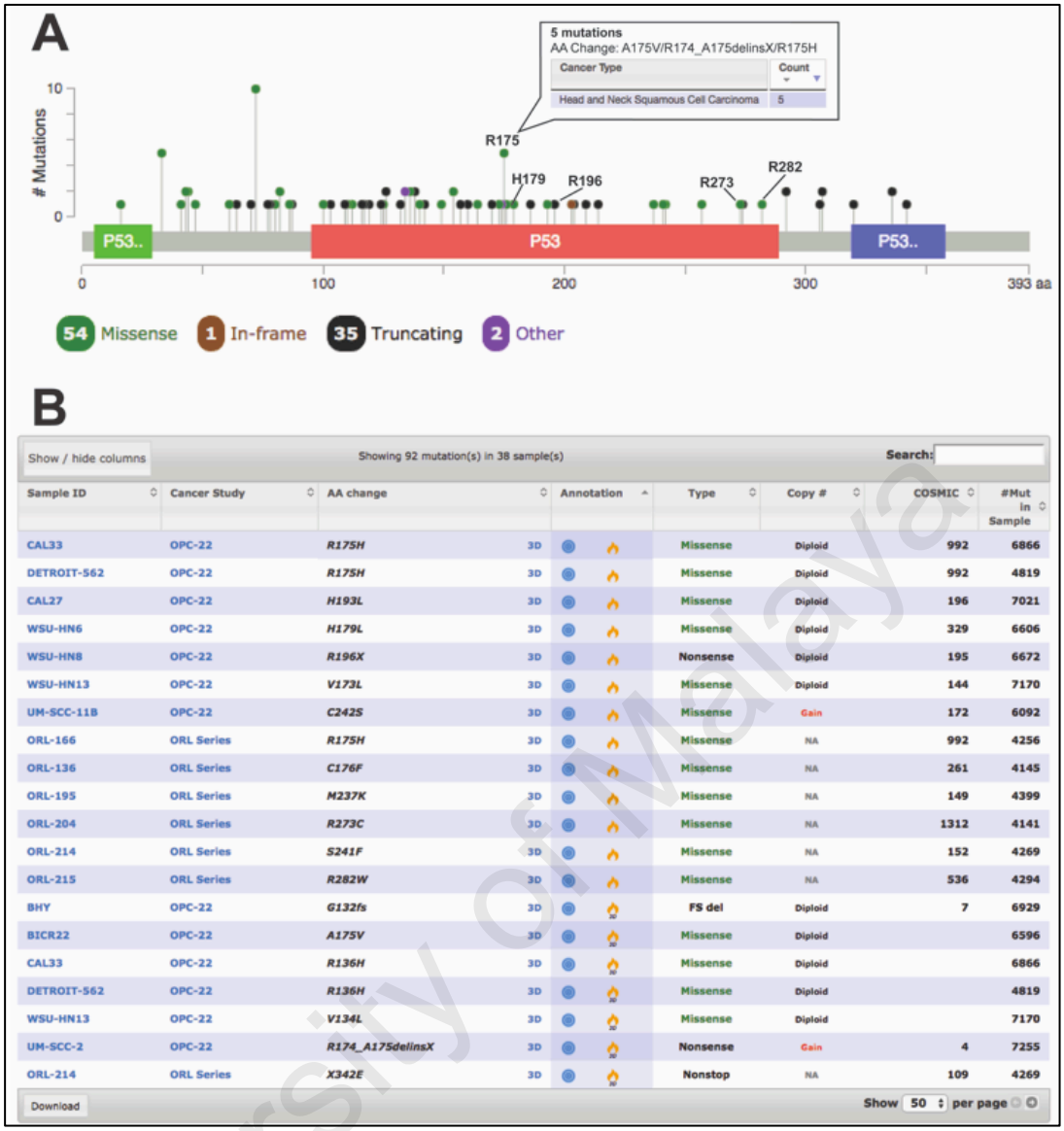

Figure 4.3: Distribution of *TP53* mutations in GENIPAC across the Pfam protein domains. (A) Hotspot mutations in *TP53* at amino acid positions R175, H179, R196, R273, and R282 are consistently shown in cell lines within GENIPAC. (B) Information about the type of mutation and its clinical implication, biological effects, and frequency reported in COSMIC are tabulated in detail.

University of Malaya

4.2 mRNA Expression and Copy Number Alterations

GENIPAC allows joint visualisation of gene expression levels with copy number variation in the HNSCC cell lines. The example of the epidermal growth factor receptor (*EGFR*) is considered here. Understanding the molecular biology of *EGFR* signaling in HNSCC pathogenesis has led to the clinical use of cetuximab as targeted therapy for HNSCC (Bonner et al., 2006; Vermorken et al., 2008) and other inhibitors targeting

EGFR, used alone or in combination with a variety of therapies that are currently being evaluated (ClinicalTrials.gov Identifier: NCT02979977, NCT00496652, and NCT00083057). In HNSCC, high *EGFR* expression levels are correlated with poor prognosis and resistance to radiation therapy (Baumann & Krause, 2004; Zimmermann et al., 2006). An examination of the *EGFR* status in 279 HNSCC tumours from TCGA (Cancer Genome Atlas Network, 2015) showed that *EGFR* is overexpressed and/or amplified in 20% of the specimens (Figure 4.4A, B). Specifically, of the 279 HNSCC tumours in TCGA, overexpression of *EGFR* is reported in 17% (47/279) of them, and amplification of copy number in 11% (30/279) of them.

GENIPAC enables the identification of cell lines that overexpress *EGFR*. Thus, the majority of the HNSCC cell lines in the ORL Series (12/16; 75%) are shown to have *EGFR* amplification concurrent with overexpression (Figure 4.4C, D). Figure 4.4D shows that the cell line in the ORL Series that has the highest level of *EGFR* expression is ORL-136, as reported previously (Fadlullah et al., 2016). Another gene that is commonly amplified and overexpressed in HNSCC is *CCND1*, located on chromosome 11q13 (Smeets et al., 2006). Overexpression of *CCND1* results in the activation of different pathways controlling cell cycle progression, migration, and differentiation (Musgrove et al., 2011). Targeting the CDK4/6-cyclin axis is currently under investigation using the drug palbociclib (Michel et al., 2016), and it is anticipated that further work to identify biomarkers of response will be an essential component in this research area. Using GENIPAC, researchers can quickly identify cell lines that contain *CCND1* amplification and those that have concomitantly high levels of mRNA (Figure 4.4C, D), so that appropriate drug testing and biomarker experiments can be designed using these cell lines.

University of Malaya

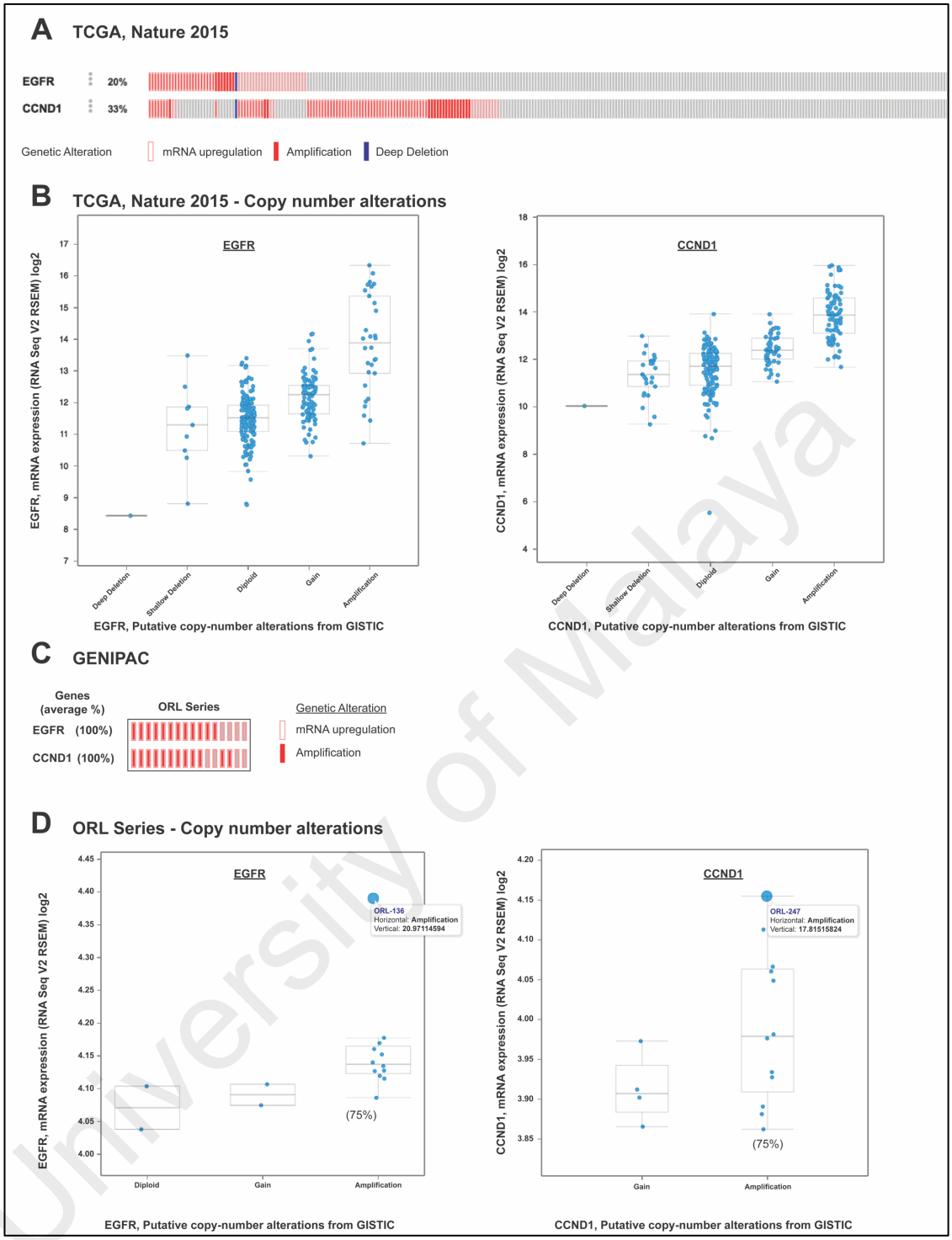

Figure 4.4: mRNA expression and copy number variations of *EGFR* and *CCND1* in TCGA and GENIPAC. (A) Genomics profiles (mRNA expression and copy number variation) of *EGFR* and *CCND1* are altered in 55 tumours (20%) and 92 tumours (33%), respectively, in TCGA. (B) Copy number alterations of *EGFR* and *CCND1* in TCGA. Similar trends were observed where amplification of copy number is correlated with overexpression of *EGFR* and *CCND1*. (C, D) *EGFR* and *CCND1* are overexpressed in all cell lines in the ORL Series, while 75% of the lines in the ORL Series have amplification in *EGFR* and *CCND1*. Cell lines with amplification tend to have higher expression of *EGFR* and *CCND1*.

University of Malaya

4.3 Visualising Genetic Alterations within Pathways using GENIPAC

Phosphatidylinositol 3-kinase (PI3K) is the most altered mitogenic signaling pathway in HNSCC, harbouring the highest percentage of mutations in patients with HNSCC (30.5%) (Lui et al., 2013). In particular, *PIK3CA* (p110) is the most mutated gene in this pathway, affecting 12.6% of patients with HNSCC, while copy number gain and mRNA overexpression of *PIK3CA* are also frequent events, occurring in 20% and 52% of patients with HNSCC, respectively (Iglesias-Bartolome et al., 2013). Genetic alterations in this pathway have clinical implications, and drugs that target specific components within this pathway are actively being studied and tested (Isaacsson Velho et al., 2015). Notably, different components of the pathway could be altered, and these have been demonstrated to modulate drug response. For example, a recent report of five HNSCC cases found that mTOR-based targeted therapy may be more effective in HNSCC tumours harbouring *PIK3CA* mutation and/or *PTEN* loss of expression (Holsinger et al., 2013).

To examine the status of the PI3K pathway in HNSCC cell lines, five representative molecules (*PIK3CA, PIK3R1, PTEN, AKT1*, and *MTOR*) in the PI3K pathway were studied. As reported in TCGA (Appendix 8), alterations in the PI3K pathway in patients with HNSCC are represented in these cell lines, and those with amplification, mRNA upregulation, and mutations in the different components are easily identifiable (Figure 4.5A). Specifically, the commonly reported *PIK3CA* activating mutations: E542K, E545K, and H1047L/R are presence in all three series of HNSCC cell lines in GENIPAC. These lines, including ORL-115 (ORL Series), H400 (H Series), CAL33 (OPC-22) could be useful models for the evaluation of therapies targeting the PI3K pathway.

University of Malaya

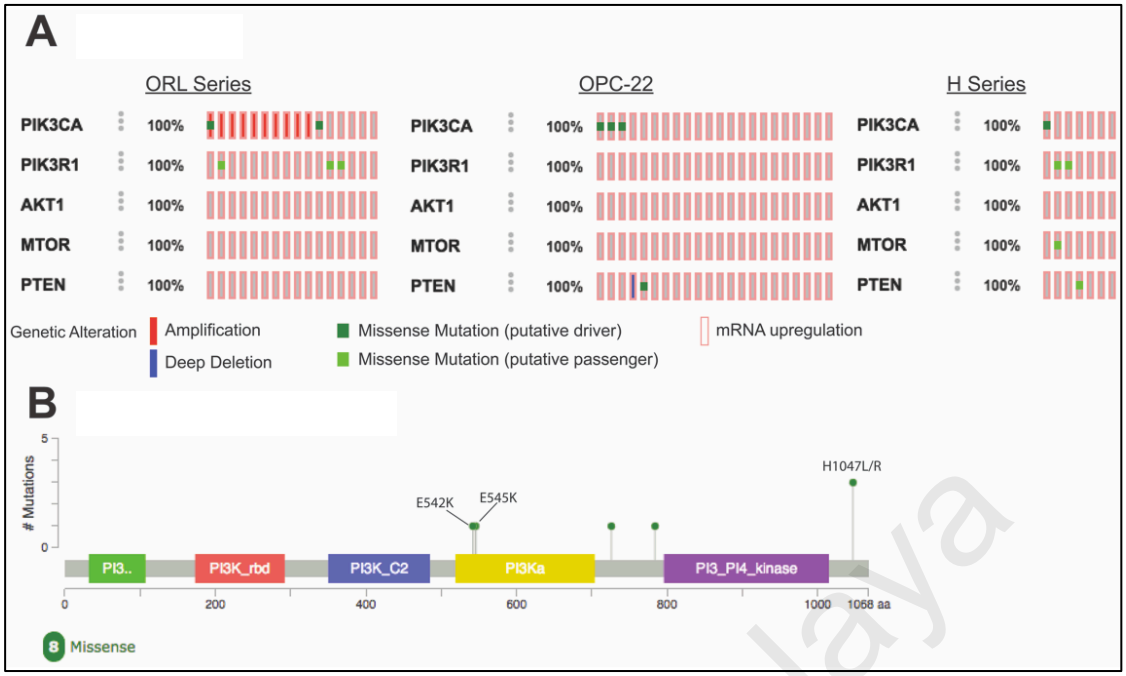

Figure 4.5: Overview of the five representative genes involved in the PI3K pathway in GENIPAC. (A) Out of the five genes involved in the PI3K pathway, *PIK3CA* has the highest frequency of alteration events in mRNA expression, mutation, and copy number variations. The percentage shown is the number of cell lines that contain any genetic change within the gene. (B) Hotspot activating mutations (E542K, E545K, H1047L/R) were present in cell lines included in GENIPAC.

Judging from the comparisons of the distribution of the most mutated genes, the mRNA expression levels, copy number amplification, and analysis of essential pathways, it is evident that the genomic information of the tumours are well-represented by the HNSCC cell lines. These well-curated genomics data therefore accentuate the relevance of using HNSCC cell lines to evaluate the efficacy of drug candidates. In the next section, the gene expression profiles of the ORL Series hosted in GENIPAC will be used to showcase the practical use of a bioinformatics tool (DeSigN; see Section 4.4) in predicting therapeutic drugs for HNSCC cell lines.

