國立臺灣大學公共衛生學院流行病學與預防醫學研究所

博士論文

Institute of Epidemiology and Preventive Medicine
College of Public Health
National Taiwan University

Doctoral Dissertation

用高通量基因資料建構不同癌症類型病患的預後指標

Development of Prognostic Models for Patients with Different Cancer Types by Using High-throughput Genomic Data

蕭亦文

YI-WEN HSIAO

指導教授：盧子彬 博士

Advisor: Tzu-Pin Lu, Ph.D.

中華民國 112 年 7 月

July 2023

doi:10.6342/NTU202300922

致謝

時光飛逝，博士班生涯也即將告一個段落了。謝謝指導老師盧子彬老師從我在臺大基因體中心擔任助理期間起的指導與鼓勵，使我不論是在基因相關研究或是統計方法上的應用都有更精進的表現。很謝謝老師給予我在研究上有自由發揮的空間，亦在我卡關時給予很多方向的指引。也謝謝老師一直以來給我很多的歷練讓我在學術的道路上更加成長茁壯，也有信心往下個階段邁進。

此外，也感謝臺大流預所提供了很優質的學習環境，在這樣的學習研究環境和優秀同儕們的相互激盪下，讓我在統計方法應用上有了很大的啟發。謝謝實驗室學弟妹，王琳、露葶與俊良等在研究上的討論與交流。也謝謝所上學長姐同學或學弟妹，偉珉、李驊、上奇、巧兒、宛融、永辰、千蘅與振宇等在學業研究或是助教課上的討論與交流，讓我收穫良多。特別感謝王琳和李驊在 2020 年末人生最低谷時候的陪伴和鼓勵。亦感謝臺大基因體中心生物資訊暨生物統計核心實驗室主持人莊曜宇老師的栽培，因為有助理期間的磨練，才能有如此扎實的研究經歷。亦感謝該實驗室的建樂學長、亮博與冠豪等同仁在生物資訊領域知識上的討論與交流。

最後，謝謝我的家人們，父母親和妹妹，願意在我追求夢想的道路上給予最無私地支持與照顧，才得以有今日所成。

doi:10.6342/NTU202300922

中文摘要

隨著高通量基因定量技術的普及以及生物資訊相關演算法的日新月異，使得大型基因體數據的取得與深入分析不再遙不可及。雖然臨床資料具有取得性以及 預後預測準確性的優勢，但仍有部分的預後預測是需要透過基因資訊來進一步解釋。所以如何有效地找尋適合不同癌症類型的生物標記來作為檢測或評估癌症疾病的發展進而設計客製化的治療方針是目前生物醫學領域的一大課題。故本論文旨在不同的基因層級上找出針對不同癌症類型病患的預後指標。本論文分成四部分: (1)比較同源重組修復缺失的分數和相對應的預後情況在非裔、歐裔和亞裔美國人之間的差異，再進一步探討哪種基因特徵類型以及那個基因集可以有效的檢測同源重組修復缺失狀態; (2)透過機器學習的方式找出有效的基因表現量圖譜來建構一個可以高靈敏度地找出高風險的卵巢癌患者風險預測模型; (3)比較 B 型肝癌和 C 型肝癌患者之間所有腫瘤浸潤淋巴細胞的差異以及其預後的情況; (4)欲建構一個可以有效找出癌症相關的競爭型核糖核酸的演算法。分別得到以下的結果: (1)找出族群特異性的同源重組修復缺失基因突變圖譜，有助於客製化建立各個族群的篩選源重組修復缺失現象病患進而提供特定的治療方針; (2)建構一個高靈敏度的高死亡風險卵巢癌患者預測模型，且選用的基因集特徵需與卵巢癌的生物機制有相關; (3)了解在腫瘤與非腫瘤組織的腫瘤浸潤淋巴細胞的總量差異在不同的肝癌亞型中是否一致，以及找出在不同

doi:10.6342/NTU202300922

肝癌亞型中哪些免疫細胞是跟預後相關的生物標記指標; (4) 建構一個穩健且可應用的找尋競爭型核糖核酸的演算法，並透過一些下游分析來進一步解讀這樣的基因調控事件。所以透過這些研究來找尋癌症特有的基因特徵，進而了解在癌症種類或是其亞型與族群間的預後差異，可以針對不同癌症或是其亞型與族群設計客製化檢測工具，以達到精準醫療的願景。

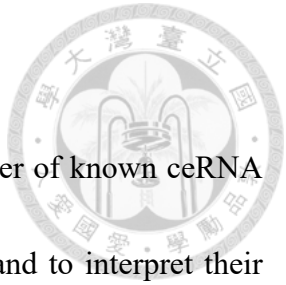
關鍵字:基因表現量、腫瘤浸潤淋巴細胞、同源重組修復缺失、卵巢癌、肝癌、泛癌症研究、競爭型核糖核酸

doi:10.6342/NTU202300922

Abstract

With the rapid development of the next-generation sequencing techniques and bioinformatics algorithms, high-throughput biological data have become less expensive and more accessible such that more studies can explore to a greater level the genetic impacts on various diseases. Although clinical data has the advantage of its accessibility and accuracy of prognostic prediction, 30-40% of patients fail to use clinical factors to evaluate their prognostic outcomes. Therefore, effectively identifying genomic biomarkers for the prognostic evaluation and the guidance of the necessary medical intervention remains the main challenge in the biomedical field. Hence, this dissertation aims to define prognostic biomarkers specifically for patients with particular cancer types using DNA- or RNA-level genomic data. There are three specific objectives in this dissertation: (1) to elucidate the racial differences in HRD scar scores in multiple cancers and to evaluate their associations with clinical outcomes; also, to assess the applicability of each HRD gene-sets for each cancer-race group. (2) to develop a bagging-based algorithm with GA-XGBoost models that predicts the high risk of death from ovarian cancer using gene expression profiles; (3) to define the cell composition of the immune response in both HBV−HCC and HCV−HCC and to investigate its relationship with clinical outcomes such as overall survival and

doi:10.6342/NTU202300922



recurrence-free survival; (4) to effectively identify the massive number of known ceRNA interactions using genome-wide transcriptome and miRNA profiles and to interpret their associations with cancers. Through the above studies, the results revealed that (1) a race-specific profile of the predisposing HRD-related genes with mutations was identified for customized design of HRD screening approaches. (2) a robust risk prediction model using selected gene expression related to ovarian cancer with a high risk of death will be constructed; (3) whether the difference of the total amount of TILs in tumor and non-tumor tissues is consistent in different liver cancer subtypes, and the biomarkers that associated with survival outcomes between two subtypes of hepatocellular carcinomas will be unveiled; (4) a computational framework for the identification of ceRNA events and their biological interpretations. In conclusion, through these studies, we would be able to identify specific genomic features of cancers, understand the prognostic difference among cancer subtypes or racial groups, and design the customized tumor testing tools for specific cancer types or racial groups, reaching the ultimate goal of precision medicine.

Keywords: Gene expression, Tumor-infiltrating lymphocytes, Homologous recombination deficiency, Ovarian cancer, Hepatocellular carcinomas, Pan-cancer study, Competing endogenous RNA

doi:10.6342/NTU202300922

CONTENTS

doi:10.6342/NTU202300922

doi:10.6342/NTU202300922

doi:10.6342/NTU202300922

doi:10.6342/NTU202300922

Chapter 1. Introduction

1.1 The current trend of cancer prognosis

Cancer prognosis prediction is one of the most important issues in cancer studies. The accuracy of such a prediction system can provide proper medical intervention, reduce unnecessary preventive treatments, and eliminate the medical burden in many countries [1]. Clinical features, like age, sex, cancer stage and TNM (tumor-node-metastasis), have been the most common prognostic factors over the past decades due to the advantage of its accessibility. With the combination of proper statistical models, such as Cox-regression models or advanced machine learning approaches, patients with specific cancer types can be classified into different prognostic levels for customized treatments [2, 3]. For example, the “predict” website supported by Public Health England and the University of Cambridge use such clinical data to predict the prognosis of cancer patients [4]. However, 30-40% of patients fail to use clinical factors to evaluate their prognostic outcomes [5]. Hence, effectively identifying genomic biomarkers for the prognostic evaluation and the guidance of the necessary medical intervention remains the main challenge in the biomedical field.

1.2 Genomic data and their clinical application

The discovery of the double-helix structure of DNA in the 1950s has revolutionized the field of biological science. In 1953, a landmark series of papers on DNA structure by many scientists,

doi:10.6342/NTU202300922

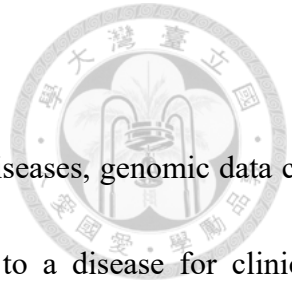
especially Franklin, Watson and Crick, was published in Nature. This work firmly established that DNA is a double helix that contains antiparallel nucleotide chains and a specific pattern for base pairings, giving birth to the new discipline of molecular biology [6, 7]. Such DNA's structure has been confirmed as the fundamental genetic information in humans and has assisted the establishment of the Central Dogma, instruction in the procedure of how DNA is converted into a functional product to maintain the biological functions in the human body. The elucidation of such biological principles provided a key to unveiling the mysteries of heredity and the mechanism of diseases. Great advances in biochemistry technology also led to an explosion of scientific research in the field of Genomics and its clinical applications. After Dr Kary Mullis invented the polymerase chain reaction (PCR) technique, which allows the quantality of nucleus acid molecules [8], more and more advanced testing approaches were developed for DNA quantification. For example, the development of the Sanger sequencing technique proposed by Dr Sanger in 1977 has accelerated the promotion of molecular biological research [9]. Recently, due to the mature and popularity of microarray and next-generation sequencing (NGS) techniques, high-throughput biological data have become less expensive and more accessible such that more studies can explore to a greater level the genetic impacts on various diseases, hoping to provide novel insights into precision medicine.

Although many modern techniques such as third-generation sequencing [10] or single-cell sequencing [11] for genomic quantification has been proposed and widely used these days,

doi:10.6342/NTU202300922

microarray and next-generation sequencing are still indispensable tools for genomic research over the past decades. The public data used in this dissertation were obtained from these two techniques, therefore it is necessary to describe each of the methodologies. Microarrays require prior knowledge of the query genome and quantify the amount of genomic information by fluorescent intensity; the most common platforms of such quantification techniques are Affymetrix [12] and Illumina [13]. In contrast to the former methods, sequencing-based platforms can directly identify the nucleic acid sequence of a given DNA or component DNA molecule; for example, the Sanger sequencing conventionally used capillary electrophoresis to read the nucleic acid sequence. Instead, NGS used the short read and massively parallel sequencing technique to empower the sequencing capabilities. Such technologies include Illumina [14], Roche 454 [15] and Ion Torrent [16] sequencing approaches. The public data used in this dissertation were mainly generated by Illumina sequencing which can simultaneously identify DNA bases by a unique fluorescent signal emitted by each base and add these bases to a nucleic acid chain. Compared to microarray, NGS-based approaches own many advantages: (1) no prior knowledge of genome required; (2) single-nucleotide resolution for genomic data; (3) higher reproducibility for scientific research [17]. Although the properties of NGS-based approaches generally outweigh those of array-based methods, many studies still apply array-based methods because of the cost-effectiveness. Therefore, in this dissertation, public data generated from these two techniques will be used for further discussion.

doi:10.6342/NTU202300922

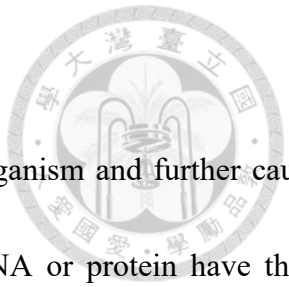


Aside from accelerating the understanding of the etiology of many diseases, genomic data can also assist clinicians to identify important biomarkers that are unique to a disease for clinical applications. By the definition, there are three types of biomarkers: diagnostic, prognostic and predictive [18, 19]. Diagnostic biomarkers can determine whether a disease/cancer occurs and even which cancer type occurs; for example, the PCA3 gene and TMPRSS2-ERG fusion gene can be served as the diagnostic guidance of prostate cancer [20]. A prognostic biomarker can determine the overall survival outcome of a patient with such genomic features no matter whether they receive the treatment or invention or not. For instance, IL-6 and VEGF-A are common genes for the prognosis prediction of ovarian cancer [21]. Regarding predictive biomarkers, such features can identify whether a specific treatment/therapy is suitable for a patient; for example, the gene expression of IGF-1R, IGF-1 and IGF-2 can guide the treatment for colon cancer patients [22]. Since these biomarkers can serve as the indicators for disease progression or prediction of drug function, such indicators have been widely applied to cancer diagnosis and the identification of potential drug targets.

1.3 Different molecular levels of genomic data

According to the principle of the central dogma of molecular biology, the DNA sequence of a gene is transcribed into an RNA molecule, i.e. message RNA (mRNA) and then translated into the amino acid sequence of a polypeptide. The proper folding structure of a polypeptide guarantees functional proteins so that they can properly modulate the biological mechanisms in an organism.

doi:10.6342/NTU202300922



The dysfunction of proteins may affect the biological functions of an organism and further cause disease or cancer. All three levels of genomic data including DNA, RNA or protein have their potential biomarkers for a specific disease/cancer. However, due to the complexity of the protein structure, the technique of protein isolation and purification is difficult to conduct than DNA/RNA isolation so that it is not easy to acquire such data. Therefore, this dissertation only focuses on the genomic data at DNA and RNA molecular level.

The human genome project has revealed that the human genome encapsulates 23 pairs of chromosomes that involve about $3*10^9$ base pairs (bps) [23]. This project also indicated that only a 0.1% difference of nucleic acid sequence between individuals without familial relationships. The main cause of such genetic difference is genetic variation triggered by either spontaneous mutations such as radiation or incorrect DNA replication or mutagens [24]. Such variation is called mutation when the mutation frequency is less than 1%; however, it turns to be the events of genetic polymorphism when the frequency is over 1% [25] and among them, about 95% of such events are single nucleotide polymorphism (SNP), that is, a single nucleotide is substituted at a specific position in the genome [26]. Since the high occurrence of SNP and the different combinations of SNPs in individual genome, many studies have applied the genomic data or biobank data to unveil the potential SNPs as biomarkers; for example, a previous study used 241 SNPs to build a risk prediction model for lung cancer patients from European population or Asian population [27]. Aside from the

doi:10.6342/NTU202300922

single nucleotide change, copy number variation (CNV) refers to large scale change, either amplification or deletion, in the genome usually over 1 kilobase pair. The relationships between such genetic features and cancers also have been widely discussed. For example, CNV occurred in BRCA1, BRCA2, TP53 and CHEK2 genes have been proved to be related to the development of breast cancer [28]. The above-mentioned genetic features both occur at the DNA sequences and the change of these genetic features show the hereditable character to the next generation.

However, such change happened on the protein-coding region will affect the expression of the gene and further generate the dysfunction of protein, leading to diseases or cancers. Gene expression, also called transcriptomic profile can be quantitated by biological experiments and has two characteristics: tissue-specific and tumor-specific [29]. Due to its tumor-specific property, more and more studies have focused on the identification of minimal geneses and their gene expression profile in the application of cancer prognosis prediction [30]. Several commercial toolkits, such as MammaPrint [31] and Oncotype DX [32], have applied particular gene expression signatures to identify the survival risk for patients with various cancer types for the fulfilment of personalized medicine. Therefore, different levels of genomic features indeed play an important role in terms of cancer prognosis.

doi:10.6342/NTU202300922

1.4 Genomic Issues

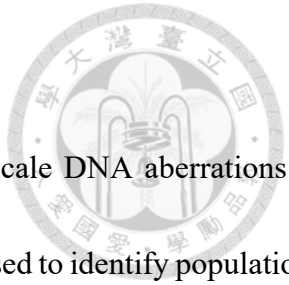
1.4.1 Genomic difference among cancer subtypes

In 2000, Perou and Sorlic et. al. [33] had proposed the concept of cancer subtypes through analyzing the microarray data from breast cancer samples. They have classified breast cancer into four subtypes based on their molecular signatures: basal-like, luminal A, luminal B and Her2+. Also, breast cancer patients with different molecular subtypes have shown a wide range of prognostic outcomes. Through the Cancer Genome Atlas (TCGA) Research Network, the genomic classification across about 25 cancer types have been verified from 2013 to 2018. Such classification has revealed that different cancer subtypes induced different molecular signatures that are involved in different biological pathways, causing a wide range of prognostic/therapeutic outcomes [34-36].

1.4.2 Genomic difference among racial groups

The difference in cancer risk among populations is better explained by the genetic landscape and frequencies of inherited variants that are predisposed to cancer than by non-genetic factors [37]. To take the genetic heterogeneity among different populations into consideration, a large-scale investigation of multiple populations is necessary. Two population-based resources (gnomAD [38, 39] and gnomAD-SV [40]) for mutational and structural changes have been developed in the last two

doi:10.6342/NTU202300922



decades to address this issue. With the curation of frequencies of large-scale DNA aberrations in different populations, advanced filtering procedures and selections can be used to identify population-specific genetic hotspots. A recent systematic study reported that a higher level of genomic instability was observed in African American patients than in European Americans with breast, head and neck, and endometrial cancers [41]. Another study showed that the occurrence of a variant of uncertain significance in genetic testing for hereditary breast and ovarian syndrome was more frequent among non-Whites than Whites [42], which further emphasizes the importance of obtaining population-specific data.

1.5 Aims

Therefore, in this dissertation, we used different levels of genetic features, including DNA level, RNA level and regulatory level, and a wide range of well-established methodologies in ovarian cancer, hepatocellular carcinomas and multiple cancers for their prognosis prediction, hoping to achieve the ultimate goals of precision medicine.

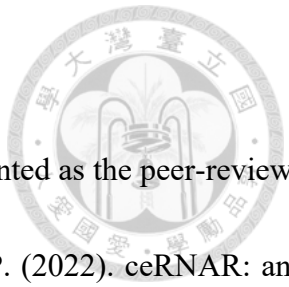
1.6 Outline

This dissertation was formed based on published articles. The supplementary data for chapter 2-5 are available in the following link: https://drive.google.com/drive/folders/1c4A4lXYGWkFWdvatlCLTJTaDwGg6RCvw?usp=sharing.

doi:10.6342/NTU202300922

This dissertation includes four main parts (Figure 1-1). Chapter 2 depicted a population-based genetic profile of HRD-related genes was identified for customized design of HRD screening approaches. This chapter also has been presented as the peer-reviewed journal article: **Hsiao, Y. W.**, & Lu, T. P. (2021). Race-Specific Genetic Profiles of Homologous Recombination Deficiency in Multiple Cancers. *Journal of Personalized Medicine*, 11(12), 1287. Chapter 3 illustrated the development of a bagging-based algorithm with GA-XGBoost models to predict the high risk of death from ovarian cancer using gene expression profiles. It extended from the master thesis of Chun-Liang Tao and all credits were shared equally. This chapter also has been presented as the peer-reviewed journal article (*: co-first author): **Hsiao, Y. W.***, Tao, C. L.*, Chuang, E. Y., & Lu, T. P. (2021). A risk prediction model of gene signatures in ovarian cancer through bagging of GA-XGBoost models. *Journal of advanced research*, 30, 113-122. Chapter 4 depicted the cell composition of the immune response in both HBV−HCC and HCV−HCC and to investigate its relationship with survival outcomes. It was extended from the master thesis of Chun-Liang Tao and all credits were shared equally. This chapter also has been presented as the peer-reviewed journal article (*: co-first author): **Hsiao, Y. W.***, Chiu, L. T.*, Chen, C. H., Shih, W. L., & Lu, T. P. (2019). Tumor-infiltrating leukocyte composition and prognostic power in hepatitis B-and hepatitis C-related hepatocellular carcinomas. *Genes*, 10(8), 630. Chapter 5 presented a computational framework for the identification of ceRNA events and their biological interpretations. It was extended from the master thesis of Lin

doi:10.6342/NTU202300922



Wang and all credits were shared equally. This chapter also has been presented as the peer-reviewed journal article (*: co-first author): **Hsiao, Y. W.***, Wang, L.*, & Lu, T. P. (2022). ceRNAR: an R package for identification and analysis of ceRNA-miRNA triplets. *PLoS computational biology*, 18(9): e1010497.

doi:10.6342/NTU202300922

1.7 Figures

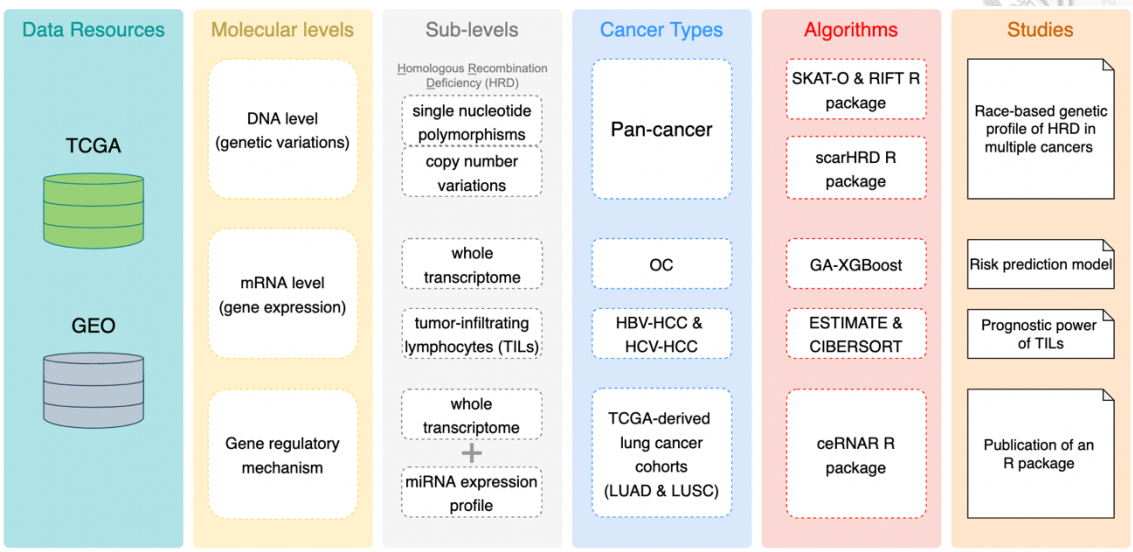


Figure 1-1. An overview of the overarching objectives of this PhD dissertation.

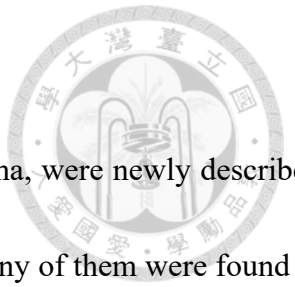
doi:10.6342/NTU202300922

Chapter 2. Race-specific genetic profiles of homologous recombination deficiency in multiple cancers

2.1 Abstract

Homologous recombination deficiency (HRD) has been used to predict both cancer prognosis and the response to DNA-damaging therapies in many cancer types. HRD has diverse manifestations in different cancers and even in different populations. Many screening strategies have been designed for detecting the sensitivity of a patient's HRD status to targeted therapies. However, these approaches suffer from low sensitivity, and are not specific to each cancer type and population group. Therefore, identifying race-specific and targetable HRD-related genes is of clinical importance. Here, we conducted analyses using genomic sequencing data that was generated by the Pan-Cancer Atlas. Collapsing non-synonymous variants with functional damage to HRD-related genes, we analyzed the association between these genes and race within cancer types using the optimal sequencing kernel association test (SKAT-O). We have identified race-specific mutational patterns of curated HRD-related genes across cancers. Overall, more significant mutation sites were found in *ATM*, *BRCA2*, *POLE,* and *TOP2B* in both the 'White' and 'Asian' populations, whereas *PTEN, EGFG*, and *RIF1* mutations were observed in both the 'White' and 'African American/Black' populations. Furthermore, supported by pathogenic tendency databases and previous reports, in the 'African American/Black' population, several associations, including *BLM* with breast invasive carcinoma, *ERCC5* with ovarian

doi:10.6342/NTU202300922



serous cystadenocarcinoma, as well as *PTEN* with stomach adenocarcinoma, were newly described here. Although several HRD-related genes are common across cancers, many of them were found to be specific to race. Further studies, using a larger cohort of diverse populations, are necessary to identify HRD-related genes that are specific to race, for guiding gene testing methods.

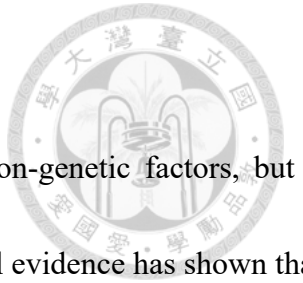
Keywords: homologous recombination deficiency; mutation; structural variation; pan-cancer; racial difference; therapeutic targets

2.2 Introduction

Homologous recombination deficiency (HRD) is a dysfunction of the homologous recombination repair (HRR) pathway, which is responsible for the repair of DNA double-strand breaks [43]. HRR pathway defects are often caused by the mutation of *BRCA1* or *BRCA2* but they can also arise from mutations in HRR-related genes, such as *AMT* or *CDK12,* that carry different mutation loads [44]. Aside from genetic mutations, structural aberrations of DNA, such as a loss of heterozygosity (LOH), large-scale state transitions (LST), and telomeric allelic imbalance (TAI), have been recognized as defining characteristics of HRD [45-47]. These genetic features are associated with an increased sensitivity to DNA-damaging agents, such as poly ADP-ribose polymerase (PARP) inhibitors and platinum-based antitumor drugs [48]. For this reason, they have served as targets of chemotherapy and immunotherapy agents for several cancers, especially in breast cancer, ovarian cancer, pancreatic cancer, and prostate cancer [49-52]. In addition to their response to drugs, the survival outcomes of many cancer patients are also associated with the degree of HRD in many cancers [53]. Thus, the genetic features related to HRD are capable of being used as biomarkers to predict patients' drug response and survival outcomes.

Several epidemiological studies have pointed to large racial differences in cancer incidence and survival of many cancer types [54, 55]. According to data from US cancer registries, African American/Black (AA/B) people have a higher incidence and lower survival of all cancers when compared to the White population, after adjusting for confounding factors, such as socioeconomic

doi:10.6342/NTU202300922



status and behavior [56]. Some of this difference may be caused by non-genetic factors, but a considerable amount may be attributed to genomic architecture. Substantial evidence has shown that the population-specific genetic background can at least partially explain the reason for unequal cancer burden among different racial groups [57, 58]. For instance, Conti et al. found that an African American/Black male has a 75% higher risk of getting prostate cancer than a White man, noting that past studies have overrepresented the White population and ignored this variation in cancer risk by race [59]. Recent research has also revealed that there is a higher prevalence of HRD events in lung cancer occurring in African Americans than in European Americans, highlighting the necessity of including underrepresented populations in genetic studies [60]. Therefore, bringing the genetic risk for people of various racial groups into focus may help to explain the role of race in the progression of cancer and HRD events, leading to better screening protocols in people of all races, as well as the earlier detection of each type of cancer.

The fundamental goal of precision medicine in cancer care is to use genetic information to prevent, diagnose, or treat cancers, and advances in this area have led to the development of targeted cancer therapies [61]. Several genetic testing strategies are available for many cancers, especially breast cancer, which is the most well-studied cancer type. For example, the Oncotype DX Breast Recurrence Score test, which is based on 21 genes [62], and the MammaPrint test, which is based on a 70-gene signature, can estimate the risk of recurrence [63], and have both been widely applied to

doi:10.6342/NTU202300922

guide the clinical treatment of breast cancer patients. Recently, an increasing number of tumor testing kits have been designed for HRD detection, such as Myriad myChoice® CDx [64] and FoundationOne® CDx [65]. Although current genetic testing still suffers from limited availability across cancer types, as well as false positive/negative issues, the role of such testing remains essential in clinical applications to achieve the ultimate goal of precision medicine.

Currently, several genetic testing methods, such as *BRCA1/2* germline testing alone, in a combination with HRR-related genes, and in a combination with genomic instability, have shown a wide range of sensitivity in HRD detection [66]. The best testing approach identifies only about 50% of women who are eligible for treatment with breast cancer drugs, such as a mixture of LYNPARZA and bevacizumab [67], highlighting that many patients are missed due to high false-negative results, such that the accurate detection of this population, using better biomarkers of HRD, is of clinical relevance. Although multiple factors underlie the racial differences in cancer prognosis and drug response, many population-based studies have pointed out that these differences may be partially attributed to inherent differences at the DNA level [68-70]. Therefore, the overarching aim of this study is to identify race-specific pathogenic HRD-related variants among cancers, providing new insight that is specific to each cancer type and population group.

doi:10.6342/NTU202300922

2.3 Materials and Methods

2.3.1. Study Cohorts and Patients

Clinical and biospecimen annotation files in txt format (n = 9125, across 33 cancer types) were downloaded from the TCGA Pan-Cancer Clinical Data Resource [71]. A full list of TCGA cancer type abbreviations can be found in the Genomic Data Commons (https://gdc.cancer.gov/resources-tcga-users/tcga-code-tables/tcga-study-abbreviations (accessed on 02/06/2021)). Samples that did not belong to one of the three racial groups according to the TCGA Pan-Cancer Clinical Data—'White', 'Asian', and 'African American/Black'—or did not have survival information were excluded. To ensure we had sufficient sample sizes to achieve adequate power in each cohort, only the cancer cohorts with >100 samples were included, and racial subgroups within each cancer cohort that had <10 samples were excluded. Consequently, 7241 patients across 24 cancer types were retained and combined (i.e., the pan-cancer cohort) for the downstream analysis. All sample selections were conducted using R software (v4.0.5).

2.3.2. The Determination of Pathogenicity for Somatic Variant Calls

A mutation annotation file (MAF; mc3.v0.2.8.PUBLIC.maf), including 33 cancer types, which was generated by the MuTect2 pipeline according to the GRCh38.d1.vd1 reference sequence, was also retrieved through the Pan-Cancer Atlas (https://gdc.cancer.gov/about-

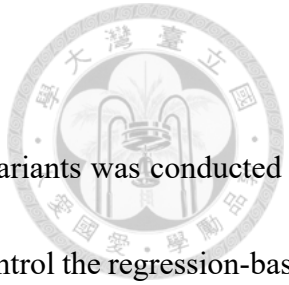
doi:10.6342/NTU202300922

data/publications/pancanatlas (accessed on 02/06/2021)). To define the causal HRD genes, gene lists from a survey of the literature were collected and named as 'DDRD_assay_42' [72], 'mutated_gene_21' [43], 'HR_PARP_132' [43], and 'DDR_276' [73]. The approximate chromosomal position of each gene in the lists, based on the GRCh38 genome version, was extracted from GeneCards (https://www.genecards.org/ (accessed on 02/06/2021)) by an in-house python web scraping script (https://github.com/ywhsiao/jpm_submission (accessed on 02/06/2021)). Through such information, the somatic variants in the MAF file within the region of these genes were then extracted. A measurement of global mutation loads, called tumor mutation burden (TMB), was calculated by directly counting the number of variants and then dividing that number by the genomic length of the target gene (unit: mutations/Mb). The above-mentioned filtering and calculation steps were performed by R software (v4.0.5).

2.3.3. Groupwise Association Test, Outlier Detection Analysis, and Variant Annotation

The previously downloaded MAF was modified and combined with the racial information from the previously downloaded clinical data for the groupwise association test. The optimal sequencing kernel association test (SKAT-O), which is a linear combination of the burden test and SKAT statistics [74], was performed to evaluate the association between the missense variants with a functional impact, which were classified as "probably/possibly damaging" or "deleterious" in two functional prediction algorithms (SIFT and PolyPhen-2), and a phenotype (i.e., a specific race) using

doi:10.6342/NTU202300922



the SKAT R package. In addition, a principal component analysis of all variants was conducted by PLINK (v2.0), and the resulting eigenvectors were used as a covariate to control the regression-based analysis. Then, we applied the rare variant influential filtering tool (RIFT), an R package which generates a delta chi-square score for each significant variant, and non-parametric outlier detection methods to identify the most influential variants that were specific to race [75]. Finally, we used pathogenicity determinations by REVEL and ClinVar to validate the pathogenic tendency of the identified variants [76, 77].

2.3.4. HRD Score Calculation

The copy number segmentation file with annotations (TCGA_mastercalls.abs_segtabs.fixed.txt) and the purity/ploidy file (TCGA_mastercalls.abs_tables_JSedit.fixed.txt) that was generated by ABSOLUTE software were downloaded from the above-mentioned Pan-Cancer Atlas portal. These two files were then compiled to generate the input format supported by the scarHRD R package [78] for the calculation of the counts of each HRD component (HRD-loss of heterozygosity (HRD-LOH), HRD-large-scale state transitions (HRD-LST), and HRD-telomeric allelic imbalance (HRD-TAI)) and then summarized the total HRD score. Here, we evaluated the global and geneset-specific HRD scores based on previously reported genesets across cancers.

doi:10.6342/NTU202300922

2.3.5. Statistical Analysis

To further compare the HRD scores and TMB values across race in each specific cancer cohort, the Wilcoxon rank-sum test or the Kruskal-Wallis test were used. When comparing such values among the three races, the false discovery rate was used for the correlation of multiple comparisons. A *p*-value of less than 0.05 is reported as statistically significant. Spearman's correlation coefficient was calculated to evaluate the correlations among the genome-wide HRD scores and global TMB values across cancers. Statistical analyses and data visualizations were performed using the 'ggpubr' [79] and 'ggsignif' [80] R packages.