4.4 Identifying Drugs through DeSigN

DeSigN (Differentially Expressed Gene Signatures – Inhibitors) (http://design.cancerresearch.my/) works by connecting gene signature to pre-defined

University of Malaya

gene expression profiles that are associated with drug response data (IC_{50}) of 140 drugs. To gauge the prediction efficiency of DeSigN, *in silico* validation using two published drug sensitivity datasets were first done. Following that, DeSigN was used to shortlist potentially efficacious inhibitors for OSCC lines. Subsequently, *in vitro* experiments were carried out to validate the efficacy of the shortlisted drug candidate on a panel of OSCC cell lines.

4.4.1 *In silico* Validation of Candidate Compounds Predicted using DeSigN

Two drug sensitivity studies with published drug response and microarray gene expression datasets deposited in the NCBI GEO database (GSE9633 and GSE4342) were used to computationally validate predictions using DeSigN. Primarily, these two studies were chosen because the tested drugs - dasatinib and gefitinib, are available in the DeSigN database. Dasatinib is an oral dual BCR/ABL and SRC family tyrosine kinase inhibitor approved for use in patients with chronic myelogenous leukemia (CML). Dasatinib has since been found to be effective as well for treating other kinds of cancer. While testing for the efficacy of dasatinib in a panel of prostatic CCL, Wang et al. (2007) chanced upon a subset of the prostatic lines that was sensitive to dasatinib treatment (GSE9633). Likewise, gefitinib is an *EGFR* inhibitor approved for non-small cell lung cancer (NSCLC) treatment. However, upon subjecting a panel of NSCLC lines to getifinib treatment, Coldren et al. (2006) reported that only a subset of the NSCLC lines was found to have a sensitive response towards gefitinib treatment (GSE4342). With respect to these two published drug sensitivity studies, both dasatinib and gefitinib are expected to be among the top-ranked inhibitors with high positive CS (indicating a sensitive response) returned by DeSigN. The DeSigN analyses for both studies were carried out using the DEGs derived from the differential analysis of sensitive versus resistant lines in response to respective drugs.

University of Malaya

4.4.1.1 GSE9633 Dataset

The GSE9633 dataset contains Affymetrix microarray gene expression profiles of a panel of 16 prostatic CCL (Wang et al., 2007). This panel of 16 prostatic CCL was used to test the efficacy of dasatinib. In this study, 11 lines were found to be sensitive to dasatinib treatment, with IC_{50} < 1μM, while another five lines had $IC_{50} \geq 1$ μM (hence defined as resistant). By querying DeSigN using 848 up-regulated and 553 down-regulated genes derived from differential gene expression analysis of sensitive versus resistant lines using limma, dasatinib was returned as one of the top-ranked (ranked eighth) drugs with CS of 0.993 (*p*-value 0.025) (Figure 4.6).

Drug	Target	KS_UP	KS_DOWN	Connectivity Score	P-Value
Vinblastine	Microtubules	0.98345	-0.99499	1.000	0.000
CGP-082996	CDK4	0.98878	-0.98840	0.999	0.005
CCT018159	HSP90	0.99084	-0.98379	0.998	0.015
AZD8055	MTORC1/2	0.99303	-0.97651	0.996	0.022
Epothilone-B	Microtubules	0.98535	-0.98566	0.996	0.014
TW-37	BCL-2, BCL-XL	0.97984	-0.99076	0.996	0.021
Parthenolide	NFKB1	0.98231	-0.98496	0.994	0.026
Dasatinib	ABL, SRC, KIT, PDGFR	0.98580	-0.97846	0.993	0.025
XMD8-85	ERK5 (MK07)	0.98948	-0.97392	0.992	0.035
Pyrimethamine	Dihydrofolate Reductase (DHFR)	0.97905	-0.98280	0.992	0.034
Bexarotene	Retinioic Acid X Family Agonist	0.96921	-0.99295	0.992	0.038
NVP-BEZ235	PI3K (Class 1) And MTORC1/2	0.96805	-0.99475	0.992	0.046
GSK-650394	SGK3	0.98486	-0.97669	0.991	0.031
Roscovitine	CDKs	0.96474	-0.99514	0.991	0.044
Sorafenib	PDGFRA, PDGFRB, KDR, KIT, FLT3	0.97017	-0.98609	0.989	0.037
Vorinostat	HDAC Inhibitor Class I, IIa, IIb, IV	-0.99279	0.97273	-0.992	0.037
PF-4708671	P70 S6KA	-0.98886	0.99270	-1.000	0.000

Figure 4.6: DeSigN prediction result for GSE9633. Using the derived DEG, 17 out of the 140 drugs were found to have *p*-value < 0.05, with which 15 of them to have positive CS and two with negative CS. The intended drug, dasatinib, was returned by DeSigN (ranked eighth), with a CS of 0.993 and *p*-value 0.025 (red line).

University of Malaya

4.4.1.2 GSE4342 Dataset

The GSE4342 dataset contains Affymetrix microarray gene expression profiles of 29 NSCLC cell lines. Coldren et al. (2006) tested the sensitivity of these NSCLC lines against gefitinib (an *EGFR*-inhibitor). Out of these 29 NSCLC lines, 18 lines were found to have $IC_{50} < 1\mu M$ (hence defined as sensitive) while the remaining 11 lines to have $IC_{50} \geq 1\mu M$ (hence defined as resistant) (Coldren et al., 2006). By querying DeSigN using 357 up-regulated and 278 down-regulated genes derived from differential gene expression analysis of sensitive versus resistant lines using limma, gefitinib was returned as one of the top-ranked inhibitors (ranked seventh) with a CS of 0.985 (p-value = 0.042) (Figure 4.7). Two additional *EGFR* drugs: BIBW2992 (afatinib; ranked first), and lapatinib (ranked third), were also returned with high positive CS (p-value < 0.05). Besides, two inhibitors, RDEA119 (refametinib; ranked fourth) and AZD6244 (selumetinib; ranked fifth) targeting the *MEK1/2*, which is the effector molecule activated by *EGFR*, were also returned as candidate drugs with high positive CS and p-value < 0.05. Against the backdrop of the GSE4342 dataset, which aimed at testing the inhibitory effects of *EGFR* on NSCLC, finding a list of efficacious candidate drugs returned by DeSigN that primarily targeted molecules in the *EGFR* pathways strengthens the confidence in the utility of DeSigN.

University of Malaya

Drug	Target	KS_UP	KS_DOWN	Connectivity Score	P-Value
BIBW2992	EGFR, ERBB2	0.98422	-0.99992	1.000	0.000
AICAR	AMPK Agonist	0.98394	-0.99765	0.999	0.013
Lapatinib	EGFR, ERBB2	0.98639	-0.98450	0.993	0.023
RDEA119	MEK1/2	0.97645	-0.99249	0.992	0.028
AZD6244	MEK1/2	0.98263	-0.98137	0.990	0.036
Pyrimethamine	Dihydrofolate Reductase (DHFR)	0.98491	-0.97067	0.986	0.046
Gefitinib	EGFR	0.98719	-0.96727	0.985	0.042
Pazopanib	VEGFR, PDGFRA, PDGFRB, KIT	-0.98868	0.98014	-0.989	0.032
Nilotinib	ABL	-0.99105	0.97699	-0.989	0.038
Elesclomol	HSP70	-0.99432	0.97354	-0.989	0.037
Camptothecin	TOP1	-0.99836	0.97425	-0.991	0.026
Vorinostat	HDAC Inhibitor Class I, IIa, IIb, IV	-0.99162	0.98349	-0.992	0.025
AZD-2281	PARP1/2	-0.99781	0.97605	-0.992	0.036
AMG-706	VEGFR, RET, C-KIT, PDGFR	-0.98424	0.99257	-0.993	0.022
Axitinib	PDGFR, KIT, VEGFR	-0.98765	0.99093	-0.994	0.013
BX-795	TBK1, PDK1, IKK, AURKB/C	-0.99123	0.98740	-0.994	0.011
AG-014699	PARP1, PARP2	-0.99859	0.99187	-1.000	0.000

Figure 4.7: DeSigN prediction result for GSE4342. DeSigN returned 17 drugs with *p*-value of < 0.05, with which seven of them were predicted to have sensitive effect against the NSCLC cell lines. The intended drug, gefitinib (ranked seventh), together with another two *EGFR* inhibitors: BIBW2992 (ranked first), and lapatinib (ranked third) were returned as one of the top–ranked inhibitors with positive CS and *p*-value < 0.05 (red lines). Another two inhibitors: RDEA119, and AZD6244, which target the *MEK1/2* (downstream molecule of *EGFR*) were also returned as top-ranked inhibitors (blue lines).

For each of the two drug sensitivity validation studies, DeSigN returned CS that correctly correlated drug response outcomes with those in the respective published GEO studies. In these studies, DeSigN successfully associated input gene signatures with the right drugs, all with *p*-values < 0.05 (Table 4.1). The list of DEG of each study used to query DeSigN is provided in Appendix 9.

University of Malaya

Table 4.1: NCBI GEO datasets validation summary.

GEO reference	Reported drug	Expected drug sensitivity	DeSigN rank	Target	CS	*p*-value
GSE9633	Dasatinib	Sensitive	8	ABL, SRC, KIT, PDGFR	0.993	0.025
GSE4342	Gefitinib	Sensitive	7	EGFR	0.985	0.042

4.4.2 Using DeSigN to Shortlist Potentially Efficacious Inhibitors for OSCC Lines

To identify potential drugs that could be efficacious in controlling the growth of OSCC cell lines, 149 up-regulated, and 251 down-regulated genes were used to query DeSigN (Appendix 10). These DEGs were derived from differential gene expression analysis between five OSCC cell lines and three NOK published previously (Fadlullah et al., 2016). The gene expression values of these OSCC and NOK lines can be retrieved from the ORL Series in GENIPAC. Five potentially effective drugs were returned by DeSigN; in addition, the HNSCC cell lines were also predicted to be resistant to three drugs (p-value < 0.05; Figure 4.8). The ranking results corroborated well with recent findings. Two of the candidates, BIBW2992 (ranked third) and bosutinib (ranked fifth), have recently been reported to have efficacy in HNSCC cell lines through computational analysis of large-scale drug screening studies (Nichols et al., 2014).

Particularly, the efficacy of bosutinib, which targets *Src*, *Abl*, and *TEC*, was further evaluated as it is a recently FDA-approved drug for treating *BCR-ABL* leukemic patients and has no known effects against HNSCC or OSCC; therefore, the efficacy of bosutinib is unanticipated when used against OSCC cell lines.

University of Malaya

Drug	Target	KS_UP	KS_DOWN	Connectivity Score	P-Value
Cyclopamine	SMO	0.98530	-0.95875	1.000	0.000
Dasatinib	ABL, SRC, KIT, PDGFR	0.96220	-0.95155	0.984	0.022
BIBW2992	EGFR, ERBB2	0.92713	-0.98647	0.984	0.045
S-Trityl-L-Cysteine	KIF11	0.93971	-0.97099	0.983	0.044
Bosutinib	SRC, ABL, TEC	0.92783	-0.97808	0.980	0.046
CEP-701	FLT3, JAK2, NTRK1, RET	-0.97413	0.94059	-0.981	0.046
WO2009093972	PI3Kb	-0.96254	0.98153	-0.996	0.013
MK-2206	AKT1/2	-0.95761	0.99515	-1.000	0.000

Figure 4.8: DeSigN prediction results for OSCC cell lines. Eight drugs were returned by DeSigN to have *p*-value of < 0.05. Bosutinib (ranked fifth), was retuned as one of the five drugs with positive CS of 0.980 and *p*-value 0.046 (red line), suggesting that OSCC lines could possibly have sensitive effects upon treatment with bosutinib.

In order to validate the efficacy of bosutinib against OSCC experimentally, three OSCC cell lines (ORL-196, ORL-204, and ORL-48) that can be cultured efficiently in the laboratory were used. The HSC-4 (mean IC_{50}: 1.82 $\pm$ 0.03 µM, *p*-value < 0.05), and MCF7 (mean IC_{50}: 12.22 $\pm$ 1.32 µM, *p*-value < 0.05) cancer lines were used as the sensitive and resistant controls respectively. The mean IC_{50} of HSC-4 and MCF7 were consistent with results of a previously reported study (Garnett et al., 2012). Notably, when exposed to bosutinib, all three OSCC cell lines had significantly lower mean IC_{50} values (between 0.75 µM and 1.19 µM; *p*-value < 0.05) compared to the sensitive control HSC-4 (Table 4.2, Figure 4.9). The raw IC_{50} values (µM) for three OSCC cell lines and the respective controls are provided in Appendix 11.

University of Malaya

Table 4.2: Mean IC_{50} (μM) of cell lines upon exposure to bosutinib treatment. HSC-4 and MCF7 were used as the sensitive and resistant control respectively. All OSCC lines (ORL-196, ORL-204, and ORL-48) had lower mean IC_{50} values compared to the sensitive control HSC-4. Statistical significance (p-value < 0.05) relative to sensitive control HSC-4 is denoted by *.

Cell lines	Mean $IC_{50} \pm$ SE (μM)
*ORL-196 (n = 4)	0.75 ± 0.03
*ORL-204 (n = 3)	0.90 ± 0.04
*ORL-48 (n = 5)	1.19 ± 0.05
HSC-4 (n = 3)	1.82 ± 0.03
MCF7 (n = 3)	12.22 ± 1.32

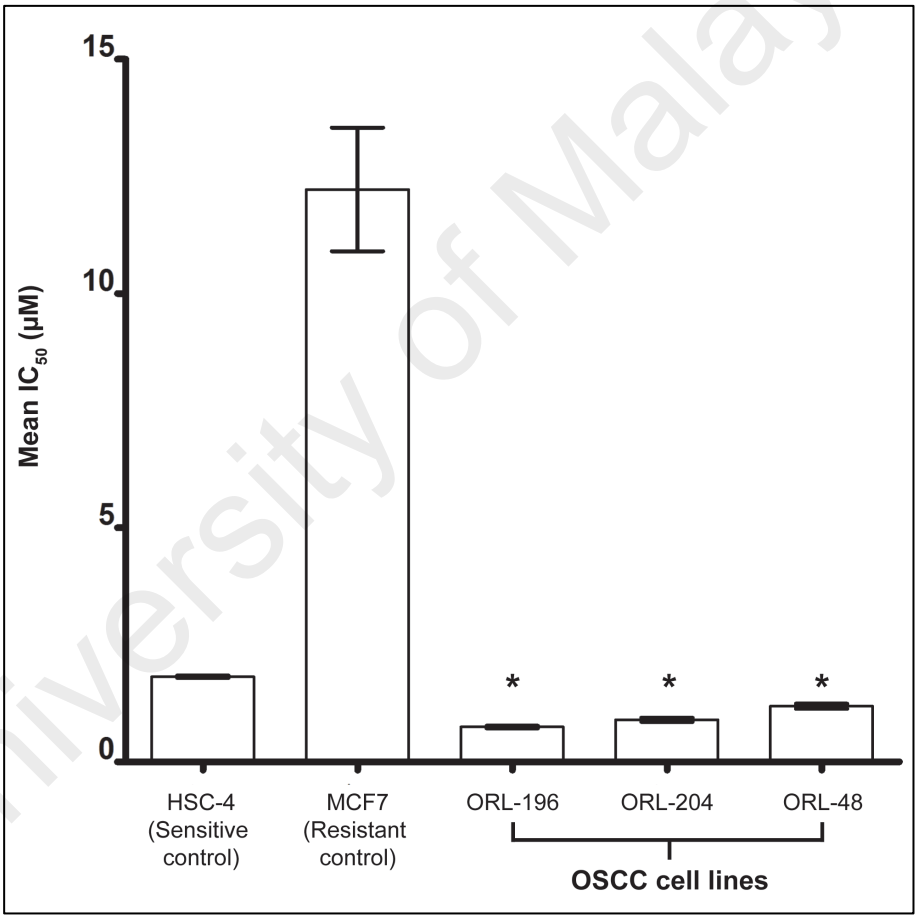

Figure 4.9: Mean IC_{50} (μM) of each OSCC cell line from MTT assay. The bars represent mean $IC_{50} \pm$ SE of at least three experiments. Statistical significance (p-value < 0.05) relative to sensitive control HSC-4 is denoted by *.