2.3.6. Survival Analysis

To investigate the effect of genomic features on survival, univariate or multivariate covariate-adjusted Cox proportional hazards models were used when the following covariates were provided in the sample's clinical information: gender, age at initial pathogenic diagnosis, TNM (tumor, nodes, and metastases) status, and stage. The performance of the survival predictors was assessed by the concordance index (c-index). The hazard ratios (HRs), along with their 95% confidence intervals (CIs) and the corresponding *p*-values of these predictors, were calculated. We also dichotomized the patients according to the second tertile value (66.7%) of the HRD score to create two groups ('HRD' versus 'not HRD') [81]. This allowed the group with the higher level of HRD, defined as HRD in this

doi:10.6342/NTU202300922

study, to always possess one-third of the patients in each dataset. The Kaplan-Meier method was used to estimate the survival endpoints and assess the significant differences in the survival outcomes between the two predefined groups through the log-rank test. The above-mentioned statistical tests were conducted by the 'ggpubr', 'survival' [82], and 'survminer' [83] R packages.

3.4 Results

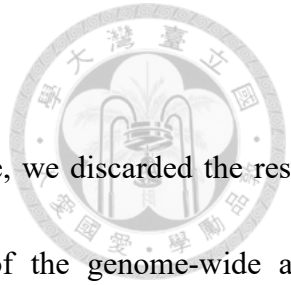
3.4.1. Distribution of the TCGA Pan-Cancer Atlas Cohort across Racial Groups

We summarize the distribution of the 7241 TGCA cases with their HRD scores across the three racial groups in Table 2-1. The 'White' group contained 83.04% (n = 6013) of the cohort, and the rest of the cohort consisted of 10.04% (n = 727) African American/Black and 6.92% (n = 501) Asian patients. Among the 24 cancer types, the largest population-specific cancer cohorts were breast invasive carcinoma (BRCA) for the 'White' group (n = 678) and the African American/Black group (n = 159), and liver hepatocellular carcinoma (LIHC) for the 'Asian' group (n = 150).

3.4.2. Association of HRD Scores with Survival across Cancer Types

We first determined the HRD scores across the Pan-Cancer Atlas cancer types by integrating the data on copy number segment and ploidy, as defined by ABSOLUTE. The HRD scores are shown in Figure 2-1A. The HRD scores, which were defined by different genesets, varied by cancer type. We do knowledge that a geneset with a lower number of genes, or which did not have a proper gene list,

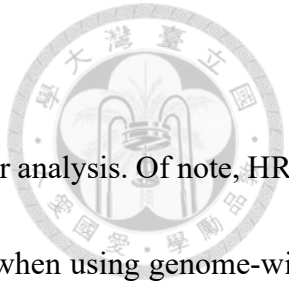
doi:10.6342/NTU202300922



had a lower HRD score estimated by the scarHRD R package. Therefore, we discarded the result defined by 'mutated_gene_21' here. The HRD distribution patterns of the genome-wide and 'DDR_276' genesets were similar, though the calculated HRD scores that were based on 'DDR_276' were generally lower than the genome-wide scores. For example, the cancer types which were ranked as having the top 5 highest HRD scores were the same in both genesets: ovarian serous cystadenocarcinoma (OV), esophageal carcinoma (ESCA), sarcoma (SARC), lung squamous cell carcinoma (LUSC), and bladder urothelial carcinoma (BLCA). Intriguingly, the HRD scores that were defined by 'DDRD_assay_42' were higher than those defined by 'HR_PARP_132' across the cancers analyzed.

We next investigated the association of the HRD scores that were defined by the different genesets with overall survival. The ability of HRD scores to predict survival was also different across cancers and even across the defined genesets; some of the genesets (e.g., 'HR_PARP_132') even showed diverse prediction outcomes when compared with the rest of the genesets (Figure 2-1B). These results show that the HRD, defined by different genesets, may affect the incidence of HRD events and their corresponding clinical outcomes across cancers. For the risk prediction (Figure S2-1A, S2-1B), we only focused on four cancer types (BRCA, pancreatic adenocarcinoma (PAAD), prostate adenocarcinoma (PRAD), and OV) that are widely discussed in the field of HRD and PARP inhibitors, along with the pan-cancer results. In general, the HRD scores were not significantly linked

doi:10.6342/NTU202300922



to survival for the individual cancers, however, they were for the pan-cancer analysis. Of note, HRD, as defined here (top tertile), was associated with the survival of OV only when using genome-wide scores to define HRD (and in a counterintuitive direction), whereas survival was associated with the HRD score in the pan-cancer cohort under HRDs from any of the genesets. The relationship between the cancer types and survival, which was determined by the Cox proportional hazards analysis (Figure 2-2), further clarified that different cancer types had different survival outcomes, suggesting that the mortality risk, as defined by HRD scores, should take cancer type into consideration.

2.4.3. Synergistic Effects between Genome-Wide HRD and Global TMB among Cancers

It is known that both the quantity of DNA mutations (global TMB value) and the changes in the copy number (genome-wide HRD scores) can be used to identify patients with HRD [84, 85]. To explore whether these two indicators were correlated, we calculated the Spearman correlation coefficients between genome-wide HRD scores and global TMB values, and the c-index was examined to evaluate their ability to predict overall survival. In Figure S2-2A, distinct correlation levels between HRD scores and TMB levels were observed in several cancer cohorts, as well as the pan-cancer cohort. These two genomic indicators were positively correlated ($p < 0.001$) in PAAD (R = 0.54), SARC (R = 0.5), and BRCA (R = 0.54), and negatively correlated in colon adenocarcinoma (COAD; R = −0.24, $p < 0.00$) and uterine corpus endometrial carcinoma (UCEC; R = −0.32, $p <$ 0.001) (Figure S2-2A, Table 2-2). As shown in Figure S2-2B, the predictive values of these indicators

doi:10.6342/NTU202300922

were generally the same in most cancers, as well as in the pan-cancer analysis; however, the predictive ability of the HRD scores in some cancer cohorts (e.g., kidney renal papillary cell carcinoma (KIRP), PAAD, PRAD, and testicular germ cell tumors (TGCT)) outweighed that of the global TMB values. Nevertheless, these results still suggest that TMB may be representative of HRD events in some cancers.

3.4.4. Racial Differences of Genome-Wide HRD and Global TMB across Cancers

To explore the racial differences in the genomic instability and mutation patterns across the 24 cancer types, we also examined the genome-wide HRD scores and TMB values, after being stratified by race in each tumor type (Figure 2-3A,B; Figure S2-3). Due to the limitations of the sample size following the racial stratification, we only considered cohorts that contained at least two populations with a subgroup sample size larger than 10. Cancers with significant racial differences ($p < 0.05$) in terms of the HRD score or TMB value are summarized in Table S2-2. Overall, six cancer types, including BLCA, BRCA, and head and neck squamous cell carcinoma (HNSC), had significant differences in terms of the genome-wide HRD scores among the populations (Kruskal-Wallis test; $p < 0.05$). Specifically, based on the pairwise comparison results (Wilcoxon rank-sum test; p-adj < 0.05), five cohorts between the 'White' and the 'Asian' groups (such as BLCA, BRCA, and LIHC), three cohorts between the 'Asian' and the 'African American/Black' groups (such as BLCA, HNSC, and UCEC), and five cohorts between the 'White' and the 'African American/Black' groups (such as

doi:10.6342/NTU202300922

BRCA, HNSC, and KIRP) were significantly different. Similar to the HRD score, racial differences were also observed in the global TMB values (Table S2-2). Generally, four cancers, including BLCA, BRCA, and LUAD, were statistically different in terms of the TMB values among all three races. According to the pairwise comparison results (Wilcoxon rank-sum test; p-adj < 0.05), five cohorts between the 'White' and the 'Asian' groups (such as BLCA, BRCA, and LIHC), two cohorts between the 'Asian' and the 'African American/Black' groups (BLCA and CESC), and two cohorts between the 'White' and the 'African American/Black' groups (BRCA and LUAD) were significantly different. Collectively, these results demonstrated that racial differences that manifest in molecular changes at the DNA level, such as TMB and HRD, should be considered in many cancers.

3.4.5. Predisposing HRD-Related Genes that Are Specific to Race

We next examined the race-specific pathogenicity of non-synonymous mutations in 364 non-repeated HRD-related genes that were extracted from the five above-mentioned genesets ('DDRD_assay_41', 'mutated_gene_21', 'HR_PARP_132', and 'DDR_276') using groupwise association tests (Table 2-3 and Figures S2-4 to S2-6). Here, prefiltering mutations with functional meaning, based on two well-known annotation databases (SIFT and PolyPhen-2) that were provided by the Pan-Cancer Atlas, allowed us to focus on those mutations with a biological impact. Four cohorts (GBM, PRAD, SKCM, and TGCT) did not meet the sample size criteria for conducting groupwise association tests, because only one population was available. The top associated

doi:10.6342/NTU202300922

predisposing genes and the numbers of their significant variants varied widely across races (Figure 2-4). The top predisposing HRD genes which were common to all three racial populations were *TP53, RIF1*, and *SMG1*. Some genes were shared only between two populations; for example, *ATM, BRCA2, POLE*, and *TOP2B* were found in both 'White' and 'Asian' populations, whereas *PTEN* and *EGFG* were observed in both the 'White' and 'African American/Black' populations. For genes with the highest variant counts in the 'White' population, we observed similar mutation levels in other populations. Nevertheless, cancer type-specific differences were apparent. For instance, in BRCA and HNSC, a highly mutated *TP53* was observed in all three populations, whereas in ESCA and PAAD, such an event was only found in the 'White' and 'Asian' populations. Together, these results highlight the race-based genetic patterns of HRD-related genes that occur in multiple cancers, which can be used to help elucidate customized targeted therapy for each population.

doi:10.6342/NTU202300922

2.5 Discussion

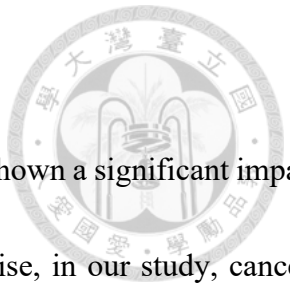
Here we report one of the most extensive multi-race investigations of HRD-predisposing genes, encompassing 7241 samples across 24 cancer types. By using a combination of bioinformatics tools and statistic methods in systematic pan-cancer analysis, we revealed novel insights into race-specific prognostic/therapeutic targets in the HRR pathway with potentially important clinical relevance.

Genome-wide HRD scores have been widely used as a gold standard to evaluate a sample's HRD status [86, 87], and current computational tools were mainly designed to simply count the number of DNA structural changes [78, 88]. Such counting results rely heavily on the selected gene lists and the number of genes in that list. Accordingly, our results showed that genome-wide HRD scores were generally higher than the scores that were defined by different gene lists. However, a larger number of genes in the gene list did not guarantee the detection of HRD events that were included in the calculation; for example, the scores that were defined by 'HR_PARP_132' were lower than those defined by 'DDRD_assay_42' across cancers, reiterating that a good selection of HRD-related genes is important to evaluate the HRD status. In addition, the distribution of HRD scores that were defined by 'DDR_276' across cancers showed a good reflection of the genome-wide pattern, but the cut-off value that assigned HRD status might need to be adjusted when using this geneset. Still, these results demonstrated that the 'DDR_276' geneset is relatively comprehensive and representative for describing HRD status.

doi:10.6342/NTU202300922

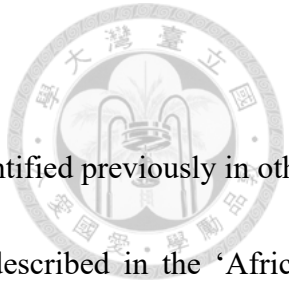
One previous study has demonstrated the synergistic effects between the number of DNA mutations and changes in the copy number [73]. Consistent with this study, our results additionally have revealed that their predictive abilities were relatively similar across many cancers, although HRD scores were associated with a slightly better prediction in terms of overall survival. These findings underscored that, to some extent, different degrees of DNA changes that are related to HRD were correlated in terms of not only their quantity, but also the patient's prognosis. The racial difference in the HRD scores and TMB values has been systematically investigated in previous studies [60]; however, few of them have considered the Asian population in their research. Hence, in this study, we included as many Asian samples as possible in each cancer cohort to compare these genetic features among racial populations. Our results firstly demonstrated that, in several cancer cohorts, both genetic indicators were significantly different between the 'White' and 'Asian' populations. This was supported by previous studies using other cancer cohorts [89]. Similar to a previous report, such a difference between the 'African American/Black' and 'Asian' populations was also observed in HNSC [60]. Intriguingly, genome-wide HRD scores showed that racial differences exist between the 'White' population and both other populations at a pan-cancer level. Collectively, these results suggest that considering racial differences, including Asian samples, for a race-based description of HRD status may provide valuable information when defining prognostic groups.

doi:10.6342/NTU202300922



Genomic instability in association with an increased HRD scores has shown a significant impact on the loss of TP53 function across multiple cancers [73, 90, 91]. Likewise, in our study, cancer-associated TP53 mutations were observed in 15 TCGA cancer types, and such mutations were not associated with the race variable, so this was used as the baseline of our groupwise association tests to support the reliability of the rest of the findings. Our analysis of groupwise association identified mutations in *EGFR* as being specific to the 'African American/Black' and 'White' populations in BRCA; EGFR is widely used in clinical drug targeting therapy [92]. Mutations in *TOP2B*, which is related to the recruitment of DNA double-strand break repair proteins [93], were exclusive to the 'Asian' and 'White' populations in PAAD. Additionally, *ATM*, which is involved in DNA damage response pathways [94], was highly mutated in Asian COAD patients. To aid the interpretation of the identified race-specific variants, evidence of pathogenic tendency from public databases provided further support. The majority of the most significant variants that were identified from the association test were likely disease-causing (Figure S2-7). Additionally, we were able to uncover many race-specific variants that are not currently presented in REVEL or ClinVar. Several associations of significant HRD-predisposing genes and cancer types in specific populations were previously reported for the same population (Table 2-4,S2-3), for example, *BAP1* with BRCA in the 'White' and 'African American/Black' populations [95] and *ATM* with STAD in the 'White' and 'Asian' populations [96, 97]. The association of *BLM* with BRCA in the 'African American/Black' and the

doi:10.6342/NTU202300922



‘White’ populations was described for the first time here, but had been identified previously in other populations [98]. While the association of *ERCC5* with OV was first described in the ‘African American/Black’ population here, *BLM* was previously found to be associated with OV in the ‘White’ population [99]. Intriguingly, the association of *MSH6* with STAD was also first identified in the ‘White’ and ’Asian’ populations, while *PTEN* was newly found to be associated with STAD in the ‘African American/Black’ population. However, such associations have been described in other populations [100, 101]. These findings, including novel associations, were further supported by older studies that evaluated individual HRD predisposition genes across populations. Overall, the knowledge of different HRD-predisposing genes and their prevalence among populations suggests the importance of incorporating race-specific interpretations into the detection of HRD for achieving the ultimate goal of personalized medicine: a tailored disease diagnosis and intervention based on an individual’s unique HRD pattern.

Leveraging the Pan-Cancer Atlas data, we found multiple significant HRD-predisposing genes for the three populations, yet there was a lack of many cancer cohorts with a sufficient sample size for the ‘Asian’ and ‘African American/Black’ populations. Even when adjusting for confounding factors in statistical tests, such small cancer cohorts likely generated false negatives. To improve the statistical power, we only included cancer–race groups that contained at least 10 samples and applied an outlier approach to stringently filter the most influential variants after the groupwise association

doi:10.6342/NTU202300922

test [102]. Sometimes the existence of more variants tended to increase the association in smaller cohorts. It is also necessary to be cautious when interpreting these variants in the context of previous reports. In addition, a racial difference in the sequencing data might affect the reliability of these findings [103]. Furthermore, evidence from population-based databases or public databases that were designed specifically for evaluating the pathogenic tendency of each variant will be needed to provide further validation. The analysis of Pan-Cancer Atlas data has potentially reached saturation in studying the racial differences in genomic features, due to the limited samples of non-White races that constitute a considerable fraction of the US population. Future cancer genomic studies should focus on racial differences in these genetic features in terms of their quantity and prognostic prediction.

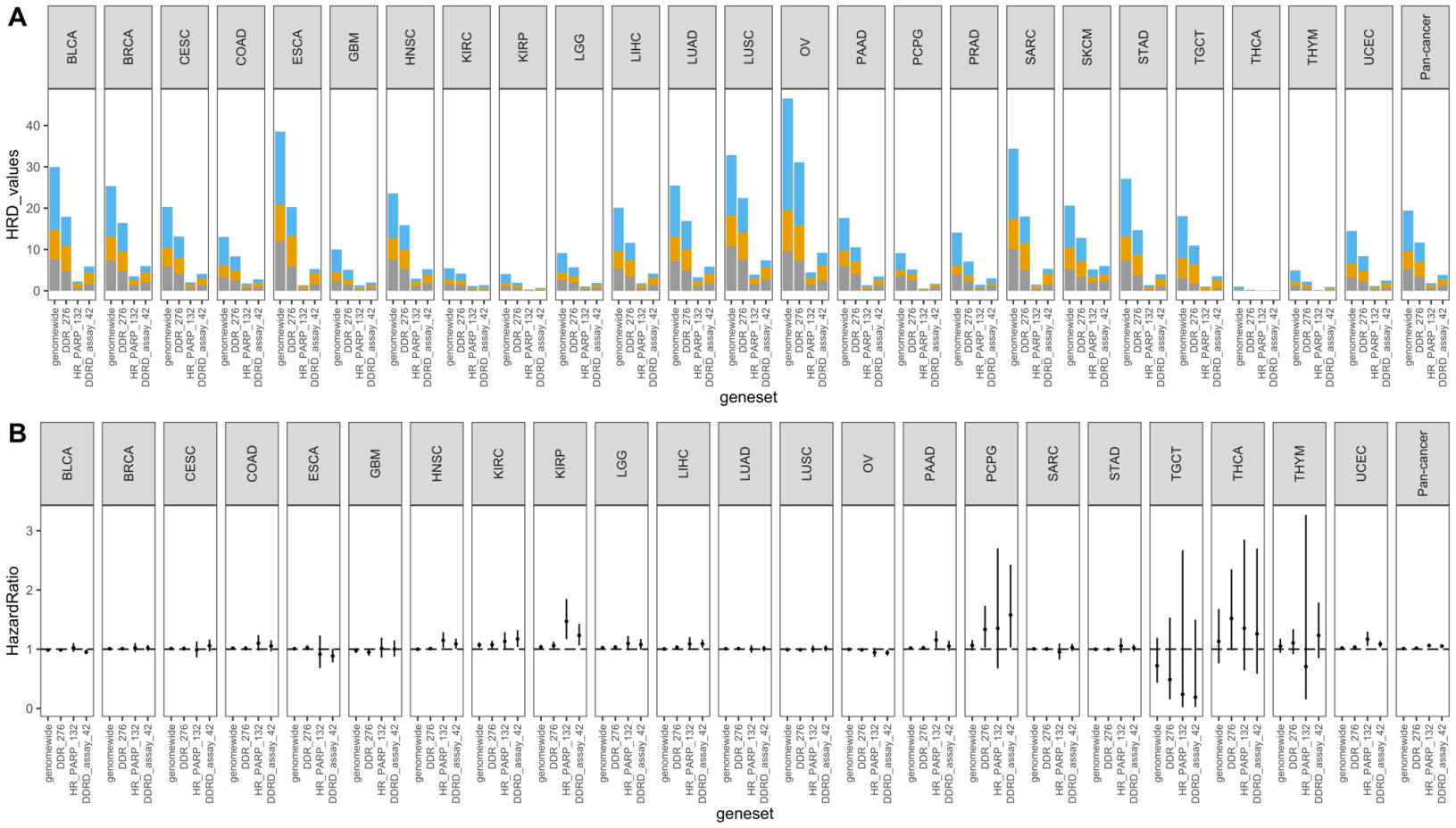
doi:10.6342/NTU202300922

2.6 Conclusions

In summary, we identified race-specific predisposing HRD genes and variants contributing to different cancer types. Our analysis of HRD events confirmed the limitations of HRD calculations that are defined by different genesets. Their syngeneic effects and racial differences between them were observed, especially for Asian populations. These results collectively reinforce the importance of considering differences in race when determining the definition, detection, and prognostic value of HRD events. Future studies in larger population-based cohorts are warranted and are prerequisite to conducting HRD-directed precision medicine in patients with distinct genetic backgrounds.

doi:10.6342/NTU202300922

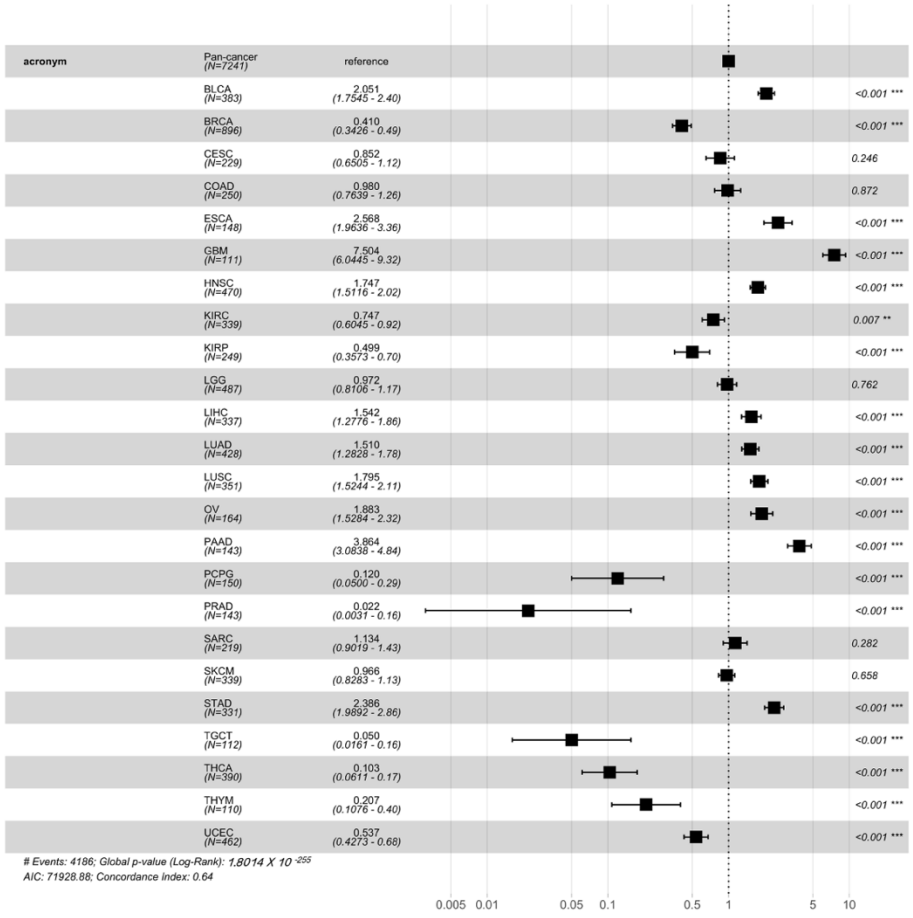
2.7 Figures

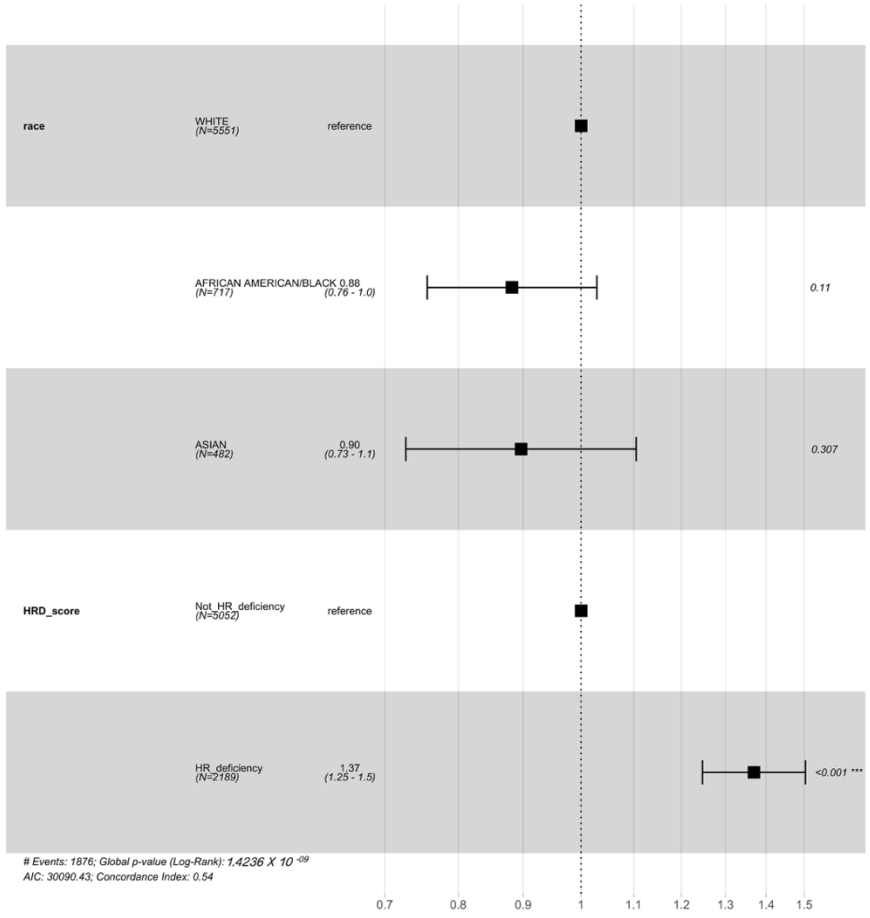


doi:10.6342/NTU202300922

Figure 2-1. The distribution of HRD scores defined by different genesets were varied in relation to their prognostic prediction. (A) Stacked bar plot for the distribution of HRD scores. Total scores were the sum of HRD-LOH (gray), HRD-TAI (orange), and HRD-LST (blue) across the different genesets for 24 individual cancer types and pan-cancer. (B) Forest plots of the association between the HRD score and overall survival. Results are shown for 24 cancer types and pan-cancer with valid outcomes data. Hazard ratios and 95% confidence intervals are shown by the dot and the line, respectively. HRD: homologous recombination deficiency; LOH: loss of heterozygosity; TAI: telomeric allelic imbalance; LST: large-scale state transitions.

doi:10.6342/NTU202300922





doi:10.6342/NTU202300922

Figure 2-2. Forest plot of the association between variable clinical characteristics (cancer type, race, and dichotomized HRD scores) and survival. HRD scores were defined at the genome-wide level. The hazard ratios (dots) and their 95% confident intervals (lines) were estimated via a Cox proportional hazards analysis." ***" represents the statistical significance.

doi:10.6342/NTU202300922

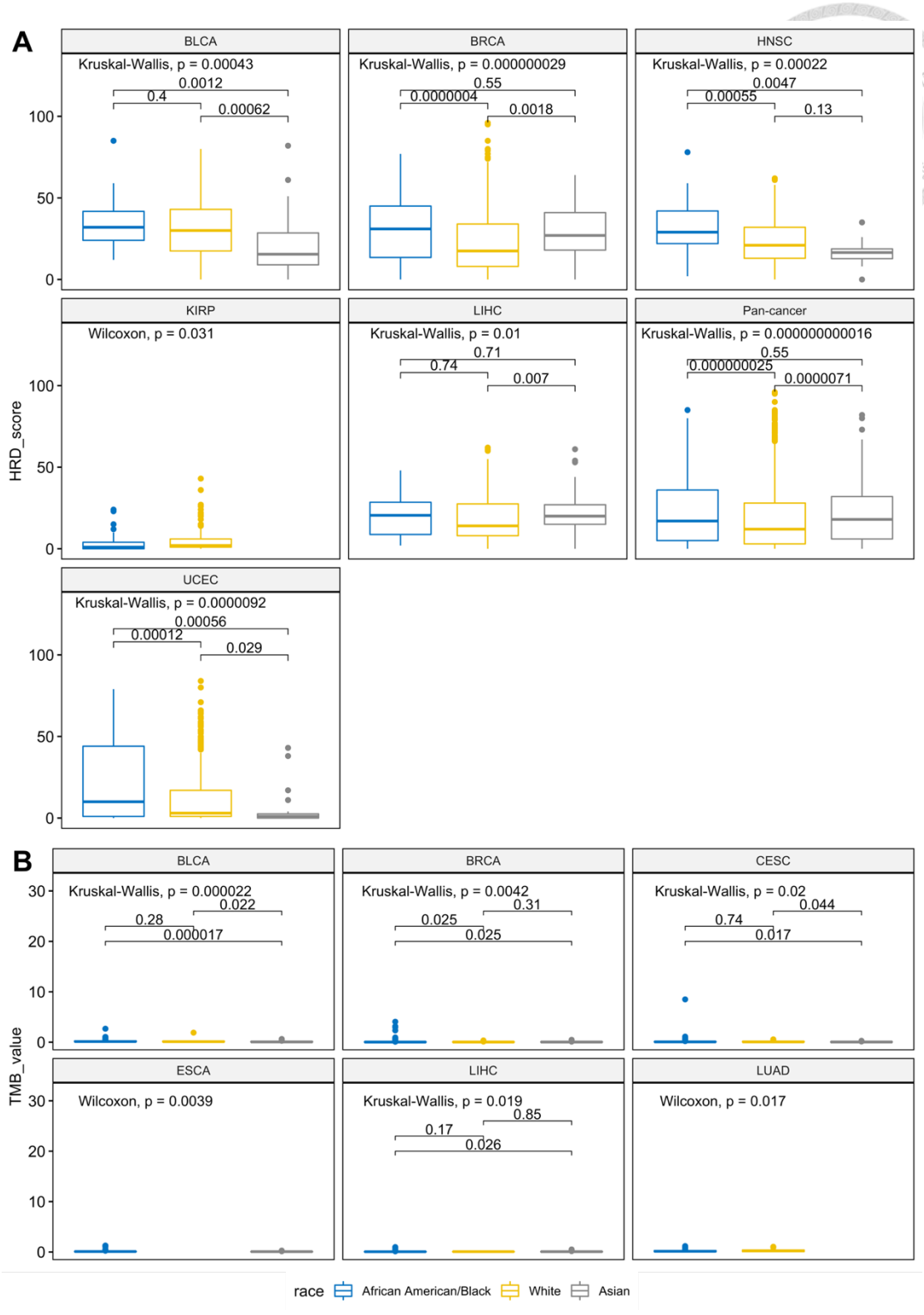


Figure 2-3. The significant interpopulation differences in the HRD scores and TMB among cancer types and pan-cancer. (A) Genome-wide HRD scores of all cancer types are stratified by racial population. (B) Global TMB values of all cancer

doi:10.6342/NTU202300922

types are stratified by racial population. The Wilcoxon rank-sum test or Kruskal-Wallis test *p* values in each cancer are displayed at the top of each plot. The groupwise Wilcoxon rank-sum test with false discovery rate adjustment *p* values are shown above the bracket for each race-specific comparison.

doi:10.6342/NTU202300922

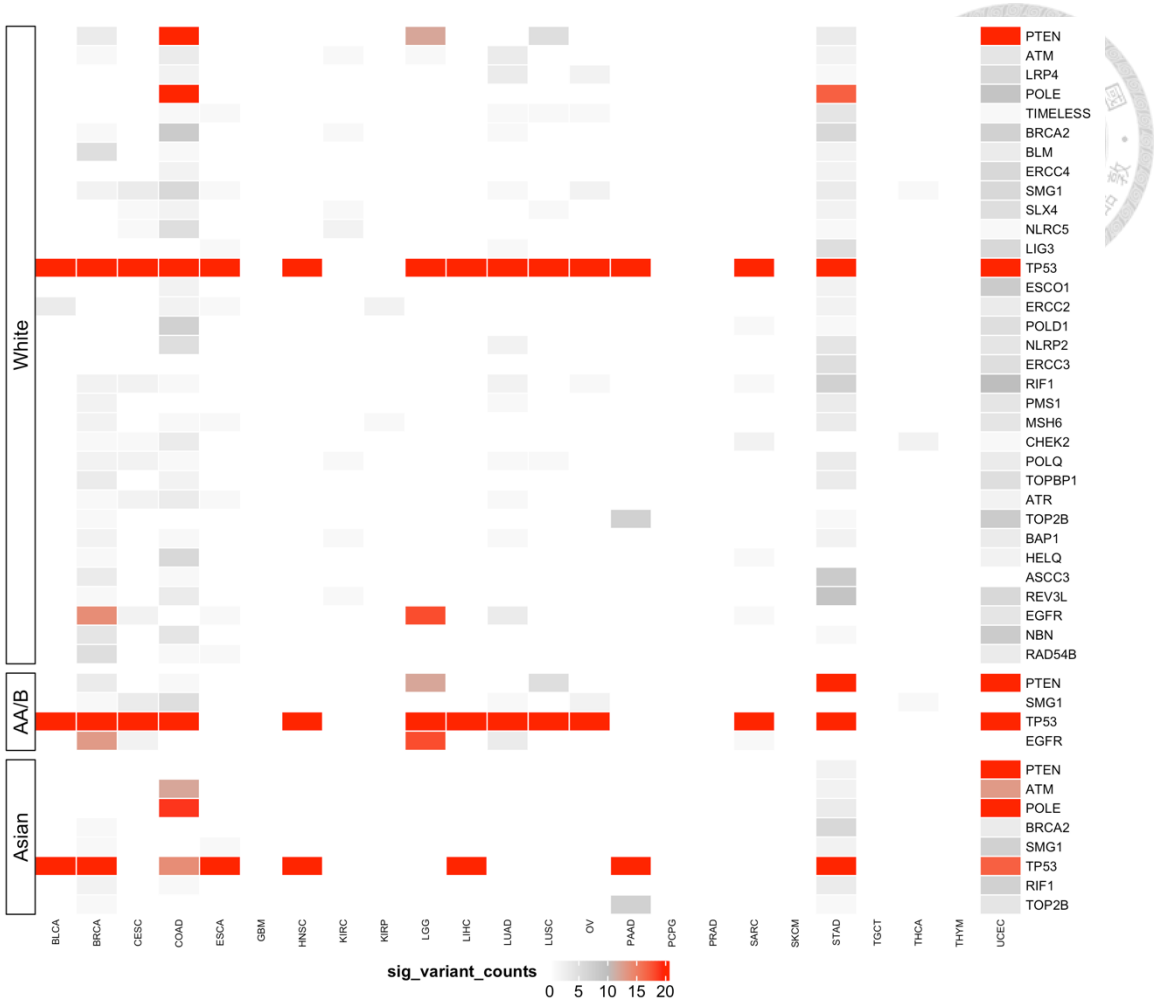


Figure 2-4. HRD-predisposing genes across 7241 TCGA cases across cancers in the 'White', 'African American/Black', and 'Asian' populations. Race-specific cancer-gene pairs containing HRD-predisposing variants as identified by SKAT-O and an outlier approach. The color scale represents the number of significant variants of predisposing genes within that cancer cohort, and only the genes with more than 10 variants summed in all cancer types are presented. AA/B stands for African American/Black.

doi:10.6342/NTU202300922

2.8 Tables

Table 2-1. Summary of demographic distribution of the Pan-Cancer Atlas.

	Asian	African American/Black	White	Total
BLCA	42	22	319	383
BRCA	59	159	678	896
CESC	18	27	184	229
COAD	11	55	184	250
ESCA	44	0	104	148
GBM	0	0	111	111
HNSC	10	46	414	470
KIRC	0	52	287	339
KIRP	0	59	190	249
LGG	0	21	466	487
LIHC	150	16	171	337
LUAD	0	52	376	428
LUSC	0	26	325	351
OV	0	18	146	164
PAAD	11	0	132	143
PCPG	0	19	131	150
PRAD	0	0	143	143
SARC	0	17	202	219
SKCM	0	0	339	339
STAD	75	11	245	331
TGCT	0	0	112	112
THCA	49	26	315	390
THYM	12	0	98	110
UCEC	20	101	341	462
Total	**501 (6.92%)**	**727 (10.04%)**	**6013 (83.04%)**	**7241 (100%)**

doi:10.6342/NTU202300922

Table 2-2. Summary of cancer types [1] with significant correlations ($p < 0.001$) between the quantity of DNA mutations (global TMB value) and the changes in the copy number (genome-wide HRD scores).

Positive Correlation	Inverse Correlation
OV ($R = 0.31$)	COAD ($R = -0.24$)
LUSC ($R = 0.33$)	UCEC ($R = -0.32$)
BLCA ($R = 0.38$)	
PAAD ($R = 0.54$)	
LUAD ($R = 0.49$)	
SARC ($R = 0.5$)	
BRCA ($R = 0.54$)	
THYM ($R = 0.32$)	
KIRC ($R = 0.25$)	
HNSC ($R = 0.23$)	
PRAD ($R = 0.39$)	
LGG ($R = 0.34$)	
PCPG ($R = 0.27$)	
Pan-cancer ($R = 0.26$)	

[1] A full list of TCGA cancer type abbreviations can be found in the Genomic Data Commons (https://gdc.cancer.gov/resources-tcga-users/tcga-code-tables/tcga-study-abbreviations(accessed on 02/06/2021)).

doi:10.6342/NTU202300922

Table 2-3. Summary of the number of significant variants identified by SKAT-O and the RIFT R package.

	White	African American/Black	Asian
BLCA	18	51	11
BRCA	135	97	56
CESC	48	41	2
COAD	200	9	160
ESCA	52	0	52
HNSC	13	20	14
KIRC	27	27	0
KIRP	9	9	0
LGG	17	17	0
LIHC	6	13	4
LUAD	118	118	0
LUSC	52	52	0
OV	18	18	0
PAAD	57	0	57
SARC	25	25	0
STAD	206	20	191
THCA	10	3	7
THYM	2	0	2
UCEC	891	43	213

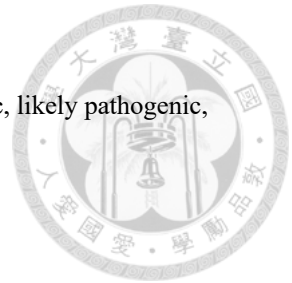
doi:10.6342/NTU202300922

Table 2-4. Several significant HRD-predisposing gene associations.

Cancer	Gene	Population	Previous Reports		Identified Variant (GRCh38)	REVEL Score [1]	Clinvar [2]
			Same Population	Other Population			
BRCA	*BAP1*	White African American/Black	Shahriyari et al. (2019)		3:52442072-T/C	0.829	pathogenic
	BLM	White African American/Black		Cybulski et al. (2019)	15:91312388-C/T	0.677	uncertain significance
OV	*ERCC5*	White African American/Black	Doherty et al. (2011)	Doherty et al. (2011)	3:103525633-G/T	0.939	pathogenic
STAD	*ATM*	White Asian	Helgason et al. (2015) Cai et al. (2015)		11:108141997-C/T	0.739	uncertain significance
	MSH6	White Asian		Karpińska-Kaczmarczyk et al. (2016)	2:48028049-G/A	0.857	uncertain significance
	PTEN	African American/Black		Nemtsova et al. (2020)	10:89692905-G/A	0.976	likely pathogenic, pathogenic

doi:10.6342/NTU202300922

[1] REVEL scores above 0.5 represent likely disease-causing variants. [2] Five levels of variants defined in ClinVar database are as follows: pathogenic, likely pathogenic, uncertain significance, likely benign, and benign.



doi:10.6342/NTU202300922

Chapter 3. A risk prediction model of gene signatures in ovarian cancer through bagging of GA-XGBoost models

3.1 Abstract

Introduction: Ovarian cancer (OC) is one of the most frequent gynecologic cancers among women, and high-accuracy risk prediction techniques are essential to effectively select the best intervention strategies and clinical management for OC patients at different risk levels. Current risk prediction models used in OC have low sensitivity, and few of them are able to identify OC patients at high risk of mortality, which would both optimize the treatment of high-risk patients and prevent unnecessary medical intervention in those at low risk.

Objectives: To this end, we have developed a bagging-based algorithm with GA-XGBoost models that predicts the risk of death from OC using gene expression profiles.

Methods: Four gene expression datasets from public sources were used as training (n=1) or validation (n=3) sets. The performance of our proposed algorithm was compared with fine-tuning and other existing methods. Moreover, the biological function of selected genetic features was further interpreted, and the response to a panel of approved drugs was predicted for different risk levels.

Results: The proposed algorithm showed good sensitivity (74% to 100%) in the validation sets, compared with two simple models whose sensitivity only reached 47% and 60%. The prognostic gene signature used in this study was highly connected to *AKT*, a key component of the PI3K/AKT/mTOR signaling pathway, which influences the tumorigenesis, proliferation, and progression of OC.

Conclusion: These findings demonstrated an improvement in the sensitivity of risk classification of OC patients with our risk prediction models compared with other methods. Ongoing effort is needed to validate the outcomes of this approach for precise clinical treatment.

doi:10.6342/NTU202300922

Keywords: Ovarian cancer, risk prediction, gene expression, machine learning, GA-XGBoost, bagging algorithm

doi:10.6342/NTU202300922

3.2 Introduction

Ovarian cancer (OC) is the seventh most common malignancy, and it causes the eighth highest mortality rate of all cancer types worldwide; 295,414 new cases were diagnosed, and 184,799 patients died of this disease in 2018 [104]. OC was initially divided into epithelial and non-epithelial types, but some recent literature has indicated that epithelial OC also has histological subtypes including high-grade serous (>70%), endometrioid (10%), clear cell (10%), mucinous (<5%), and low-grade serous (<5%) [105]. These histologically distinct tumor types have shown a wide range of different prognoses. For example, epithelial tumors classified as low-grade serous, endometrioid, mucinous, or clear cell usually present themselves at an early stage and have a good prognosis, while the high-grade serous type mostly presents itself at an advanced stage with a poor prognosis [106]. It has been revealed that the five-year survival of OC patients diagnosed at an early stage is about 90%, whereas that of patients at a late stage is less than 30% after surgery [107, 108]. However, most OC patients are diagnosed at the advanced stage due to the asymptomatic features of the early stage. As a result, more sophisticated research into both the diagnostic and predictive aspects of OC is urgently needed.

According to a clinical guideline from the National Comprehensive Cancer Network (NCCN), whether OC patients should receive post-surgery chemotherapy mainly depends on their clinical features, such as tumor stage and tumor grade [109]. In general, it is recommended for OC patients at stages II to IV to receive chemotherapy after surgery. OC patients at stages IA or IB with grade 1 tumors are recommended to have follow-up tests after surgery, while those with grade 2 tumors are suggested for either follow-up with the regular investigative tests or post-surgery chemotherapy. However, there is still some controversy regarding which OC patients, especially advanced-stage patients, will obtain the most clinical benefit from post-surgery adjuvant chemotherapy. National cancer statistics from the Taiwan Cancer Registry

doi:10.6342/NTU202300922

reported that 72.56% of OC patients had received post-surgery chemotherapy in 2016, and 60.65% of these patients were diagnosed at stage I [110], revealing that decisions regarding medical intervention do not always follow the NCCN guideline. To date, there are no entirely acceptable criteria to guide treatment decisions, especially in terms of post-surgery treatment in patients with low risk.

Due to the complexity and heterogeneity of cancer, gene expression profiling can provide biological insights into cancer prognosis, over and above the use of clinical features [111]. Hence, more and more cancer-related studies are taking these molecular indicators into account [112-115]. For example, a commercially available 70-gene signature test (MammaPrint) has been able to distinguish breast cancer patients at high versus low risk of recurrence, based on their 5- or 10-year recurrence rate [116], which can assist with clinical decision-making for early-stage patients [117]. Oncotype DX is another example of a genomic test that uses a clinically validated set of 21 genes to assess the risk of breast cancer recurrence [118]. Nevertheless, this kind of test may only be applicable to a particular set of patients (e.g., those with a particular hormone expression pattern) and may not fully explain the eventual clinical outcome, suggesting that unbiased approaches with a full prognostic gene signature are needed for accurate cancer risk assessment [119].

Prior studies on OC [120, 121] have proposed models for predicting survival and have discussed hazard ratios (HRs) based on gene expression data. However, very few classifiers have been built to predict high risk of mortality in OC patients with high sensitivity. Several recent studies have extensively investigated robust machine learning-based methods for the identification of prognostic molecules in breast cancer, which shares many standard pathological features with OC [122-125]. However, few of these novel approaches have been applied to OC [126]. Therefore, the purpose of our study was to incorporate a bagging-based algorithm with GA-XGboost models into a comprehensive risk prediction model, using

doi:10.6342/NTU202300922

prognosis-related genes to arrive at a clinically meaningful classification of OC patients. The accuracy of our prediction model was evaluated in comparison with that of other conventional methods. The primary objective of this study was to effectively identify high-risk OC patients, with the long-term goal of reducing unnecessary preventive treatments in low-risk patients.

doi:10.6342/NTU202300922

3.3 Materials and Methods

An overview of the workflow is illustrated in Fig. 3-1. With the aim of identifying high-risk patients with OC, we constructed a complex set of procedures, including data preprocessing, dimensional reduction, a bagging-based algorithm with GA-XGBoost models, and external validation, to construct a comprehensive prediction model.

3.3.1 Datasets and data preprocessing

For the evaluation of the predictive model, four gene expression datasets (GSE26193 [127, 128], GSE30161 [129], GSE19829 [130], GSE63885 [131]) that had OC outcomes were collected in this work (Table 3-1). All datasets used were from the publicly available Gene Expression omnibus (GEO) database (https://www.ncbi.nlm.nih.gov/geo/), and the platform used for these datasets was the Affymetrix Human Genome U133 Plus 2.0 Array (GPL570). In each dataset, patients who lacked 3-year follow-up information were excluded. Then, the GSE26193 dataset (n = 106) was divided into a training set and a validation set for building the prediction model. The remaining datasets, including GSE63885 (n = 73), GSE30161 (n = 50) and GSE19829 (n = 23), were used for external validation. Based on the clinical data, we further stratified patients in each dataset into two groups. Two previous studies have suggested that around 50% of OC patients suffer from recurrence within 1.5 to 2 years [132, 133]; hence, we set three years as a cut-off to ensure most patients with recurrence were included in the following analyses. The low-risk group was defined as patients with overall survival of three years or more, whereas the high-risk group was defined as patients with overall survival less than three years.

For the minimization of batch effects among different datasets, raw intensity-level data merged from all datasets were first normalized using robust multichip averaging (RMA) and then by quantile normalization with default parameters using the *affy* (version 1.62.0) [134] and *preprocessCore* (version 1.46.0) [135] R packages.

doi:10.6342/NTU202300922

3.3.2 Variable selection of gene expression patterns for dimension reduction

For feature reduction of the training dataset from GSE26193, the t-test and fold change method was used as a criterion to identify differentially expressed genes between low- and high-risk OC patients. An absolute $\log_2$ fold change ≥ 2 and a P-value < 0.05 were set as the cut-off values to screen for these probes.

3.3.3 XGBoost

The XGBoost (extreme gradient boosting) algorithm is a learning framework based on gradient boosted decision trees [136]. Compared with traditional boosting tree models implemented with only first order derivative information [137], this boosting model uses a second-order Taylor expansion for calculating the loss function and its scalability to enhance not only computational speed but also the model performance. Therefore, XGBoost was used for risk prediction classifiers for OC patients in this paper.

3.3.4 Genetic algorithm for the most suitable combination of selected gene expression patterns

Genetic algorithms (GAs) have been designed to replicate the concept of natural selection by searching for an available combination of gene expression profiling probes which will produce a predictive model with superior performance [138, 139]. Therefore, in terms of feature selection, the XGBoost algorithm could be further improved by using a GA, a process that we call GA-XGBoost. As shown in Fig. 3-2, a GA involves five main phases: initial population, fitness function, selection, crossover, and mutation. In the GA, genetic coding segments of a chromosome are represented using a string of zeros and ones. Therefore, in this study, in order to correspond to expression being either on or off, significantly expressed probes were denoted as 1, whereas the rest were assigned as 0 (Figure S3-1).

doi:10.6342/NTU202300922

First, we randomly sampled a combination of probes from those significant probes that were determined in the previous step to be a chromosome (i.e., a string of zeros and ones corresponding to the expression status of each gene expression profiling probe), and then repeated this procedure to generate a population of chromosomes defined as the first generation. Second, the fitness values (i.e., sensitivity and specificity in this study) of each chromosome was calculated by the fitness function (i.e., the XGBoost model), and only the ones with the highest fitness were retained in the next generation. In the roulette wheel selection, the wheel is divided according to the fitness values; that is, the fittest chromosome has the largest share, whereas the weakest chromosome has the smallest percentage (Figure S3-2A). The underlying assumption of this step is that the fitter chromosomes will tend to have a better chance of survival among the whole population, and then will mate to create the next generation. As a result, the fittest individuals will be stochastically selected from a particular population to form the next generation.

The chromosome showing the best fitness value (i.e., highest sensitivity) among the models (i.e., XGBoost model) from the first generation is either passed directly to the next generation, or crossover and mutation operators are performed to generate the next generation. In the crossover step, two parental chromosomes with the best fitness are selected from the original population, and a random threshold (for example, 20% of genetic information from parent chromosomes) is defined to determine the proportion of values within the chromosome that should be swapped to form two offspring chromosomes (Figure S3-2B). For emulating the dispersion of a mutation in a population, a proportion of the values (such as 10 %) in a chromosome should be flipped, which means if it is a zero it now becomes a one and vice versa (Figure S3-2C and S3-3).

Finally, the conditions for evaluating when the GA should be stopped are defined as follows:

doi:10.6342/NTU202300922

$$training\ sensitivity - validation\ sensitivity \leq 0.05$$
$$training\ specificity - validation\ specificity \leq 0.05 \quad (1)$$

The process of crossover and mutation will be repeated until these criteria are met or the final generation is reached (a predetermined number), unless the chromosomes with the best fitness of all generations meet the criteria and are outputted as a prediction model. By this process, the outputted model with the highest sensitivity for the classification of OC patients into risk response groups (high risk/low risk) is developed. The GA-XGBoost model was performed using R (version 3.5.2) and the *xgboost* R package (version 0.82.1) [140].

3.3.5 Bagging-based algorithm and external validation

To construct a robust bagging algorithm [141], GSE26193 was first divided into training and internal validation datasets, with a 2:1 split. Then, 70% of the training data was randomly selected to perform variable selection which was mentionedas described above, which generated 15 GA-XGBoost models for bagging. Those with a specificity < 0.3 were dropped, and based on each model's performance in both internal and external validation sets, the voting system was used to further identify OC patients with high risk.

3.3.6 Other existing methods

Other proposed methods can also be used in risk prediction based on gene expression values. Two traditional methods used in this study were least absolute shrinkage and selection operator (LASSO) regression [142] and forward stepwise logistic regression [143]. The performance evaluation was conducted by comparison of the predictive results, including accuracy, specificity, sensitivity, and F1-score, between GA-XGBoost and these two methods.

doi:10.6342/NTU202300922

3.3.7 Survival analysis

Through this bagging algorithm of GA-XGBoost models, the common differentially expressed genes from all models were identified for the classification of two risk groups. Survival analyses were performed by the *survival* R package [144], and Kaplan-Meier survival curves were plotted to compare whether those expression profiles could distinguish between high- and low-risk groups of OC patients in internal and external validation sets. A Cox proportional hazards model was also used to compare the difference between survival curves for different risk groups.

3.3.8 Drug prediction for the identification of effective drugs

To further identify potential drugs effectively targeting each risk group, the dataset GSE36133, including gene expression profiles and a drug sensitivity indicator represented by activity area in the Cancer Cell Line Encyclopedia (CCLE) project, was used [42]. This project collected the drug response of 44 OC cell lines exposed to 24 commercially available drugs. The values of the activity area quantify the drug responses of each cell line. For this analysis, the expression profiles of the 44 OC cell lines were used as the inputs of our model to identify their potential risk level (high or low). Then, the Wilcoxon rank-sum test was used to evaluate which drugs have a significant difference in the activity area between high- and low-risk groups.

3.3.9 Functional analysis

To understand the relationship between the respective differentially expressed genes obtained from this bagging algorithm of GA-XGBoost models and OC, we also used the Ingenuity® Pathway Analysis (IPA®) software program (QIAGEN Inc.,

doi:10.6342/NTU202300922

https://www.qiagenbio-informatics.com/products/ingenuity-pathway-analysis) to identify their potential functional role in biological processes.

3.3.10 Statistical analysis

Categorical variables, such as stages, grades, clinical signatures, and subtypes, were reported as counts and percentages. Between-group comparisons (i.e., high- and low-risk groups) were performed by a Fisher's exact test. A P-value below 0.05 was defined as statistically significant. The analyses were conducted using R (version 3.5.2).

doi:10.6342/NTU202300922

3.4 Results

3.4.1 Clinical characteristics for the training set

Table 2 presents the clinical characteristics of the training set (GSE26193; n = 106). The majority of samples in this dataset were from patients with stage III and grade III, constituting 55.7% and 63.2% of the samples, respectively. However, there were no significant differences, in terms of stage (P = 0.2828) and grade (P = 0.2665), between the two risk groups. Similarly, clinical signatures (P = 0.2515) and subtypes (P = 0.8113) also showed no difference between the two groups. Therefore, these clinical variables do not account for the risk of death from OC.

3.4.2 Parameter optimization

After dimension reduction of gene expression features, 507 differentially expressed probes (i.e., 406 genes) were extracted and used to inform the bagging-based algorithm that uses GA-XGBoost models. The bagging results using an internal validation set, three individual external validation sets, and a combined external validation set for different combinations of parameters are displayed in Table 3. To determine the best combination of parameters for our model, it is possible to fix all the settings except one and then decide which one has the strongest effect on model performance. The optimum combination of parameters has moderate specificity when the maximum sensitivity is reached, and these outcomes need to be supported by at least two external validation sets. First, we adjusted the number of GA-XGBoost models used in the bagging algorithm, and it can be seen that using 15 models showed the best performance, in terms of both specificity and sensitivity. Using more than 15 models may cause overfitting, while it may not be stable due to the small sample size when the number of models is less than 15. Second, the proportion of the GSE26193 dataset used for training (50%, 70%, or 90%) was adjusted, and 70% was optimal. Fewer training samples (50%) may

doi:10.6342/NTU202300922

generate an unstable bagging algorithm; on the contrary, a larger sample size may have an overfitting issue due to less variation among the models. Then, four tunable parameters used in GA-XGBoost were adjusted: the number of chromosomes in a generation, the number of generations, the mutation rate, and the number of tree layers. It can be observed that the combination of 300 chromosomes, 500 generations, a 50% mutation rate, and three tree layers are the best conditions. Lastly, imposing a requirement for high specificity (> 70%), led to a less robust model with extremely low sensitivity and accuracy in many validation sets; for example, in GSE63885, the specificity, sensitivity, and accuracy were 0.784, 0.417, and 0.603, respectively.

3.4.3 Validation of the bagging-based algorithm that uses GA-XGBoost models

Table 4 presents the prediction ability of the 15 GA-XGBoost models used in the bagging algorithm using the internal validation set (GSE26193; n = 35). The range of the number of selected gene expression patterns among these models was 24 to 150 based on the 15 cycles of variable selections 15 times using fold-change and P-value as the cut-offs after randomly selecting 70 % of the training dataset from entire data, and the sensitivity of each model was over 0.8. A patient was considered as “high risk” when there were over seven models supporting this. As shown in Table 3-5, the bagging algorithm also maintains high sensitivity (100%) and specificity (52.4%) in the internal validation set. Among 35 patients in this validation set, 24 of them were predicted as high risk, and 11 were low risk. Kaplan-Meier survival analysis was also performed to determine the prognostic outcome, and the result indicated a significant difference (P = 0.0024) between high-risk and low-risk groups (Figure 3-3A). Notably, the survival time of the high-risk group decreased, while that of the low-risk group was maintained as time passed.

doi:10.6342/NTU202300922

In order to confirm that the high sensitivity of our bagging algorithm in predicting the risk level was not caused by model overfitting to the training set, we also tested the same bagging algorithm using a combined external validation set (GSE30161, GSE19829, and GSE63885; n = 146) (Table 3-5). The sensitivity and specificity values of the bagging algorithm in this combined set were 82.4% and 38.5%. The Kaplan-Meier survival analysis was performed after combining all external validation sets, displaying that there is a significant difference between the two risk groups (P = 0.014; Figure 3-3B). The individual external validation sets (GSE30161, n = 50; GSE19829, n = 23; GSE63885, n = 73) were also tested (Table S3-1). The sensitivity values of the bagging algorithm in these sets were 73.9%, 100%, and 83.3%, respectively, while the specificity values were 44.4%, 14.3% and 43.2%. The Kaplan-Meier survival analysis also showed a distinct difference in the survival time between the two groups in GSE63885 (P = 0.035), while the other two datasets (GSE30161, P = 0.2; GSE19829, P = 0.29) did not have a significant difference, likely due to the small sample size. The values of the HR and corresponding 95% confidence interval for each validation set were further visualized using forest plots, except for one individual external validation set (GSE19829) with an extremely large HR because no events happened in the low-risk group during the observed period (Figure 3-3C). The HR point estimates of GSE63885 and GSE30161 were 2.502 (1.037, 6.033) and 2.156 (1.154, 4.027). Overall, the pooled HR of these external validation sets was 2.161 (1.144, 4.082).

3.4.4 Performance comparison

To verify the necessity and effectiveness of constructing a highly sensitive prediction model using such complicated GA-XGBoost models in the bagging algorithm, we replaced the GA-XGBoost model with two simple models: forward stepwise logistic regression and LASSO regression. The performance of these two models in both internal and external validation sets

doi:10.6342/NTU202300922

is displayed in Table 3-5. It can be seen that the results of the GA-XGBoost model in both internal and external validation sets achieved higher sensitivity and accuracy than the other two models, showing that the GA-XGBoost model is superior for risk prediction using gene expression values.

3.4.5 Functional analysis

To identify the biological function associated with the differentially expressed probes, we uploaded our probe list to the Ingenuity® Pathway Analysis (IPA®) server. The top disease/function annotations were significantly enriched in female genital tract serous carcinoma (P = 2.33E-34), and 64 differentially expressed genes (DEGs) were involved (Table S3-2 and S3-3). Additionally, based on a network analysis, it is noteworthy that the top regulatory network constructed by the DEGs was mainly regulated by the hub gene, *AKT*, which is implicated in many cancers (Figure 4).

3.4.6 Effective drug prediction

For the identification of drugs effective in either high- or low-risk OC patients, the expression profiles and drug responses of 44 OC cell lines in the CCLE project were used. Through our bagging-based algorithm with GA-XGBoost models, 35 cell lines were defined as high risk, whereas the others were classified as low risk. Regarding the drug response, however, only 17-AAG (17-N-allylamino-17-demethoxygeldanamycin/ Tanespimycin), an kind of antitumor antibiotics, and RAF265, a novel RAF/VEGFR2 inhibitor, had a slight difference in treatment efficacy at killing tumor cells between high-risk and low-risk cell lines, with P-values of 0.08 and 0.055, respectively (Table S3-4).

doi:10.6342/NTU202300922

3.5 Discussion

As previously mentioned, few risk prediction models can predict high-risk OC patients with excellent sensitivity, and most of these models are not machine learning-based approaches. Therefore, we coupled t-tests and 2-fold changes to select features that were fitted by a GA-XGBoost model within a bagging algorithm. This method exhibited high sensitivity and moderate specificity in identifying high-risk patients who qualify for chemotherapy. Also, the combined HR point estimate of external validation sets indicated that the selected predictors are effective to distinguish the two risk groups.

Although our bagging algorithm successfully showed a high sensitivity for detecting high-risk OC patients, the low specificity of 38.5% in the external validation sets inferred a low accuracy for identifying the low-risk groups. Yet, few studies have focused on risk prediction models using gene expression for OC, so it is not feasible to compare the performance of our method with other models. However, two prediction models for breast cancer were amenable to comparison. Naderi et al. [145] used a Cox-ranked classifier with a prognostic signature of 70 genes and found it to have sensitivities of 77% and 63% in two external datasets, suggesting it may tend to ignore some high-risk patients who need to take chemotherapy. Similarly, another breast cancer study [146] developed three predictive models with good sensitivities (0.97 to 1) but low specificity (30%), suggesting that the issue of low specificity in current risk prediction models remains a challenge in these female-specific cancers.

The specificity of the bagging algorithm was lower in the combined external validation sets than in the internal validation set, showing that some overfitting issues may exist in this approach. GAs themselves tend to overfit the training set, and unfortunately, there is no solution to this problem in GAs [147]. Overfitting may also arise from the complexity of the GA-XGBoost model [148] or from model diversity (Table S3-5), which limits the prediction performance [149]. We tried various strategies to avoid overfitting, including random sampling

doi:10.6342/NTU202300922

of the training set to increase the variety of each GA-XGBoost model, combining the GA with XGBoost via the shrinkage method [150], and comparison with forward stepwise logistic regression and LASSO regression. These methods produced an improved but still suboptimal prediction model, showing that the process of risk prediction is imperfect and iterative. Future research should balance the complexity and diversity of the prediction model with the performance of the bagging algorithm.

Regarding the biological evidence of significantly differentially expressed genes involved in our bagging algorithm, the network analysis from Ingenuity® Pathway Analysis (IPA®) revealed that the *AKT* (AKT serine/threonine kinase) gene is a hub for many of these genes. This gene is a crucial molecule in the PI3K/AKT/mTOR signaling pathway, which is vital in regulating cell proliferation, survival, and migration [151]. It has been reported that this pathway is frequently deregulated and associated with poor prognosis at advanced tumor stage in OC [152]; as a result, this pathway has become one of the famous anticancer targets in OC [153, 154]. Both *PAK* (P21 activated kinase) and *INHBA* (inhibin subunit beta A) showed direct interactions with AKT, but only the latter was presented in our real dataset. A recent study revealed that higher expression of *INHBA* was connected to higher risk of death in patients with late-stage OC; hence, it is a potential target by for blocking this gene to suppress tumor progression [155]. The IPA results also revealed that *AKT* has many indirect interactions with insulin-like growth factor binding protein family members (i.e., IGFBP-4 and IGFBP-5; based on IPA output), which can modulate insulin-like growth factors that have endocrine, autocrine, or paracrine functions [156]. *IGFBP-4* expression is elevated in the early tumor stage [157], and *IGFBP-5* is known to be a tumor suppressor by inhibiting expression of *AKT* [158]. In addition, several other genes connected to *AKT* in this network result also play important roles ion the prognosis of OC. For example, *GHR* (growth hormone receptor), including estrogen (ER) or progesterone receptors (PR), has been widely known their association with better

doi:10.6342/NTU202300922

survival outcomes [159]. Also, the potential of *VAV3* (vav guanine nucleotide exchange factor 3) overexpression in cancer stem cells being a biomarker for poor survival outcomes in OC has been proved [160]. Moreover, *PCSK5* is a member of proprotein convertases (PCs), and the increased expression of this protein family was related to poor survival outcomes in OC [161]. These findings suggest that the function of selected genes in this study is highly associated with the survival of OC patients.

Several factors may explain can lead to only the slight difference in the response of high-risk and low-risk cell lines to 17-AAG and RAF265. Firstly, only 55 of the 1457 cell lines (3.77%) in the CCLE data resources are ovary cell lines, illustrating that a small number of samples may produce biased performance estimates when performing cross-validation of such high- dimensional data but with a small sample size [162]. Secondly, although it has been reported that various types of OC, such as clear cell carcinoma (CCC), serous carcinoma (SC), and endometrioid and mucinous carcinoma (EM), showed different drug responses [163] but the cell lines in CCLE were not be classified into these subtypes, suggesting that the heterogeneity of the ovary cell lines may also affect the drug response results. Lastly, the expression patterns in tumor tissues and normal cell lines are still not the same [164], resulting in the so a model trained by tumor tissue samples may not be generalized to cell line samples.