University of Malaya

This finding is supported by fluorescence-activated cell sorting (FACS) analysis of the cells where bosutinib induced cell death in OSCC cell lines in a time-dependent manner (Figure 4.10A, Appendix 12). In particular, ORL-196 cells were found to be most susceptible to bosutinib amongst the three tested OSCC lines, as about 35% of apoptotic cells were detected as early as 24 hours following treatment with bosutinib. By 72 hours, a significant proportion of apoptotic cells (35 – 95%) were detected in all the OSCC cell lines (p-value < 0.01), indicating the cytotoxic effect of bosutinib in these OSCC cells.

Further confirmation from the Click-iT EdU cell proliferation assay showed that bosutinib inhibited the proliferation of ORL-48, ORL-196 and ORL-204 cells as demonstrated by the significant reduction in the number of proliferating cells (red-stained cells) compared to the non-treated cells (Figure 4.10B). ORL-196 and ORL-204 demonstrated growth inhibition of ~70 – 80% (p-value = 0.03, n = 3; p-value = 0.049, n = 2 respectively) whilst ORL-48 showed growth inhibition of ~40% following bosutinib treatment at 1 μM at 72 hours (p-value = 0.04, n = 2) (Figure 4.10C, Appendix 13, and Appendix 14). The level of inhibition in the OSCC cell lines corroborated well with their mean IC_{50} value for bosutinib. Taken together, these biological observations showed that the anti-proliferative and cytotoxic properties of bosutinib against OSCC cell lines are real.

University of Malaya

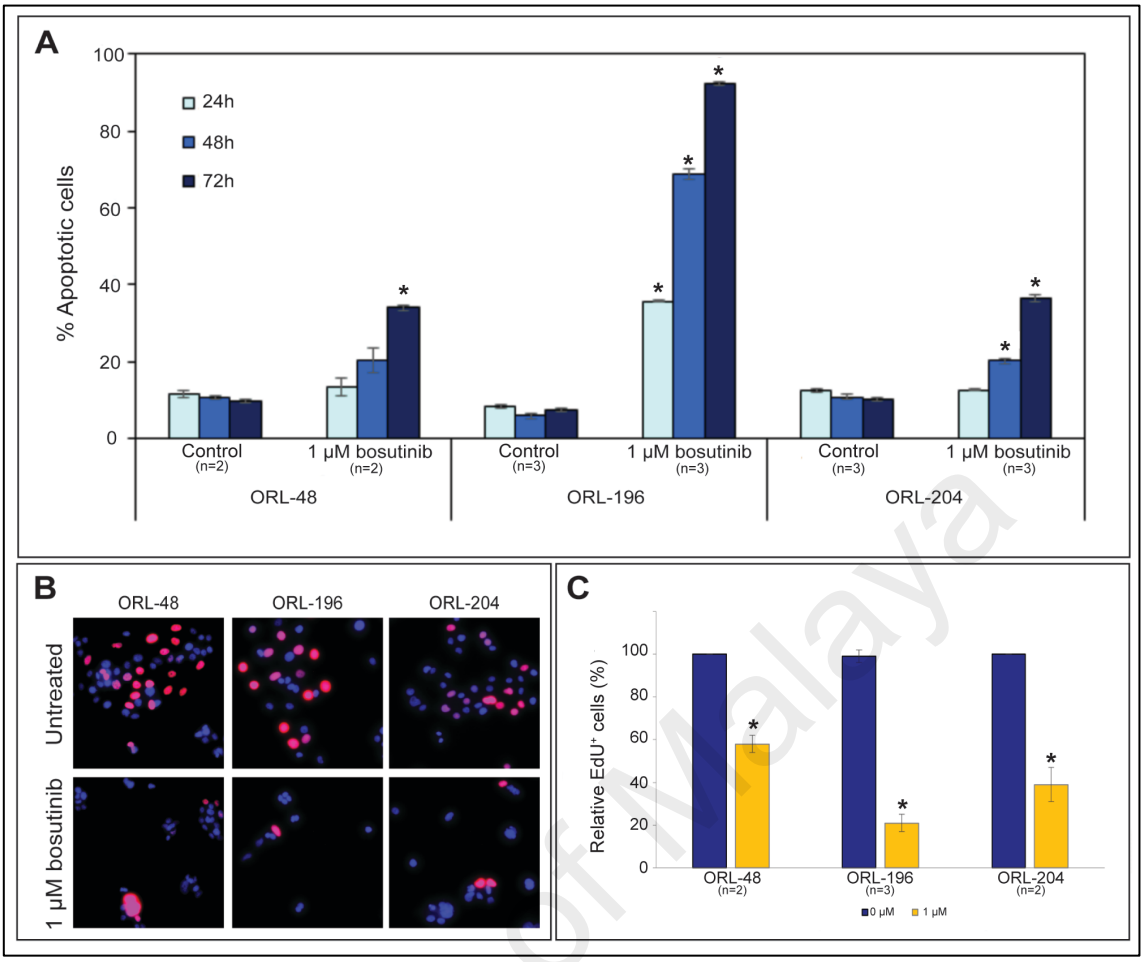

Figure 4.10: Differential sensitivity of OSCC cell lines, ORL-48, ORL-196 and ORL-204 to bosutinib. (A) Bosutinib induced apoptosis in OSCC cell lines. ORL-48, ORL-196, and ORL-204 cells were treated with 1 μM of bosutinib for 24, 48 and 72 hours followed by Annexin V/PI staining coupled with flow cytometry analysis. The bars represent the mean percentage of apoptotic cells ± SE of each cell line of at least two experiments. (B) Bosutinib inhibited the proliferation of OSCC cells as demonstrated by the reduced number of proliferating cells (red stained cells) following 72 hours treatment at 1 μM. The blue-stained nuclei represent the total number of cells in a field while the red-stained nuclei represent proliferating cells that have incorporated the EdU label. (C) OSCC cell proliferation was significantly inhibited by bosutinib with ORL-196 showing the greatest sensitivity (~80% inhibition) followed by ORL-204 (~70% inhibition) and ORL-48 (~50% inhibition) after bosutinib treatment at 1 μM for 72 hours. Statistical significance (p-value < 0.05) relative to control cells is denoted by *.

4.5 Evaluation of Different Gene Signature Similarity Scoring Algorithms for Optimal Drug Sensitivity Prediction

The ability to associate the right drugs with patterns of perturbations in gene expression requires a robust gene signature similarity scoring algorithm. Thus far, the KS statistic is the most commonly-used gene signature similarity scoring algorithm to

associate gene expression to the drug response phenotype. More recently, however, several other newer computational algorithms for detecting gene signature similarity have been developed to make use of the perturbagen-induced signatures contained in CMap. Yet, few systematic evaluations have been done to evaluate the performance of these more recent algorithms against the KS statistic (Cheng et al., 2014; Musa et al., 2017). Thus, a systematic evaluation of the strengths and weaknesses for these algorithms (KS, sscMap ordered, sscMap unordered, XCos, XSum, and WTCS) was conducted.

The Ushijima dataset (Ushijima et al., 2013), which consists of 39 query signatures, was used to illustrate how different gene signature similarity scoring algorithms can affect the prediction performance. Except for the ranking analysis, algorithm performance evaluation using PPV was limited to the top 50 of the ranked list of drugs returned by each scoring algorithm. The cut-off of top 50 was chosen because, in a real-case drug-repurposing scenario, the correctly predicted drugs should theoretically rank high up in the ranked list of inhibitors returned by each algorithm.

Additionally, due to the implementation of XSum and XCos algorithms that subset the CMap reference profile to top 500 and bottom 500 of differentially expressed genes for each drug instance, a signature that does not have overlapping genes with these 1000 genes would be discarded. In such a scenario, the connectivity score for XSum and XCos could not be computed. The 39 gene signatures from this dataset were reviewed based on the criteria described above to ensure that the gene signatures could be evaluated across all gene signature similarity scoring algorithms. Thus, 22 of the 39 Ushijima signatures were deemed to be suitable for downstream analysis.

In summary, there are altogether 6100 drug instances in the CMap reference database, in which these drug instances consist of 1309 unique small molecule inhibitors. Specifically, every small molecule inhibitor is associated with different number of drug

University of Malaya

instances, being derived from different drug treatment concentration, the time points captured as well as the cell lines tested.

4.5.1 Ranking Analysis

To carry out the performance evaluation, the first metric used was ranking analysis. To ease understanding, for a given gene signature derived from a drug, an algorithm is said to perform well if it could predict that same particular drug with a relatively higher ranking (approaching rank 1). Figure 4.11 shows the cluster heat map of the highest drug instance ranking ($\log_{10}$ transformed) returned by each algorithm for the respective 22 Ushijima signatures. The dendrogram shows that the five algorithms: sscMap ordered, sscMap unordered, XCos, XSum, and WTCS returned the expected ranks relatively better than the KS method. In addition, sscMap unordered and sscMap ordered had similar ranking profiles, as is the case of WTCS and XCos. In particular, WTCS registered the highest overall median ranking of 1.5, closely followed by the other four algorithms (median ranking in the range of 2 to 6). KS, meanwhile, did not perform well in this case. It returned the lowest median ranking of 69, more than 45-fold lower than the best performing algorithm, WTCS. Additional information about the rankings returned by each algorithm for the respective 22 signatures is available at Appendix 15.

University of Malaya

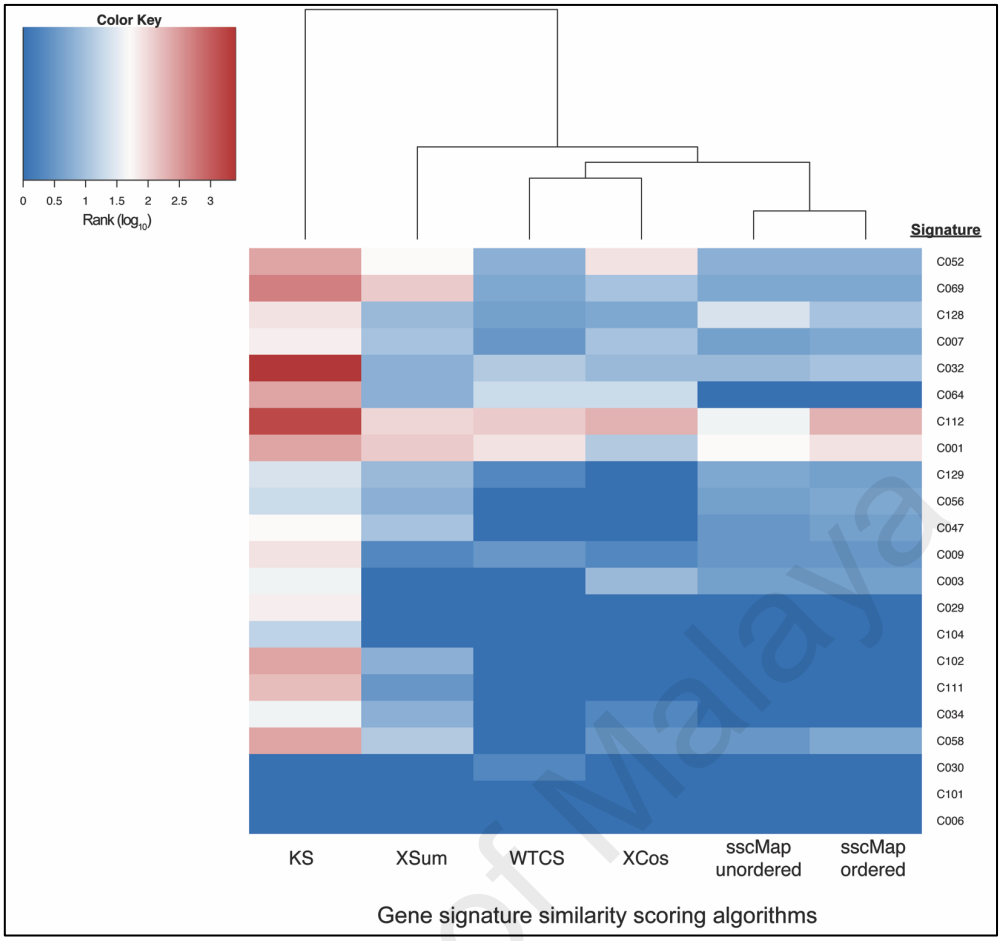

Figure 4.11: Heat map of the highest drug instance ranking ($\log_{10}$ transformed) returned by each algorithm for the respective 22 Ushijima signatures. Hierarchical clustering of these six algorithms based on the drug instance ranking revealed the presence of two major subgroups. Cells in blue indicate drug instance rankings that are highly ranked (approaching rank 1) while cells in red indicate lower ranking of the drug instances returned by the respective algorithm. The heat map was made using the Euclidean distance as the distance metric, and the Ward algorithm as the clustering algorithm (Warnes et al., 2019).

4.5.2 Positive Predictive Value

Having demonstrated that all algorithms apart from the KS method performed relatively well in returning the highest ranking of the intended drug instance, the next performance evaluation metric considered was the PPV. The PPV analysis across all six algorithms gradually increased from interval of $K = 1$ until 50. At each interval of K, the PPV for the 22 Ushijima signatures was computed and compared across the six gene signature similarity scoring algorithms.

University of Malaya

Figure 4.12 shows the PPV profile for the six gene signature similarity scoring algorithms as a function of *K*. Similar to the findings observed in the previous ranking analysis, all algorithms had better PPV performance than the KS method. When considering only the drug instances that fall within top ten (*K* = 10), WTCS, XCos, sscMap ordered, and sscMap unordered generally had better PPV performance than the rest; WTCS, meanwhile, recorded the highest mean PPV at *K* interval of 1 (mean PPV = 0.50). Within the cut-off of *K* = 10, XSum showed a relatively intermediate PPV performance. The cut-off of *K* = 15 seems to be a reasonable cut-off point because the rate of change in mean PPV is quite small for values of *K* after this cut-off, for all methods (Figure 4.12). To summarise, except KS, all other methods had similar mean PPV profiles, particularly for interval of *K* after 15.