Some drawbacks exist in this study. First, the insufficient number of samples may influence the accuracy of this method, and expanding the sample size of OC patients is still necessary. Second, unmeasured and residual confounders may exist that affect the results. Finally, quantile normalization did not remove the batch effect across these datasets. Moreover, the feature combinations from the GA were different even when the same parameters were set. Because the risk prediction approach we present here is not comprehensive enough to extend into other cancers, further research is required to fully develop a risk prediction model that considers cancer heterogeneity, cancer subtypes, and functional pathways. In the future, we

doi:10.6342/NTU202300922

will work to apply our method to other data sources, such as gene expression profiles from next-generation sequencing data in OC.

doi:10.6342/NTU202300922

3.5 Conclusion

Predictive modeling using gene signatures for the early identification of high-risk individuals has shed light on personalized medicine, especially in stratified prevention strategies and clinical management. Considering that there are few risk-prediction models of OC using gene expression, we developed a bagging-based algorithm with GA-XGBoost models to predict the mortality risk of OC patients based on their gene expression patterns. Our method accurately predicted high-risk OC patients and has the potential to reduce unnecessary healthcare for those with low risk. However, several limitations still need to be addressed. Therefore, in the future, further investigations are necessary and warranted to validate the outcomes before clinical application.

doi:10.6342/NTU202300922

3.6 Figures

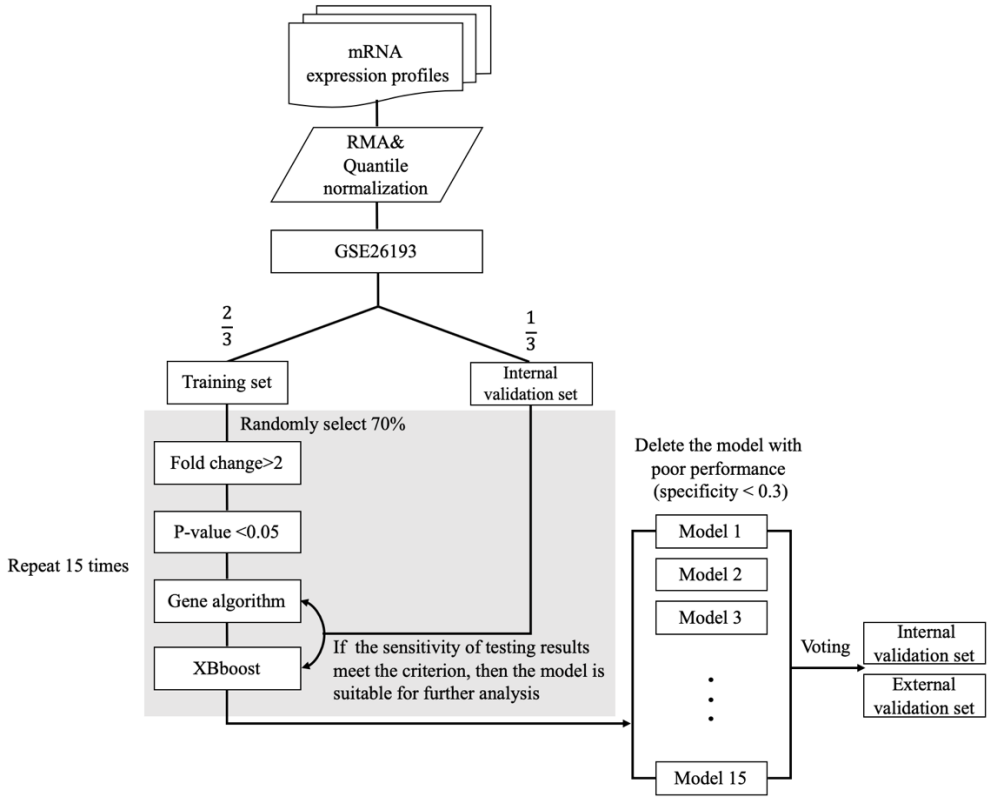


Figure 3-1. The pipeline of our bagging-based algorithm with GA-XGBoost models.

doi:10.6342/NTU202300922

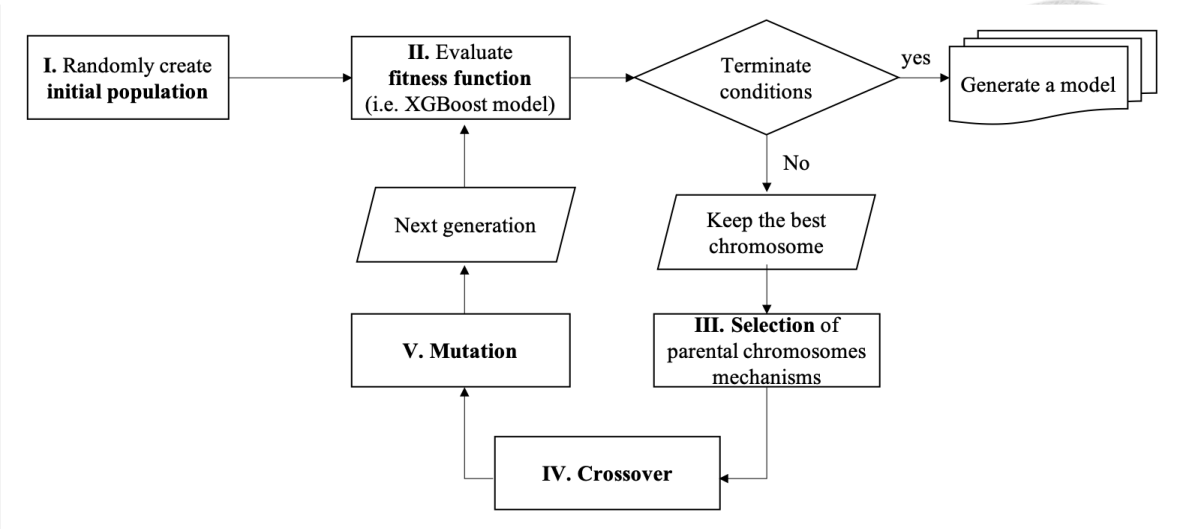


Figure 3-2. Genetic Algorithm flowchart. GA algorithm includes five main steps: initial population, fitness function, selection, crossover, and mutation.

doi:10.6342/NTU202300922

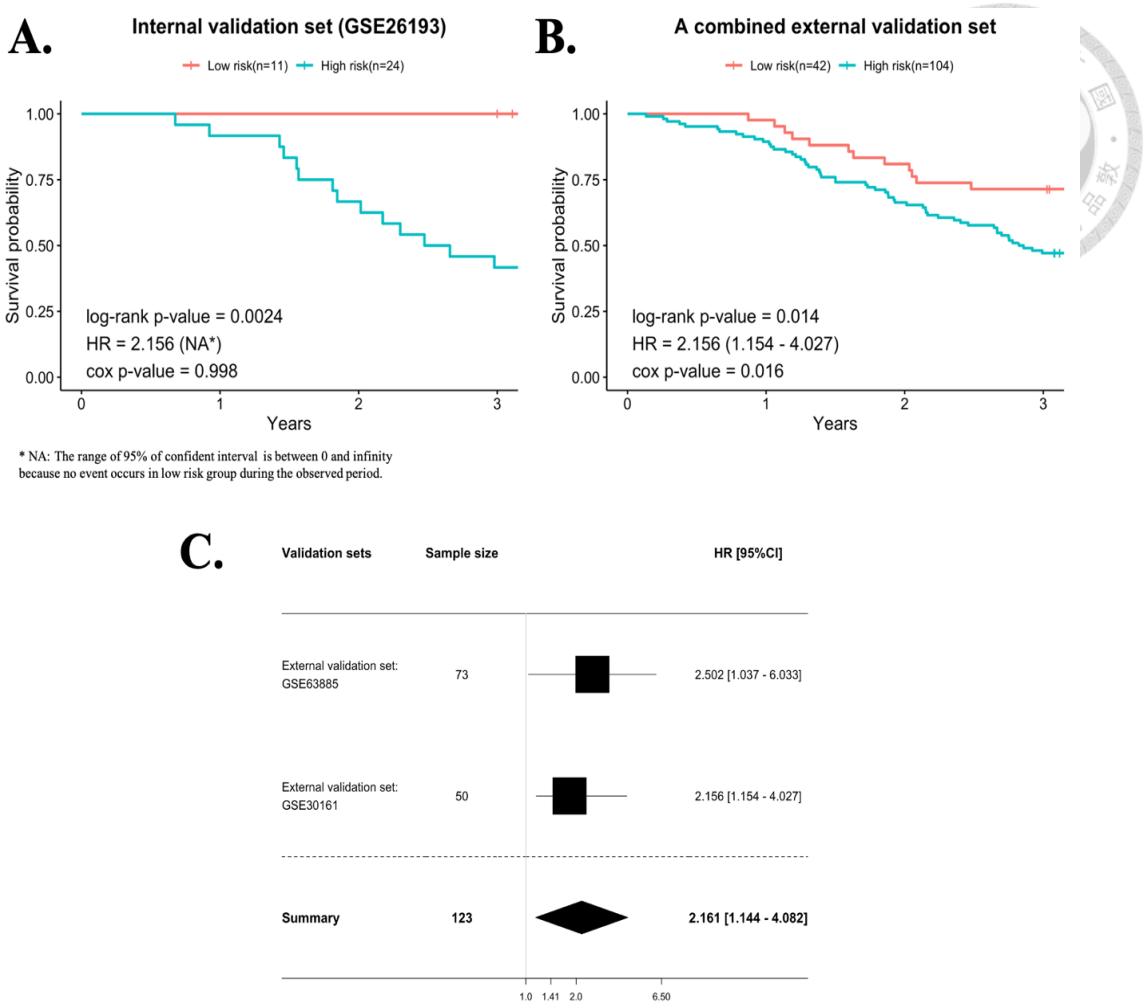


Figure 3-3. Survival analysis of the internal validation set and external validation sets. (A) Kaplan-Meier analysis was conducted for the internal validation set (GSE26193; n =25), and the patients were divided into two groups based on their risk scores. Significant differences ($P < 0.05$) were identified between the two groups over time. (B) Similar results are shown in the combined external validation set (GSE30161, GSE19829 and GSE63885; n = 146). (C) Forest plot of the hazard ratio (HR) and corresponding 95% confidence interval (CI) for individual external validation sets, excepting GSE19829. The vertical line indicates the null value (HR = 1). Each box indicates an individual study point estimate of the HR, and horizontal lines crossing these boxes indicate the 95% confident intervals. The diamond denotes the overall summary estimate of pooled studies.

doi:10.6342/NTU202300922

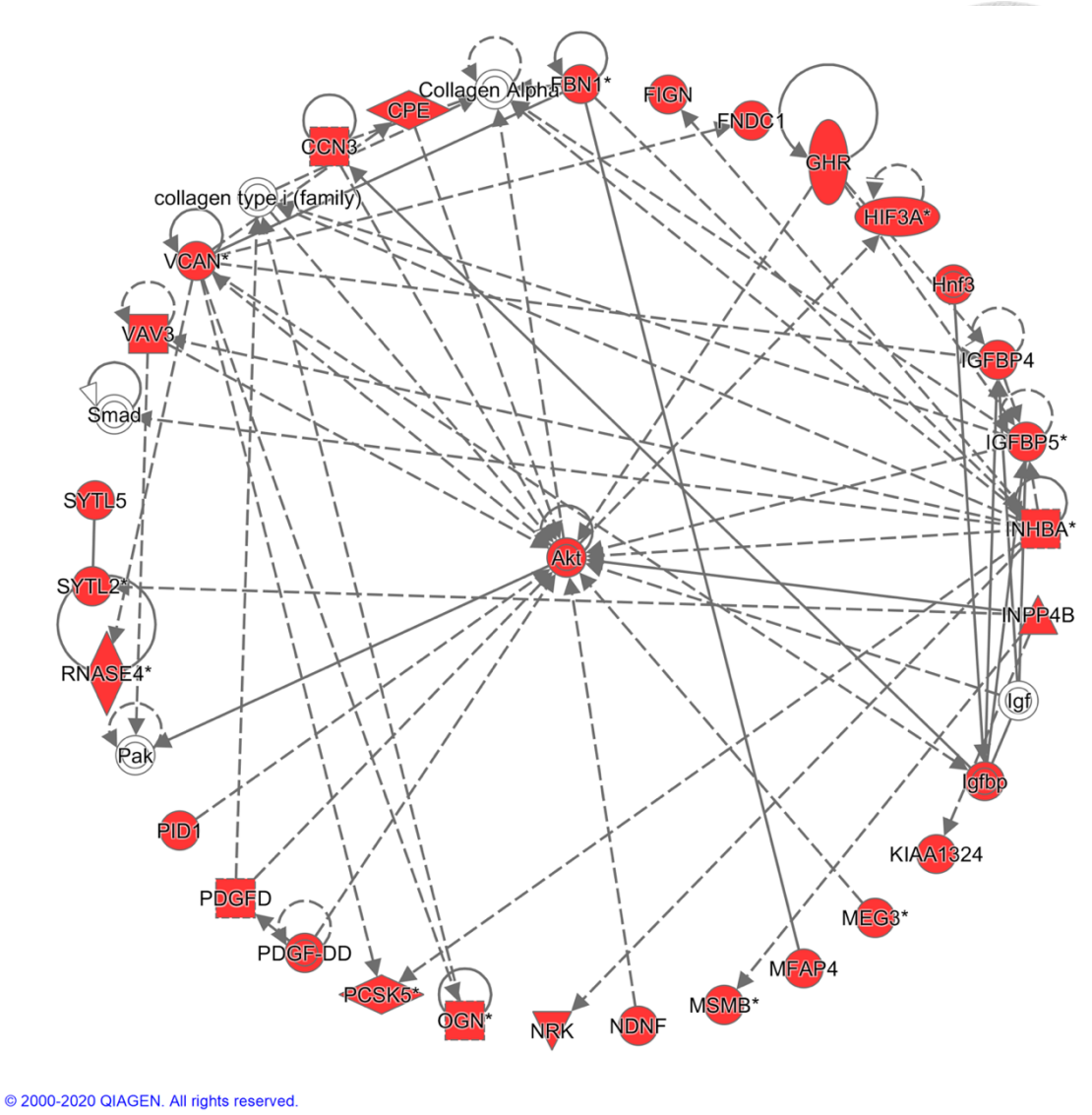


Figure 3-4. The top network result from the Ingenuity® Pathway Analysis (IPA®) program. Red molecules represent the respective differentially expressed genes in our dataset, while white molecules indicate the putative genes that may be possibly involved in this network based on the IPA® database. Solid lines infer a direct interaction while dashed lines infer an indirect interaction.

doi:10.6342/NTU202300922

3.7 Tables

Table 3-1. Summary of GEO datasets used in this study.

Datasets (GEO Accession)	Year	Country	Number of total samples	Number of used samples	Chemotherapy
GSE26193	2011	France	107	106	Yes
GSE30161	2012	United States	58	50	Yes
GSE19829	2010	United States	28	23	Yes
GSE63885	2014	Poland	75	73	Yes

doi:10.6342/NTU202300922

Table 3-2. Statistical analysis of clinical variables in the GSE26193 dataset.

		Overall survival < 3 years (N=52) No. (%)	Overall survival ≥ 3 years (N=54) No. (%)	P-values
Stage	I	15(14.15)	22(20.76)	
	II	4(3.77)	6(5.66)	0.2828
	III	33(31.13)	26(24.53)	
Grade	I	2(1.89)	5(4.72)	
	II	19(17.92)	13(12.26)	0.2665
	III	31(29.25)	36(33.96)	
Signature	Oxidative stress	22(20.75)	29(27.36)	0.2515
	Fibrosis	30(28.30)	25(23.59)	
Subtype	Adenocarcinoma	1(0.94)	2(1.88)	
	Brenner Tumor	1(0.94)	0(0)	
	Carcinosarcoma	2(1.88)	0(0)	
	Clear Cell	3(2.83)	3(2.83)	0.8113
	Endometrioid	3(2.83)	5(4.72)	
	Mucinous	5(4.72)	3(2.83)	
	Serous	37(34.91)	41(38.69)	

doi:10.6342/NTU202300922

Table 3-3. Parameter tuning of bagging-based algorithm with GA-XGBoost models.

	GSE26193 (internal validation set)			Combined external validation set*		
	Accuracy	Sensitivity	Specificity	Accuracy	Sensitivity	Specificity
1. Number of GA-XGBoost models for bagging						
10 models/1 model deleted	0.829	1	0.667	0.5	0.853	0.192
15 models/1 model deleted	0.714	1	0.524	0.589	0.824	0.385
20 models/1 model deleted	0.829	1	0.684	0.541	0.588	0.5
2. The size of training data (Proportion of samples in GSE26193*)*						
50%/3 models deleted	0.857	0.95	0.733	0.623	0.544	0.692
70%/1 model deleted	0.714	1	0.524	0.589	0.824	0.385
90%/1 model deleted	0.857	1	0.667	0.603	0.544	0.654
3. Number of chromosomes in each generation						
100 chromosomes/0 model deleted	0.857	1	0.737	0.582	0.72	0.462
300 chromosomes/1 model deleted	0.714	1	0.524	0.589	0.824	0.385
500 chromosomes/1 model deleted	0.943	1	0.894	0.555	0.632	0.487
4. Number of generations in GA-XGBoost						
300 generations/0 model deleted	0.886	1	0.789	0.596	0.544	0.641
500 generations/1 model deleted	0.714	1	0.524	0.589	0.824	0.385
1000 generations/0 model deleted	0.857	1	0.737	0.562	0.456	0.654
5. Mutation rates in GA-XGBoost						
0.3/1 model deleted	0.8	1	0.632	0.527	0.765	0.321
0.5/1 model deleted	0.714	1	0.524	0.589	0.824	0.385
0.7/2 models deleted	0.771	1	0.579	0.514	0.735	0.321
6. Number of tree layers used in GA-XGBoost						
two layers/1 model deleted	0.714	1	0.524	0.589	0.824	0.385

doi:10.6342/NTU202300922

three layers/0 model deleted	0.914	1	0.85	0.589	0.706	0.487
five layers/0 model deleted	0.914	1	0.842	0.562	0.574	0.551
7. Model with high specificity						
High specificity	0.943	1	0.895	0.582	0.485	0.667

*This combined dataset included three external validation sets (GSE30161, GSE19829, and GSE63885; n = 146).

doi:10.6342/NTU202300922

Table 3-4. Results of individual models for an internal validation set based on bagging of GA-XGBoost models.

Model No.	Number of variables	GSE26193 (internal validation set)			
		Sensitivity	Specificity	Accuracy	F1-score
1	81	0.929	0.429	0.629	0.667
2	117	0.929	0.81	0.857	0.839
3	39	1	0.143	0.486	0.609
4	140	0.929	0.619	0.743	0.743
5	52	0.929	0.524	0.686	0.703
6	25	1	0.81	0.886	0.875
7	53	1	0.524	0.714	0.737
8	70	1	0.476	0.686	0.718
9	129	1	0.333	0.600	0.667
10	73	1	0.333	0.600	0.667
11	59	1	0.524	0.714	0.737
12	87	1	0.476	0.686	0.718
13	30	1	0.476	0.686	0.718
14	24	0.929	0.524	0.686	0.703
15	150	0.857	0.619	0.714	0.706

doi:10.6342/NTU202300922

Table 3-5. Performance comparison between the bagging of GA-XGBoost models and two existing models.

	Accuracy	Sensitivity	Specificity	F1-score
GA-XGBoost				
GSE26193 (internal validation set)	0.714	1	0.524	0.737
Combined external validation set*	0.589	0.824	0.385	0.651
GSE30161	0.580	0.739	0.444	0.618
GSE19829	0.478	1	0.143	0.600
GSE63885	0.630	0.833	0.432	0.690
Forward logistic regression				
GSE26193 (internal validation set)	0.600	0.533	0.650	0.533
Combined external validation set*	0.514	0.456	0.564	0.466
GSE30161	0.620	0.652	0.593	0.612
GSE19829	0.478	0.556	0.429	0.455
GSE63885	0.452	0.306	0.595	0.355
LASSO regression				
GSE26193 (internal validation set)	0.543	0.643	0.476	0.529
Combined external validation set*	0.555	0.544	0.564	0.532
GSE30161	0.520	0.478	0.556	0.478
GSE19829	0.565	0.889	0.357	0.615
GSE63885	0.575	0.500	0.649	0.537

*This combined dataset included three external validation sets (GSE30161, GSE19829, and GSE63885; n = 146).

doi:10.6342/NTU202300922

Chapter 4. The Comparisons of Prognostic Power and Expression Level of Tumor Infiltrating Leukocytes in Hepatitis B- and Hepatitis C-related Hepatocellular Carcinomas

4.1 Abstract

Background: Tumor-infiltrating lymphocytes (TILs) are immune cells surrounding tumor cells, and several studies have shown that TILs are potential survival predictors in different cancers. However, the challenge arises; few studies have been performed for dissecting the differences between hepatitis B- and hepatitis C-related hepatocellular carcinoma (HBV-HCC and HCV-HCC). Therefore, we aim to determine whether the expression levels of the TILs are potential predictors for survival outcomes in hepatocellular carcinomas and which TILs are the most significant ones.

Methods: Two bioinformatics algorithms, including ESTIMATE and CIBERSORT, were utilized to analyze the gene expression profiles from 6 datasets. The ESTIMATE algorithm examined the total expression level of the TILs, whereas the CIBERSORT algorithm reported the expression levels of 22 different TILs. Both subtypes of hepatocellular carcinoma, including HBV-HCC and HCV-HCC, were analyzed accordingly.

doi:10.6342/NTU202300922

Results: The results indicated that the total expression level of TILs was higher in non-tumor regardless of the HCC types. Alternatively, the significant TILs associated with survival outcome and recurrence pattern varied from subtypes to subtypes. For example, in HBV-HCC, plasma cells (hazard ratio [HR]=1.05; 95% CI 1.00-1.10; p=0.034) and activated dendritic cells (HR=1.08; 95% CI 1.01-1.17; p=0.03) were significantly associated with the overall survival, whereas in HCV-HCC disease, monocyte (HR=1.13) were significantly associated with the overall survival. Furthermore, for the recurrence-free survival (RFS), CD8+ T cells (HR=0.98) and M0 macrophages (HR=1.02) were potential biomarkers in HBV-HCC, whereas neutrophil (HR=1.01) was an independent predictor in HCV-HCC. Lastly, in HCC including HBV-HCC and HCV-HCC, CD8+ T cells (HR=0.97) and activated dendritic cells (HR=1.09) have significant association with overall survival (OS); gamma delta T cells (HR=1.04), monocytes (HR=1.05), M0 macrophages (HR=1.04), M1 macrophages (HR=1.02) and activated dendritic cells (HR=1.15) are highly associated with RFS.

Conclusions: These findings demonstrated that the TILs are potential survival predictors in HCC and different kinds of TILs are observed according to the virus types. Therefore, further investigations are warranted to elucidate the etiology of TILs in HCC, which may improve the immunotherapy outcomes.

doi:10.6342/NTU202300922

Keywords: hepatocellular carcinoma; hepatitis B virus; hepatitis C virus; tumor-infiltrating lymphocytes; immune cell; ESTIMATE; CIBERSORT

doi:10.6342/NTU202300922

4.2 Introduction

Hepatocellular carcinoma (HCC) is the most frequent liver malignancy and ranked as the second leading cause of cancer death in the world [165]. Although the majority of liver malignancy is related to viral hepatitis B (HBV) or C (HCV) [166], there are many differences between two types of HCC in terms of activated pathways [167], gene expression profile [168], immunologic responses [169] and clinical prognosis [170]. Recently, a meta-analysis of the effects of targeted cancer drugs, Sorafenib, was conducted to evaluate that whether there is a difference in the overall survival between HBV- and HCV-induced HCC patients after receiving this drug [171]. This study discovered that the overall survival of HBV(+)HCV(-) patients after Sorafenib treatment was significantly improved. On the contrary, insufficient pieces of evidence to support a similar outcome on HBV(-)HCV(+) patients. Therefore, new treatment strategies should be developed to treat virus-driven HCCs effectively.

Immune checkpoint therapy is widely used to treat melanoma, such as squamous-cell lung carcinoma [172], renal cell carcinoma [173], and bladder cancer [174]. However, the local inhibition of the anti-tumor immune responses in the microenvironment may make immuno-therapy challenging to implement [175]. Tumor-infiltrating lymphocytes (TILs) are white blood cells surrounding in the tumor stroma and inside the tumor [176]. It reported that this kind of lymphocytes has a vital

doi:10.6342/NTU202300922

role in the prognosis and prediction of many malignancies [177, 178]. The prognostic association of hepatocellular carcinoma has widely investigated. For example, the higher amount of macrophages discovered in HCC patients associated with poor clinical outcome [179, 180] and the proportion of different kinds of macrophages (M1 and M2) also affected patient's prognosis [181]. These findings implied that different types of TILs or its subtypes in HCC might play essential roles in carcinogenesis or tumor inhibition. For instance, HCC patients who had more dendritic cells [182-184], natural killer cells [185], T lymphocytes [186, 187] or B cells [188-190] had a better prognosis; while those who had more neutrophils [191, 192], monocytes [193, 194], regulatory T lymphocytes [195] and CXCR3+ subtype B cells [196] had the poorest prognosis. Hence, the understanding of the role of TILs in tumor immunology may overcome the current barrier in the anti-cancer field.

The different gene expression patterns between HBV- and HCV-induced HCC patients suggested [197]. Norio Iizuka et al. found that 31 differentially expressed genes (DEGs) highly shows in HBV-HCC case and involved in signal transduction, transcription, and metastasis. While 52 DEGs significantly increased in HCV-HCC case were related to the immune system and detoxification [198]. This study concluded that the pathogenesis of HCC triggered by viral hepatitis B or C might be different. Moreover, those DEGs would be the critical diagnostic markers, implying that distinct

doi:10.6342/NTU202300922

treatments may be applied to HBV-HCC and HCV-HCC patients, respectively. Many research studies have focused on the association of HCC patients and TILs [199, 200]. Nevertheless, very few studies have addressed the issues such as the different immune patterns caused by those two virus-infected conditions, and the relationship of different types of immune cells and the prognostic outcomes.

Although many studies have widely investigated the TILs composition in HCC, the majority pay more attention to HCC in general [201], or individual HBV-/HCV-infected case [202]. Some even only focus on one specific immune cell [203]. Therefore, this study aims to define the compositions of the immune response in both HBV-HCC and HCV-HCC diseases to investigate its relationships with a clinical annotation such as overall survival and recurrence-free survival. We applied two well-established algorithms (ESTIMATE and CIBERSORT) to ascertain the relationships between the immune infiltrates and the clinical prognosis, hoping to have a better understanding of these virus-driven HCC diseases.

doi:10.6342/NTU202300922

4.3 Materials and Methods

4.3.1 Identification and selection of included studies

To focus on the TILs influence on HCC with different virus status, we divided the samples into three groups: HBV-related HCC, HCV-related HCC, and HCC. Specifically, the HCC HBeAg-positive or HBcAb-positive carriers were defined as HBV-related HCC, whereas the type of hepatitis C virus (HCV) infected patients positively detected by the anti-HCV agent characterized as HCV-related HCC. To further identify the gene expression datasets in NCBI GEO related to three HCC types with distinct virus status, we used the following search terms: “hepatitis B virus”, “hepatitis C virus”, “hepatocellular carcinoma”, “survival”, “clinical” and “recurrence”, which yields 10 datasets. Regarding TCGA data, we directly used liver cancer dataset (n=371), including both the gene expression profile and the clinical information.

Primary inclusion criteria for the datasets in this study were: (1) virus status is clearly stated for HBV-related HCC and HCV-related HCC samples and samples having no specific information are defined as HCC samples; (2) contains expression profiles of both tumor and its corresponding adjacent normal tissues for each sample; (3) includes at least one reliable prognostic factor such as overall survival rate in clinical information; (4) is the most recently published or has a larger sample size when the dataset is used repeatedly in different studies. However, some datasets must be

doi:10.6342/NTU202300922

excluded because of the following reasons: (1) two HCC types have coexisted; (2) the cause of hepatocellular carcinoma was by metastasis; (3) the sample size was smaller than 30; (4) array/NGS platforms used in those datasets were not suitable for the downstream pipeline. The selection of these datasets gave six datasets (GEO10143, GSE76427, GSE54236, GSE14520, GSE14520, and TCGA) meets the criteria in the data extraction step.

4.3.2 Statistical analysis

The analysis pipeline used in this study is illustrated in Figure 4-1. The total TIL quantity and the proportion of immune cells of tumor tissue or adjacent normal tissue from the same patient were calculated based on the expression profiles by ESTIMATE [204] and CIBERSORT [205] separately. The ESTIMATE algorithm uses gene expression data (tumor/non-tumor) to infer the expression level of infiltrating immune cells in tumor tissues and is based on the calculation of the single sample Gene Set Enrichment Analysis (ssGSEA). For the ssGSEA algorithm, 171 marker genes are required to generate an immune score. This score is in proportion to the real quantity of immune cells; as a result, the number of immune cells among samples can be directly compared using this value. The CIBERSORT algorithm is a tool for estimating the abundances of immune cells using transcriptomic data. This software applied linear support vector regression (SVR) to deconvolve the relative fractions of immune cells

doi:10.6342/NTU202300922

of given transcriptional profile from a bulk tumor based on the signature matrix as reference.

For the different batches and/or platforms used in the analyzed datasets, the original pipelines of the ESTIMATE algorithm and the CIBERSORT algorithm contained a normalization step. In order to minimize the batch effects, we used the quantile normalization to in both the RNA-Seq expression data and microarray data within one dataset. For multiple datasets, the CIBERSORT algorithm reported the proportions of different TILs within one dataset, which means the proportions has already been normalized within one dataset and the total sum of the proportions equals 100. Therefore, it is feasible to use the proportions from different datasets to do the comparisons. However, for the ESTIMATE algorithm, different datasets may result in different levels of immune scores and thus we compared the immunes scores within one dataset instead of doing the comparisons across different datasets.

The paired Wilcoxon signed-rank test was applied to identify the difference in quantification and characterization of TILs according to the virus types. P-values were corrected for multiple testing by the Benjamin-Hochberg method [206]. To evaluate whether immune cells are associated with survival outcomes, Cox proportional hazards regression model was used (the coxph function in the survival R package). The statistical results (cox coefficients, hazard ratios with 95% confidence interval, and p-

doi:10.6342/NTU202300922

values) for each immune cell type obtained. We used univariate Cox regression model to determine the effect of clinical factors on survival. Additionally, the association of each immune cell types and significant clinical factors were assessed via multivariate Cox regression model using the *statistical* R package.

doi:10.6342/NTU202300922

4.4 Results

4.4.1 Selection of included datasets

The characteristics of the selected datasets has shown in Table 4-1. The described patients in these datasets were from Japan, Singapore, China, Italy, and America, and all the data were collected in 2008 or later. Two datasets included all HBV-HCC, HCV-HCC and HCC patients, and four restricted to at least one HCC types of patients. In total, 1276 patients finally included in this study: 313 were HBV-HCC patients, 135 were HCV-HCC patients, and 828 were HCC patients. The median follow-up ranged from 0.99 to 7.8 years.

4.4.2 Estimation of infiltrating cells

ESTIMATE method was firstly used to compare the quantities of the immune cells between intra-tumor and non-tumor of the same patient. In the five datasets (GSE10143, GSE76427, GSE54236, GSE14520, and GSE17856) used in the estimation of the number of immune cells, the immune scores were significantly reduced in intra-tumor (Figure 4-2). The p-values of each Wilcoxon singed-rank test was lower than 0.0001.

doi:10.6342/NTU202300922

4.4.3 Composition of TILs

To systematically investigate the compositional difference of TILs between intra-tumor and non-tumor from the same HCC patient, we then applied CIBERSORT method to calculate the proportion of TILs in terms of distinct tissue type.

With respect to HBV-HCC group (GSE14520, n=204), the signatures of CD4+ resting memory T-cells (q-value<0.001), CD4+ activated memory T-cells (q-value<0.001), natural killer activated cells (q-value<0.001), dendritic resting cells (q-value<0.001) and resting mast cells (q-value<0.001) were found to predominate in tumor. On the contrary, plasma cells (q-value=0.001), CD8+ T-cells (q-value<0.001), gamma detla T-cell (q-value<0.001), M1 macrophages (q-value<0.001), M2 macrophages (q-value<0.001) and activated mast cells (q-value<0.001) highly dominated in non-tumor.

Regarding HCV-HCC group, two datasets (GSE10143, n=46; GSE17856, n=40) were investigated (Table S4-1). In GSE10143 dataset, the proportion of M0 macrophage and neutrophil cells were higher in the intra-tumor part (2.9%±3.8%; 5.1%±3.4%) than in non-tumor (1.2%±2%; 3.4%±2.3%); on the contrast, relatively higher number of both memory B cells and CD8+ T cells were presented in non-tumor. In GSE17856 dataset, more gamma delta T-cells (q-value=0.05) and M0 macrophages (q-value=0.002) were counted in the intra-tumor part while more monocytes (q-

doi:10.6342/NTU202300922

value<0.001) were existed in non-tumor. After enlarging the sample size (GSE10143 and GSE17856, n=86) in HCV-HCC group for the statistical power, the results revealed that there was a significant increase in the proportion of M0 macrophages (q-value<0.001) and neutrophil cells (q-value=0.028) infiltrating in the intra-tumor part; however, there were higher percentage of CD8+ cells (q-value=0.04) and monocytes (q-value<0.001) in non-tumor (Figure 4-3).

In respect of the HCC group, either individual datasets or a pooled dataset (GSE74627, n=52; GSE54236, n= 78; GSE10143, n=62; GSE14520, n=204; GSE17584, n=40; total n = 437) used for examination (Figure 4-3 and Table S4-2). The analytical results of the pooled dataset demonstrated that there were five immune cell types significantly expressed in the intra-tumor part: CD4+ memory activated T cells (q-value=0.038), NK activated cells (q-value<0.001), M0 macrophages (q-value<0.001), dendritic activated cells (q-value<0.001) and mast resting cells (q-value<0.001); whereas, the immune cells detected more in non-tumor were as follows: plasma cells (q-value<0.001), CD8+ activated T cells (q-value<0.001), gamma delta T cells (q-value<0.001), monocytes (q-value=0.005), M1 macrophages (q-value<0.001), M2 macrophages (q-value<0.003) and mast activated cells (q-value<0.001) in the pooled datasets (Figure 4-3). The TILs with same infiltrating pattern among three HCC groups further summarized in Additional file 3. Taken together, these results indicate

doi:10.6342/NTU202300922

that immune cells in different virus-driven HCC groups might play an important role in the process of carcinogenesis.

4.4.4 Prognostic associations of clinical diagnose and immune cells in tumor tissue

To determine the prognostic effect of immune cells in tumor tissue, we utilized overall survival (OS) and recurrence-free survival (RFS) as the indicators.

In regard to HBV-HCC group, two datasets (GSE14520, n=204; TCGA, n=95) were analyzed using univariate, adjusted univariate, and multivariate analysis (Table S4-4 and S4-7). The clinical factors that affected patients' OS and RFS in the two datasets were shown in Table S4-5. In GSE14520, there were five clinical factors affecting patients' OS: TNM stage, tumor size, AFP concentration and multinodular characteristic; whereas the factors significantly associated with patients' RFS were as follows: gender, TNM stage, and BCLC stage. After adjusting the above factors, aside from dendritic activated cells (HR=1.10; p=0.028), dendritic resting cells (HR=1.09; p=0.042) also negatively affected the OS of patients in this dataset. The univariate Cox regression of OS using a pooled dataset (GSE14520 and TCGA; total n=299) revealed that plasma cells (HR=1.06; p=0.005), M1 macrophage (HR=0.96; p=0.021) and dendritic activated cells (HR=1.12; p=0.006) all contributed to the model. However, there was a slight difference when we excluded non-Asian patients (n=287). According

doi:10.6342/NTU202300922

to the univariate analysis, plasma cells (HR=1.06; p=0.008), dendritic activated cells (HR=1.04; p=0.05) and dendritic resting cells (HR=1.12; p=0.007) were negatively associated with OS and M1 macrophage (HR=0.96; p=0.03) was positively associated with OS in Asian case; whereas M0 macrophage (HR=1.02; p=0.005) was negatively associated with RFS and M1 macrophage (HR=0.96; p=0.033) was positively associated with RFS.

Regarding HCV-HCC group, three datasets (GSE10143, GSE17856, and TCGA) were investigated (Table S4-8) using only univariate analysis, because there was either no clinical annotation or no significant association between known clinical factors and OS [207]. The unadjusted univariate analysis of a pooled dataset (GSE10143 and TCGA; total n=95) revealed that lymphocytes negatively associated with OS were NK resting cells (HR=1.13; p=0.021) and monocytes (HR=1.21; p=0.012). Another unadjusted Cox regression analysis of a pooled dataset (GSE10143 and GSE17856; n=89) showed lymphocytes negatively associated with RSF were NK resting cells (HR=1.11; p=0.035), M2 macrophages (HR=1.04; p=0.003) and neutrophils (HR=1.10; p=0.014).

For the HCC group, four datasets (GSE10143, n=62; GSE76427, n=52; GSE542365, n=78; TCGA, n=329) were analyzed using univariate, adjusted univariate, and multivariate analysis (Table S4-9). The adjusted univariate analysis of a pooled

doi:10.6342/NTU202300922

dataset (GSE10143, GSE76427, GSE54236, GSE17856, GSE14520, TCGA; total n=793) displayed that CD8+ T cells (HR=0.97; p=0.015) was positively associated with OS and dendritic activated cells (HR=1.09; p=0.01) was negatively associated with OS; while another pooled dataset (GSE10143, GSE76427, GSE14520, GSE17856; total n=418) showed that 4 immune cells (plasma cells, gamma delta T cells, NK resting cells and monocytes) have a negative impact on RFS.

We further identified whether the same immune cells are associated with survival among the three HCC groups. Table S4-10 shows that RFS was negatively affected by neutrophils in both HBV-HCC (HR=1.11) and HCV-HCC patients (HR=1.20). Moreover, in HBV-HCC and HCC tumor, OS had a negative association with dendritic activated cells, and RFS showed similar trend against M0 macrophages. Meanwhile, the proportion of NK resting cells and M2 macrophages in HCV-HCC and HCC groups had a negative impact on RFS. Unfortunately, no common immune cells had a similar impact across three HCC groups. Also, the TILs that both affected the survival and involved in the carcinogenesis were summarized in Table 4-3. For instance, plasma cells (HR=1.05), M1 macrophage (HR=0.95) and dendritic activated cells (HR=1.08) were the potential indicators of the survival in HBV-HCC whereas monocytes (HR=1.21) was the one in HCV-HCC. These results suggest that the TILs served as the survival predictors in HCC may be varied based on its subtypes.

doi:10.6342/NTU202300922

4.5 Discussion

4.5.1 Different immune responses in virus-driven HCCs

The hepatitis B virus (HBV) is a small DNA virus with the ability to insert into the host genome, causing tumorigenesis. HBV generates the regulatory protein HBx involved in cell growth and carcinogenesis. Similarly, in HCC disease, this protein has been reported to regulate the Wnt pathway [208] and interfere with the function of innate immunity [209]. While the hepatitis C virus (HCV) is a positive-stranded RNA virus, which produces nuclear protein to inhibit NK cells' ability [210]. To our knowledge, the distributions and proportions of the TILs in HBV and HCV have not been comprehensively investigated. Therefore, different viruses might induce different TILs. Our results demonstrated that the immune infiltrates associated with overall survival in HCV-HCC patients were differed from those found in HBV-HCC and HCC groups.

Consistent with previous studies [211, 212], this evidence proved that different virus status activated diverse immune responses and further influenced the patient's survival rate. However, because of the small sample size, there were also limitations associated with the statistical analysis. Regarding HBV-HCC and HCC groups, the results revealed that both plasma cells and dendritic activated cells were negatively associated with overall survival. Dong-Ming Kuang et al. pointed out that plasma cells

doi:10.6342/NTU202300922

secrete IgG molecule to enhance M2 polarization and M2-polarized macrophage push plasma cells to secrete more IgG, forming a positive-feedback loop in hepatoma [213]. Additionally, other studies indicated that environmental semimature dendritic cells might activate extra FcγRIIlow/− B cells in HCC tumors to suppress cytotoxic T-cell function [214] and these semimature dendritic cells also induce immune tolerance through enhancing the production of regulatory T cells [215]. Therefore, these results found in HBV-HCC and HCC cases are in line with those of previous studies.

Similarly, we also found that the comparative proportions of plasma cells, M0 macrophages, M1 macrophages, CD8+ T-cells and dendritic activated cells significantly varied between intra-tumor and non-tumor in HBV-HCC group; meanwhile, those differences were also directly related to the survival. While neutrophils and monocytes were identified in HCV-HCC group, and CD8+ T cells and activated dendritic cells were recognized in the HCC group. These results indicated that immune infiltrates may participate in the process of oncogenesis, and the proportion of these would further affect the survival time in each patient, suggesting these immune cells may be the potential targets in immunotherapy. For example, in HBV-HCC disease, corruption of dendritic cells or enhancement of M1 macrophages' proliferation and function could act as anti-cancer strategies. Besides, both activated dendritic cells and M0 macrophages were differentially expressed in HBV-HCC and HCC diseases,

doi:10.6342/NTU202300922

indicating that the oncogenic mechanism of non-virus-infected HCC might be similar to those of HBV-HCC.

4.5.2 Limitations and future prospects

The prevalence of HCC in many Asian countries is generally higher than in western countries [216]. Highly infected with HBV is the leading risk in Eastern and South-Eastern countries like China, South Korea, and Malaysia [217]; whereas the incidence rate of HCV-HCC is high in other Asian countries like Japan and Singapore [218]. In this study, most of the datasets with relatively larger sample size in the public domains were also from Asia. The ratios of Asian samples to American samples in both HBV-HCC and HCV-HCC were around 2 to 1; whereas the ratio of Asian samples to other populations (European and American) was about 1:1.2. Therefore, it was not feasible to conduct a population-specific analysis by dividing the samples into different populations.

Few gene expression datasets of HCCs are currently available in the public databases and our exclusion criteria further restricted the number of analyzable datasets. In addition, since all datasets retrieved from the public domain, we can only analyze the variables provided in each dataset. Such limited clinical information, for example, a wide range of follow-up time among datasets, might affect the analytical results.

doi:10.6342/NTU202300922

Therefore, further investigations are required and warranted to validate the results of this study.

Since both the total quantity and compositions of TILs were different from cancer to cancer, the cytokines and the microenvironment centering on the TILs were also different. So, it is possible that different TILs were observed in the adjacent normal tissue, even in the same organ. Furthermore, the tumor purity varies in different samples, and thus, it might be another reason why we observed different TILs in the non-tumor cells. Lastly, although these adjacent normal cells were defined as non-tumor tissue, it is difficult to ensure that no lesion or tumor cells exist in them.

Lastly, the algorithms applied in this study might not be the most feasible one to identify which immune infiltrates increasingly or decreasingly expressed, or even to evaluate the exact expressed values changed in each TIL type. Although many pieces of researches still have applied the relative expressed values in survival prediction [219, 220], future work should be undertaken to improve ESTIMATE algorithm to compute an absolute score which is not affected by technical issues such as different platforms.

doi:10.6342/NTU202300922

4.6 Conclusions

In this study, we have shown that the density of infiltrating leukocytes intra-tumor is higher than that of the tumor part, and the fractions of TILs among HBV-HCC, HCV-HCC, and HCC groups are rather diverse. Additionally, our results also revealed that different HCC group would present different immune cells affecting the overall survival in patients. While for the datasets which contained limited clinical factors were avoid. Therefore, further validation using a most significant number of samples is necessary, and future prospective improvements in the predicting algorithms could also be accessed to minimized the platform effect and explore the absolute quantity of each immune cell in non-tumor or the tumor.

doi:10.6342/NTU202300922

4.7 Figures

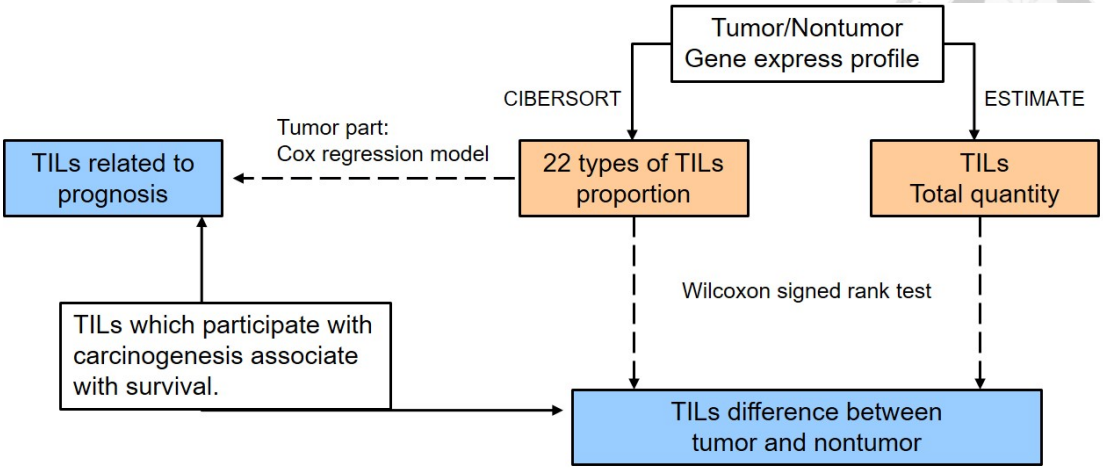


Figure 4-1. Overview of this study.

doi:10.6342/NTU202300922

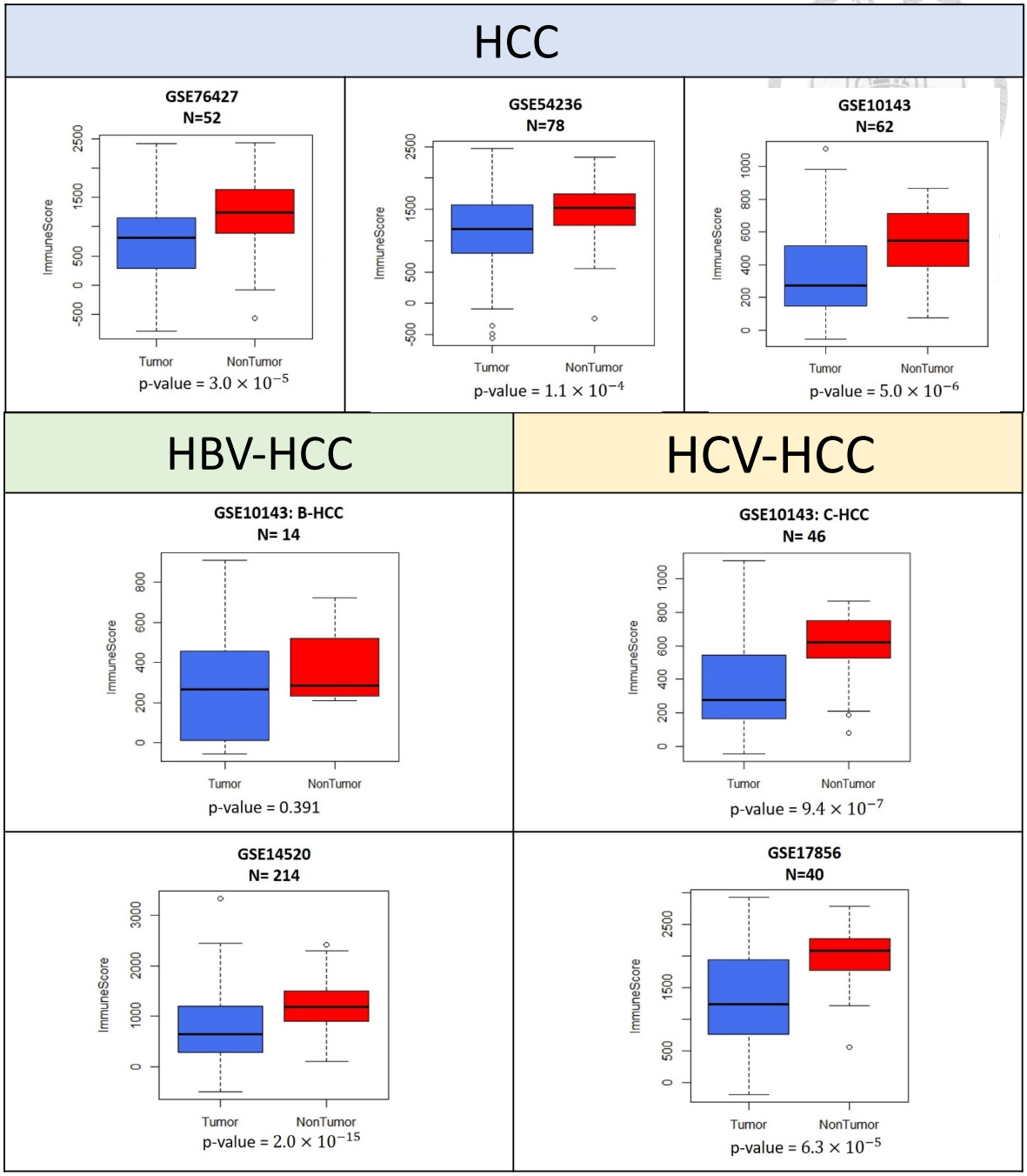


Figure 4-2. Immune scores for intra-tumor and non-tumor parts of hepatocellular carcinoma samples. These five datasets with paired data were quantile normalized and the normalized data were then inputted into ESTIMATE software for the calculation of immune scores comparing the intra-tumor part and its matched non-tumor part. P-values were tested by Wilcoxon signed-rank test.

doi:10.6342/NTU202300922

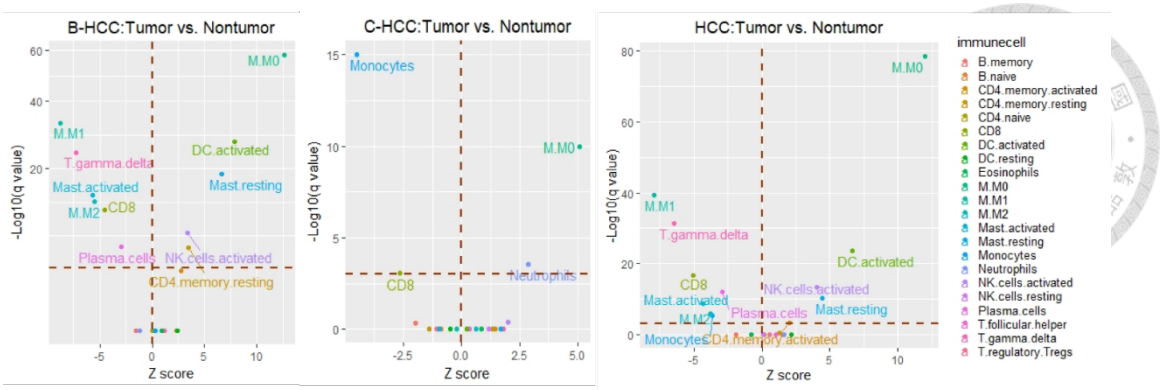


Figure 4-3. The relative compositions of TILs population between the intra-tumor and non-tumors using CIBERSORT. The datasets used for these visual outputs: (1) HBV-HCC: GSE14520; (2) HCV-HCC GSE17856 and GSE10143; (3) HCC: GSE74627, GSE54236, GSE10143, GSE14520, and GSE17584. All immune cell types were evaluated, but only the statistically significant are labeled on the plot. P-value was calculated by Wilcoxon rank-sum test and then adjusted by Bonferroni correction (q-value). The red dotted line on the y-axis indicates q value of 0.05 whereas the one on the x-axis indicates a Z score of 0.

doi:10.6342/NTU202300922

4.8 Tables

Table 4-1. Characteristics of 6 public datasets and the classifications of HCC types.

Datasets	Year	Country	Sex (M/F)	Mean Age	TNM stage	No. of patients		
						HBV-HCC	HCV-HCC	HCC*
Paired data								
GSE10143	2008	Japan	-	-	-	14	46	62
GSE76427	2018	Singapore	45/7	59.5±12.5	I: 28 II: 12 III: 12	-	-	52
GSE54236	2016	Italy	61/17	-	-	-	-	78
GSE14520	2010	China	178/26	50±10.6	I: 88 II: 74 III: 42	204	-	204
GSE17856	2010	Japan	-	-	-	-	40	40
Tumor-only data								
GSE76427	2018	Singapore	93/22	63.5±12.7	I: 55 II: 35 III,IV: 24	-	-	115
TCGA	2014	American	107/222	55±11.7	I:161 II: 82 III,IV:86	95	49	329
Total						313	135	880

* HBV-HCC: hepatitis B-related hepatocellular carcinoma; HCV-HCC: hepatitis C-related hepatocellular carcinoma; HCC: hepatocellular carcinoma caused by either virus or other reasons in clinical feature.

doi:10.6342/NTU202300922

Table 4-2. The comparative composition of TILs in HBV-HCC.

GSE No.	Group	Tissue	Immune Scores	P value	Cibersort		p value	q value
					Types	Mean of Proportion		
GSE14520	HBV-HCC (n=204)	Tumor	782.7±643.6	<0.001	T.cells.CD4.memory.resting	6.4±8.4/3.9±6	<0.001	0.003
					T.cells.CD4.memory.activated	0.4±1.4/0.1±0.6	<0.001	0.038
					NK.cells.activated	7.1±4.4/5.7±4	<0.001	<0.001
					Macrophages.M0	12.1±11.1/1.6±4	<0.001	<0.001
					Dendritic.cells.activated	1.4±2.3/0.1±0.4	<0.001	<0.001
					Mast.cells.resting	3.8±4.8/1.3±2.4	<0.001	<0.001
		Nontumor	1215.4±425.9		Plasma.cells	4.7±3.4/5.6±2.8	<0.001	0.001
					T.cells.CD8	11.2±8.5/14.8±7.4	<0.001	<0.001
					T.cells.gamma.delta	5.7±5.2/10±6.5	<0.001	<0.001
					Macrophages.M1	10.9±5.7/15.7±5.4	<0.001	<0.001
					Macrophages.M2	9.7±6.2/13.4±7.5	<0.001	<0.001
					Mast.cells.activated	1.9±3/4±4.4	<0.001	<0.001

* HBV-HCC: hepatitis B-related hepatocellular carcinoma; T: Tumor; NT: Non-tumor; p value tested by Wilcoxon signed-rank test and q value was adjusted by the Bonferroni correction; Proportion values were expressed as mean±SEM.

doi:10.6342/NTU202300922

Table 4-3. Summary of the TILs that both participated in carcinogenesis and affected the survival of the patient.

Groups	Overall survival (HR)	Recurrence-free survival (HR)
HBV-HCC	Plasma cell (1.05) Macrophages.M1 (0.95) Dendritic.cells.activated (1.08)	T.cells.CD8 (0.98) Macrophages.M0 (1.02)
HCV-HCC	Monocytes (1.21)	Neutrophils (1.01)
HCC	T.cells.CD8 (0.97) Dendritic.cells.activated (1.09)	Plasma.cells (1.05) T.cells.gamma.delta (1.04) Monocytes (1.05) Macrophages.M0 (1.04) Macrophages.M2 (1.02) Dendritic.cells.activated (1.15)

*HR: Hazard Ratio

doi:10.6342/NTU202300922

Chapter 5. ceRNAR: an R package for identification and analysis of ceRNA-miRNA triplets

5.1 Abstract

Competitive endogenous RNA (ceRNA) represents a novel mechanism of gene regulation that controls several biological and pathological processes. Recently, an increasing number of in silico methods have been developed to accelerate the identification of such regulatory events. However, there is still a need for a tool supporting the hypothesis that ceRNA regulatory events only occur at specific miRNA expression levels. To this end, we present an R package, ceRNAR, which allows identification and analysis of ceRNA-miRNA triplets via integration of miRNA and RNA expression data. The ceRNAR package integrates three main steps: (i) identification of ceRNA pairs based on a rank-based correlation between pairs that considers the impact of miRNA and a running sum correlation statistic,

doi:10.6342/NTU202300922

(ii) sample clustering based on gene-gene correlation by circular binary segmentation, and (iii) peak merging to identify the most relevant sample patterns. In addition, ceRNAR also provides downstream analyses of identified ceRNA-miRNA triplets, including network analysis, functional annotation, survival analysis, external validation, and integration of different tools. The performance of our proposed approach was validated through simulation studies of different scenarios. Compared with several published tools, ceRNAR was able to identify true ceRNA triplets with high sensitivity, low false-positive rates, and acceptable running time. In real data applications, the ceRNAs common to two lung cancer datasets were identified in both datasets. The bridging miRNA for one of these, the ceRNA for MAP4K3, was identified by ceRNAR as hsa-let-7c-5p. Since similar cancer subtypes do share some biological patterns, these results demonstrated that our proposed algorithm was able to identify potential ceRNA targets in real patients. In summary, ceRNAR offers

doi:10.6342/NTU202300922

a novel algorithm and a comprehensive pipeline to identify and analyze ceRNA regulation. The package is implemented in R and is available on GitHub (https://github.com/ywhsiao/ceRNAR).

Keywords

Gene regulation, microRNAs, competing endogenous RNAs (ceRNAs), mRNA, expression profiles

doi:10.6342/NTU202300922

5.2 Introduction

Regulation of gene expression can occur at multiple levels via both transcriptional and post-transcriptional mechanisms [221]. Many non-coding RNAs have critical roles in post-transcriptional regulation of protein-coding genes [222]. MicroRNAs (miRNAs) are short, non-coding, single-stranded RNAs with ~22 nucleotides. They usually bind protein-coding genes via partial complementarity with many miRNA response elements (MREs) to repress gene expression by inhibiting translation. Previous studies have shown that miRNAs are involved in a broad range of cancer-associated biological processes, including apoptosis, proliferation, metastasis, and angiogenesis [223]. Similar to gene expression, miRNA expression has cancer-specific patterns that can be used to detect cancers. Therefore, the expression values of RNA can serve as diagnostic, prognostic, or therapeutic biomarkers in a diverse range of cancers [224].

doi:10.6342/NTU202300922

The concept of competing endogenous RNAs (ceRNAs), also called miRNA sponges or miRNA decoys, has revolutionized our knowledge of miRNA regulatory mechanisms. Such RNAs include canonical protein-coding messenger RNAs (mRNAs), long non-coding RNAs (lncRNAs), circular RNAs (circRNAs), and pseudogenes [225]. Their mechanism is to compete with miRNAs for binding their regulatory sequences. There are two primary hypotheses regarding the regulatory function of ceRNAs, based on their expression level or their number of MREs [226]. Taking miRNA-mRNA regulation for example (i.e., where two mRNAs act as ceRNAs that can bind to the same miRNA), the first hypothesis is that the miRNA tends to be sequestered by the mRNA with the higher expression level, leading to weakened inhibitory effects of the miRNA on the other mRNA and thus increasing the expression of the other mRNA under the assumption of equal MREs on the two RNAs. The

doi:10.6342/NTU202300922

second hypothesis is that the miRNA has a greater affinity for the mRNA with more MREs in its sequence.

Some ceRNAs have been identified in multiple cancers; for instance, PTEN is an important tumor suppressor gene that was also reported to encode ceRNA in prostate cancer [227], glioblastoma [228], and melanoma [229]. This suggests that elucidating ceRNAs can improve the understanding of biological mechanisms in regulating cancer cells. However, using biological experiments to identify ceRNAs is time-consuming and labor-intensive. To address this issue, an increasing number of computational methods have been developed for identifying ceRNAs.

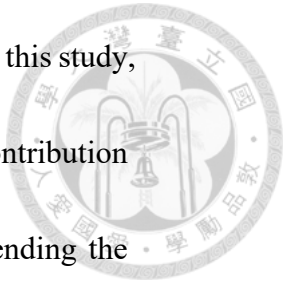
The traditional algorithm is based on the probability theory that two mRNAs share miRNAs and their binding sites (i.e., MREs) [10, 11]. A hypergeometric test is applied to find out if the probability of binding of an mRNA to a given miRNA is larger than that which would occur by chance. However, such an approach usually uses a selected

doi:10.6342/NTU202300922

threshold to choose significant genes and takes only these genes into consideration; it may not extend to the whole-genome level and fails to consider the correlation among genes because this approach treats each gene independently [230].

Recently, because of the popularity of and advances in genomic sequencing technology, more and more mRNA and miRNA gene expression profiles at the whole-genome level have been released publicly [231-233]. However, the analytical results of such continuous data may tend to be sensitive to the outliers in the sample and to the size of the dataset [234]. Therefore, a second approach has been developed based on the observation of linear correlations between pairs of mRNAs that suggest they have a higher chance of competing with a specific miRNA [235-237]. Unfortunately, such a method ignores the contribution of the expression of the miRNA in a ceRNA binding event. Additionally, it also uses a permutation test to estimate mRNA pairs with significant correlation results; sometimes, it has a

doi:10.6342/NTU202300922



higher computational cost. To overcome these drawbacks, in this study, we present a novel rank-based algorithm considering the contribution of miRNA expression in a ceRNA binding event and extending the pairwise correlation approach to identify ceRNA-miRNA triplets. All the steps in this algorithm have been incorporated into a user-friendly R package called ceRNAR, which also provides several downstream analyses to further interpret the biological meaning of identified ceRNA events for its users.

doi:10.6342/NTU202300922

5.3 Results

5.3.1 Simulation results

To evaluate the performance of our proposed method, simulations of mRNA and miRNA expression data in 100 samples were performed in different scenarios (S1 Table). Notably, we focused on the sensitivity and the positive predictive value (PPV) because the number of ceRNA triplets was small among all possible combinations of triplets.

In the beginning, we presumed the sample distribution of each gene follows normal distribution because about 98% of genes' sample distributions passed the normality test based on 9,835 samples in The Cancer Genome Atlas (TCGA) pan-cancer atlas (Figure S5-1). Therefore, we firstly presumed synthetic expression data were generated from a multivariate normal distribution with a mean value of 0 and a covariance matrix whose entries are 0.9. This ensured that the ceRNAR algorithm supports the hypothesis that pairs of ceRNA

doi:10.6342/NTU202300922

binding partners of a specific miRNA are highly correlated and that it works well to sensitively identify them among a pool of target pairs. However, it is unclear whether such an event between two target genes would occur at the lower or higher expression of a specific miRNA. Therefore, five scenarios were designed to capture how the molecular elements within a triplet interact with each other, and simulations of different parameters were performed. Notably, the number of identified ceRNAs dropped as the window size increased (Table S5-2 and S5-3, Figure 5-1). This is because more uncorrelated samples were included in the analysis when longer window sizes were used, especially larger than 30%. In other words, higher noise was included in the analysis, resulting in difficulty in identifying true ceRNA triplets. Next, we simulated different scenarios in which the ceRNA peak was located at different miRNA expression levels (scenarios 1 to 4). As shown in Table S5-2 and Figure 5-3A, the performance of our proposed algorithm was highly similar across the four scenarios with

doi:10.6342/NTU202300922

true ceRNA signals. Such performance also was shown when removing the permutation test (Figure 5-3B and Table S5-3). Lastly, to reduce the calculation complexity and computation time, we evaluated the performance of the algorithm without the random walk step, which we called the “fast” version. Only minor differences were observed in these two versions (0.1 to 0.25 in terms of PPV value, Tables S5-2 and S5-3), suggesting that both versions were able to identify true ceRNA triplets without reporting high false positive results (Figure 5-3).

5.3.2 Application to TCGA cancer cohort datasets

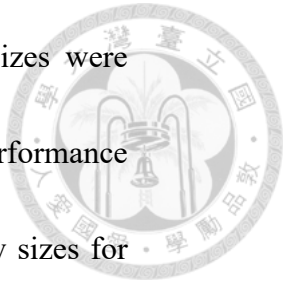
In addition to the analytical parameters for the algorithm, a simulation of different proportions of the correlated samples was performed. Because no major differences were observed in the simulated scenarios and running versions, only scenario 3 was evaluated in this round of testing. As shown in Table **S5-4**, four

doi:10.6342/NTU202300922

settings for the proportion of correlated samples were analyzed (10%, 20%, 30%, and 40%). Notably, the proposed algorithm showed similar performance in the three highest settings (Figure 5-**2**A).

To further examine whether more false positive ceRNA triplets are reported in the fast version, a null scenario without a true ceRNA signal was simulated (scenario 5). The results showed that the negative predictive values of these versions were all higher than 0.9 (S1 and S2 Tables), suggesting there is a low chance of identifying false positive signals in both versions of the algorithm (Figure 5-**2**B). Subsequently, we examined how much time can be saved by using the fast version of the proposed algorithm, which omitted the random walk step. On average, the fast version accelerated the algorithm by approximately 76 times (283.967 seconds versus 21,600 seconds), and only slightly higher false positive rates were reported (Figure 5-4B). Lastly, to evaluate whether different window sizes for merging peaks are critical to the performance of the proposed algorithm, four settings of the

doi:10.6342/NTU202300922



proportion of correlated samples and merging window sizes were analyzed. As shown in Figure 5-4C and Table **S5-5**, the performance of the proposed algorithm was not sensitive to the window sizes for merging peaks.

In addition, to ensure the algorithm only sensitively identified ceRNA events under higher correlation values between them, we also generated simulated data from a multivariate normal distribution with the same mean value but a covariance matrix whose entries are 0.6 (Figure S5-2A and Table S5-6) or 0.3 (Figure S5-2 and Table S5-7). Although the ceRNAR algorithm was still able to detect ceRNA pairs with moderate sensitivity values (0.4 to 0.5) when the correlation values between target genes within a pair were 0.6, it did not work well (sensitivity values were all below 0.25) when the correlation values between target genes within a pair were 0.3. Together, these results support our correlation-based hypothesis that ceRNA events tend to occur when they are highly correlated. However, the above-mentioned

doi:10.6342/NTU202300922

results were based on simple and naïve simulations. We also mimicked the distribution of expression from pan-cancer data (9,835 samples), allowing the synthetic data to be generated from a multivariate normal distribution with a randomly selected mean value and a covariance matrix whose entries are also randomly selected. The simulation studies were also performed under different correlation levels. As shown in Figure S5-3 and Tables S5-8 to S5-11, the performance of the ceRNAR algorithm was determined by the level of the correlation values. Nevertheless, it still performed well (sensitivity values > 0.75) when the correlation between genes was high and when the window size was 10 throughout all scenarios (Figure S5-3B), suggesting the ceRNAR algorithm is efficient to identify most potential ceRNA pairs when their correlation pattern is relatively high (0.8-0.9) to compete with a specific miRNA, and such a pattern exists in at least 20% of the sample. The optimal parameter settings for real data were also

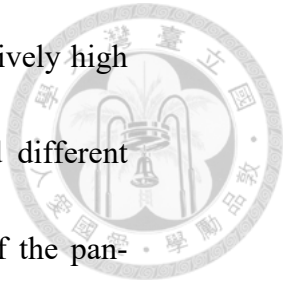
doi:10.6342/NTU202300922

observed. For 100 samples, the best window size was 10, and the best cutoff correlation value for selecting the most significant event was 0.7.

5.3.3 Comparison with other tools

We have compared our tool with other state-of-the-art tools, including SPONGE, JAMI, GDCRNATools, and CERNIA, in terms of their performances using synthetic data and their running time using real data. Figure 5-5A and Table S5-12 illustrate that ceRNAR workflow generally outperformed the other tools in terms of sensitivity and PPV in all scenarios when the window size is set to 10 and the correlation cutoff is set to 0.7. It can be mainly observed when the correlation values among correlated genes are relatively high. However, all of the tools could identify valid ceRNA triplets without reporting high false-positive results except JAMI (Figure 5-5B and S12 Table). GDCRNATools generally had a high sensitivity compared to the other tools, and a lower specificity than ceRNAR, suggesting

doi:10.6342/NTU202300922



that it is good for catching ceRNAs but comes with a relatively high rate of false positives. Regarding running time, we used different sample numbers (250, 500, and 100) on a small subset of the pan-cancer dataset with 15 genes which form 105 triplets; we also used different triplet numbers (105, 1,225, and 4,950) on a small subset of the pan-cancer dataset with 250 samples (Figure S5-4 and Table S5-13). Although the ceRNA algorithm was not fast with a large sample size and a large number of triplets compared with SPONGE, GDCRNATools and CERNIA, it was slightly faster than JAMI.

5.3.4 Application to TCGA cancer cohort datasets

To further validate the applicability and robustness of the ceRNAR algorithm, we also applied the algorithm to two TCGA-derived lung cancer cohorts – TCGA-LUAD and TCGA-LUSC (Table S5-14). The top bridging miRNAs and the hub genes among ceRNA triplets are shown in Table S5-15 and S5-16. Intriguingly, the two

doi:10.6342/NTU202300922

cancer cohorts (LUAD and LUSC) shared some common triplets involving 53 miRNAs and 905 ceRNAs, which allowed construction of a miRNA-modulated ceRNA regulatory network (Figure S5-5). Among them, PLEKHG6 had the largest number of co-expressed ceRNAs, and of the 53 common miRNAs, the top three miRNAs that bridged over 20 ceRNA pairs were hsa-miR-183-5p, hsa-miR-133a-3p, and hsa-miR-142-5p. MAP4K3 was another common hub gene in both datasets, and its bridging miRNA was hsa-let-7c-5p, around which a regulatory network of corresponding ceRNAs was built (Figure S5-6A), and the expression level related to the regulatory occurrence of its bridging ceRNAs is displayed in Figure S5-6B. Since lung cancers share some common molecular characteristics, such results demonstrate the applicability and robustness of the ceRNAR algorithm in multiple cancers or diseases.