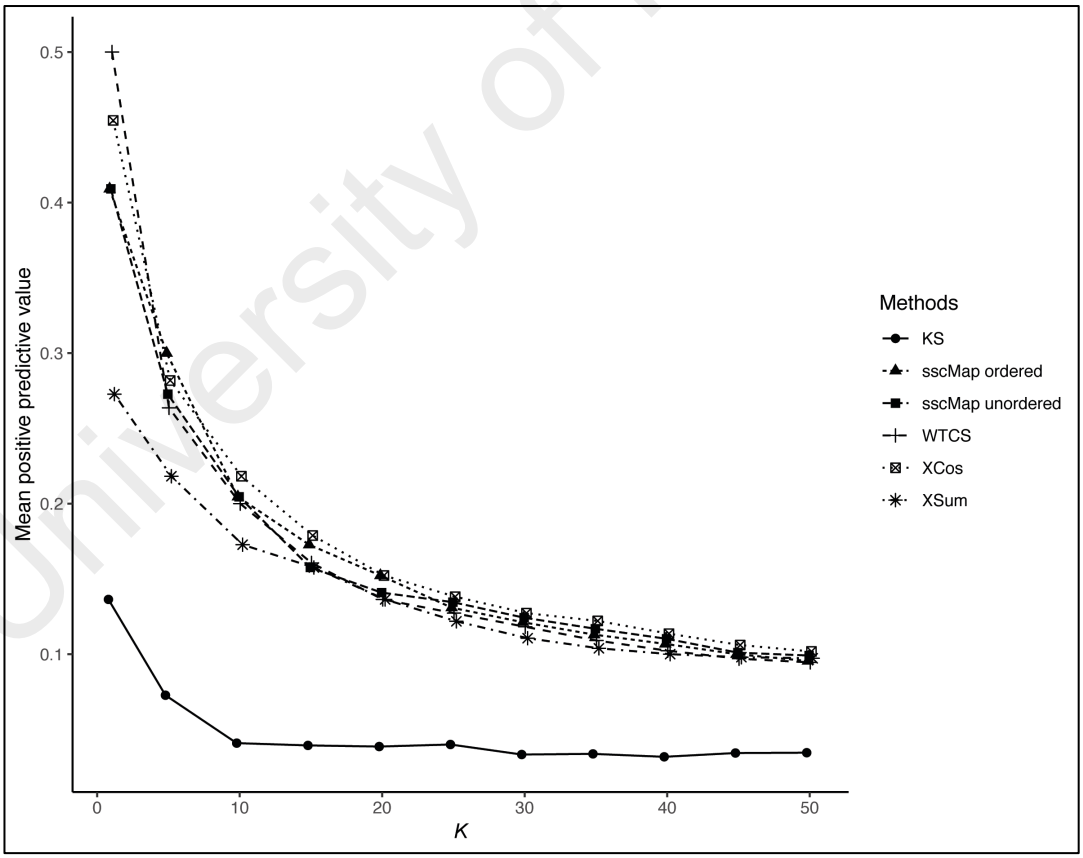

Figure 4.12: Mean PPV analysis of the six gene signature similarity scoring algorithms, with the cut-off for interval of *K* gradually increasing from 1 to 50. All algorithms performed better than KS method at all interval of *K*. WTCS scored the highest mean PPV at *K* interval of 1.

University of Malaya

4.5.3 Mechanism of Action Enrichment Analysis

Besides PPV, the ability of gene signature similarity scoring algorithm to pick up drug instances of similar mechanism of action (MoA) given a query signature was also assessed. For example, if a gene signature (up-regulated and down-regulated genes) associated with gefitinib (*EGFR* small molecule inhibitor) is used to query the CMap reference database, a practical algorithm would be able to pick up other drug instances of the same MoA to gefitinib, in this case, drug instances that are involved in EGFR signaling. Thus, an algorithm that is more efficient in clustering drug instances of similar MoA towards the top of the rank-ordered list of drug prediction would generate a higher positive enrichment score (ES) value and vice versa. Figure 4.13 shows the cluster heat map of MoA analysis in terms of ES returned by each algorithm using the 22 Ushijima signatures. All algorithms except KS had a mean MoA ES of at least 0.75, suggesting these five algorithms are more efficient in retrieving drug instances of similar MoA compared to the KS method.

University of Malaya

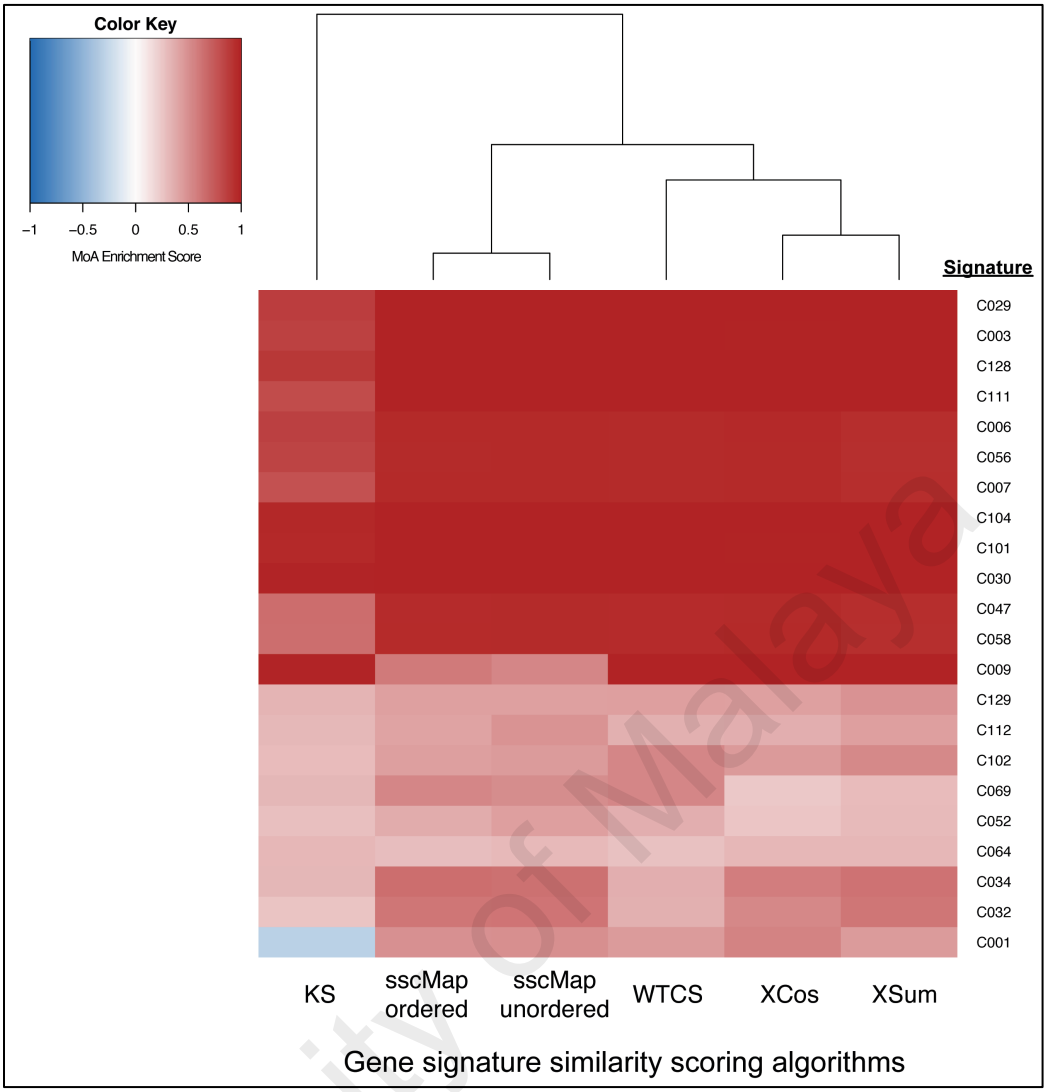

Figure 4.13: Heat map of the ES of MoA for the 22 Ushijima signatures returned by six different scoring algorithms. Generally, all algorithms but KS had a mean ES for MoA of at least 0.75. Cells in red indicate positive ES for MoA analysis, while cells in blue indicate negative value for ES of MoA. The heat map was made using the Euclidean distance as the distance metric, and the Ward algorithm as the clustering algorithm (Warnes et al., 2019).

4.5.4 Stability Analysis

The next evaluation metric is the stability analysis of scoring algorithms under varying query sizes. Here, we say that an algorithm is consistent, if, given different permutations of size N' of a query signature of size N where $N' < N$, the mean rank of the intended drug instance is close to the rank obtained using the query signature. The ranks of the intended drug instance returned by a consistent algorithm are robust against changes in the size of query signatures.

University of Malaya

To carry out the stability analysis, two Ushijima signatures: C006 and C058 were used. The first signature contained 716 up-regulated and 580 down-regulated genes; the second contained 1021 up-regulated and 953 down-regulated genes. The minimum size, each for up-regulated and down-regulated genes, was set at 50. The maximum for Signature C006 was set at 400; and for Signature C058, at 800.

Figure 4.14 shows the mean rankings returned by each algorithm for Signature C006 at four different sampling sizes. It can be seen that XSum, sscMap ordered, sscMap unordered, and WTCS performed relatively stable under varying query sizes. Variation in mean ranking due to query signature size was clear for KS and XCos algorithm. Amongst the tested six algorithms, XCos is probably most prone to variation in query signature size (Figure 4.15). For both stability analyses, the original ranking of Signature C006 and C058 (indicated by the black bars) were used for benchmarking purpose. Judging from these two analyses, it seems that WTCS, sscMap ordered, sscMap unordered and XSum are relatively more robust to variations in query signature size compared to KS and XCos.

University of Malaya

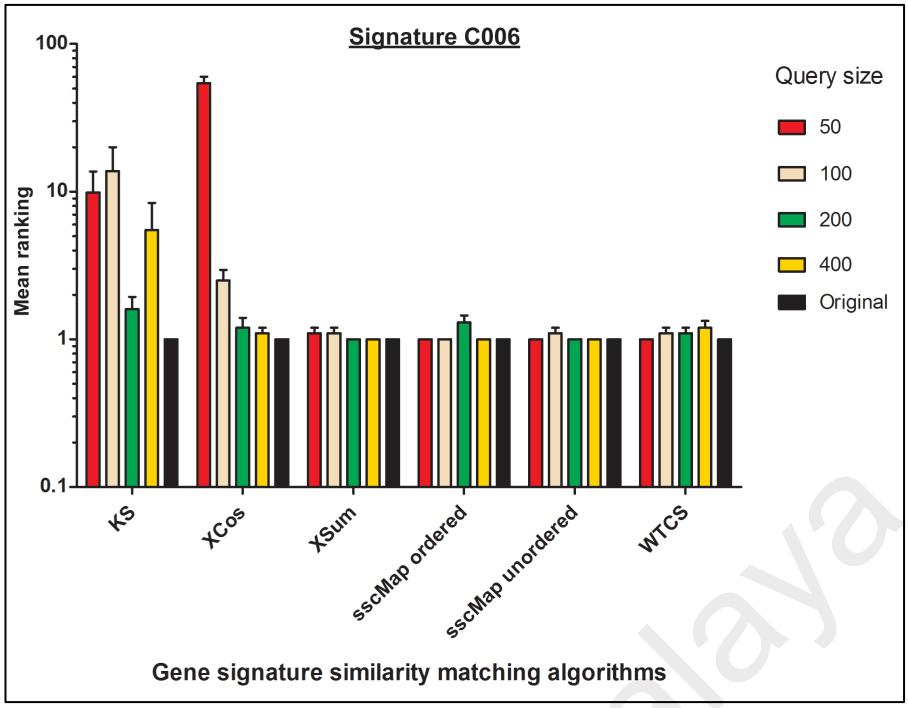

Figure 4.14: The stability analysis of different scoring algorithms under varying query sizes for the Signature C006. By examining the mean ranking returned by each algorithm, XSum, sscMap ordered, sscMap unordered and WTCS performed relatively more stable than KS and XCos under all query sizes. Black bars represent the original ranking of Signature C006 returned by each algorithm.

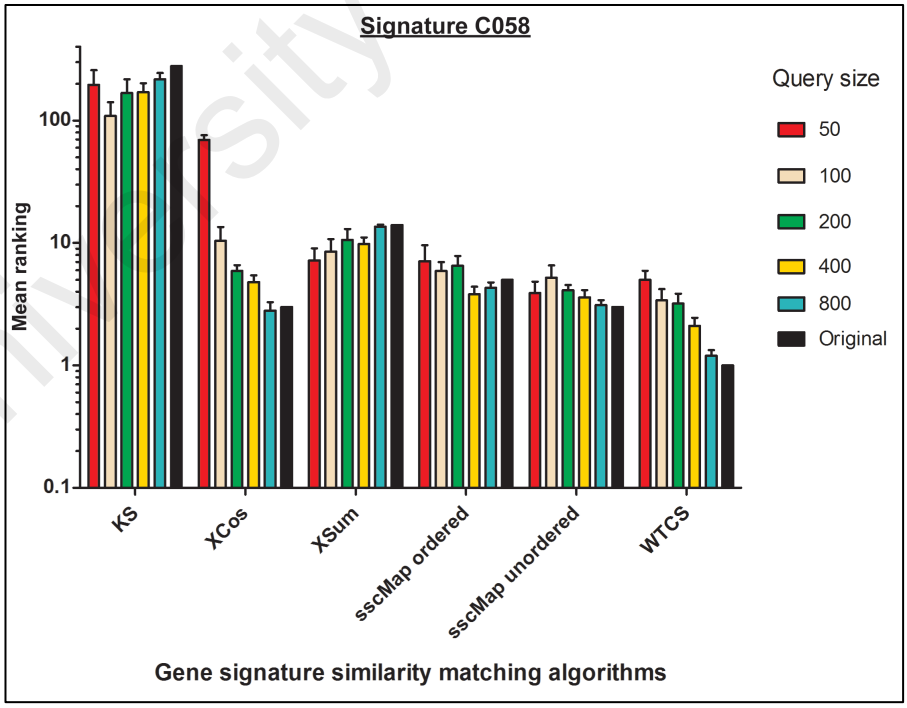

Figure 4.15: The stability analysis of different scoring algorithms under varying query sizes for the Signature C058. KS, XSum, sscMap ordered, sscMap unordered and WTCS performed relatively more stable than XCos under all query sizes. The fluctuation of ranking was observed under query size of 50 for XCos. Black bars depicted the original ranking of Signature C006 returned by each algorithm.

University of Malaya

Summarising from the performance evaluation results with respect to rankings, PPV, ES of similar MoA and stability test on varying query sizes, it appears that KS performs poorly in almost all aspects, and XCos as well, to some extend. Table 4.3 summarises the performance of these algorithms under four evaluation metrics, where every algorithm apart from KS gave similar performance across different performance evaluation metrics, and consistently scored better than the KS algorithm. All associated performance evaluation metrics analyses of 22 Ushijima signatures are provided in Appendix 16.

Table 4.3: Summary of the performance evaluation metrics for the 22 Ushijima signatures. Abbreviation: SE = standard error.

Algorithms	Median ranking	Mean PPV ± SE			Mean MoA ES	Stability
		$K = 1$	$K = 5$	$K = 10$		
WTCS	1.5	0.50 ± 0.11	0.26 ± 0.05	0.20 ± 0.04	0.75	Stable
XCos	2.0	0.45 ± 0.11	0.28 ± 0.06	0.22 ± 0.05	0.75	Not stable
sscMap unordered	3.0	0.41 ± 0.11	0.27 ± 0.05	0.20 ± 0.04	0.76	Stable
sscMap ordered	4.0	0.41 ± 0.11	0.30 ± 0.05	0.20 ± 0.04	0.76	Stable
XSum	6.0	0.27 ± 0.10	0.22 ± 0.07	0.17 ± 0.04	0.76	Stable
KS	69.0	0.14 ± 0.07	0.07 ± 0.04	0.04 ± 0.03	0.61	Not stable

University of Malaya

CHAPTER 5: DISCUSSION

5.1 GENIPAC

GENIPAC (Genomic Information Portal on Cancer Cell Lines) uses the functionality of cBioPortal, a powerful open access platform for exploring multidimensional cancer genomics data of tumour samples. Due to cBioPortal's ease of use, novel discoveries across multiple samples have been made, such as the identification of *APOBEC3A* as a potential oral cancer prognostic biomarker (Chen et al., 2017) and the identification of distinct subtypes and suggestions of new drivers of esophageal cancer (Lin et al., 2018). For HNSCC cell lines, several resources are currently available online for users to access their genomics profiles, including COSMIC (Forbes et al., 2011), GDSC (Garnett et al., 2012) and CCLE (Barretina et al., 2012). These databases serve specific purposes. For example, the COSMIC database documents somatic mutations found in cancers, while GDSC and CCLE host gene expression and copy number variation data. Several novel gene-drug associations are available in the latter two databases, such as the amplification of *CCND1* or loss of *SMAD4*, which are associated with sensitivity to multiple *EGFR* family inhibitors (Garnett et al., 2012). Paradoxically, these disparate databases are evidence of a lack of a centralised genomic database that puts all available genomic information on cell lines in a single platform. Such a resource would be invaluable for the exploration of genetic alterations across samples to facilitate biological discoveries and to validate hypothesis-driven research questions.

In line with the objective of developing a user-friendly web resource for exploring, visualising, and analysing genomics information of commonly-used head and neck CCL, GENIPAC offers the HNSCC research community a consolidated resource that shares genomic information of many commonly-used HNSCC cell lines

University of Malaya

(http://genipac.cancerresearch.my/). In total, GENIPAC currently hosts genomic information of three HNSCC studies, comprising of 44 HNSCC cell lines. Notably, GENIPAC provides an additional 33 HNSCC lines not currently available on any other online databases. With this inclusion, the total amount of HNSCC cell line data across four databases: GENIPAC, COSMIC, GDSC, CCLE stands at 98 (Figure 5.1). For additional information about the HNSCC cell lines, such as the name of the HNSCC cell lines available in these four databases, please refer to Appendix 17 and Appendix 18.

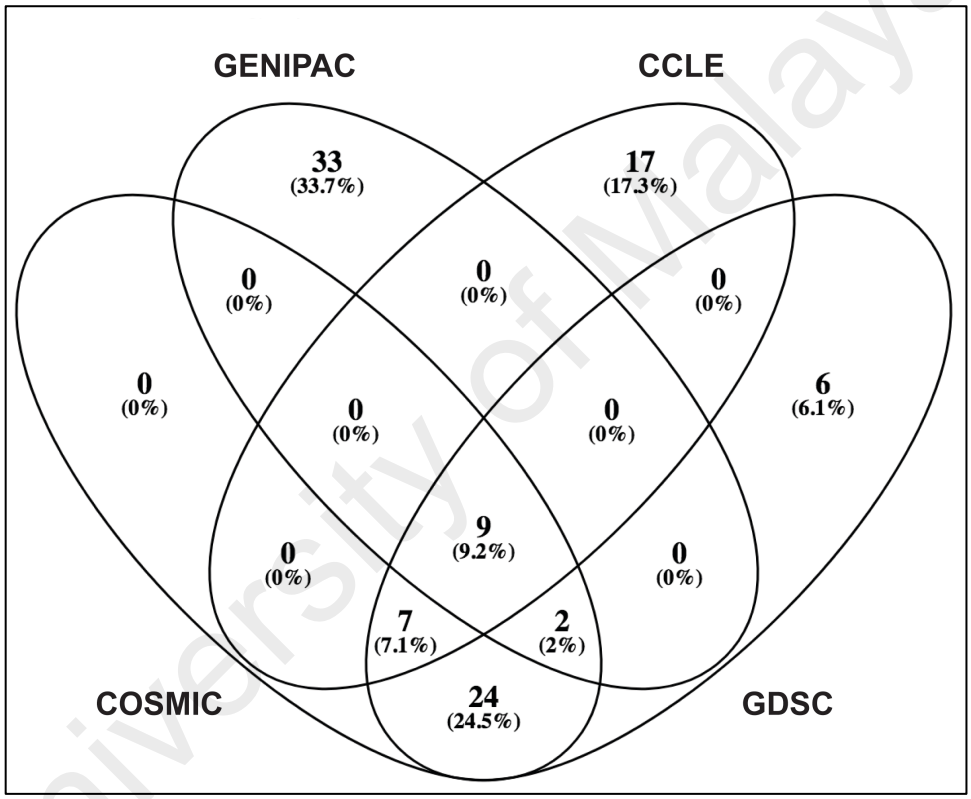

Figure 5.1: Venn diagram of HNSCC cell lines distribution in GENIPAC, COSMIC, CCLE, and GDSC. Among the 44 HNSCC cell lines available in GENIPAC, 33 of them are not available on any other databases. Notably, a total number of 98 HNSCC lines are hosted across these four databases, where nine lines are shared across all the four databases.

Among the three HNSCC studies currently available in GENIPAC, the ORL Series is the main focus of this thesis. This is mainly because the HNSCC cell lines in ORL Series were derived from Asian patients with diverse etiology that is relevant to this part of the

University of Malaya

region, in particular for HNSCC patients in the Malaysian setting. Specifically, the gene expression of the OSCC cell lines in ORL Series was mined to derive gene signatures that were used to query DeSigN. On the other hand, GENIPAC was also used to identify ORL-115 and ORL-150 lines in ORL Series that harbour *PIK3CA* mutation. Subsequent *in vitro* and *in vivo* experiments confirmed the association of *PIK3CA* mutation (H1047L and Q546R) with the resistance of OSCC cell lines to palbociclib (CDK4/6 inhibitor) (Zainal et al., 2019).

The few examples stated above demonstrate the utility of GENIPAC to perform data exploration and visualisation from integrated genomics data sources, such as the gene expression profile as well as the mutational data. It also underscores the value of sharing genomic information of HNSCC cell lines under one single platform, which is critical in driving research in the postgenomic era. One of the unique features of GENIPAC is that it meant to be a dynamic web resource that will evolve and expand continuously as and when new data available. For example, one of the studies that are planned to be included in future GENIPAC implementation is the head and neck pre-cancer cell culture model (de Boer et al., 2019). De Boer et al. (2019) recently reported their efforts in establishing and genetically-characterised 29 margins and five (erythro)leukoplakia from head and neck mucosal lining. This study is of particular interest because these established pre-cancer cells could be suitable *in vitro* model to develop targeted treatments to prevent HNSCC formation.

Nasopharyngeal carcinoma (NPC) is another aspect of head and neck study that is planned to be included in the upcoming GENIPAC implementation; particularly nasopharyngeal cancer is the third most common cancer in Malaysian male, and approximately 60% of the nasopharyngeal cancer cases were detected at a late stage (Stage 3 & 4). Forming part of the head and neck cancer, currently there is no targeted

University of Malaya

therapy for NPC treatment, highlighting the significant unmet need to new and effective treatment options for NPC patients.

As a whole, GENIPAC represents a unique tool for the HNSCC community, as it provides a valuable research platform to generate data-driven hypothesis from integrative genomics data. The availability and continuous evolution of GENIPAC is envisioned to help accelerate discoveries in the field of cancer research.

5.2 DeSigN

The aim of developing DeSigN is to identify potentially efficacious drugs that can be used to treat oral cancer patients, by leveraging on the well-curated genomic information of HNSCC hosted in GENIPAC portal. In particular, the genomic information of gene expression profiles in ORL Series was mined to derive the OSCC gene signature, which was then used to query DeSigN for potential drug candidates. Guided by the OSCC query signature, one of the drug candidates, bosutinib, was returned by DeSigN as a potential candidate drug against the OSCC cell lines. Subsequent *in vitro* experiments validated the efficacy of bosutinib in controlling the tumour growth in OSCC cell lines. Emerging evidence, meanwhile, supports the possible use of bosutinib for the treatment of HNSCC. First, the molecular target of bosutinib, *Src* has been reported to be a frequently altered gene in HNSCC and has been identified as a promising drug target (Pickering et al., 2013). Second, an analysis of gene expression data from 42 HNSCC cell lines also predicted that bosutinib has an anti-tumour effect on HNSCC (Nichols et al., 2014). To the best of our knowledge, this is the first time bosutinib was shown experimentally to have potency in OSCC cell lines.

Using gene expression changes as an attribute to guide small molecule inhibitor selection was first demonstrated by Lamb et al. (2006) in their CMap seminal work.

University of Malaya

Generally, the CMap concept can be regarded as a functional lookup table (analogous to the periodic table in chemistry), by which large-scale public database of CCL gene expression and drug response data such as CMap and GDSC can be mined to repurpose drugs for diseases that would otherwise have limited therapeutic options. Leveraging on this CMap approach, several types of cancers, such as the ovarian and lung cancers (Jahchan et al., 2013; Raghavan et al., 2016) have successfully expanded their treatment options. Likewise, the drug sensitivity prediction algorithms challenge organised by the National Cancer Institute (NCI) and the Dialogue on Reverse Engineering Assessment and Methods (DREAM) project (Costello et al., 2014) reported that gene expression patterns have the best predictive power for drug response prediction, thus further accentuating the relevance of using gene expression changes to guide candidate drugs selection.

Given a robust gene expression signature, results returned by DeSigN could lead to the identification of other more relevant candidate drugs. Take the drug sensitivity study of gefitinib in a panel of NSCLC cell lines (GSE4342) as an example. Originally intended to predict the ranking of gefitinib based on the gene signature derived from GSE4342, DeSigN additionally returned two other *EGFR* inhibitor: BIBW2992, and lapatinib (ranked first and third). In particular, BIBW2992 is a relevant drug candidate because it is currently approved for treating NSCLC patients who are refractory to gefitinib and erlotinib. Being a newer generation of *EGFR* inhibitor, BIBW2992 could potentially replace gefitinib, a first-generation *EGFR* tyrosine kinase inhibitor that is increasingly becoming a non-viable solution. Clinical studies have shown that cancer cells of NSCLC patients treated with gefitinib inevitably develop resistance and relapse, with 8 – 10 months of median time to progression (Maemondo et al., 2010; Sequist et al., 2011; Stinchcombe, 2014).

University of Malaya

While DeSigN and other drug repurposing tools such as NFFinder, DMAP, and FMCM similarly adopted the CMap concept, DeSigN has its uniqueness whereby it capitalises on the large panel of 707 human CCL in GDSC that have well-characterised gene expression and drug response data. In comparison to CMap, DeSigN constructs drug-associated baseline gene expression profile of resistant and sensitive cell lines from these 707 cell lines, whereas CMap associates response to the drug by constructing the gene expression profiles of pre- and post-treatment conditions using five cell lines (HL60, MCF7, PC3, SKMEL5, and ssMCF7). NFFinder, meanwhile, explores the relationship of transcriptomic data to drugs, diseases, and experts using three databases: CMap, GEO (gene expression), and DrugMatrix (drug toxicology profiles) (Setoain et al., 2015), while DMAP builds its protein/gene-drug response database using STITCH (chemical-protein interaction networks) and HAPPI (protein-protein interactions) (Huang et al., 2015). The FMCM tool, on the other hand, extends the utility of CMap by allowing users to query the CMap using a module of functional genes instead of a set of individual up-regulated and down-regulated genes (Chung et al., 2014). The characteristics of DeSigN and other drug repurposing tools are shown in Table 5.1.

Table 5.1: Comparison of drug repurposing tools that utilised the CMap concept.

Tools	Relationship feature	Reference database
DeSigN	Baseline DEGs to drug response	GDSC
NFFinder	Transcriptomic data to drugs, diseases, and experts	GEO, CMap, and DrugMatrix
DMAP	Protein/gene to drug response	STITCH and HAPPI
FMCM	Pre- and post-treatment gene expression to drug response	CMap

The new leads derived from DeSigN are useful for accelerating the discovery of new drugs for HNSCC treatment, which is currently limited to cetuximab, nivolumab and pembrolizumab (Bauml et al., 2017; Ferris et al., 2016; Vermorken et al., 2008).

University of Malaya

Importantly, it should be emphasised that all candidates with positive and significant CS should be equally considered for validation instead of considering just a few top-ranked candidates, since factors such as cost of the drug, ease of availability, method of administering, side effects and other factors, are essential practical considerations in the clinical setting.

Perhaps more compelling findings were that DeSigN predicted two tyrosine kinase inhibitors, palbociclib (CDK4/6 inhibitor) (Zainal et al., 2019) and afatinib (pan-EGFR inhibitor) (Yee et al., 2019) to be sensitive on a panel of head and neck CCL and these are successfully validated (*in vitro* and *in vivo*). Furthermore, DeSigN was also recently used by scientists from the Roswell Park Cancer Institute to prioritise drugs for basal breast cancers with a dysregulation of the citric acid cycle (CAC) pathway. Interestingly, none of the shortlisted 11 drugs such as 681640, GDC-0449, and NU-7441 that were predicted by DeSigN were initially designed to target the CAC pathway directly (Rosario et al., 2018) and therefore, without bioinformatics tools such as DeSigN 1.0, scientists might not have thought of testing drugs of different disease indications as possible treatment options.

5.2.1 Limitations and Future Implementation Work on DeSigN

Despite successes, the current implementation of DeSigN has a few shortcomings. First, it focuses on using differentially expressed genes as the starting points to associate gene signatures with drug response phenotype, ignoring the gene function redundancy and gene-gene interactions. This kind of input may not be necessarily optimal, for a disease is thought to rarely be a consequence of an abnormality in a single gene, but is instead reflected by a disruption of a complex gene network (Barabási et al., 2011). Furthermore, the parameters for generating gene signatures may be influenced by sample

University of Malaya

size, the cell line being studied, disease severity, or differential gene expression methods (Patil et al., 2015). Moreover, gene expression signatures are often defined as an alteration in the expression of a gene (or genes) with validated specificity in terms of diagnosis, prognosis or prediction of therapeutic response, ignoring the possibility of gene interactions (Chibon, 2013). In other words, this definition focuses on the quantitative gene expression alteration but ignores the interconnections relationship.

Additionally, it is noteworthy that genes that are involved in dysregulated pathways in the pathogenesis of cancer may not always have their expression substantially altered (de la Fuente, 2010), rendering them not to be listed as differentially expressed genes. A possible solution is to refine the DEG input genes using a set of genes from a subnetwork that is shown to be strongly associated with a disease condition. NetDecoder, for example, could dissect the subnetworks of input genes to identify key genes significantly impacting the cell behaviour specific to a given disease context (da Rocha et al., 2016).

Despite all these shortcomings, drug sensitivity prediction algorithms challenge posed by the DREAM project reassures us that gene expression data carries the most weight in predicting drug efficacy. The issues mentioned above regarding the optimal usage of gene expression signatures may be resolved by integrating the gene expression signatures with network biology methods, which consider not only the expression but also the modular function of the genes (Liu et al., 2018). One of the avenues that could be explored is the Weighted Gene Co-Expression Network Analysis (WGCNA) method proposed by Zhang and Horvath (2005) by which nodes represent genes and nodes are connected if the corresponding genes are significantly co-expressed across samples. Recently, the WGCNA method was employed to re-analyse the transcriptional profiles of CMap data, and it was observed that CMap data could be clustered into seven gene set modules (Liu et al., 2018; B. Zhang & Horvath, 2005). A further Gene Ontology analysis found that

University of Malaya

these modules were associated with molecular functions such as cell adhesion, extracellular matrix organisation, mRNA splicing, and translational initiation. The same WGCNA approach can, therefore, be applied to the differentially expressed genes before DeSigN analysis, thereby solving the gene function redundancy as well as functional annotation problems mentioned earlier.

DeSigN analyses might not yield highly accurate results because it does not yet fully take into consideration context-dependence of a particular disease (e.g., tissue specificity). Busby et al. (2018), for example, recently reported their failure in identifying medications that can alter breast cancer risk using the combination methods of CMap and pharmaco-epidemiology. One of their arguments is that it might be due to the heterogeneous nature of breast cancers that are made up of different subtypes that could potentially mask the crucial signals for specific breast cancer subtypes. Hence, for future implementation of DeSigN reference database, perhaps an option of choosing which subtypes of tissue one is interested in querying would improve predictive accuracy. On top of that, such kind of drug prediction analysis should also bear in mind the biological background of the disease in order to have a more relevant drug prediction.

Another way of evolving DeSigN is to integrate machine learning strategies in the framework of DeSigN analyses. One of the machine learning strategies that DeSigN can explore is the implementation of Learning to Rank (LTR) method to improve drug-target prediction. Yuan et al. (2016), for example, implemented LTR in DrugE-Rank, a machine learning approach they proposed that showed significant improvement in predicting FDA approved new and experimental drugs using the drug-target interaction obtained from DrugBank as the training datasets. Perhaps instead of using drug-target interactions data, DeSigN can make use of the 140 unique drug-gene signatures in the reference database as the training dataset to check for drug candidate prediction improvement. Another

University of Malaya

essential feature implemented by DrugE-Rank is the ensemble learning to rank method, in which it integrates six different similarity-based machine learning methods to improve their drug-target prediction performance. Perhaps the same concept can be applied for future DeSigN implementation, in which the six gene signature similarity scoring algorithms discussed in Section 3.4.2 can be integrated as an ensemble method to improve drug candidate prediction.