In addition, we compared our findings against experimentally validated miRNA-gene pairs in the miRSponge database to endorse the

doi:10.6342/NTU202300922

potential ceRNAs identified by ceRNAR. As shown in Figure S5-7, around 1% (ceRNA pairs that can be validated among total ceRNA founded based on ceRNAR algorithm) of experimentally validated ceRNA triplets were identified in LUAD and LUSC. The low proportion may be attributed to the fact that only 158 ceRNA triplets were analyzed and that those ceRNA triplets may not be expressed in lung tissues. Furthermore, approximately 13-14% of ceRNA triplets with at least one experimentally validated miRNA-target interaction were identified by using the ceRNAR algorithm. A Chi-square test was performed to examine whether the findings from the ceRNAR package were significantly enriched in the TCGA data, and the results showed that the P-values obtained from the LUAD and LUSC datasets were both less than 2.2e-16, suggesting our ceRNAR package can successfully identify previously reported ceRNA triplets.

doi:10.6342/NTU202300922

5.4 Discussion

ceRNA events are a newly discovered type of post-transcriptional regulation, and the identification of ceRNA-miRNA triplets using *in silico* methods is an emerging research area. Therefore, we developed a novel computational algorithm to explore such regulatory events for further biomedical interpretation and application. Our proposed method is based on a simple pairwise correlation approach that considers the miRNA-modulated ceRNA interaction. First, we ranked samples based on their miRNA expression value to include the contribution of miRNA expression and identify which miRNA expression intervals tend to have a higher correlation with pairs of mRNA targets. Secondly, we used a sliding window approach to form more correlation values in a triplet to improve the performance and outcomes of the subsequent statistical approach. Lastly, we applied a cumulation-like approach to sum up the slight changes in correlation values across samples. We used segment clustering to understand the

doi:10.6342/NTU202300922

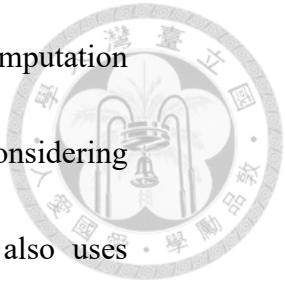
sample clustering in terms of the gene-gene correlations and the miRNA expression intervals so that we could also use sample proportion to support our findings. Several simulations have been conducted for the optimization of the parameters subject to specific ranges of settings, and the robustness of our approach when it does not involve a permutation test has also been evaluated through a simulation study. Connecting with six downstream analyses, our R package may assist researchers to have a deeper understanding of the disease-specific biological regulation and prognostic application for each identified ceRNA-miRNA triplet.

Recently, more and more tools have been developed to identify potential ceRNA events. It is important to systematically evaluate the ceRNAR package in comparison with the five other published ceRNA prediction tools—SPONGE [238], CERNIA [239], GDCRNATools [240], JAMI [241], and CUPID [242]—that are expression-based (rather than sequence-based, i.e., spongeScan [243]). We have

doi:10.6342/NTU202300922

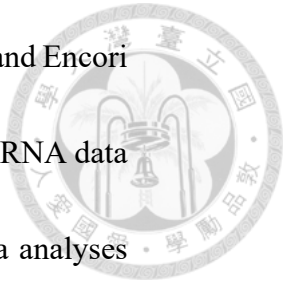
compared them in terms of several features, such as miRNA-target data sources, study design, ceRNA classes, ceRNA prediction algorithm, and language for implementation (Table S5-17). Noting that JAMI is the multi-threading version of CUPID, we decided to keep only JAMI for further analyses. Thus, we used four algorithms, including SPONGE, CERNIA, GDCRNATools, and JAMI, for the comparisons. Notably, the four algorithms and our ceRNAR all used a similar strategy, which is utilizing miRNA-target data sources to identify potential miRNA-gene/lncRNA/pseudogene pairs from other databases. All four of these algorithms are implemented in R. JAMI is based on conditional mutual information, which is particularly useful to capture non-linear associations by estimating the effect of a miRNA on its target pairs through a permutation test [244]. Excepting JAMI, the rest algorithms consider the correlations between miRNA-mediated genes/lncRNA. Although the majority of such algorithms are correlation-based, some differences still exist. For examples,

doi:10.6342/NTU202300922



GDCRNATools is based on sensitivity correlation computation through effectively estimating covariate matrices and also considering the impact of a miRNA on its target pairs. SPONGE also uses sensitivity correlation to quantify the impact of a miRNA on its target pairs (i.e., linear partial correlation), but further applies a null model-based p-value computation to estimate potential ceRNA pairs. It is worth mentioning that CERNIA and JAMI consider both MRE- and expression-based data, whereas SPONGE, GDCRNATools and ceRNAR only analyze genome-wide expression data. Since several studies [236, 245, 246] indicate that ceRNA triplets may be observed in a specific range of miRNA expression, such an approach can help to focus on the true positive region with high signal-to-noise ratio instead of missing the ceRNA triplets due to signal dilution by global noise. Our ceRNAR algorithm showed the highest sensitivity in identifying potential ceRNA triplets (Figure 5-5).

doi:10.6342/NTU202300922



Regarding the two online servers, miRTissue_ce [247] and Encori (i.e., starBase v2) [248] are two web servers that integrate ceRNA data sources, ceRNA prediction algorithms, and even some data analyses and visualizations that can be easily accessed by the users. Fiannaca *et al.* [247] have compared miRTissue_ce and Encori in terms of many features. Here, we further compared these web servers with our ceRNAR based on these features to see whether there is any add-on value that ceRNAR can provide these web services. First, the interactions between miRNAs and target genes in ceRNAR are supported by nine databases, including two experimentally validated miRNA-target databases and seven computationally predicted miRNA-target databases. But Encori and miRTissue_ce are supported by 4 and 8 computationally predicted miRNA-target databases, respectively. Second, Encori uses a hypergeometric test to predict ceRNA, and miRTissue_ce integrates that method with a global test and SPONGE. However, one major disadvantage of the

doi:10.6342/NTU202300922

hypergeometric test is that the test requires a predefined p-value threshold to select significant genes. When using differentially expressed genes, such an approach is not suitable to be applied to the whole genome, because the hypergeometric test fails to consider the interactions among genes due to its independence assumption about genes. This is why we present a novel rank-based algorithm considering the contribution of miRNA expression in a ceRNA event and extend the pair-wise correlation approach to identify ceRNA-miRNA triplets using whole genomic information. Lastly, these two web servers can predict many types of ceRNA events, but ceRNAR only focuses on one ceRNA event class (i.e., mRNA-miRNA).

In real application, we utilized non-small cell lung cancer (NSCLC) data to evaluate the applicability of ceRNAR. NSCLC accounts for 85% of lung cancer and is one of the most common malignant tumors worldwide [244]. Although there has been progress in successful treatment for NSCLC patients these past several decades,

doi:10.6342/NTU202300922

the 5-year survival rate for NSCLC is still relatively low (25%) [249]. Also, the molecular networks involved in NSCLC remain incompletely described in terms of their roles in etiology, progression, and metastasis. Hence, we applied the ceRNAR algorithm to two NSCLC-related cancer datasets in TCGA lung cancer cohorts. Several common miRNAs and ceRNAs identified by the ceRNAR algorithm have also been previously reported by other studies. For example, hsa-miR-183-5p was found to inversely regulate *PTPN4*, serving as a therapeutic target to suppress the metastatic potential in NSCLC patients [250], and hsa-let-7c-5p was verified to prevent cancer metastasis by degrading its bridging hub ceRNA, *MAP4K3* [251]. Although our simulation results suggest that the majority of performance indicators have only slight differences in the four scenarios using both complete and fast versions, and the best parameter setting for window size is 10 and for peak threshold is 0.7, and the fine-tuning of appropriate parameters for non-TCGA datasets still needs to

doi:10.6342/NTU202300922

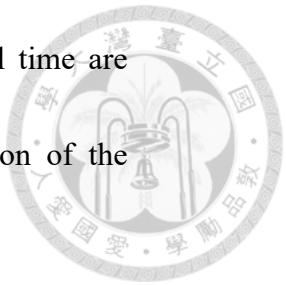
be tested. Nevertheless, these results still demonstrated our proposed method is robust and potentially applicable, allowing it to be extended to studies of other diseases.

However, some limitations still existed in our study. First, a smaller sample size of cancer cohorts (i.e., a smaller window size in our case) may lead to less statistical power of the findings. Second, we presumed a linear relationship between the two ceRNAs in each triplet, but in reality, they were not always linearly correlated. Although we have implemented a sliding window approach to capture such relationships, other methods such as mutual information [252] can also be applied. Moreover, the accuracy of the miRNA target prediction databases we used may have affected the definition of putative ceRNA-miRNA triplets and the outcomes of the ceRNAR algorithm because the mechanisms of some miRNA targeting systems have not been fully understood. It is also worth mentioning that our simulation results were based on a predefined covariance matrix. That is the true

doi:10.6342/NTU202300922

positive events were from the correlation-based approaches, and thus such events only showed linear relationships among the elements. Notably, such design may lead to the poor performance of JAMI because their algorithm was developed by using the mutual information strategy, which was able to capture non-linear relationships among ceRNA pairs in addition to the linear ones. Lastly, the majority of our findings from the real case study were novel compared to the miRSponge database, although some of the miRNA-targets contained an interaction that was previously experimentally validated. Perhaps further experimental validation of those triplets that contained one experimentally validated miRNA-target interaction should be prioritized to increase the robustness of our algorithm and the reliability of the novel findings. Therefore, the consideration of all types of miRNA sponges, the amount of MREs, multiple miRNAs that may compete for the same pair of target genes, nontrivial correlations which involve the comparison of pairwise correlation and pairwise

doi:10.6342/NTU202300922



partial correlation, and the minimization of computational time are important key areas for further optimization and extension of the ceRNAR algorithm.

In summary, ceRNAR is a promising tool for the recognition of ceRNA-miRNA triplets and ceRNA-ceRNA interaction networks in many human diseases, and hence will speed up our knowledge of the regulatory mechanisms and functions of ceRNA-miRNA triplets in the pathogenesis of disease, including cancers.

doi:10.6342/NTU202300922

5.5 Materials and Methods

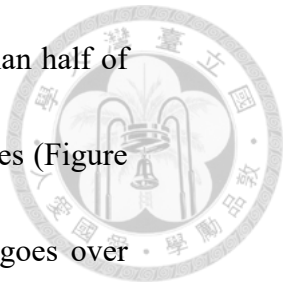
5.5.1 Pipeline of ceRNAR

The ceRNAR package is written in R (version 4.0.5) and is available in the Github repository. The main pipeline of ceRNAR is illustrated in Figure 5-4 and contains three major components for the identification and analysis of ceRNA-miRNA triplets:

- Data preprocessing
- Identification of ceRNA-miRNA triplets
- Downstream analyses

To reduce the computational complexity and time cost, the interactions between miRNAs and target genes were based on two experimentally validated miRNA-target databases (miRTarBase [253] and miRecords [254]) and seven computationally predicted miRNA-target databases (DIANA-micro T-CDS [255], EIMMO [256], miRDB[257], miRanda [258], PITA [259], RNA22 [260] and TargetScan [261]). In the default settings, only those interactions that

doi:10.6342/NTU202300922

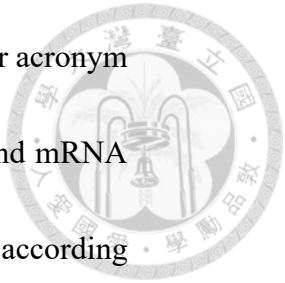


were validated by experiments and/or predicted by more than half of the databases are retained as target miRNAs and target genes (Figure S5-8A). Conceptually, the ceRNAR algorithm iteratively goes over each miRNA-target list and runs through each mRNA pair in a list to evaluate the chance of the potential ceRNA event involved. For a specific triplet (i.e., a miRNA and its two targets), their expression vectors are extracted from the original expression matrix. Therefore, a miRNA expression vector, miRNAm = [mm1, mm2, …], and two mRNA expression vectors, mRNAi = [genei1, genei2, …] and mRNAj = [genej1, genej2, …], are used as inputs into the ceRNAR algorithm to iteratively evaluate whether each mRNA pair is a potential ceRNA event (Figure S5-8B).

5.5.2 Data preprocessing

To prepare expression data for further analyses, ceRNAR can automatically retrieve TCGA data, including mRNA expression,

doi:10.6342/NTU202300922



miRNA expression, and survival data, by entering the cancer acronym [262], but it also supports the use of customized miRNA and mRNA expression matrices that are pre-normalized and formatted according to the instructions. In ceRNAR, we implement two functions to fulfill these two approaches: ceRNATCGA and ceRNACustomize.

5.5.3 Identification of ceRNA-miRNA triplets

To identify miRNA-ceRNA triplets (defined here as a miRNA and two target genes) from expression profiles at a specific miRNA expression level, the ceRNAMethod function can be used, and it contains three modules sequentially: ceRNApairFiltering, SegmentClustering, and PeakMerging (Figure 5-5). We have two assumptions in this study: (1) the expression levels of two target genes tend to be highly correlated when a possible ceRNA event occurs; (2) such events between target genes of a certain miRNA occur at a specific expression interval of that miRNA. However, for each ceRNA

doi:10.6342/NTU202300922

triplet from the real data, it is difficult to know which levels of miRNA expression (low, middle, or high expression intervals) will lead to a high correlation between a pair of target genes. Notably, for one miRNA, the high correlation values between two target genes can only be observed in a specific range of the miRNA expression. Definitely, the expression level of one miRNA can be regulated by many factors, such as compensation and/or other interactors. However, with our current understanding of all miRNAs, it is not feasible to consider all potential regulators of one specific miRNA at the same time. Therefore, for one single miRNA, we adapted the approach of utilizing its expression level as the final output instead of considering all possible confounding factors, which can be regarded as the hidden layers of the miRNA expression value.

doi:10.6342/NTU202300922

5.5.4 The ceRNApairFiltering method

We adopted the sliding window approach to identify the correlation patterns of the two target genes within a specific range of expression values for the miRNA. That is the reason why we ranked the samples based on their miRNA expression value from each triplet to identify the correlation patterns and provided the number of samples that meet the criteria. Therefore, the purpose of this function is to identify ceRNA-miRNA triplets based on the Pearson correlation coefficient through the sliding window approach (Figure S5-9) and a running sum statistic for such values by the random walk approach (Figure S5-10). First, the samples are sorted based on their miRNA expression levels. For a putative triplet ($gene_i$, $miRNA_m$, $gene_j$) among N samples, the correlation coefficients of gene expression values between mRNA and miRNA are calculated within each window (i.e., a length of sample size that is always less than N) with predefined window size (w) as follows:

doi:10.6342/NTU202300922

$$r_k^{m,i,j} =$$

$$corr\left(\left[gene_k^i, gene_{k+1}^i, \ldots, gene_{k+(w-1)}^i\right], \left[gene_k^j, gene_{k+1}^j, \ldots, gene_{k+(w-1)}^j\right]\right)$$

(1)

Here k is the number of windows and a predefined integer. Technically, the correlation value between two genes is calculated using the gene expression values of all samples. By using a sliding window approach [263], we can artificially create different varieties of a real dataset to increase its size, which is a sort of data augmentation technique [264].

Because the accumulated changes in terms of correlation values between two genes among samples may tend to increase the chance that a gene pair is highly correlated and, further, will affect a specific phenotype, we borrowed the concept of gene set enrichment analysis (GSEA) [265], which captures the accumulated changes in the expression of all genes within a pathway through a random-walk method, to identify a significant triplet according to the number of the samples enriched in such an event. The main idea of the GSEA

doi:10.6342/NTU202300922

algorithm is to understand whether the differentially expressed genes are significantly enriched in the samples belonging to the same phenotype (case or control). First, the differentially expressed genes were ranked by using the differences in expression level between the two phenotypes. Next, the GSEA algorithm gave a positive score to the genes located in the pathway, whereas a negative score was assigned to the genes outside of the pathway. One possible scenario to obtain a high score is that the differentially expressed genes were clustered and enriched in one phenotype instead of being randomly distributed. This scenario is the same as what we want to identify for the ceRNA triplet. That is, the highest correlation values were observed in a specific range of the miRNA expression and thus we used the two statistics, $P_{obey}(k)$ and $P_{violate}(k)$ to identify the range (Figure S5-11). Conceptually, to calculate a score (S) to represent the enriched correlation levels among samples against a specific phenotype, we first rank ordered the k windows to form $L=\{miRNA_1,$

doi:10.6342/NTU202300922

$miRNA_2$, …, $miRNA_k$}based on the average miRNA expression within each window:

$$miRNA_k = \sum_{n=0}^{w-1} m_{k+n}^{m,i,j} \Big/ w \qquad (2)$$

Next, the S is computed by walking down the window list to evaluated $P_{obey}(k)$, which is the proportion of samples whose gene correlations are over 0.3 (i.e., $r_k^{m,i,j} > 0.3$), weighted by their corresponding correlation and $P_{violate}(k)$, which is the proportion of the rest of the samples at a given ranking window position k across all samples. The formulas are as follows:

$$P_{obey}(k) = \sum_{\substack{r_k \in condition \\ l \le k}} \frac{|r_l|}{N_R}, \ where\ N_R = \sum_{r_k \in condition} |r_l| \qquad (3)$$

$$P_{violate}(k) = \sum_{\substack{r_k \notin condition \\ l \le k}} \frac{1}{N_{violate}}, \ where\ N_{violate} = \sum_{r_k \notin condition} I(r_k \notin condition) \qquad (4)$$

doi:10.6342/NTU202300922

S is then defined as the maximum distance from zero of the substations of $P_{obey}(k)$ from $P_{violate}(k)$. If the samples with similar correlation values are not enriched at a particular miRNA expression interval, it means those gene pairs are not biologically relevant to compete with miRNA in a miRNA-gene triplet; that is, there is no ceRNA event observed in this triplet (Figure S5-11). Finally, we accessed the significance of an observed S by comparing it with the theoretical S computed by randomly permutating the expression of the candidate $miRNA_m$ 1,000 times to provide assessment of significance. When an entire sample of potential triplets is evaluated, these p-values will be adjusted by the false discovery rate, which provides the estimated probability of whether a triplet within an entire set of potential triplets is a false positive finding or not. After that, we report the ceRNA pairs with statistical significance of the observed S (e.g., adjusted p-value <0.05) as significant ceRNA interactions. For one miRNA, a low score S will be observed if the two target genes have no correlation within a

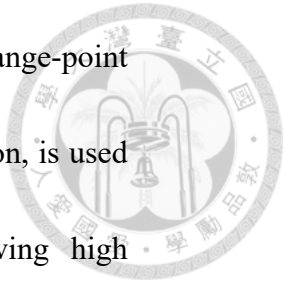
doi:10.6342/NTU202300922

specific range of miRNA expression. On the other hand, the score S will be high if a large proportion of the samples showed high correlation among the two target genes ($P_{obey}(k)$ is high). Consequently, we can use the score S to identify a ceRNA triplet when S is high, and a statistical approach was performed to ensure that S cannot be identified randomly.

5.5.5 The SegmentClustering method

The motivation of the segment clustering is to group the samples showing high correlation between the expression levels of the two target genes into one single cluster. In our analysis, we divided the samples into small groups (i.e., a window) to calculate the correlation values of the two target genes. Accordingly, we may identify several different groups showing high correlations that actually can be clustered into one single cluster because of their similar correlation values. Therefore, in this method, the concept of a circular binary

doi:10.6342/NTU202300922



segment algorithm, which was originally designed for change-point problems such as the identification of copy number variation, is used [266] to evaluate whether those small windows showing high correlation values should be clustered into a larger group. This method has been widely used in the analysis of copy number variations (CNVs) by many algorithms [267, 268]. Since several previous studies [236, 245, 246] indicated that ceRNA triplets may be observed in a specific range, the purpose of such an approach is to explore the clustering patterns of samples in terms of their gene-gene correlation values per triplet and then group samples with similar correlation values so that a certain miRNA expression interval with the highest correlation within a ceRNA pair can be observed. This algorithm starts with all segments (i.e., several intervals of rank-ordered miRNA expression) identified from the whole dataset. Similar to the previous method, the samples are sorted by the miRNA expression values. A recursive test for the change-points is calculated based on the correlation values of each

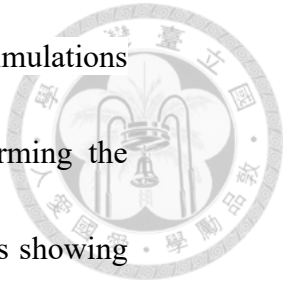
doi:10.6342/NTU202300922

gene-miRNA pair between each set of two neighboring regions of rank-ordered miRNA expression, and it stops when no significant changes can be found in any two segments. The maximal t-statistic with a permutation reference distribution is chosen to obtain the corresponding p-value. Let $X_1, \ldots, X_n$ be the correlation coefficients of two genes, which are indexed by the corresponding miRNA expression of the n samples being studied. The test statistic is given by $Z_c = max_{1 \le i < j \le n} |Z_{ij}|$, where Z_{ij} is the two-sample t-statistic to compare the mean of the correlation of two genes with the index, which refers to the position within the rank-ordered miRNA expression from i+1 to j, to the mean of the correlation of the rest of the genes. That is,

$$Z_{ij} = \{\frac{1}{j-i} + \frac{1}{n-j+i}\}^{-1/2} \left\{\frac{S_j - S_i}{j-i} - \frac{S_n - S_j + S_i}{n-j+i}\right\} \quad (5)$$

where $S_i = X_1 + \cdots + X_i$, $S_j = X_1 + \cdots + X_j$ $(1 \le i < j \le n)$ are the partial sums. If the p-values are smaller than a threshold level α (typically 0.01), a change is declared to be statistically significant. The locations of the change-points as the i and j that maximize the test

doi:10.6342/NTU202300922

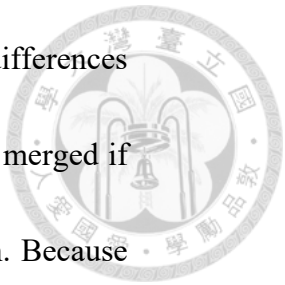


statistic are also estimated by using either Monte Carlo simulations [269] or the approximation approach [270]. After performing the segment clustering approach, only a few groups of samples showing high correlation remain for further analyses. We have a basic assumption here for the ceRNA triplet: the correlation values should be stable across the sliding window approach and thus the correlation values should not be bouncing up and down within a small sample size. Following this assumption, we performed the peak merging step to ensure two nearby peaks were merged into one under certain criteria.

5.5.6 The PeakMerging method

This method is designed to prevent smashed segments resulting from noise. First, two segments are merged when the difference of their correlation values is smaller than a predefined value, and whether the finalized segments are peaks or troughs is defined compared to the baseline. Then, the Fisher transformation is performed to test the

doi:10.6342/NTU202300922



difference between two adjacent peaks by comparing their differences against the null hypothesis. Two adjacent peaks are further merged if there is no statistically significant difference between them. Because we presumed it is less likely that two genes are highly correlated to compete for a miRNA at more than two miRNA expression intervals, the triplet was abandoned when more than two peaks occurred in such a relationship. Lastly, candidate ceRNAs are finally selected when their correlation pattern contains a peak with a correlation value over a predefined threshold value (0.3, 0.5, or 0.7).

The final output of the ceRNAMethod function provides information on each miRNA, its candidate ceRNA pairs, its peak position within the rank-ordered miRNA expression interval, and the number of samples involved in such ceRNA interactions (Figure S5-12).

doi:10.6342/NTU202300922

5.5.7 Downstream functional analyses

To characterize the biological functions and pathways of identified ceRNA pairs, we implemented the *ceRNAModule* function to generate ceRNA modules from their interaction networks. For the ceRNA modules, the *ceRNAFunction* function is used to perform functional enrichment analysis based on two ontology databases, the Gene Ontology database (GO, http://geneontology.org/) [271] and the Kyoto Encyclopedia of Genes and Genomes Pathway Database (KEGG, https://www.genome.jp/kegg/) [272]. Survival analysis has been widely used in biomedical fields to indicate whether the ceRNAs in the discovered interaction module are associated with the survival of cancer patients. Hence, in ceRNAR, we implemented the *ceRNASurvival* function to perform survival analysis. First, the risk score of each sample is calculated using a multivariate Cox model. All the samples (i.e., different patients) are divided into high or low risk groups based on their median risk scores. The Kaplan-Meier method

doi:10.6342/NTU202300922

with a log-rank test is used to test and visualize the difference between the high and low risk groups. Additionally, the *ceRNALocation* function allows users to further visualize the expression level of a specific miRNA when modulated by a specific ceRNA. We also implemented an integrated function (*ceRNAIntegrate*) to combine our results with other state-of-the-art tools, such as SPONGE [238] and JAMI [241], and a validation function (*ceRNAValidate*) based on the miRSponge database. The graphical outputs of the above-mentioned functions are presented in Figure S5-13 to S5-15.

5.5.8 Simulation study

In the simulation study, two types of expression data (ceRNA and miRNA) of 100 samples were generated. In each triplet, we presumed 10-40% of samples have correlated genes whereas the rest of the samples do not. The simulated expression profile of 100 samples of miRNA was distributed uniformly between 0 and 1 and was used to

doi:10.6342/NTU202300922

sort our samples. We simulated the gene expression profiles of these target gene pairs under two situations. The first is a naïve profile of the correlated genes simply generated from a multivariate normal distribution with a mean value of 0 and a covariance matrix whose entries are 0.9, while the expression distribution of the uncorrelated genes was specified by setting its mean and variance to zero. Another simulation scenario is mimicking a real-data profile (9,835 pan-cancer samples) generated from a multivariate normal distribution with a randomly selected mean value (±2, ±1 and 0), and a covariance matrix whose entries are also randomly selected from 0.3 to 0.9 (correlation level: all), while the expression distribution of the uncorrelated genes was specified by setting its mean (also randomly selected) at ±2, ±1, and 0, and its variance (randomly selected) at 0 to 0.2 (Figure S5-16). We also specified three correlation levels (high = 0.8-0.9, mid = 0.5-0.7, and low = 0.3-0.4) and randomly selected values within them to construct the covariance matrix. The normality test for sample

doi:10.6342/NTU202300922

distribution of each gene was conducted by the Anderson-Darling method [273].

Five scenarios were considered and are summarized in Table S5-1. The first three scenarios were designed for a single peak (i.e., the highest correlation value of the target genes compared to the baseline) occurring at different locations (i.e., different miRNA expression intervals): lower miRNA expression (left, scenario 3), higher miRNA expression (right, scenario 2), and average miRNA expression (center, scenario 1), which represents specific correlation coefficient values existing at different expression levels of miRNA. The fourth scenario was designed for considering the occurrence of two peaks. A null scenario was also compared to test the validity of the other conditions. Two parameters were set under each scenario: window size (10%, 20%, 30%, and 40%) and the threshold at which to define a peak (0.3, 0.5, and 0.7). Each scenario was simulated 1,000 times. We conducted these simulations using either the complete or fast versions of the

doi:10.6342/NTU202300922

algorithm, the latter of which reduced computational complexity by omitting the random walk step. The proportion of correlated samples (10%, 20%, 30%, and 40%) was also analyzed.

5.5.9 Real data application for tools comparison and validation

In order to compare many aspects, including the computational cost and the quality of the results, of ceRNAR with the other tools (SPONGE [238], CERNIA [239] and JAMI [241]), which similarly used miRNA and paired target gene expression to identify gene-miRNA-gene triplets, we used subsets of the TCGA pan-cancer atlas with different sample sizes and gene/triplet numbers. We ran these tools with default parameter settings in parallel processing with 8 cores with varying sample sizes and numbers of genes.

Regarding real data application, we retrieved two sequencing datasets (TCGA-LUAD and TCGA-LUSC) from TCGA [274] to demonstrate the potential application of our proposed algorithm. These

doi:10.6342/NTU202300922

datasets are all retrieved from public domains (GDG data portal: https://portal.gdc.cancer.gov/). The TCGA-LUAD dataset contained 510 samples from lung adenocarcinoma patients, and the TCGA-LUSC dataset contained 476 samples from squamous cell carcinoma patients. To validate the results, the potential ceRNAs identified by our tool were validated experimentally using the miRsponge database [275], and the ceRNAs shared by these two datasets were further analyzed.

doi:10.6342/NTU202300922

5.7 Figures

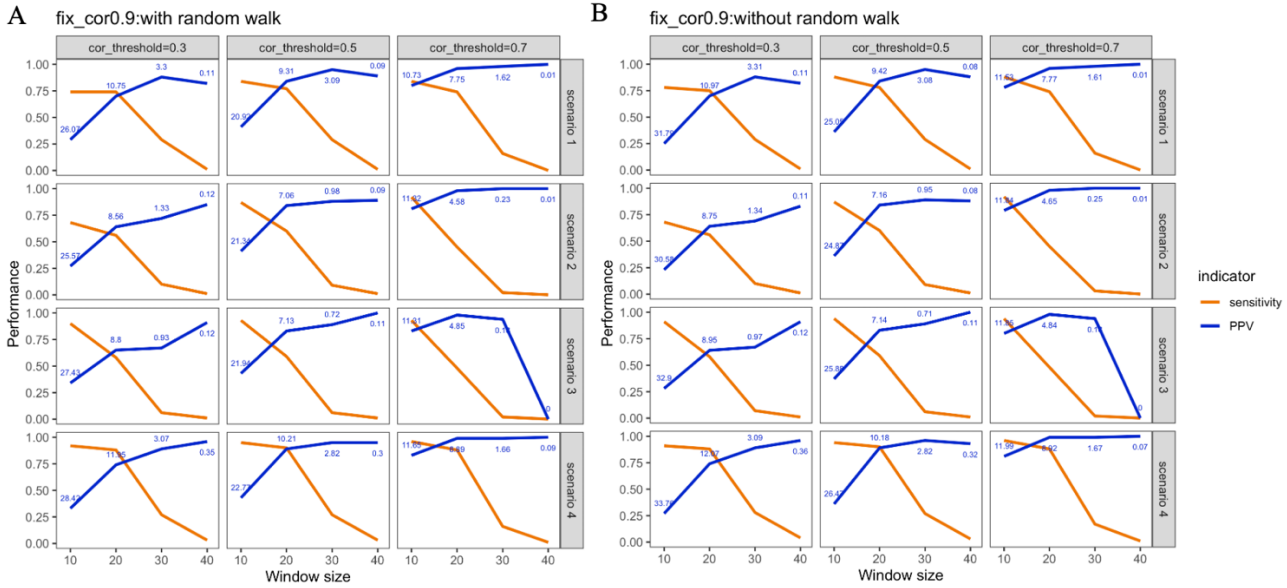
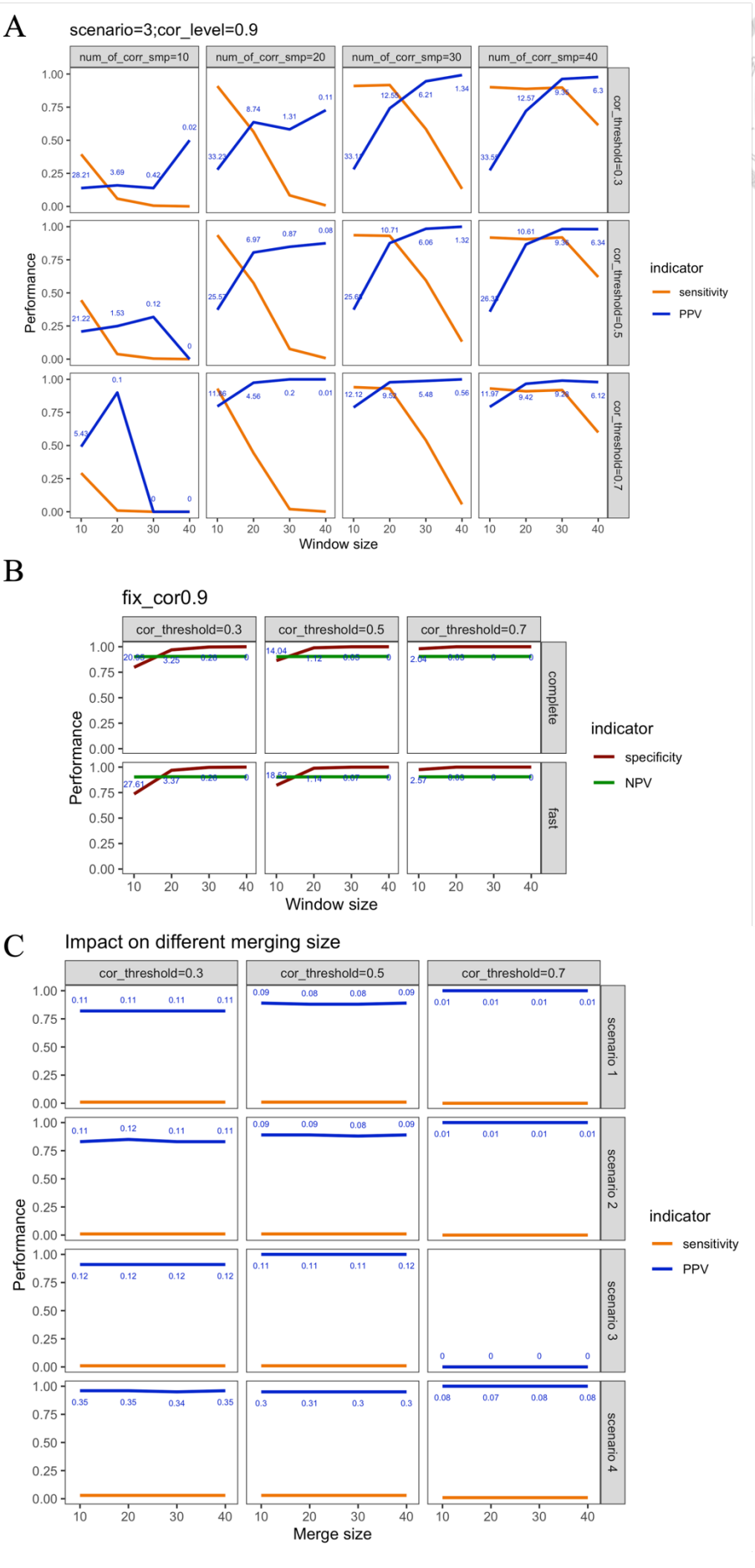


Figure 5-1. Performance of our proposed method in four scenarios (1 to 4). (A) The complete version. (B) The fast version. The numbers in blue represent the average number of identified ceRNA-miRNA triplets after 100 simulations.