To ensure DeSigN stay relevant and competitive with newly-launched pharmacogenomics studies and drug candidates, the drug coverage of DeSigN reference profiles will continue to be enhanced by incorporating more clinically relevant experimental drugs and small molecule inhibitors from single-agent databases such as the GDSC version 2 (Iorio et al., 2016), Cancer Cell Line Encyclopedia (CCLE) (Barretina et al., 2012), Cancer Therapeutics Response Portal (CTRP) (version 1 and 2) (Basu et al., 2013), and Genentech Cell Line Screening Initiative (gCSI) (Haverty et al., 2016). Plans are also in place of incorporating drug combinations datasets from Merck Research Laboratories (O'Neil et al., 2016) and The National Cancer Institute ALMANAC project (Holbeck et al., 2017). Importantly, this will enable the collection of reference profiles in DeSigN to span across more than 700 drugs and more than a thousand CCL.

One significant improvement will be made for future DeSigN implementation, which is the inclusion of WTCS and sscMap as the new gene signature similarity scoring algorithms. This is based on the findings that both WTCS and sscMap perform better than standard KS statistic in terms of ranking, PPV, ES of similar MoA, and stability test under varying query sizes. Table 5.2 shows the comparison of current and future implementation of DeSigN.

Table 5.2: Comparison of current and future DeSigN implementation.

Features	Current	Future

University of Malaya

Number of drugs	140	> 700
Number of cell lines	707	> 1000
Gene signature similarity scoring algorithm	KS statistic	WTCS and sscMap
Pharmacogenomic databases	GDSC version 1	• GDSC version 2 • CCLE • CTRP version 1 • CTRP version 2 • Merck Research Laboratories • NCI ALMANAC

Lastly, DeSigN is envisioned in the future to be a practical biomarker-driven tumour type-agnostic bioinformatics prediction tool for better patient stratification and clinical trial design. This is to echo with the current development of the tumour type-agnostic clinical trial approach, whereby patients are enrolled in a clinical trial based on the affected target gene rather than the conventional tumour histological site. Indeed, some drugs work on all tumours regardless of the site of a tumour, however other drugs only work for some, but not all, tumours that have the same biomarker. For example, the *TRK* inhibitor larotrectinib appears to work for all adult and pediatric cancers with *TRK* fusions (Taylor et al., 2018), and pembrolizumab and nivolumab, for adult and pediatric unresectable or metastatic solid tumours harbouring high microsatellite instability, or deficiency in DNA mismatch repair mechanism (Yan & Zhang, 2018). In contrast, patients with *BRAF* V600E mutation-positive metastatic melanoma showed a response of approximately 50% to vemurafenib (*BRAF* inhibitor) (Chapman et al., 2011), but only circumstantial responses were observed in anaplastic thyroid cancer, cholangiocarcinoma, salivary duct cancer, ovarian cancer and colorectal cancer (Hyman et al., 2015). This highlights that tumour type-agnostic approach is not universal, and it is critical to identify new and combinatory treatment approaches to address drug resistance. It is indeed exciting if DeSigN could be used to run drug sensitivity prediction

University of Malaya

before the commencement of actual clinical trials where responders and non-responders could be identified, thus resulting to savings in time and cost.

5.3 Gene Signature Similarity Scoring Algorithms Evaluation for Optimal Drug Sensitivity Prediction

Having adopted the CMap concept, DeSigN successfully shortlisted bosutinib as the candidate drug that could be efficacious against OSCC cell lines. One key component for the successful development of DeSigN is the implementation of the KS statistic as the gene signature similarity scoring algorithm. The KS statistic has been the standard algorithm used to associate gene signatures given by the users to the drug perturbed gene expression profile in the CMap reference database. Newer algorithms have since been developed, yet few systematic evaluations have been done to evaluate their performance against the KS statistic.

To ensure optimal drug sensitivity prediction, six scoring algorithms (KS, WTCS, sscMap unordered, sscMap ordered, XSum, and XCos) were chosen to evaluate their performance using 22 Ushijima signatures against the CMap reference database. This comprehensive performance evaluation of multiple gene signature similarity scoring algorithms can, therefore, serve as a benchmark to assess any new methodologies in the future. Table 5.3 shows the different characteristics of these six gene signature similarity scoring algorithms. Amongst the six algorithms, KS requires the gene signature to be transformed to rank-ordered, while sscMap ordered takes on the signed rank-ordered for gene signature. In terms of requiring fold change value to be provided, only XCos and sscMap ordered require this input. With respect to the reference profile, all algorithms use the fold change ranking in descending order. However, KS and WTCS require one additional step, which is to transform the fold change values into rank-ordered forms.

University of Malaya

Table 5.3: Different characteristics of gene signature similarity scoring algorithms.

Methods	Gene signature	Fold change value	Reference profile
KS	Rank-ordered	X	Rank-ordered
WTCS	X	X	Rank-ordered
XSum	X	X	Fold change-ordered
XCos	X	√	Fold change-ordered
sscMap ordered	Signed rank-ordered	√	Fold change-ordered
sscMap unordered	X	X	Fold change-ordered

Having carried out the performance evaluation concerning the rankings, PPV, ES of similar MoA, and stability test on varying query sizes, it appears that KS performs poorly in almost all aspects as compared to the other five algorithms. As shown in Table 5.3, the KS method has quite a distinct characteristic as compared to the other five algorithms. These features, i.e., gene signature ranking, the requirement of fold change value, and ordering of the reference profile that lies in the nature of the technical implementation of KS execution might have contributed to the poor performance. More technical analyses that evaluate these features need to be done in order to conclude this finding, which is out of the scope of this current thesis.

On the other hand, additional drug perturbed reference databases aside from the CMap database should be used for performance evaluation for a more conclusive evaluation. This is because bias might have arisen from the use of CMap alone due to the limited number of CCL (n = 5) used to derive the profiles of transcriptional responses to 1309 small molecule inhibitors. Table 5.4 shows the breakdown of the number of gene expression profiles derived for each cell line, which showed that the majority of the gene expression profiles in CMap are derived from MCF7, PC3, and HL60. With more than 99% of the gene expression profiles derived from three cell lines, perhaps the fraction of

University of Malaya

all possible induced cellular states represented in the CMap reference database might probably be minimal.

Table 5.4: Breakdown of the number of transcriptional profile derived for each cell line in the CMap reference database.

Cell line	Tissue type	Number of transcriptional profile derived
MCF7	Breast	3095
PC3	Prostate	1741
HL60	Leukemia	1229
ssMCF7	Breast	18
SKMEL5	Melanoma	17

Judging from the performance of these six algorithms, it seems that every algorithm apart from KS gave quite a similar performance across different performance evaluation metrics, and consistently scored better than the KS algorithm. However, both XSum and XCos have one technical limitation, whereby when the gene signatures provided by users do not match with the top 500 up-regulated, and bottom 500 down-regulated genes in the reference profile, the similarity scoring for XSum and XCos could not be executed. Taking into these considerations, both WTCS and sscMap could, therefore, be a better choice of gene signature similarity scoring algorithm for drug-disease prediction as their performance is similar and they do not have clear technical limitations.

University of Malaya

CHAPTER 6: CONCLUSION

6.1 GENIPAC

In this study, a user-friendly web resource called GENIPAC (Genomic Information Portal on Cancer Cell Lines) was developed for exploring, visualising, and analysing genomics information of commonly-used HNSCC cell lines. In total, 44 HNSCC cell lines from three different studies (ORL Series, OPC-22, and H Series) are hosted in GENIPAC. The aim of implementing GENIPAC is to create easy access for head and neck cancer research community to mine genomic information of HNSCC cell lines, hence accelerate better understanding of HNSCC and lastly to develop new precision therapeutic options for HNSCC treatment.

Importantly, the functional utility of GENIPAC was demonstrated with some of the genomic alterations that are commonly reported in HNSCC, such as *TP53, EGFR, CCND1,* and *PIK3CA*. These alterations as reported in 530 head and neck tumour samples in The Cancer Genome Atlas (TCGA) were shown to recapitulate in the HNSCC cell lines in GENIPAC. Additionally, GENIPAC also enables the visualisation of pathways within which these genomic alterations fall in.

With the addition of more head and neck cancer data, hopefully, the cancer research community would find GENIPAC a valuable research platform to generate data-driven hypotheses for integrative analysis to accelerate discoveries in the field of HNSCC.

6.2 DeSigN

DeSigN (Differentially Expressed Gene Signatures - Inhibitors) is the drug repurposing tool that was developed to identify novel drugs that have good potential to

University of Malaya

be repurposed for head and neck cancer therapy. The gene signature similarity scoring algorithm, KS statistic implemented within DeSigN was used to correlate oral squamous cell carcinoma (OSCC) gene signatures with the pre-defined gene expression profiles associated with 140 drug response data (IC_{50}) available in Genomics of Drug Sensitivity in Cancer (GDSC).

DeSigN predicted bosutinib, an Src/Abl kinase inhibitor as a sensitive inhibitor for OSCC cell lines. Bosutinib is recently approved by the FDA for treating BCR-ABL leukemic patients and have no known effects no against OSCC. Subsequent experimental validation demonstrated that indeed, OSCC cell lines were sensitive to bosutinib with IC_{50} of 0.8–1.2 μM. As further confirmation, bosutinib was also shown to have anti-proliferative activity in OSCC cell lines, demonstrating that DeSigN was able to predict drug that could control the growth of cancer cells.

Moving forward, both WTCS and sscMap will be incorporated in the future implementation of DeSigN framework to increase the drug sensitivity prediction accuracy of DeSigN.

6.3 Concluding Remarks

The works presented in this thesis show that cancer genomics data mining and integration through the development of GENIPAC and DeSigN could be a viable approach in accelerating drug development process. It demonstrates an alternative route of developing new therapeutic candidates for cancers with limited therapeutic options, such as oral cancer. Importantly, clinical studies have shown that the current approved targeted therapies are only effective in less than 20% of HNSCC patients, which means the remaining 80% HNSCC patients are in urgent needs of new therapeutics options. The

University of Malaya

drug repurposing approach through GENIPAC and DeSigN should, therefore, be able to expand the repertoire of therapeutic options for HNSCC patients in the near future.

University of Malaya

REFERENCES

Ang, K. K., Berkey, B. A., Tu, X., Zhang, H. Z., Katz, R., Hammond, E. H., . . . Milas, L. (2002). Impact of epidermal growth factor receptor expression on survival and pattern of relapse in patients with advanced head and neck carcinoma. *Cancer Research, 62*(24), 7350-7356.

Azizah, A. M., Nor Saleha, I. T., Noor Hashimah, A., Asmah, Z. A., & Mastulu, W. (2016). *Malaysian National Cancer Registry Report 2007-2011*. Putrajaya, Malaysia: National Cancer Institute.

Barabási, A. L., Gulbahce, N., & Loscalzo, J. (2011). Network medicine: a network-based approach to human disease. *Nature Reviews Genetics, 12*(1), 56-68.

Barretina, J., Caponigro, G., Stransky, N., Venkatesan, K., Margolin, A. A., Kim, S., . . . Garraway, L. A. (2012). The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. *Nature, 483*(7391), 603-607.

Basu, A., Bodycombe, N. E., Cheah, J. H., Price, E. V., Liu, K., Schaefer, G. I., . . . Schreiber, S. L. (2013). An interactive resource to identify cancer genetic and lineage dependencies targeted by small molecules. *Cell, 154*(5), 1151-1161.

Baumann, M., & Krause, M. (2004). Targeting the epidermal growth factor receptor in radiotherapy: radiobiological mechanisms, preclinical and clinical results. *Radiotherapy and Oncology, 72*(3), 257-266.

Bauml, J., Seiwert, T. Y., Pfister, D. G., Worden, F., Liu, S. V., Gilbert, J., . . . Haddad, R. (2017). Pembrolizumab for platinum- and cetuximab-refractory head and neck cancer: results from a single-arm, phase II study. *Journal of Clinical Oncology, 35*(14), 1542-1549.

Bonner, J. A., Harari, P. M., Giralt, J., Azarnia, N., Shin, D. M., Cohen, R. B., . . . Ang, K. K. (2006). Radiotherapy plus cetuximab for squamous-cell carcinoma of the head and neck. *The New England Journal of Medicine, 354*(6), 567-578.

Bosschieter, J., Nieuwenhuijzen, J. A., van Ginkel, T., Vis, A. N., Witte, B., Newling, D., . . . van Moorselaar, R. J. A. (2018). Value of an immediate intravesical instillation of mitomycin C in patients with non-muscle-invasive bladder cancer: a prospective multicentre randomised study in 2243 patients. *European Urology, 73*(2), 226-232.

Boucher, B. J., & Mannan, N. (2002). Metabolic effects of the consumption of *Areca catechu*. *Addiction Biology, 7*(1), 103-110.

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

Cancer Genome Atlas Network. (2015). Comprehensive genomic characterization of head and neck squamous cell carcinomas. *Nature, 517*(7536), 576-582.

University of Malaya

Carlson, M. (2017). org.Hs.eg.db: genome wide annotation for human. R package version 3.5.0 [accessed 2018 Jan 23]. Available at: https://bioconductor.org/packages/release/data/annotation/html/org.Hs.eg.db.html.

Cerami, E., Gao, J., Dogrusoz, U., Gross, B. E., Sumer, S. O., Aksoy, B. A., . . . Schultz, N. (2012). The cBio cancer genomics portal: an open platform for exploring multidimensional cancer genomics data. *Cancer Discovery, 2*(5), 401-404.

Chang, M. T., Asthana, S., Gao, S. P., Lee, B. H., Chapman, J. S., Kandoth, C., . . . Taylor, B. S. (2016). Identifying recurrent mutations in cancer reveals widespread lineage diversity and mutational specificity. *Nature Biotechnology, 34*(2), 155-163.

Chapman, P. B., Hauschild, A., Robert, C., Haanen, J. B., Ascierto, P., Larkin, J., . . . BRIM-3 Study Group. (2011). Improved survival with vemurafenib in melanoma with BRAF V600E mutation. *The New England Journal of Medicine, 364*(26), 2507-2516.

Chen, T. W., Lee, C. C., Liu, H., Wu, C. S., Pickering, C. R., Huang, P. J., . . . Chang, Y. S. (2017). *APOBEC3A* is an oral cancer prognostic biomarker in Taiwanese carriers of an APOBEC deletion polymorphism. *Nature Communications, 8*(1), 465.

Cheng, J., Xie, Q., Kumar, V., Hurle, M., Freudenberg, J. M., Yang, L., & Agarwal, P. (2013). Evaluation of analytical methods for connectivity map data. In R. B. Altman, A. K. Dunker, L. Hunter, T. A. Murray, & T. E. Klein (Eds.), *Pacific Symposium on Biocomputing 2013* (pp. 5-16). Kohala Coast, Hawaii: USA.

Cheng, J., Yang, L., Kumar, V., & Agarwal, P. (2014). Systematic evaluation of connectivity map for disease indications. *Genome Medicine, 6*(12), 540.

Chher, T., Hak, S., Kallarakkal, T. G., Durward, C., Ramanathan, A., Ghani, W. M. N., . . . Zain, R. B. (2018). Prevalence of oral cancer, oral potentially malignant disorders and other oral mucosal lesions in Cambodia. *Ethnicity & Health, 23*(1), 1-15.

Chibon, F. (2013). Cancer gene expression signatures - the rise and fall? *European Journal of Cancer, 49*(8), 2000-2009.

Choi, Y., Sims, G. E., Murphy, S., Miller, J. R., & Chan, A. P. (2012). Predicting the functional effect of amino acid substitutions and indels. *PLoS One, 7*(10), e46688.

Chung, F. H., Chiang, Y. R., Tseng, A. L., Sung, Y. C., Lu, J., Huang, M. C., . . . Lee, H. C. (2014). Functional Module Connectivity Map (FMCM): a framework for searching repurposed drug compounds for systems treatment of cancer and an application to colorectal adenocarcinoma. *PLoS One, 9*(1), e86299.