doi:10.6342/NTU202300922



doi:10.6342/NTU202300922

Figure 5-2. Performance of our proposed method in three extended cases. (A) Different proportions of correlated samples in scenario 3 using the fast version. (B) Comparison between complete and fast versions in the null scenario (i.e., scenario 5). (C) Different distances for merging peaks when window size is equal to 40 using the complete version. The numbers in blue represent the number of identified ceRNA-miRNA triplets.

doi:10.6342/NTU202300922

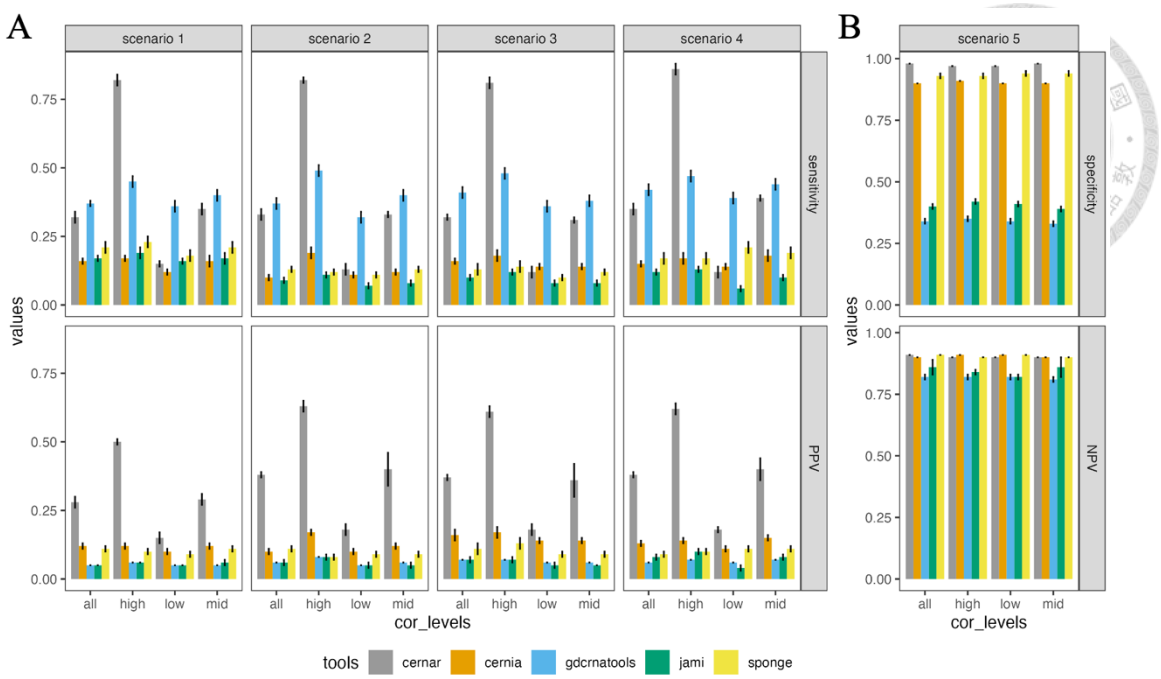


Figure 5-3. Performance comparisons between ceRNIA, JAMI, SPONGE, GDCRNATools and ceRNAR packages using synthetic data that mimic the real-world case. (A) Sensitivity and PPV of various tools at different correlation levels for 100 samples of 105 pairs of target genes under four scenarios (1 to 4). (B) Specificity and NPV of various tools at different correlation levels for 100 sample of 105 pairs of target genes under null scenario (scenario 5). "cor_levels" represents four different correlation levels (all: 0.3-0.9; high: 0.8-0.9; low: 0.3-0.4; mid: 0.5-0.7).

doi:10.6342/NTU202300922

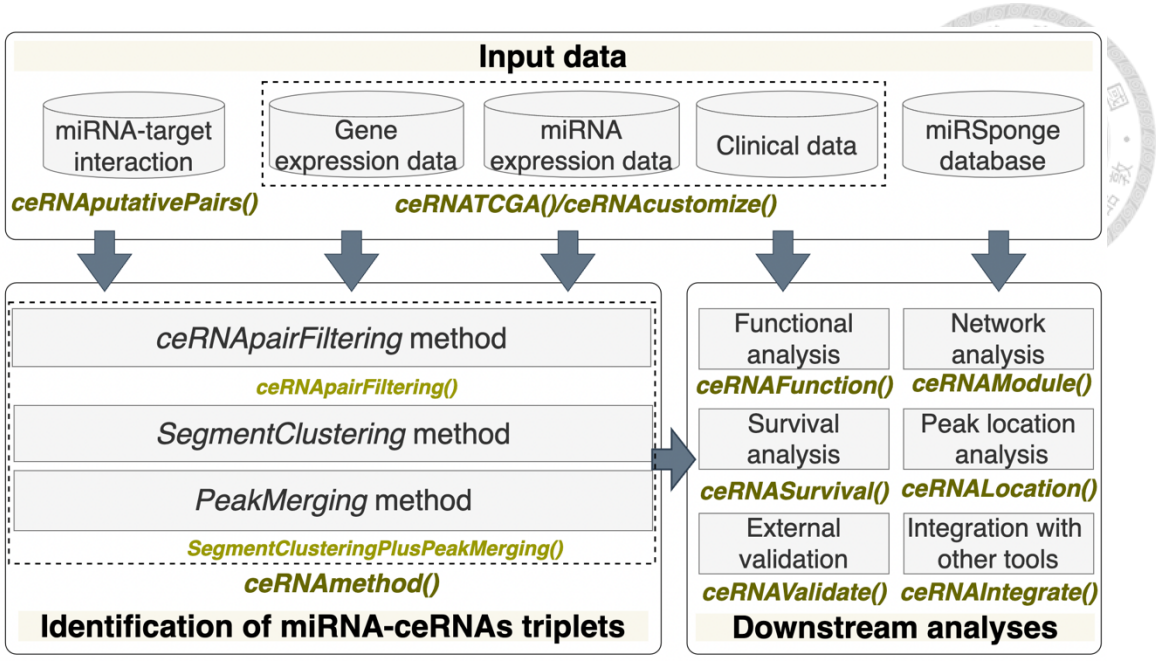


Figure 5-4. An overview of our ceRNAR package. This package includes three main parts. First, "Input data" has three functions: *ceRNAputativePairs()* for extracting curated miRNA-target lists, *ceRNATCGA()* for retrieving TCGA data, and ceRNAcustomize() for importing customized data. Second, "Identification of miRNA-ceRNA triplets" involves *ceRNApairFiltering()* and *SegmentClusteringPlusPeakMerging()* and is wrapped into a *ceRNAmethod()* function. Lastly, "Downstream analyses" contains six functions for different tasks, that is, *ceRNAFunction()*, *ceRNAModule()*, *ceRNASurvival()*, *ceRNALocation()*, *ceRNAValidate()* and *ceRNAIntegrate()*.

doi:10.6342/NTU202300922

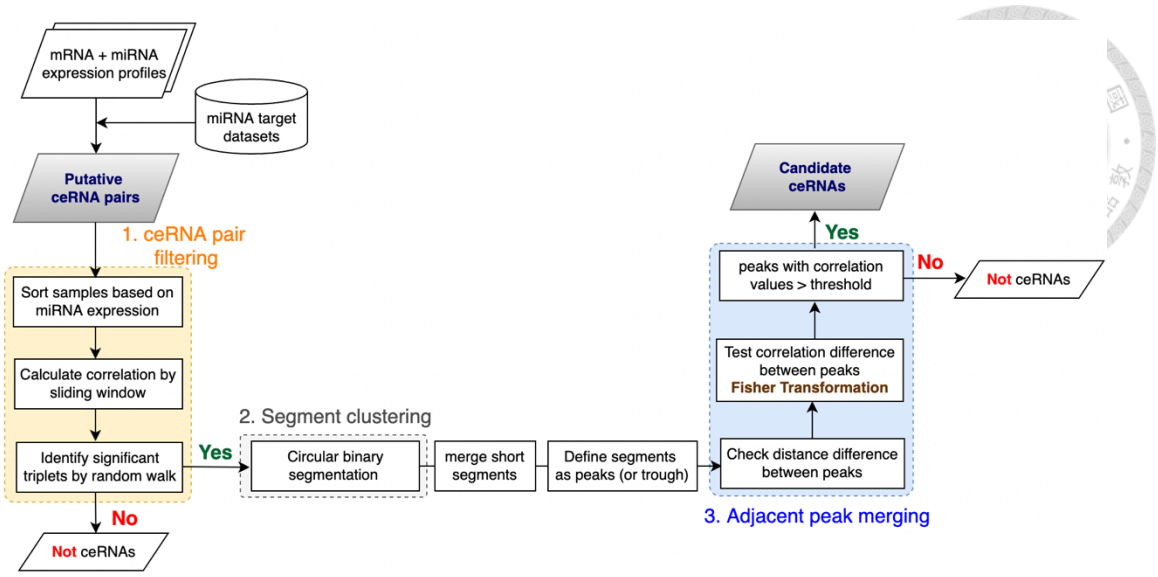


Figure 5-5. Three main modules involved in the identification of ceRNA-miRNA triplets. First, 'ceRNA pair filtering' contains sample sorting based on a miRNA expression vector, data augmentation in terms of correlation calculation through a sliding window, and permutation test by random walk. Next, 'Segment clustering' is based on circular binary segmentation to identify certain miRNA expression intervals (i.e., segments) with the highest correlation between potential ceRNA pairs. Short segments defined by our criteria are further merged and whether a segment is a peak or a trough is defined. Lastly, 'Adjacent peak merging' checks whether a triplet is a candidate ceRNA through two filtering conditions.

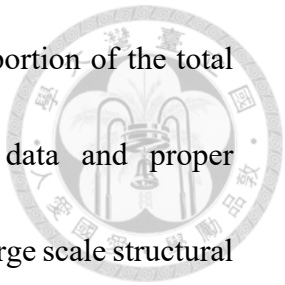
doi:10.6342/NTU202300922

6. Discussion

With the rapid development of the next-generation sequencing techniques and bioinformatics algorithms, high-throughput biological data have become less expensive and more accessible such that more studies can explore to a greater level of the genetic impacts on various diseases. The public accessibility of high-throughput genomic data further fulfils the possibility of large-scale and integrative genomic studies to deeply address many biological issues. Currently, several public genomic data repositories have been established and are freely assessable for the public; for example, GEO created and maintained by NCBI have stored thousands of genomic data generated by any kind of sequencing techniques [231]; and TCGA is the one of larger pan-cancer genomics project which contains multi-omics data and clinical information in 33 cancer type [276]. Leveraging the power of these public data resources, in this dissertation, we have identified several genetic features at different genomic levels as prognostic/therapeutic targets in multiple cancers to pilot the clinical applications, hoping to achieve the ultimate goals of precision medicine.

The huge number of genetic markers at different genetic levels may hamper the process of identifying prognostic/therapeutic biomarkers. Hence, in this dissertation, we used two different approaches to address this issue: (1) summarized values for tumor-specific features, and (2) statistical methods including machine learning

doi:10.6342/NTU202300922



approaches for feature selection. The former, for example, the proportion of the total amount of TILs can be estimated through gene expression data and proper bioinformatics tools [277]; and HRD score can be counted through large scale structural variation data and its corresponding tools [278]. Also, through the definition of functional damage and curated databases, we can narrow down the highly pathogenic variant locus for further biological interpretation [279]. As for the latter one, traditional both parametric or non-parametric statistical tests like t-test or the Wilcoxon signed-rank test have been widely applied into the identification of differentially expressed genes that can be considered as the unique signatures for specific disease patterns [280]. Several regression-based methods, such as SKAT-O, have also been developed to identify phenotype-specific variants [281]. Recently, the machine learning approach has been revolutionized the fields of the discovery of potential biomarkers that are of clinical relevance. For example, tree-based approaches such as random forest and XGBoost, provide feature selection strategies to remain the most relevant features for either subtype classification or prognostic prediction [282]. Although such a quantitative or statistical approach does accelerate the speed of the biomarker discover, it still needs lots of supports from either biological experiments or curated databases such as QIAGEN Ingenuity Pathway Analysis (IPA) to warrant the applicability of the identified biomarkers [283]. Nevertheless, along with the more large scale of genomic

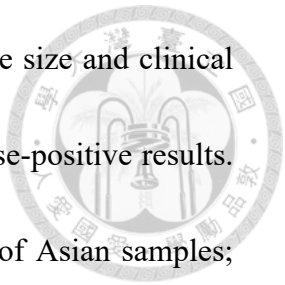
doi:10.6342/NTU202300922

data, the development of advanced statistical approaches also plays an important role in the understanding of cancer genetics.

Aside from considering only single omics data, more and more studies now have a focus on at least two layers of genomic features or even multi-omics in their study to gain comprehensive insights into the whole genomic regulation system and mechanism that cause the diseases. Over the past decades, many non-coding RNAs have identified and confirmed their role in post-transcriptional regulation of protein-coding genes and such RNAs include microRNAs, long non-coding RNAs, circular RNAs [284]. Although they have no direct impact on protein function, they do regulate the expression of protein-coding genes by interfering with the physical binding of translation-related proteins and further affecting the amount of protein that is translated [285]. Therefore, in this dissertation, we also considered the regulatory events to understand the etiology of diseases, including cancers, hoping to unveil the novel pathogenic mechanism for the better design of the medical intervention.

However, several drawbacks exist in this dissertation. First, batch effects across data from different studies and technical variance across different platforms may affect the robustness of the model building and the accuracy of the results. Although several new statistical approaches, such as COMBAT and XPN, have been developed to minimize such effects [286], the result interpretation still needs to be cautioned for

doi:10.6342/NTU202300922



further clinical application. Second, public data with limited sample size and clinical information for specific cancer types may tend to provide more false-positive results. For instance, our HRD study has suffered from a limited number of Asian samples; hence, a larger scale of population-based studies is necessary for comprehensive investigation for Asian unique HRD patterns. Lastly, some limitations have been observed in the well-established methodologies used in this dissertation. GA-XGBoost algorithm tended to generate different optimal genetic combinations which may not be a robust result. ESTIMATE algorithm can only provide a total amount of TILs while CIBERSORT can only offer the proportion of each TILs not the absolute quantity of each of them. Therefore, the improvement of current methodologies is warranted for the accurate estimation of genomic features and robust results for biological interpretation and clinical usage.

With the advance of techniques and computational knowledge, future works should focus on three aspects to speed up the field of cancer prognosis. First, the resolution of the data quality can be improved through single-cell techniques [287]. Second, considering all genomic features, including DNA, RNA and protein, in cancer prognosis is a current trend to provide deep insights into the understanding of cancer mechanism and the discovery the novel prognostic markers [288]. Lastly, many new artificial intelligence frames, such as CNN and RNN, can further assist the

doi:10.6342/NTU202300922

classification and biomarker discovery through multi-omics data [289]. In addition, advanced computational techniques, such as GPU or quantum computers, can accelerate the analytic procedure in the field of medical genomics [290, 291].

In conclusion, in this dissertation, we used different levels of genetic features: (1) DNA level includes homologous recombination deficiency (HRD), (2) RNA level includes whole transcriptomes and tumor-infiltrating lymphocytes (TILs) and (3) post-transcription regulation (the identification of ceRNA); and used a wide range of well-established or novel methodologies to study cancer prognosis prediction in multiple cancer types including ovarian cancer, hepatocellular carcinomas (cancer subtypes) and pan-cancer (racial difference), hoping to achieve the ultimate goals of precision medicine.

doi:10.6342/NTU202300922

7. References

1. Mackillop WJ. The importance of prognosis in cancer medicine. TNM Online. 2003.

2. Parikh RB, Manz C, Chivers C, Regli SH, Braun J, Draugelis ME, et al. Machine learning approaches to predict 6-month mortality among patients with cancer. JAMA network open. 2019;2(10):e1915997-e.

3. Cox DR, Oakes D. Analysis of survival data: Chapman and Hall/CRC; 2018.

4. Dos Reis FJC, Wishart GC, Dicks EM, Greenberg D, Rashbass J, Schmidt MK, et al. An updated PREDICT breast cancer prognostication and treatment benefit prediction model with independent validation. Breast Cancer Research. 2017;19(1):1-13.

5. Stone P, Lund S. Predicting prognosis in patients with advanced cancer. Annals of oncology. 2007;18(6):971-6.

6. Watson JD, Crick FHJN. Molecular structure of nucleic acids: a structure for deoxyribose nucleic acid. 1953;171(4356):737-8.

7. Watson J. and FHC Crick. 1953b. Genetical implications of the structure of deoxyribonucleic acid. Nature, 171: 964-967. 1953.

8. Mullis K, Faloona F, Scharf S, Saiki R, Horn G, Erlich H, editors. Specific enzymatic amplification of DNA in vitro: the polymerase chain reaction. Cold Spring

doi:10.6342/NTU202300922

Harbor symposia on quantitative biology; 1986: Cold Spring Harbor Laboratory Press.

9. Sanger F, Nicklen S, Coulson ARJPotnaos. DNA sequencing with chain-terminating inhibitors. 1977;74(12):5463-7.

10. Wee Y, Bhyan SB, Liu Y, Lu J, Li X, Zhao M. The bioinformatics tools for the genome assembly and analysis based on third-generation sequencing. Briefings in functional genomics. 2019;18(1):1-12.

11. Olsen TK, Baryawno N. Introduction to single-cell RNA sequencing. Current protocols in molecular biology. 2018;122(1):e57.

12. Dalma-Weiszhausz DD, Warrington J, Tanimoto EY, Miyada CG. [1] The Affymetrix GeneChip® Platform: An Overview. Methods in enzymology. 2006;410:3-28.

13. Slatko BE, Gardner AF, Ausubel FM. Overview of next-generation sequencing technologies. Current protocols in molecular biology. 2018;122(1):e59.

14. Sheridan C. Erratum: Illumina claims $1,000 genome win. Nature biotechnology. 2014;32(2):115.

15. Balzer S, Malde K, Lanzén A, Sharma A, Jonassen I. Characteristics of 454 pyrosequencing data—enabling realistic simulation with flowsim. Bioinformatics. 2010;26(18):i420-i5.

doi:10.6342/NTU202300922

16. Diekstra A, Bosgoed E, Rikken A, van Lier B, Kamsteeg E-J, Tychon M, et al. Translating sanger-based routine DNA diagnostics into generic massive parallel ion semiconductor sequencing. Clinical chemistry. 2015;61(1):154-62.

17. Serratì S, De Summa S, Pilato B, Petriella D, Lacalamita R, Tommasi S, et al. Next-generation sequencing: advances and applications in cancer diagnosis. OncoTargets and therapy. 2016;9:7355.

18. Carlomagno N, Incollingo P, Tammaro V, Peluso G, Rupealta N, Chiacchio G, et al. Diagnostic, predictive, prognostic, and therapeutic molecular biomarkers in third millennium: a breakthrough in gastric cancer. 2017;2017.

19. Shaw A, Bradley MD, Elyan S, Kurian KMJB. Tumour biomarkers: diagnostic, prognostic, and predictive. 2015;351.

20. Salagierski M, Schalken JAJTJou. Molecular diagnosis of prostate cancer: PCA3 and TMPRSS2: ERG gene fusion. 2012;187(3):795-801.

21. Alain PJJoL, Medicine P. IL-6 and VEGF-A, novel prognostic biomarkers for ovarian cancer? 2018. 2018;3(5).

22. Peters G, Gongoll S, Langner C, Mengel M, Piso P, Klempnauer J, et al. IGF-1R, IGF-1 and IGF-2 expression as potential prognostic and predictive markers in colorectal-cancer. 2003;443(2):139-45.

doi:10.6342/NTU202300922

23. Makałowski WJABP. The human genome structure and organization. 2001;48(3):587-98.

24. Davisson MT, Bergstrom DE, Reinholdt LG, Donahue LRJCpimb. Discovery genetics: the history and future of spontaneous mutation research. 2012;2(2):103-18.

25. Karki R, Pandya D, Elston RC, Ferlini CJBmg. Defining "mutation" and "polymorphism" in the era of personal genomics. 2015;8(1):1-7.

26. Wakeley J, Nielsen R, Liu-Cordero SN, Ardlie KJTAJoHG. The discovery of single-nucleotide polymorphisms—and inferences about human demographic history. 2001;69(6):1332-47.

27. Cheng Y, Jiang T, Zhu M, Li Z, Zhang J, Wang Y, et al. Risk assessment models for genetic risk predictors of lung cancer using two-stage replication for Asian and European populations. 2017;8(33):53959.

28. Kuiper RP, Ligtenberg MJ, Hoogerbrugge N, van Kessel AGJCoig, development. Germline copy number variation and cancer risk. 2010;20(3):282-9.

29. Kosti I, Jain N, Aran D, Butte AJ, Sirota MJSr. Cross-tissue analysis of gene and protein expression in normal and cancer tissues. 2016;6(1):1-16.

30. Talhouk A, George J, Wang C, Budden T, Tan TZ, Chiu DS, et al. Development and validation of the gene expression predictor of high-grade serous ovarian

doi:10.6342/NTU202300922

carcinoma molecular subTYPE (PrOTYPE). Clinical Cancer Research. 2020;26(20):5411-23.

31. Mook S, Van't Veer LJ, Rutgers EJ, Piccart-Gebhart MJ, Cardoso FJCG-P. Individualization of therapy using MammaPrint® ì: From development to the MINDACT Trial. 2007;4(3):147-55.

32. Kaklamani VJEromd. A genetic signature can predict prognosis and response to therapy in breast cancer: Onco type DX. 2006;6(6):803-9.

33. Perou CM, Sørlie T, Eisen MB, Van De Rijn M, Jeffrey SS, Rees CA, et al. Molecular portraits of human breast tumours. 2000;406(6797):747-52.

34. Sunakawa Y, Lenz H-J. Molecular classification of gastric adenocarcinoma: translating new insights from the cancer genome atlas research network. Current treatment options in oncology. 2015;16(4):17.

35. Collisson E, Campbell J, Brooks A, Berger A, Lee W, Chmielecki J, et al. Comprehensive molecular profiling of lung adenocarcinoma: The cancer genome atlas research network. Nature. 2014;511(7511):543-50.

36. Ricketts CJ, De Cubas AA, Fan H, Smith CC, Lang M, Reznik E, et al. The cancer genome atlas comprehensive molecular characterization of renal cell carcinoma. Cell reports. 2018;23(1):313-26. e5.

doi:10.6342/NTU202300922

37. Oak N, Cherniack AD, Mashl RJ, Hirsch FR, Ding L, Beroukhim R, et al. Ancestry-specific predisposing germline variants in cancer. Genome medicine. 2020;12:1-15.

38. Karczewski K, Francioli LJML. The genome aggregation database (gnomAD). 2017.

39. Koch LJNRG. Exploring human genomic diversity with gnomAD. 2020;21(8):448-.

40. Collins RL, Brand H, Karczewski KJ, Zhao X, Alföldi J, Francioli LC, et al. A structural variation reference for medical and population genetics. 2020;581(7809):444-51.

41. Yuan J, Hu Z, Mahal BA, Zhao SD, Kensler KH, Pi J, et al. Integrated analysis of genetic ancestry and genomic alterations across cancers. Cancer cell. 2018;34(4):549-60. e9.

42. Caswell-Jin JL, Gupta T, Hall E, Petrovchich IM, Mills MA, Kingham KE, et al. Racial/ethnic differences in multiple-gene sequencing results for hereditary cancer risk. Genetics in Medicine. 2018;20(2):234-9.

43. Lord CJ, Ashworth A. BRCAness revisited. Nature Reviews Cancer. 2016;16(2):110-20.

doi:10.6342/NTU202300922

44. Thompson LH, Schild D. Homologous recombinational repair of DNA ensures mammalian chromosome stability. Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis. 2001;477(1-2):131-53.

45. Abkevich V, Timms K, Hennessy B, Potter J, Carey M, Meyer L, et al. Patterns of genomic loss of heterozygosity predict homologous recombination repair defects in epithelial ovarian cancer. 2012;107(10):1776-82.

46. Birkbak NJ, Wang ZC, Kim J-Y, Eklund AC, Li Q, Tian R, et al. Telomeric allelic imbalance indicates defective DNA repair and sensitivity to DNA-damaging agents. Cancer discovery. 2012;2(4):366-75.

47. Popova T, Manié E, Rieunier G, Caux-Moncoutier V, Tirapo C, Dubois T, et al. Ploidy and large-scale genomic instability consistently identify basal-like breast carcinomas with BRCA1/2 inactivation. Cancer research. 2012;72(21):5454-62.

48. Ang JE, Gourley C, Powell CB, High H, Shapira-Frommer R, Castonguay V, et al. Efficacy of chemotherapy in BRCA1/2 mutation carrier ovarian cancer in the setting of PARP inhibitor resistance: a multi-institutional study. Clinical Cancer Research. 2013;19(19):5485-93.

49. Audeh MW, Carmichael J, Penson RT, Friedlander M, Powell B, Bell-McGuinn KM, et al. Oral poly (ADP-ribose) polymerase inhibitor olaparib in patients with

doi:10.6342/NTU202300922

BRCA1 or BRCA2 mutations and recurrent ovarian cancer: a proof-of-concept trial. The lancet. 2010;376(9737):245-51.

50. Tutt A, Robson M, Garber JE, Domchek SM, Audeh MW, Weitzel JN, et al. Oral poly (ADP-ribose) polymerase inhibitor olaparib in patients with BRCA1 or BRCA2 mutations and advanced breast cancer: a proof-of-concept trial. The Lancet. 2010;376(9737):235-44.

51. Mateo J, Carreira S, Sandhu S, Miranda S, Mossop H, Perez-Lopez R, et al. DNA-repair defects and olaparib in metastatic prostate cancer. New England Journal of Medicine. 2015;373(18):1697-708.

52. Yarchoan M, Myzak MC, Johnson III BA, De Jesus-Acosta A, Le DT, Jaffee EM, et al. Olaparib in combination with irinotecan, cisplatin, and mitomycin C in patients with advanced pancreatic cancer. Oncotarget. 2017;8(27):44073.

53. Jeggo PA, Pearl LH, Carr AM. DNA repair, genome stability and cancer: a historical perspective. Nature Reviews Cancer. 2016;16(1):35-42.

54. Ellis L, Canchola AJ, Spiegel D, Ladabaum U, Haile R, Gomez SL. Racial and ethnic disparities in cancer survival: the contribution of tumor, sociodemographic, institutional, and neighborhood characteristics. Journal of Clinical Oncology. 2018;36(1):25.

doi:10.6342/NTU202300922

55. Siegel R, Miller K, Fuchs H, Jemal A. Cancer statistics, 2021 (vol 71, pg 7, 2021). CA-A CANCER JOURNAL FOR CLINICIANS. 2021.

56. Özdemir BC, Dotto G-P. Racial differences in cancer susceptibility and survival: more than the color of the skin? Trends in cancer. 2017;3(3):181-97.

57. Wallace TA, Martin DN, Ambs S. Interactions among genes, tumor biology and the environment in cancer health disparities: examining the evidence on a national and global scale. Carcinogenesis. 2011;32(8):1107-21.

58. Daly B, Olopade OI. A perfect storm: how tumor biology, genomics, and health care delivery patterns collide to create a racial survival disparity in breast cancer and proposed interventions for change. CA: a cancer journal for clinicians. 2015;65(3):221-38.

59. Conti DV, Darst BF, Moss LC, Saunders EJ, Sheng X, Chou A, et al. Trans-ancestry genome-wide association meta-analysis of prostate cancer identifies new susceptibility loci and informs genetic risk prediction. Nature genetics. 2021;53(1):65-75.

60. Sinha S, Mitchell KA, Zingone A, Bowman E, Sinha N, Schäffer AA, et al. Higher prevalence of homologous recombination deficiency in tumors from African Americans versus European Americans. Nature Cancer. 2020;1(1):112-21.

doi:10.6342/NTU202300922

61. Malone ER, Oliva M, Sabatini PJ, Stockley TL, Siu LL. Molecular profiling for precision cancer therapies. Genome medicine. 2020;12(1):1-19.

62. Syed YY. Oncotype DX Breast Recurrence Score®: A Review of its Use in Early-Stage Breast Cancer. Molecular Diagnosis & Therapy. 2020;24(5):621-32.

63. Slodkowska EA, Ross JS. MammaPrint™ 70-gene signature: another milestone in personalized medical care for breast cancer patients. Expert review of molecular diagnostics. 2009;9(5):417-22.

64. Arora S, Balasubramaniam S, Zhang H, Berman T, Narayan P, Suzman D, et al. FDA approval summary: olaparib monotherapy or in combination with Bevacizumab for the maintenance treatment of patients with advanced ovarian cancer. The Oncologist. 2021;26(1):e164-e72.

65. Takeda M, Takahama T, Sakai K, Shimizu S, Watanabe S, Kawakami H, et al. Clinical Application of the FoundationOne CDx Assay to Therapeutic Decision-Making for Patients with Advanced Solid Tumors. The oncologist. 2021;26(4):e588-e96.

66. Tung NM, Garber JE. BRCA 1/2 testing: therapeutic implications for breast cancer management. British journal of cancer. 2018;119(2):141-52.

doi:10.6342/NTU202300922

67. Ray-Coquard I, Pautier P, Pignata S, Pérol D, González-Martín A, Berger R, et al. Olaparib plus bevacizumab as first-line maintenance in ovarian cancer. New England Journal of Medicine. 2019;381(25):2416-28.

68. Li JZ, Absher DM, Tang H, Southwick AM, Casto AM, Ramachandran S, et al. Worldwide human relationships inferred from genome-wide patterns of variation. science. 2008;319(5866):1100-4.

69. Barbujani G, Colonna V. Human genome diversity: frequently asked questions. Trends in Genetics. 2010;26(7):285-95.

70. Phan VH, Tan C, Rittau A, Xu H, McLachlan AJ, Clarke SJ. An update on ethnic differences in drug metabolism and toxicity from anti-cancer drugs. Expert opinion on drug metabolism & toxicology. 2011;7(11):1395-410.

71. Liu J, Lichtenberg T, Hoadley KA, Poisson LM, Lazar AJ, Cherniack AD, et al. An integrated TCGA pan-cancer clinical data resource to drive high-quality survival outcome analytics. Cell. 2018;173(2):400-16. e11.

72. Mulligan JM, Hill LA, Deharo S, Irwin G, Boyle D, Keating KE, et al. Identification and validation of an anthracycline/cyclophosphamide–based chemotherapy response assay in breast cancer. Journal of the National Cancer Institute. 2014;106(1):djt335.

doi:10.6342/NTU202300922

73. Knijnenburg TA, Wang L, Zimmermann MT, Chambwe N, Gao GF, Cherniack AD, et al. Genomic and molecular landscape of DNA damage repair deficiency across The Cancer Genome Atlas. Cell reports. 2018;23(1):239-54. e6.

74. Lee S, Wu MC, Lin X. Optimal tests for rare variant effects in sequencing association studies. Biostatistics. 2012;13(4):762-75.

75. Blumhagen RZ, Schwartz DA, Langefeld CD, Fingerlin TE. Identification of Influential Variants in Significant Aggregate Rare Variant Tests. Human Heredity. 2020;85(1):11-23.

76. Ioannidis NM, Rothstein JH, Pejaver V, Middha S, McDonnell SK, Baheti S, et al. REVEL: an ensemble method for predicting the pathogenicity of rare missense variants. The American Journal of Human Genetics. 2016;99(4):877-85.

77. Iacocca MA, Chora JR, Carrié A, Freiberger T, Leigh SE, Defesche JC, et al. ClinVar database of global familial hypercholesterolemia-associated DNA variants. Human mutation. 2018;39(11):1631-40.

78. Sztupinszki Z, Diossy M, Krzystanek M, Reiniger L, Csabai I, Favero F, et al. Migrating the SNP array-based homologous recombination deficiency measures to next generation sequencing data of breast cancer. NPJ breast cancer. 2018;4(1):1-4.

79. Kassambara A, Kassambara MA. Package 'ggpubr'. 2020.

doi:10.6342/NTU202300922

80. Ahlmann-Eltze C, Patil I. ggsignif: R Package for Displaying Significance Brackets for'ggplot2'. 2021.

81. Da Costa AA, Do Canto LM, Larsen SJ, Ribeiro ARG, Stecca CE, Petersen AH, et al. Genomic profiling in ovarian cancer retreated with platinum based chemotherapy presented homologous recombination deficiency and copy number imbalances of CCNE1 and RB1 genes. BMC cancer. 2019;19(1):1-10.

82. Therneau TM, Lumley T. Package ‘survival’. R Top Doc. 2015;128(10):28-33.

83. Kassambara A, Kosinski M, Biecek P, Fabian S. Package ‘survminer’. Drawing Survival Curves using ‘ggplot2’(R package version 03 1). 2017.

84. Jiang Y, Dang S, Yang L, Han Y, Zhang Y, Mu T, et al. Association between homologous recombination deficiency and tumor mutational burden in lung cancer. American Society of Clinical Oncology; 2020.

85. Raymond C, Hernandez J, Brobey R, Wang Y, Potts K, Garg K, et al. Detection of HRD gene mutations and copy number changes in cfDNA from prostate cancer patients. American Society of Clinical Oncology; 2017.

86. Chao A, Lai C-H, Wang T-H, Jung S-M, Lee Y-S, Chang W-Y, et al. Genomic scar signatures associated with homologous recombination deficiency predict adverse clinical outcomes in patients with ovarian clear cell carcinoma. Journal of molecular medicine. 2018;96(6):527-36.

doi:10.6342/NTU202300922

87. Patel JN, Braicu I, Timms KM, Solimeno C, Tshiaba P, Reid J, et al. Characterisation of homologous recombination deficiency in paired primary and recurrent high-grade serous ovarian cancer. British journal of cancer. 2018;119(9):1060-6.