Cohen, E. E., Lingen, M. W., & Vokes, E. E. (2004). The expanding role of systemic therapy in head and neck cancer. *Journal of Clinical Oncology, 22*(9), 1743-1752.

University of Malaya

Coldren, C. D., Helfrich, B. A., Witta, S. E., Sugita, M., Lapadat, R., Zeng, C., . . . Bunn, P. A., Jr. (2006). Baseline gene expression predicts sensitivity to gefitinib in non-small cell lung cancer cell lines. *Molecular Cancer Research, 4*(8), 521-528.

Costello, J. C., Heiser, L. M., Georgii, E., Gonen, M., Menden, M. P., Wang, N. J., . . . Stolovitzky, G. (2014). A community effort to assess and improve drug sensitivity prediction algorithms. *Nature Biotechnology, 32*(12), 1202-1212.

da Rocha, E. L., Ung, C. Y., McGehee, C. D., Correia, C., & Li, H. (2016). NetDecoder: a network biology platform that decodes context-specific biological networks and gene activities. *Nucleic Acids Research, 44*(10), e100.

Dastur, A., Choi, A., Costa, C., Yin, X., Williams, A., McClanaghan, J., . . . Benes, C. H. (2019). NOTCH1 represses MCL-1 levels in GSI-resistant T-ALL, making them susceptible to ABT-263. *Clinical Cancer Research, 25*(1), 312-324.

de Boer, D. V., Brink, A., Buijze, M., Stigter-van Walsum, M., Hunter, K. D., Ylstra, B., . . . Brakenhoff, R. H. (2019). Establishment and genetic landscape of precancer cell model systems from the head and neck mucosal lining. *Molecular Cancer Research, 17*(1), 120-130.

de la Fuente, A. (2010). From 'differential expression' to 'differential networking' - identification of dysfunctional regulatory networks in diseases. *Trends in Genetics, 26*(7), 326-333.

DePristo, M. A., Banks, E., Poplin, R., Garimella, K. V., Maguire, J. R., Hartl, C., . . . Daly, M. J. (2011). A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nature Genetics, 43*(5), 491-498.

Dong, H., Strome, S. E., Salomao, D. R., Tamura, H., Hirano, F., Flies, D. B., . . . Chen, L. (2002). Tumor-associated B7-H1 promotes T-cell apoptosis: a potential mechanism of immune evasion. *Nature Medicine, 8*(8), 793-800.

Duan, Q., Flynn, C., Niepel, M., Hafner, M., Muhlich, J. L., Fernandez, N. F., . . . Ma'ayan, A. (2014). LINCS Canvas Browser: interactive web app to query, browse and interrogate LINCS L1000 gene expression signatures. *Nucleic Acids Research, 42*(Web Server issue), W449-460.

Fadlullah, M. Z., Chiang, I. K., Dionne, K. R., Yee, P. S., Gan, C. P., Sam, K. K., . . . Cheong, S. C. (2016). Genetically-defined novel oral squamous cell carcinoma cell lines for the development of molecular therapies. *Oncotarget, 7*(19), 27802-27818.

Ferris, R. L., Blumenschein, G., Jr., Fayette, J., Guigay, J., Colevas, A. D., Licitra, L., . . . Gillison, M. L. (2016). Nivolumab for recurrent squamous-cell carcinoma of the head and neck. *The New England Journal of Medicine, 375*(19), 1856-1867.

Forbes, S. A., Bindal, N., Bamford, S., Cole, C., Kok, C. Y., Beare, D., . . . Futreal, P. A. (2011). COSMIC: mining complete cancer genomes in the Catalogue of Somatic Mutations in Cancer. *Nucleic Acids Research, 39*(Database issue), D945-950.

University of Malaya

Gao, J., Aksoy, B. A., Dogrusoz, U., Dresdner, G., Gross, B., Sumer, S. O., . . . Schultz, N. (2013). Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal. *Science Signaling, 6*(269), pl1.

Garnett, M. J., Edelman, E. J., Heidorn, S. J., Greenman, C. D., Dastur, A., Lau, K. W., . . . Benes, C. H. (2012). Systematic identification of genomic markers of drug sensitivity in cancer cells. *Nature, 483*(7391), 570-575.

Ghani, W. M. N., Ramanathan, A., Prime, S. S., Yang, Y. H., Rahman, Z. A. A., Ismail, S. M., . . . Zain, R. B. (2019). Survival of oral cancer patients in different ethnicities. *Cancer Investigation, In Press*.

Gilyoma, J. M., Rambau, P. F., Masalu, N., Kayange, N. M., & Chalya, P. L. (2015). Head and neck cancers: a clinico-pathological profile and management challenges in a resource-limited setting. *BMC Research Notes, 8*, 772.

Goodspeed, A., Heiser, L. M., Gray, J. W., & Costello, J. C. (2016). Tumor-derived cell lines as molecular models of cancer pharmacogenomics. *Molecular Cancer Research, 14*(1), 3-13.

Goodspeed, A., Jean, A., Theodorescu, D., & Costello, J. C. (2018). A gene expression signature predicts bladder cancer cell line sensitivity to EGFR inhibition. *Bladder Cancer, 4*(3), 269-282.

Guo, X. E., Ngo, B., Modrek, A. S., & Lee, W. H. (2014). Targeting tumor suppressor networks for cancer therapeutics. *Current Drug Targets, 15*(1), 2-16.

Haverty, P. M., Lin, E., Tan, J., Yu, Y., Lam, B., Lianoglou, S., . . . Bourgon, R. (2016). Reproducible pharmacogenomic profiling of cancer cell line panels. *Nature, 533*(7603), 333-337.

Herbst, R. S., & Hong, W. K. (2002). IMC-C225, an anti-epidermal growth factor receptor monoclonal antibody for treatment of head and neck cancer. *Seminars in Oncology, 29*(5 Suppl 14), 18-30.

Holbeck, S. L., Camalier, R., Crowell, J. A., Govindharajulu, J. P., Hollingshead, M., Anderson, L. W., . . . Doroshow, J. H. (2017). The National Cancer Institute ALMANAC: a comprehensive screening resource for the detection of anticancer drug pairs with enhanced therapeutic activity. *Cancer Research, 77*(13), 3564-3576.

Holsinger, F. C., Piha-Paul, S. A., Janku, F., Hong, D. S., Atkins, J. T., Tsimberidou, A. M., & Kurzrock, R. (2013). Biomarker-directed therapy of squamous carcinomas of the head and neck: targeting PI3K/PTEN/mTOR pathway. *Journal of Clinical Oncology, 31*(9), e137-140.

Huang, H., Nguyen, T., Ibrahim, S., Shantharam, S., Yue, Z., & Chen, J. Y. (2015). DMAP: a connectivity map database to enable identification of novel drug repositioning candidates. *BMC Bioinformatics, 16 Suppl 13*, S4.

University of Malaya

Hyman, D. M., Puzanov, I., Subbiah, V., Faris, J. E., Chau, I., Blay, J. Y., . . . Baselga, J. (2015). Vemurafenib in multiple nonmelanoma cancers with BRAF V600 mutations. *The New England Journal of Medicine, 373*(8), 726-736.

Iglesias-Bartolome, R., Martin, D., & Gutkind, J. S. (2013). Exploiting the head and neck cancer oncogenome: widespread PI3K-mTOR pathway alterations and novel molecular targets. *Cancer Discovery, 3*(7), 722-725.

Iorio, F., Knijnenburg, T. A., Vis, D. J., Bignell, G. R., Menden, M. P., Schubert, M., . . . Garnett, M. J. (2016). A landscape of pharmacogenomic interactions in cancer. *Cell, 166*(3), 740-754.

Irizarry, R. A., Hobbs, B., Collin, F., Beazer-Barclay, Y. D., Antonellis, K. J., Scherf, U., & Speed, T. P. (2003). Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics, 4*(2), 249-264.

Isaacsson Velho, P. H., Castro, G., Jr., & Chung, C. H. (2015). Targeting the PI3K pathway in head and neck squamous cell carcinoma. *American Society of Clinical Oncology Educational Book*, 123-128.

Jahchan, N. S., Dudley, J. T., Mazur, P. K., Flores, N., Yang, D., Palmerton, A., . . . Sage, J. (2013). A drug repositioning approach identifies tricyclic antidepressants as inhibitors of small cell lung cancer and other neuroendocrine tumors. *Cancer Discovery, 3*(12), 1364-1377.

Joshi, P., Dutta, S., Chaturvedi, P., & Nair, S. (2014). Head and neck cancers in developing countries. *Rambam Maimonides Medical Journal, 5*(2), e0009.

Kampangsri, W., Vatanasapt, P., Kamsa-Ard, S., Suwanrungruang, K., & Promthet, S. (2013). Betel quid chewing and upper aerodigestive tract cancers: a prospective cohort study in Khon Kaen, Thailand. *Asian Pacific Journal of Cancer Prevention, 14*(7), 4335-4338.

Kimmel, K. A., & Carey, T. E. (1986). Altered expression in squamous carcinoma cells of an orientation restricted epithelial antigen detected by monoclonal antibody A9. *Cancer Research, 46*(7), 3614-3623.

Lamb, J., Crawford, E. D., Peck, D., Modell, J. W., Blat, I. C., Wrobel, M. J., . . . Golub, T. R. (2006). The connectivity map: using gene-expression signatures to connect small molecules, genes, and disease. *Science, 313*(5795), 1929-1935.

Lawrence, M., Huber, W., Pagès, H., Aboyoun, P., Carlson, M., Gentleman, R., . . . Carey, V. J. (2013). Software for computing and annotating genomic ranges. *PLoS Computational Biology, 9*(8), e1003118.

Lee, B. K. B., Gan, C. P., Chang, J. K., Tan, J. L., Fadlullah, M. Z., Abdul Rahman, Z. A., . . . Cheong, S. C. (2018). GENIPAC: a genomic information portal for head and neck cancer cell systems. *Journal of Dental Research, 97*(8), 909-916.

Lee, B. K. B., Tiong, K. H., Chang, J. K., Liew, C. S., Abdul Rahman, Z. A., Tan, A. C., . . . Cheong, S. C. (2017). DeSigN: connecting gene expression with therapeutics for drug repurposing and development. *BMC Genomics, 18*(Suppl 1), 934.

University of Malaya

Leemans, C. R., Braakhuis, B. J., & Brakenhoff, R. H. (2011). The molecular biology of head and neck cancer. *Nature Reviews Cancer, 11*(1), 9-22.

Li, H., Wawrose, J. S., Gooding, W. E., Garraway, L. A., Lui, V. W., Peyser, N. D., & Grandis, J. R. (2014). Genomic analysis of head and neck squamous cell carcinoma cell lines and human tumors: a rational approach to preclinical model selection. *Molecular Cancer Research, 12*(4), 571-582.

Liao, Y., Smyth, G. K., & Shi, W. (2014). featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics, 30*(7), 923-930.

Lin, D. C., Dinh, H. Q., Xie, J. J., Mayakonda, A., Silva, T. C., Jiang, Y. Y., . . . Phillip Koeffler, H. (2018). Identification of distinct mutational patterns and new driver genes in oesophageal squamous cell carcinomas and adenocarcinomas. *Gut, 67*(10), 1769-1779.

Liu, W., Tu, W., Li, L., Liu, Y., Wang, S., Li, L., . . . He, H. (2018). Revisiting connectivity map from a gene co-expression network analysis. *Experimental and Therapeutic Medicine, 16*(2), 493-500.

Love, M. I., Huber, W., & Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology, 15*(12), 550.

Loyha, K., Vatanasapt, P., Promthet, S., & Parkin, D. M. (2012). Risk factors for oral cancer in northeast Thailand. *Asian Pacific Journal of Cancer Prevention, 13*(10), 5087-5090.

Lui, V. W., Hedberg, M. L., Li, H., Vangara, B. S., Pendleton, K., Zeng, Y., . . . Grandis, J. R. (2013). Frequent mutation of the PI3K pathway in head and neck cancer defines predictive biomarkers. *Cancer Discovery, 3*(7), 761-769.

Maemondo, M., Inoue, A., Kobayashi, K., Sugawara, S., Oizumi, S., Isobe, H., . . . North-East Japan Study Group. (2010). Gefitinib or chemotherapy for non-small-cell lung cancer with mutated EGFR. *The New England Journal of Medicine, 362*(25), 2380-2388.

Martin, D., Abba, M. C., Molinolo, A. A., Vitale-Cross, L., Wang, Z., Zaida, M., . . . Gutkind, J. S. (2014). The head and neck cancer cell oncogenome: a platform for the development of precision molecular therapies. *Oncotarget, 5*(19), 8906-8923.

Marur, S., & Forastiere, A. A. (2016). Head and neck squamous cell carcinoma: update on epidemiology, diagnosis, and treatment. *Mayo Clinic Proceedings, 91*(3), 386-396.

Mehra, R., Seiwert, T. Y., Gupta, S., Weiss, J., Gluck, I., Eder, J. P., . . . Haddad, R. (2018). Efficacy and safety of pembrolizumab in recurrent/metastatic head and neck squamous cell carcinoma: pooled analyses after long-term follow-up in KEYNOTE-012. *British Journal of Cancer, 119*(2), 153-159.

University of Malaya

Mermel, C. H., Schumacher, S. E., Hill, B., Meyerson, M. L., Beroukhim, R., & Getz, G. (2011). GISTIC2.0 facilitates sensitive and confident localization of the targets of focal somatic copy-number alteration in human cancers. *Genome Biology, 12*(4), R41.

Michel, L., Ley, J., Wildes, T. M., Schaffer, A., Robinson, A., Chun, S. E., . . . Adkins, D. (2016). Phase I trial of palbociclib, a selective cyclin dependent kinase 4/6 inhibitor, in combination with cetuximab in patients with recurrent/metastatic head and neck squamous cell carcinoma. *Oral Oncology, 58*, 41-48.

Monks, A., Scudiero, D., Skehan, P., Shoemaker, R., Paull, K., Vistica, D., . . . Boyd, M. (1991). Feasibility of a high-flux anticancer drug screen using a diverse panel of cultured human tumor cell lines. *Journal of the National Cancer Institute, 83*(11), 757-766.

Morgan, M., Pagès, H., Obenchain, V., Hayden, N. (2017). Rsamtools: binary alignment (BAM), FASTA, variant call (BCF), and tabix file import. R package version 1.30.0 [accessed 2018 Jan 23]. Available at: https://rdrr.io/bioc/Rsamtools/.

Murata, S., Yamamoto, H., Shimizu, T., Naitoh, H., Yamaguchi, T., Kaida, S., . . . Tani, M. (2018). 5-fluorouracil combined with cisplatin and mitomycin C as an optimized regimen for hyperthermic intraperitoneal chemotherapy in gastric cancer. *Journal of Surgical Oncology, 117*(4), 671-677.

Musa, A., Ghoraie, L. S., Zhang, S. D., Glazko, G., Yli-Harja, O., Dehmer, M., . . . Emmert-Streib, F. (2017). A review of connectivity map and computational approaches in pharmacogenomics. *Briefings in Bioinformatics, 18*(5), 903.

Musa, A., Tripathi, S., Kandhavelu, M., Dehmer, M., & Emmert-Streib, F. (2018). Harnessing the biological complexity of big data from LINCS gene expression signatures. *PLoS One, 13*(8), e0201937.

Musgrove, E. A., Caldon, C. E., Barraclough, J., Stone, A., & Sutherland, R. L. (2011). Cyclin D as a therapeutic target in cancer. *Nature Reviews Cancer., 11*(8), 558-572.

Needle, M. N. (2002). Safety experience with IMC-C225, an anti-epidermal growth factor receptor antibody. *Seminars in Oncology, 29*(5 Suppl 14), 55-60.

Ng, C. J., Teo, C. H., Abdullah, N., Tan, W. P., & Tan, H. M. (2015). Relationships between cancer pattern, country income and geographical region in Asia. *BMC Cancer, 15*, 613.

Nichols, A. C., Black, M., Yoo, J., Pinto, N., Fernandes, A., Haibe-Kains, B., . . . Barrett, J. W. (2014). Exploiting high-throughput cell line drug screening studies to identify candidate therapeutic agents in head and neck cancer. *BMC Pharmacology & Toxicology, 15*(66).

O'Neil, J., Benita, Y., Feldman, I., Chenard, M., Roberts, B., Liu, Y., . . . Shumway, S. D. (2016). An unbiased oncology compound screen to identify novel combination strategies. *Molecular Cancer Therapeutics, 15*(6), 1155-1162.

University of Malaya

Olshen, A. B., Venkatraman, E. S., Lucito, R., & Wigler, M. (2004). Circular binary segmentation for the analysis of array-based DNA copy number data. *Biostatistics, 5*(4), 557-572.

Patil, P., Bachant-Winner, P. O., Haibe-Kains, B., & Leek, J. T. (2015). Test set bias affects reproducibility of gene signatures. *Bioinformatics, 31*(14), 2318-2323.

Pau, G., Fuchs, F., Sklyar, O., Boutros, M., & Huber, W. (2010). EBImage-an R package for image processing with applications to cellular phenotypes. *Bioinformatics, 26*(7), 979-981.

Peck, D., Crawford, E. D., Ross, K. N., Stegmaier, K., Golub, T. R., & Lamb, J. (2006). A method for high-throughput gene expression signature analysis. *Genome Biology, 7*(7), R61.

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

Powers, D. M. W. (2011). Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation. *Journal of Machine Learning Technologies, 2*(1), 37-63.

Prime, S. S., Nixon, S. V., Crane, I. J., Stone, A., Matthews, J. B., Maitland, N. J., . . . Scully, C. (1990). The behaviour of human oral squamous cell carcinoma in cell culture. *The Journal of Pathology, 160*(3), 259-269.

Prime, S. S., Eveson, J. W., Stone, A. M., Huntley, S. P., Davies, M., Paterson, I. C., & Robinson, C. M. (2004). Metastatic dissemination of human malignant oral keratinocyte cell lines following orthotopic transplantation reflects response to TGF-β1. *The Journal of Pathology, 203*(4), 927-932.

Qiu, W., Schonleben, F., Li, X., Ho, D. J., Close, L. G., Manolidis, S., . . . Su, G. H. (2006). *PIK3CA* mutations in head and neck squamous cell carcinoma. *Clinical Cancer Research, 12*(5), 1441-1446.

R Core Team. (2015). R: a language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing.

R Core Team. (2018). R: a language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing.

Raghavan, R., Hyter, S., Pathak, H. B., Godwin, A. K., Konecny, G., Wang, C., . . . Fridley, B. L. (2016). Drug discovery using clinical outcome-based connectivity mapping: application to ovarian cancer. *BMC Genomics, 17*(1), 811.

Ritchie, M. E., Phipson, B., Wu, D., Hu, Y., Law, C. W., Shi, W., & Smyth, G. K. (2015). limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Research, 43*(7), e47.

University of Malaya

Rosario, S. R., Long, M. D., Affronti, H. C., Rowsam, A. M., Eng, K. H., & Smiraglia, D. J. (2018). Pan-cancer analysis of transcriptional metabolic dysregulation using The Cancer Genome Atlas. *Nature Communications, 9*(1), 5330.

Sankaranarayanan, R., Ramadas, K., Thara, S., Muwonge, R., Thomas, G., Anju, G., & Mathew, B. (2013). Long term effect of visual screening on oral cancer incidence and mortality in a randomized trial in Kerala, India. *Oral Oncology, 49*(4), 314-321.

Schneider-Merck, T., Lammerts van Bueren, J. J., Berger, S., Rossen, K., van Berkel, P. H., Derer, S., . . . Dechant, M. (2010). Human IgG2 antibodies against epidermal growth factor receptor effectively trigger antibody-dependent cellular cytotoxicity but, in contrast to IgG1, only by cells of myeloid lineage. *The Journal of Immunology, 184*(1), 512-520.

Secretan, B., Straif, K., Baan, R., Grosse, Y., El Ghissassi, F., Bouvard, V., . . . WHO International Agency for Research on Cancer Monograph Working Group. (2009). A review of human carcinogens-Part E: tobacco, areca nut, alcohol, coal smoke, and salted fish. *The Lancet Oncology, 10*(11), 1033-1034.

Seiwert, T. Y., Burtness, B., Mehra, R., Weiss, J., Berger, R., Eder, J. P., . . . Chow, L. Q. (2016). Safety and clinical activity of pembrolizumab for treatment of recurrent or metastatic squamous cell carcinoma of the head and neck (KEYNOTE-012): an open-label, multicentre, phase 1b trial. *The Lancet Oncology, 17*(7), 956-965.

Sequist, L. V., Waltman, B. A., Dias-Santagata, D., Digumarthy, S., Turke, A. B., Fidias, P., . . . Engelman, J. A. (2011). Genotypic and histological evolution of lung cancers acquiring resistance to EGFR inhibitors. *Science Translational Medicine, 3*(75), 75ra26.

Seshan, V., & Olshen, A. (2017). DNAcopy: DNA copy number data analysis. R package version 1.52.0.

Setoain, J., Franch, M., Martinez, M., Tabas-Madrid, D., Sorzano, C. O., Bakker, A., . . . Pascual-Montano, A. (2015). NFFinder: an online bioinformatics tool for searching similar transcriptomics experiments in the context of drug repositioning. *Nucleic Acids Research, 43*(W1), W193-199.

Sheth, S., & Weiss, J. (2018). Pembrolizumab and its use in the treatment of recurrent or metastatic head and neck cancer. *Future Oncology, 14*(16), 1547-1558.

Shoemaker, R. H. (2006). The NCI60 human tumour cell line anticancer drug screen. *Nature Reviews Cancer, 6*(10), 813-823.

Siegel, R. L., Miller, K. D., & Jemal, A. (2017). Cancer Statistics, 2017. *CA: A Cancer Journal for Clinicians, 67*(1), 7-30.

Smeets, S. J., Braakhuis, B. J., Abbas, S., Snijders, P. J., Yistra, B., van de Wiel, M. A., . . . Brakenhoff, R. H. (2006). Genome-wide DNA copy number alterations in head and neck squamous cell carcinomas with or without oncogene-expressing human papillomavirus. *Oncogene, 25*(17), 2558-2564.

University of Malaya

Smirnov, N. V. (1939). Estimate of deviation between empirical distribution functions in two independent samples. (in Russian). *Bulletin of Moscow University, 2*(2), 3-16.

Smyth, G. K. (2004). Linear models and empirical bayes methods for assessing differential expression in microarray experiments. *Statistical Applications in Genetics and Molecular Biology, 3*, Article3.

Sokolova, M., Japkowicz, N., & Szpakowicz, S. (2006). Beyond accuracy, F-score and ROC: a family of discriminant measures for performance evaluation. In A. Sattar & B. Kang (Eds.), *AI 2006: advances in artificial intelligence* (pp. 1015-1021). Berlin, Heidelberg: Springer.

Sparano, J. A., Gray, R. J., Makower, D. F., Pritchard, K. I., Albain, K. S., Hayes, D. F., . . . Sledge, G. W. (2015). Prospective validation of a 21-gene expression assay in breast cancer. *The New England Journal of Medicine, 373*(21), 2005-2014.

Sreeramareddy, C. T., Pradhan, P. M., Mir, I. A., & Sin, S. (2014). Smoking and smokeless tobacco use in nine South and Southeast Asian countries: prevalence estimates and social determinants from demographic and health surveys. *Population Health Metrics, 12*, 22.

Stinchcombe, T. E. (2014). Recent advances in the treatment of non-small cell and small cell lung cancer. *F1000 Prime Reports, 6*, 117.

Strome, S. E., Dong, H., Tamura, H., Voss, S. G., Flies, D. B., Tamada, K., . . . Chen, L. (2003). B7-H1 blockade augments adoptive T-cell immunotherapy for squamous cell carcinoma. *Cancer Research, 63*(19), 6501-6505.

Subramanian, A., Narayan, R., Corsello, S. M., Peck, D. D., Natoli, T. E., Lu, X., . . . Golub, T. R. (2017). A next generation connectivity map: L1000 platform and the first 1,000,000 profiles. *Cell, 171*(6), 1437-1452.e1417.

Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S., Ebert, B. L., Gillette, M. A., . . . Mesirov, J. P. (2005). Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences of the United States of America, 102*(43), 15545-15550.

Taylor, J., Pavlick, D., Yoshimi, A., Marcelus, C., Chung, S. S., Hechtman, J. F., . . . Abdel-Wahab, O. (2018). Oncogenic TRK fusions are amenable to inhibition in hematologic malignancies. *The Journal of Clinical Investigation, 128*(9), 3819-3825.

Trapnell, C., Roberts, A., Goff, L., Pertea, G., Kim, D., Kelley, D. R., . . . Pachter, L. (2012). Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nature Protocols, 7*(3), 562-578.

Ushijima, M., Mashima, T., Tomida, A., Dan, S., Saito, S., Furuno, A., . . . Matsuura, M. (2013). Development of a gene expression database and related analysis programs for evaluation of anticancer compounds. *Cancer Science, 104*(3), 360-368.

University of Malaya

Vargas, A. J., & Harris, C. C. (2016). Biomarker development in the precision medicine era: lung cancer as a case study. *Nature Reviews Cancer, 16*(8), 525-537.

Vatanasapt, P., Suwanrungruang, K., Kamsa-Ard, S., Promthet, S., & Parkin, M. D. (2011). Epidemiology of oral and pharyngeal cancers in Khon Kaen, Thailand: a high incidence in females. *Asian Pacific Journal of Cancer Prevention, 12*(10), 2505-2508.

Vermorken, J. B., Mesia, R., Rivera, F., Remenar, E., Kawecki, A., Rottey, S., . . . Hitt, R. (2008). Platinum-based chemotherapy plus cetuximab in head and neck cancer. *The New England Journal of Medicine, 359*(11), 1116-1127.

Vermorken, J. B., Trigo, J., Hitt, R., Koralewski, P., Diaz-Rubio, E., Rolland, F., . . . Baselga, J. (2007). Open-label, uncontrolled, multicenter phase II study to evaluate the efficacy and toxicity of cetuximab as a single agent in patients with recurrent and/or metastatic squamous cell carcinoma of the head and neck who failed to respond to platinum-based therapy. *Journal of Clinical Oncology, 25*(16), 2171-2177.

Vidović, D., Koleti, A., & Schürer, S. C. (2014). Large-scale integration of small molecule-induced genome-wide transcriptional responses, kinome-wide binding affinities and cell-growth inhibition profiles reveal global trends characterizing systems-level drug action. *Frontiers in Genetics, 5*, 342.

Wang, K., Li, M., & Hakonarson, H. (2010). ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Research, 38*(16), e164.

Wang, X. D., Reeves, K., Luo, F. R., Xu, L. A., Lee, F., Clark, E., & Huang, F. (2007). Identification of candidate predictive and surrogate molecular markers for dasatinib in prostate cancer: rationale for patient selection and efficacy monitoring. *Genome Biology, 8*(11), R255.

Warnakulasuriya, S. (2009). Global epidemiology of oral and oropharyngeal cancer. *Oral Oncology, 45*(4-5), 309-316.

Warnes, G. R., Bolker, B., Bonebakker, L., Gentleman, R., Huber, W., Liaw, A., . . . Venables, B. (2019). gplots: Various R programming tools for plotting data. R package version 3.0.1.1. Available at: https://cran.r-project.org/web/packages/gplots/index.html.

Wei, G. G., Gao, L., Tang, Z. Y., Lin, P., Liang, L. B., Zeng, J. J., . . . Zhang, L. C. (2019). Drug repositioning in head and neck squamous cell carcinoma: an integrated pathway analysis based on connectivity map and differential gene expression. *Pathology, Research and Practice*, *215*(6), 152378.

Wu, C. L., Roz, L., McKown, S., Sloan, P., Read, A. P., Holland, S., . . . Thakker, N. (1999). DNA studies underestimate the major role of *CDKN2A* inactivation in oral and oropharyngeal squamous cell carcinomas. *Genes, Chromosomes & Cancer, 25*(1), 16-25.

University of Malaya

Wu, T. D., & Nacu, S. (2010). Fast and SNP-tolerant detection of complex variants and splicing in short reads. *Bioinformatics, 26*(7), 873-881.

Xiao, Y., Hsiao, T. H., Suresh, U., Chen, H. I., Wu, X., Wolf, S. E., & Chen, Y. (2014). A novel significance score for gene selection and ranking. *Bioinformatics, 30*(6), 801-807.

Xu, L., Tan, A. C., Naiman, D. Q., Geman, D., & Winslow, R. L. (2005). Robust prostate cancer marker genes emerge from direct integration of inter-study microarray data. *Bioinformatics, 21*(20), 3905-3911.

Yan, L., & Zhang, W. (2018). Precision medicine becomes reality - tumor type-agnostic therapy. *Cancer Communications (London, England), 38*(1), 6.

Yang, W., Soares, J., Greninger, P., Edelman, E. J., Lightfoot, H., Forbes, S., . . . Garnett, M. J. (2013). Genomics of Drug Sensitivity in Cancer (GDSC): a resource for therapeutic biomarker discovery in cancer cells. *Nucleic Acids Research, 41*(Database issue), D955-961.

Yee, P. S., Zainal, N. S., Gan, C. P., Lee, B. K. B., Mun, K. S., Abraham, M. T., . . . Cheong, S. C. (2019). Synergistic growth inhibition by afatinib and trametinib in preclinical oral squamous cell carcinoma models. *Targeted Oncology, 14*(2), 223-235.

Yeudall, W. A., Paterson, I. C., Patel, V., & Prime, S. S. (1995). Presence of human papillomavirus sequences in tumour-derived human oral keratinocytes expressing mutant p53. *European Journal of Cancer Part B Oral Oncology, 31B*(2), 136-143.

Zainal, N. S., Lee, B. K. B., Wong, Z. W., Chin, I. S., Yee, P. S., Gan, C. P., . . . Cheong, S. C. (2019). Effects of palbociclib on oral squamous cell carcinoma and the role of *PIK3CA* in conferring resistance. *Cancer Biology & Medicine, 16*(2), 264-276.

Zandberg, D. P., & Strome, S. E. (2014). The role of the PD-L1:PD-1 pathway in squamous cell carcinoma of the head and neck. *Oral Oncology, 50*(7), 627-632.

Zhang, B., & Horvath, S. (2005). A general framework for weighted gene co-expression network analysis. *Statistical Applications in Genetics and Molecular Biology, 4*, Article17.

Zhang, S. D., & Gant, T. W. (2008). A simple and robust method for connecting small-molecule drugs using gene-expression signatures. *BMC Bioinformatics, 9*, 258.

Zhao, M., Sano, D., Pickering, C. R., Jasser, S. A., Henderson, Y. C., Clayman, G. L., . . . Myers, J. N. (2011). Assembly and initial characterization of a panel of 85 genomically validated cell lines from diverse head and neck tumor sites. *Clinical Cancer Research, 17*(23), 7248-7264.

Zimmermann, M., Zouhair, A., Azria, D., & Ozsahin, M. (2006). The epidermal growth factor receptor (EGFR) in head and neck cancer: its role and treatment implications. *Radiation Oncology, 1*, 11.

University of Malaya

LIST OF PUBLICATIONS AND PAPERS PRESENTED

1. Lee, B. K. B., Tiong, K. H., Chang, J. K., Liew, C. S., Abdul Rahman, Z. A., Tan, A. C., . . . Cheong, S. C. (2017). DeSigN: connecting gene expression with therapeutics for drug repurposing and development. *BMC Genomics*, *18*(Suppl 1), 934.
2. Lee, B. K. B., Gan, C. P., Chang, J. K., Tan, J. L., Fadlullah, M. Z., Abdul Rahman, Z. A., . . . Cheong, S. C. (2018). GENIPAC: a genomic information portal for head and neck cancer cell systems. *Journal of Dental Research*, *97*(8), 909-916.

University of Malaya