88. Wang Y, Ung MH, Cantor S, Cheng C. Computational investigation of homologous recombination DNA repair deficiency in sporadic breast cancer. Scientific reports. 2017;7(1):1-15.

89. Byun JS, Singhal SK, Park S, Yi DI, Yan T, Caban A, et al. Racial differences in the association between luminal master regulator gene expression levels and breast cancer survival. Clinical Cancer Research. 2020;26(8):1905-14.

90. Chien J, Sicotte H, Fan J-B, Humphray S, Cunningham JM, Kalli KR, et al. TP53 mutations, tetraploidy and homologous recombination repair defects in early stage high-grade serous ovarian cancer. Nucleic acids research. 2015;43(14):6945-58.

91. Chen Y, Cui L, Zhang B, Zhao X, Xu B. Efficacy of platinum-based chemotherapy in advanced triple-negative breast cancer in association with homologous recombination deficiency. Wolters Kluwer Health; 2021.

92. Vallböhmer D, Lenz H-J. Epidermal growth factor receptor as a target for chemotherapy. Clinical colorectal cancer. 2005;5:S19-S27.

doi:10.6342/NTU202300922

93. Ju B-G, Lunyak VV, Perissi V, Garcia-Bassets I, Rose DW, Glass CK, et al. A topoisomerase IIß-mediated dsDNA break required for regulated transcription. science. 2006;312(5781):1798-802.

94. Jin MH, Oh D-Y. ATM in DNA repair in cancer. Pharmacology & therapeutics. 2019;203:107391.

95. Shahriyari L, Abdel-Rahman M, Cebulla C. BAP1 expression is prognostic in breast and uveal melanoma but not colon cancer and is highly positively correlated with RBM15B and USP19. PLoS One. 2019;14(2):e0211507.

96. Helgason H, Rafnar T, Olafsdottir HS, Jonasson JG, Sigurdsson A, Stacey SN, et al. Loss-of-function variants in ATM confer risk of gastric cancer. Nature genetics. 2015;47(8):906-10.

97. Cai H, Jing C, Chang X, Ding D, Han T, Yang J, et al. Mutational landscape of gastric cancer and clinical application of genomic profiling based on target next-generation sequencing. Journal of translational medicine. 2019;17(1):1-12.

98. Cybulski C, Kluźniak W, Huzarski T, Wokołorczyk D, Kashyap A, Rusak B, et al. The spectrum of mutations predisposing to familial breast cancer in Poland. International journal of cancer. 2019;145(12):3311-20.

doi:10.6342/NTU202300922

99. Doherty JA, Weiss NS, Fish S, Fan W, Loomis MM, Sakoda LC, et al. Polymorphisms in nucleotide excision repair genes and endometrial cancer risk. Cancer Epidemiology and Prevention Biomarkers. 2011;20(9):1873-82.

100. Karpińska-Kaczmarczyk K, Lewandowska M, Ławniczak M, Białek A, Urasińska E. Expression of mismatch repair proteins in early and advanced gastric cancer in Poland. Medical science monitor: international medical journal of experimental and clinical research. 2016;22:2886.

101. Nemtsova MV, Kalinkin AI, Kuznetsova EB, Bure IV, Alekseeva EA, Bykov II, et al. Clinical relevance of somatic mutations in main driver genes detected in gastric cancer patients by next-generation DNA sequencing. Scientific reports. 2020;10(1):1-11.

102. Howell DC. Median absolute deviation. Encyclopedia of statistics in behavioral science. 2005.

103. Sherman RM, Forman J, Antonescu V, Puiu D, Daya M, Rafaels N, et al. Assembly of a pan-genome from deep sequencing of 910 humans of African descent. Nature genetics. 2019;51(1):30-5.

104. Bray F, Ferlay J, Soerjomataram I, Siegel RL, Torre LA, Jemal A. Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA: a cancer journal for clinicians. 2018;68(6):394-424.

doi:10.6342/NTU202300922

105. Gilks CB, Prat J. Ovarian carcinoma pathology and genetics: recent advances. Human pathology. 2009;40(9):1213-23.

106. Prat J. Pathology of cancers of the female genital tract. International Journal of Gynecology & Obstetrics. 2015;131:S132-S45.

107. Rauh-Hain JA, Krivak TC, del Carmen MG, Olawaiye AB. Ovarian cancer screening and early detection in the general population. Reviews in obstetrics and gynecology. 2011;4(1):15.

108. Holland JF, Pollock RE. Holland-Frei cancer medicine 8: PMPH-USA; 2010.

109. Morgan RJ, Armstrong DK, Alvarez RD, Bakkum-Gamez JN, Behbakht K, Chen L-m, et al. Ovarian cancer, version 1.2016, NCCN clinical practice guidelines in oncology. Journal of the National Comprehensive Cancer Network. 2016;14(9):1134-63.

110. HEALTH PROMOTION ADMINISTRATION MINISTRY OF HEALTH AND WELFARE TAIWAN. CANCER REGISTRY ANNUAL REPORT, 2016 TAIWAN. Taiwan: HEALTH PROMOTION ADMINISTRATION MINISTRY OF HEALTH AND WELFARE TAIWAN; 2018.

111. Szalat R, Avet-Loiseau H, Munshi NC. Gene expression profile in clinical practice. Clinical cancer research: an official journal of the American Association for Cancer Research. 2016;22(22):5434.

doi:10.6342/NTU202300922

112. Marisa L, de Reyniès A, Duval A, Selves J, Gaub MP, Vescovo L, et al. Gene expression classification of colon cancer into molecular subtypes: characterization, validation, and prognostic value. PLoS medicine. 2013;10(5).

113. Yasui W, Oue N, Aung PP, Matsumura S, Shutoh M, Nakayama H. Molecular-pathological prognostic factors of gastric cancer: a review. Gastric cancer. 2005;8(2):86-94.

114. Markert EK, Mizuno H, Vazquez A, Levine AJ. Molecular classification of prostate cancer using curated expression signatures. Proceedings of the National Academy of Sciences. 2011;108(52):21276-81.

115. Stratford JK, Bentrem DJ, Anderson JM, Fan C, Volmar KA, Marron J, et al. A six-gene signature predicts survival of patients with localized pancreatic ductal adenocarcinoma. PLoS medicine. 2010;7(7).

116. Van't Veer LJ, Dai H, Van De Vijver MJ, He YD, Hart AA, Mao M, et al. Gene expression profiling predicts clinical outcome of breast cancer. nature. 2002;415(6871):530-6.

117. Pohl H, Kotze MJ, Grant KA, van der Merwe L, Pienaar FM, Apffelstaedt JP, et al. Impact of MammaPrint on Clinical Decision-Making in South African Patients with Early-Stage Breast Cancer. The breast journal. 2016;22(4):442-6.

doi:10.6342/NTU202300922

118. Paik S, Shak S, Tang G, Kim C, Baker J, Cronin M, et al. A multigene assay to predict recurrence of tamoxifen-treated, node-negative breast cancer. New England Journal of Medicine. 2004;351(27):2817-26.

119. Ming C, Viassolo V, Probst-Hensch N, Chappuis PO, Dinov ID, Katapodi MC. Machine learning techniques for personalized breast cancer risk prediction: comparison with the BCRAT and BOADICEA models. Breast Cancer Research. 2019;21(1):75.

120. Crijns AP, Fehrmann RS, de Jong S, Gerbens F, Meersma GJ, Klip HG, et al. Survival-related profile, pathways, and transcription factors in ovarian cancer. PLoS medicine. 2009;6(2).

121. Bonome T, Levine DA, Shih J, Randonovich M, Pise-Masison CA, Bogomolniy F, et al. A gene signature predicting for survival in suboptimally debulked patients with ovarian cancer. Cancer research. 2008;68(13):5478-86.

122. Tabl AA, Alkhateeb A, ElMaraghy W, Rueda L, Ngom A. A machine learning approach for identifying gene biomarkers guiding the treatment of breast cancer. Frontiers in Genetics. 2019;10.

123. Xu X, Zhang Y, Zou L, Wang M, Li A, editors. A gene signature for breast cancer prognosis using support vector machine. 2012 5th International Conference on BioMedical Engineering and Informatics; 2012: IEEE.

doi:10.6342/NTU202300922

124. Ferroni P, Zanzotto FM, Riondino S, Scarpato N, Guadagni F, Roselli M. Breast cancer prognosis using a machine learning approach. Cancers. 2019;11(3):328.

125. Abdou Y, Baird A, Dolan J, Lee S, Park S, Lee S. 176O Machine learning-assisted prognostication based on genomic expression in the tumour microenvironment of estrogen receptor positive and HER2 negative breast cancer. Annals of Oncology. 2019;30(Supplement_5):mdz240. 002.

126. Gao Y-C, Zhou X-H, Zhang W. An ensemble strategy to predict prognosis in ovarian cancer based on gene modules. Frontiers in genetics. 2019;10:366.

127. Gentric G, Kieffer Y, Mieulet V, Goundiam O, Bonneau C, Nemati F, et al. PML-regulated mitochondrial metabolism enhances chemosensitivity in human ovarian cancers. Cell metabolism. 2019;29(1):156-73. e10.

128. Mateescu B, Batista L, Cardon M, Gruosso T, De Feraudy Y, Mariani O, et al. miR-141 and miR-200a act on ovarian tumorigenesis by controlling oxidative stress response. Nature medicine. 2011;17(12):1627.

129. Ferriss JS, Kim Y, Duska L, Birrer M, Levine DA, Moskaluk C, et al. Multi-gene expression predictors of single drug responses to adjuvant chemotherapy in ovarian carcinoma: predicting platinum resistance. PloS one. 2012;7(2).

130. Konstantinopoulos PA, Spentzos D, Karlan BY, Taniguchi T, Fountzilas E, Francoeur N, et al. Gene expression profile of BRCAness that correlates with

doi:10.6342/NTU202300922

responsiveness to chemotherapy and with outcome in patients with epithelial ovarian cancer. Journal of clinical oncology. 2010;28(22):3555.

131. Lisowska KM, OIbryt M, Dudaladava V, Pamuła-Piłat J, Kujawa K, Grzybowska E, et al. Gene expression analysis in ovarian cancer–faults and hints from DNA microarray study. Frontiers in oncology. 2014;4:6.

132. Colombo N, Lorusso D, Scollo P. Impact of recurrence of ovarian cancer on quality of life and outlook for the future. International Journal of Gynecologic Cancer. 2017;27(6).

133. Ushijima K. Treatment for recurrent ovarian cancer—at first relapse. Journal of oncology. 2010;2010.

134. Irizarry RA, Gautier L, Bolstad BM, Miller C. Methods for affymetrix oligonucleotide arrays. R package version 112. 2006;1.

135. Bolstad BM. preprocessCore: A collection of pre-processing functions. R package version. 2013;1(0).

136. Chen T, Guestrin C, editors. Xgboost: A scalable tree boosting system. Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining; 2016.

137. Friedman JH. Stochastic gradient boosting. Computational statistics & data analysis. 2002;38(4):367-78.

doi:10.6342/NTU202300922

138. Holland JH. Genetic algorithms. Scientific american. 1992;267(1):66-73.

139. Goldberg DE. Genetic algorithms: Pearson Education India; 2006.

140. Chen T, He T, Benesty M, Khotilovich V, Tang Y. Xgboost: extreme gradient boosting. R package version 04-2. 2015:1-4.

141. Breiman L. Bagging predictors. Machine learning. 1996;24(2):123-40.

142. Tibshirani R. Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society: Series B (Methodological). 1996;58(1):267-88.

143. In Lee K, Koval JJ. Determination of the best significance level in forward stepwise logistic regression. Communications in Statistics-Simulation and Computation. 1997;26(2):559-75.

144. Lin H, Zelterman D. Modeling survival data: Extending the cox model. Taylor & Francis; 2002.

145. Naderi A, Teschendorff A, Barbosa-Morais N, Pinder S, Green A, Powe D, et al. A gene-expression signature to predict survival in breast cancer across independent data sets. Oncogene. 2007;26(10):1507-16.

146. Haibe-Kains B, Desmedt C, Piette F, Buyse M, Cardoso F, Van't Veer L, et al. Comparison of prognostic gene expression signatures for breast cancer. BMC genomics. 2008;9(1):394.

doi:10.6342/NTU202300922

147. Allen F, Karjalainen R. Using genetic algorithms to find technical trading rules. Journal of financial Economics. 1999;51(2):245-71.

148. Myung IJ. The importance of complexity in model selection. Journal of mathematical psychology. 2000;44(1):190-204.

149. Kuncheva LI, Whitaker CJ. Measures of diversity in classifier ensembles and their relationship with the ensemble accuracy. Machine learning. 2003;51(2):181-207.

150. Liao X, Cao N, Li M, Kang X, editors. Research on short-term load forecasting using XGBoost based on similar days. 2019 International Conference on Intelligent Transportation, Big Data & Smart City (ICITBS); 2019: IEEE.

151. Owusu-Brackett N, Shariati M, Meric-Bernstam F. Role of PI3K/AKT/mTOR in Cancer Signaling. In: Badve S, Kumar GL, editors. Predictive Biomarkers in Oncology: Applications in Precision Medicine. Cham: Springer International Publishing; 2019. p. 263-70.

152. Musa F, Schneider R. Targeting the PI3K/AKT/mTOR pathway in ovarian cancer. Translational Cancer Research. 2015;4(1):97-106.

153. Cheaib B, Auguste A, Leary A. The PI3K/Akt/mTOR pathway in ovarian cancer: therapeutic opportunities and challenges. Chinese journal of cancer. 2015;34(1):4-16.

doi:10.6342/NTU202300922

154. Ghoneum A, Said N. PI3K-AKT-mTOR and NFκB Pathways in Ovarian Cancer: Implications for Targeted Therapeutics. Cancers. 2019;11(7):949.

155. Li X, Yang Z, Xu S, Wang Z, Jin P, Yang X, et al. Targeting INHBA in Ovarian Cancer Cells Suppresses Cancer Xenograft Growth by Attenuating Stromal Fibroblast Activation. Disease Markers. 2019;2019.

156. Grimberg A, Cohen P. Role of insulin-like growth factors and their binding proteins in growth control and carcinogenesis. Journal of cellular physiology. 2000;183(1):1-9.

157. Mosig R. IGFBP-4 is a candidate serum biomarker for detection and surveillance of early stage epithelial ovarian cancer. Research. 2015.

158. Hwang JR, Cho Y-J, Lee Y, Park Y, Han HD, Ahn HJ, et al. The C-terminus of IGFBP-5 suppresses tumor growth by inhibiting angiogenesis. Scientific reports. 2016;6:39334.

159. Chen S, Dai X, Gao Y, Shen F, Ding J, Chen Q. The positivity of estrogen receptor and progesterone receptor may not be associated with metastasis and recurrence in epithelial ovarian cancer. Scientific Reports. 2017;7(1):1-7.

160. Kwon A-Y, Kim G-I, Jeong J-Y, Song J-Y, Kwack K-B, Lee C, et al. VAV3 overexpressed in cancer stem cells is a poor prognostic indicator in ovarian cancer patients. Stem cells and development. 2015;24(13):1521-35.

doi:10.6342/NTU202300922

161. Page RE, Klein-Szanto AJ, Litwin S, Nicolas E, Al-Jumaily R, Alexander P, et al. Increased expression of the pro-protein convertase furin predicts decreased survival in ovarian cancer. Analytical Cellular Pathology. 2007;29(4):289-99.

162. Vabalas A, Gowen E, Poliakoff E, Casson AJ. Machine learning algorithm validation with a limited sample size. PloS one. 2019;14(11):e0224365.

163. Kim S, Han Y, Kim SI, Kim H-S, Kim SJ, Song YS. Tumor evolution and chemoresistance in ovarian cancer. NPJ precision oncology. 2018;2(1):1-9.

164. Ochs MF, Ertel A, Verghese A, Byers SW, Tozeren A. Pathway-specific differences between tumor cell lines and normal and tumor tissue cells. 2006.

165. Porta M. Textbook of cancer epidemiology. Journal of Epidemiology and Community Health. 2003;57(7):543-. doi: 10.1136/jech.57.7.543. PubMed PMID: PMC1732521.

166. El-Serag HB, Rudolph KL. Hepatocellular carcinoma: epidemiology and molecular carcinogenesis. Gastroenterology. 2007;132(7):2557-76. Epub 2007/06/16. doi: 10.1053/j.gastro.2007.04.061. PubMed PMID: 17570226.

167. Lee SY, Song KH, Koo I, Lee K-H, Suh K-S, Kim B-Y. Comparison of pathways associated with hepatitis B-and C-infected hepatocellular carcinoma using pathway-based class discrimination method. Genomics. 2012;99(6):347-54.

doi:10.6342/NTU202300922

168. Iizuka N, Oka M, Yamada-Okabe H, Mori N, Tamesa T, Okada T, et al. Comparison of gene expression profiles between hepatitis B virus-and hepatitis C virus-infected hepatocellular carcinoma by oligonucleotide microarray data on the basis of a supervised learning method. Cancer research. 2002;62(14):3939-44.

169. Ringelhan M, Pfister D, O'Connor T, Pikarsky E, Heikenwalder M. The immunology of hepatocellular carcinoma. Nature immunology. 2018:1.

170. Ng J, Wu J. Hepatitis B-and hepatitis C-related hepatocellular carcinomas in the United States: similarities and differences. Hepatitis monthly. 2012;12(10 HCC).

171. Jackson R, Psarelli E-E, Berhane S, Khan H, Johnson P. Impact of viral status on survival in patients receiving sorafenib for advanced hepatocellular cancer: a meta-analysis of randomized phase III trials. J Clin Oncol. 2017;35(6):622-8.

172. Pakkala S, Owonikoko TK. Immune checkpoint inhibitors in small cell lung cancer. Journal of thoracic disease. 2018;10(Suppl 3):S460-S7. doi: 10.21037/jtd.2017.12.51. PubMed PMID: 29593891.

173. Ross K, Jones RJ. Immune checkpoint inhibitors in renal cell carcinoma. Clinical science (London, England : 1979). 2017;131(21):2627-42. doi: 10.1042/CS20160894. PubMed PMID: 29079639.

174. Massari F, Di Nunno V, Cubelli M, Santoni M, Fiorentino M, Montironi R, et al. Immune checkpoint inhibitors for metastatic bladder cancer. Cancer Treatment Reviews. 2018;64:11-20. doi: https://doi.org/10.1016/j.ctrv.2017.12.007.

175. Flavell RA, Sanjabi S, Wrzesinski SH, Licona-Limón P. The polarization of immune cells in the tumour environment by TGFβ. Nature Reviews Immunology. 2010;10(8):554.

176. Chiou S-H, Sheu B-C, Chang W-C, Huang S-C, Hong-Nerng H. Current concepts of tumor-infiltrating lymphocytes in human malignancies. Journal of Reproductive Immunology. 2005;67(1):35-50. doi: https://doi.org/10.1016/j.jri.2005.06.002.

177. Salgado R, Denkert C, Demaria S, Sirtaine N, Klauschen F, Pruneri G, et al. The evaluation of tumor-infiltrating lymphocytes (TILs) in breast cancer: recommendations by an International TILs Working Group 2014. Annals of oncology : official journal of the European Society for Medical Oncology. 2015;26(2):259-71. Epub 2014/09/11. doi: 10.1093/annonc/mdu450. PubMed PMID: 25214542.

178. Galon J, Costes A, Sanchez-Cabo F, Kirilovsky A, Mlecnik B, Lagorce-Pagès C, et al. Type, Density, and Location of Immune Cells Within Human Colorectal Tumors

doi:10.6342/NTU202300922

Predict Clinical Outcome. Science. 2006;313(5795):1960. doi: 10.1126/science.1129139.

179. Zhu X-D, Zhang J-B, Zhuang P-Y, Zhu H-G, Zhang W, Xiong Y-Q, et al. High expression of macrophage colony-stimulating factor in peritumoral liver tissue is associated with poor survival after curative resection of hepatocellular carcinoma. Journal of clinical oncology. 2008;26(16):2707-16.

180. Budhu A, Forgues M, Ye Q-H, Jia H-L, He P, Zanetti KA, et al. Prediction of venous metastases, recurrence, and prognosis in hepatocellular carcinoma based on a unique immune response signature of the liver microenvironment. Cancer cell. 2006;10(2):99-111.

181. Li J-Q, Yu X-J, Wang Y-C, Huang L-Y, Liu C-Q, Zheng L, et al. Distinct patterns and prognostic values of tumor-infiltrating macrophages in hepatocellular carcinoma and gastric cancer. Journal of translational medicine. 2017;15(1):37.

182. Cai X-Y, Gao Q, Qiu S-J, Ye S-L, Wu Z-Q, Fan J, et al. Dendritic cell infiltration and prognosis of human hepatocellular carcinoma. Journal of cancer research and clinical oncology. 2006;132(5):293-301.

183. Ninomiya T, Akbar SMF, Masumoto T, Horiike N, Onji M. Dendritic cells with immature phenotype and defective function in the peripheral blood from patients with hepatocellular carcinoma. Journal of hepatology. 1999;31(2):323-31.

doi:10.6342/NTU202300922

184. Matsui M, Machida S, Itani-Yohda T, Akatsuka T. Downregulation of the proteasome subunits, transporter, and antigen presentation in hepatocellular carcinoma, and their restoration by interferon-γ. Journal of gastroenterology and hepatology. 2002;17(8):897-907.

185. Sung PS, Jang JW. Natural Killer Cell Dysfunction in Hepatocellular Carcinoma: Pathogenesis and Clinical Implications. International journal of molecular sciences. 2018;19(11):3648.

186. Subleski JJ, Wiltrout RH, Weiss JM. Application of tissue-specific NK and NKT cell activity for tumor immunotherapy. Journal of autoimmunity. 2009;33(3-4):275-81.

187. Thompson ED, Enriquez HL, Fu Y-X, Engelhard VH. Tumor masses support naive T cell infiltration, activation, and differentiation into effectors. Journal of Experimental Medicine. 2010;207(8):1791-804.

188. Shi J-Y, Gao Q, Wang Z-C, Zhou J, Wang X-Y, Min Z-H, et al. Margin-infiltrating CD20+ B cells display an atypical memory phenotype and correlate with favorable prognosis in hepatocellular carcinoma. Clinical Cancer Research. 2013;19(21):5994-6005.

doi:10.6342/NTU202300922

189. Garnelo M, Tan A, Her Z, Yeong J, Lim CJ, Chen J, et al. Interaction between tumour-infiltrating B cells and T cells controls the progression of hepatocellular carcinoma. Gut. 2017;66(2):342-51.

190. Mossanen JC, Tacke F. Role of lymphocytes in liver cancer. Oncoimmunology. 2013;2(11):e26468.

191. Kuang D-M, Zhao Q, Wu Y, Peng C, Wang J, Xu Z, et al. Peritumoral neutrophils link inflammatory response to disease progression by fostering angiogenesis in hepatocellular carcinoma. Journal of hepatology. 2011;54(5):948-55.

192. Li Y-W, Qiu S-J, Fan J, Zhou J, Gao Q, Xiao Y-S, et al. Intratumoral neutrophils: a poor prognostic factor for hepatocellular carcinoma following resection. Journal of hepatology. 2011;54(3):497-505.

193. Kuang D-M, Peng C, Zhao Q, Wu Y, Zhu L-Y, Wang J, et al. Tumor-Activated Monocytes Promote Expansion of IL-17–Producing CD8+ T Cells in Hepatocellular Carcinoma Patients. The Journal of Immunology. 2010;185(3):1544-9.

194. Ji J, Eggert T, Budhu A, Forgues M, Takai A, Dang H, et al. Hepatic stellate cell and monocyte interaction contributes to poor prognosis in hepatocellular carcinoma. Hepatology. 2015;62(2):481-95.

doi:10.6342/NTU202300922

195. Fu J, Xu D, Liu Z, Shi M, Zhao P, Fu B, et al. Increased regulatory T cells correlate with CD8 T-cell impairment and poor survival in hepatocellular carcinoma patients. Gastroenterology. 2007;132(7):2328-39.

196. Liu RX, Wei Y, Zeng QH, Chan KW, Xiao X, Zhao XY, et al. Chemokine (C-X-C motif) receptor 3–positive B cells link interleukin-17 inflammation to protumorigenic macrophage polarization in human hepatocellular carcinoma. Hepatology. 2015;62(6):1779-90.

197. Honda M, Kaneko S, Kawai H, Shirota Y, Kobayashi K. Differential gene expression between chronic hepatitis B and C hepatic lesion. Gastroenterology. 2001;120(4):955-66. doi: https://doi.org/10.1053/gast.2001.22468.

198. Iizuka N, Oka M, Yamada-Okabe H, Mori N, Tamesa T, Okada T, et al. Comparison of Gene Expression Profiles between Hepatitis B Virus- and Hepatitis C Virus-infected Hepatocellular Carcinoma by Oligonucleotide Microarray Data on the Basis of a Supervised Learning Method. Cancer Research. 2002;62(14):3939.

199. Foerster F, Hess M, Gerhold-Ay A, Marquardt JU, Becker D, Galle PR, et al. The immune contexture of hepatocellular carcinoma predicts clinical outcome. Scientific Reports. 2018;8(1):5351. doi: 10.1038/s41598-018-21937-2.

200. Rohr-Udilova N, Klinglmüller F, Schulte-Hermann R, Stift J, Herac M, Salzmann M, et al. Deviations of the immune cell landscape between healthy liver

doi:10.6342/NTU202300922

and hepatocellular carcinoma. Scientific Reports. 2018;8(1):6220. doi: 10.1038/s41598-018-24437-5.

201. Ding W, Xu X, Qian Y, Xue W, Wang Y, Du J, et al. Prognostic value of tumor-infiltrating lymphocytes in hepatocellular carcinoma: A meta-analysis. Medicine. 2018;97(50).

202. Kim G-A, Lim Y-S, Han S, Choi J, Shim JH, Kim KM, et al. High risk of hepatocellular carcinoma and death in patients with immune-tolerant-phase chronic hepatitis B. Gut. 2018;67(5):945-52.

203. Tu J-F, Ding Y-H, Ying X-H, Wu F-Z, Zhou X-M, Zhang D-K, et al. Regulatory T cells, especially ICOS+ FOXP3+ regulatory T cells, are increased in the hepatocellular carcinoma microenvironment and predict reduced survival. Scientific reports. 2016;6:35056.

204. Yoshihara K, Shahmoradgoli M, Martínez E, Vegesna R, Kim H, Torres-Garcia W, et al. Inferring tumour purity and stromal and immune cell admixture from expression data. Nature communications. 2013;4:2612-. doi: 10.1038/ncomms3612. PubMed PMID: 24113773.

205. Newman AM, Liu CL, Green MR, Gentles AJ, Feng W, Xu Y, et al. Robust enumeration of cell subsets from tissue expression profiles. Nature methods.

2015;12(5):453-7. Epub 2015/03/30. doi: 10.1038/nmeth.3337. PubMed PMID: 25822800.

206. Benjamini Y, Hochberg Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. Journal of the Royal Statistical Society Series B (Methodological). 1995;57(1):289-300.

207. Hoshida Y, Villanueva A, Kobayashi M, Peix J, Chiang DY, Camargo A, et al. Gene expression in fixed tissues and outcome in hepatocellular carcinoma. The New England journal of medicine. 2008;359(19):1995-2004. Epub 2008/10/15. doi: 10.1056/NEJMoa0804525. PubMed PMID: 18923165.

208. Yang P, Markowitz GJ, Wang X-F. The hepatitis B virus-associated tumor microenvironment in hepatocellular carcinoma. National science review. 2014;1(3):396-412. doi: 10.1093/nsr/nwu038. PubMed PMID: 25741453.

209. Wei C, Ni C, Song T, Liu Y, Yang X, Zheng Z, et al. The Hepatitis B Virus X Protein Disrupts Innate Immunity by Downregulating Mitochondrial Antiviral Signaling Protein. The Journal of Immunology. 2010;185(2):1158. doi: 10.4049/jimmunol.0903874.

210. Castello G, Costantini S, Scala S. Targeting the inflammation in HCV-associated hepatocellular carcinoma: a role in the prevention and treatment. Journal of

doi:10.6342/NTU202300922

translational medicine. 2010;8:109-. doi: 10.1186/1479-5876-8-109. PubMed PMID: 21047421.

211. Cancer Genome Atlas Research Network. Electronic address wbe, Cancer Genome Atlas Research N. Comprehensive and Integrative Genomic Characterization of Hepatocellular Carcinoma. Cell. 2017;169(7):1327-41.e23. doi: 10.1016/j.cell.2017.05.046. PubMed PMID: 28622513.

212. Arzumanyan A, Reis HM, Feitelson MA. Pathogenic mechanisms in HBV- and HCV-associated hepatocellular carcinoma. Nature reviews Cancer. 2013;13(2):123-35. Epub 2013/01/25. doi: 10.1038/nrc3449. PubMed PMID: 23344543.

213. Kuang D-M. Positive Feedback Loop between B cells/plasma cells and Macrophages Promote M2 Macrophage-Elicited Hepatoma Progression. The Journal of Immunology. 2016;196(1 Supplement):211.6.

214. Ouyang F-Z, Wu R-Q, Wei Y, Liu R-X, Yang D, Xiao X, et al. Dendritic cell-elicited B-cell activation fosters immune privilege via IL-10 signals in hepatocellular carcinoma. Nature communications. 2016;7:13453-. doi: 10.1038/ncomms13453. PubMed PMID: 27853178.

215. Kuang D-M, Zhao Q, Xu J, Yun J-P, Wu C, Zheng L. Tumor-Educated Tolerogenic Dendritic Cells Induce CD3ε Down-Regulation and Apoptosis of T Cells

doi:10.6342/NTU202300922

through Oxygen-Dependent Pathways. The Journal of Immunology. 2008;181(5):3089. doi: 10.4049/jimmunol.181.5.3089.

216. Ashtari S, Pourhoseingholi MA, Sharifian A, Zali MR. Hepatocellular carcinoma in Asia: Prevention strategy and planning. World journal of hepatology. 2015;7(12):1708.

217. Yuen MF, Hou JL, Chutaputti A. Hepatocellular carcinoma in the Asia pacific region. Journal of gastroenterology and hepatology. 2009;24(3):346-53.

218. Nguyen L, Nguyen M. Systematic review: Asian patients with chronic hepatitis C infection. Alimentary pharmacology & therapeutics. 2013;37(10):921-36.

219. Lué A, Serrano MT, Bustamante FJ, Iñarrairaegui M, Arenas JI, Testillano M, et al. Neutrophil-to-lymphocyte ratio predicts survival in European patients with hepatocellular carcinoma administered sorafenib. Oncotarget. 2017;8(61):103077-86. doi: 10.18632/oncotarget.21528. PubMed PMID: 29262546.

220. Tu J-F, Ding Y-H, Ying X-H, Wu F-Z, Zhou X-M, Zhang D-K, et al. Regulatory T cells, especially ICOS(+) FOXP3(+) regulatory T cells, are increased in the hepatocellular carcinoma microenvironment and predict reduced survival. Scientific reports. 2016;6:35056-. doi: 10.1038/srep35056. PubMed PMID: 27725696.

221. Ideker T, Krogan NJ. Differential network biology. Molecular systems biology. 2012;8(1):565.

222. Bartel DP. MicroRNAs: target recognition and regulatory functions. cell. 2009;136(2):215-33.

223. Lou W, Liu J, Gao Y, Zhong G, Chen D, Shen J, et al. MicroRNAs in cancer metastasis and angiogenesis. Oncotarget. 2017;8(70):115787.

224. Hayes J, Peruzzi PP, Lawler S. MicroRNAs in cancer: biomarkers, functions and therapy. Trends in molecular medicine. 2014;20(8):460-9.

225. Salmena L, Poliseno L, Tay Y, Kats L, Pandolfi PP. A ceRNA hypothesis: the Rosetta Stone of a hidden RNA language? Cell. 2011;146(3):353-8.

226. Su X, Xing J, Wang Z, Chen L, Cui M, Jiang B. microRNAs and ceRNAs: RNA networks in pathogenesis of cancer. Chinese journal of cancer research. 2013;25(2):235.

227. Tay Y, Kats L, Salmena L, Weiss D, Tan SM, Ala U, et al. Coding-independent regulation of the tumor suppressor PTEN by competing endogenous mRNAs. Cell. 2011;147(2):344-57.

228. Wang Q, Cai J, Fang C, Yang C, Zhou J, Tan Y, et al. Mesenchymal glioblastoma constitutes a major ceRNA signature in the TGF-β pathway. Theranostics. 2018;8(17):4733.

doi:10.6342/NTU202300922

229. Karreth FA, Tay Y, Perna D, Ala U, Tan SM, Rust AG, et al. In vivo identification of tumor-suppressive PTEN ceRNAs in an oncogenic BRAF-induced mouse model of melanoma. Cell. 2011;147(2):382-95.

230. Yan J, Du L, Yao X, Shen L. Machine learning in brain imaging genomics. Machine learning and medical imaging: Elsevier; 2016. p. 411-34.

231. Clough E, Barrett T. The gene expression omnibus database. Statistical genomics: Springer; 2016. p. 93-110.

232. Tomczak K, Czerwinska P, Wiznerowicz M. The Cancer Genome Atlas (TCGA): an immeasurable source of knowledge, Współczesna Onkologia, vol. 19, no. 1A, pp. A68-A77. 2015.

233. Brazma A, Parkinson H, Sarkans U, Shojatalab M, Vilo J, Abeygunawardena N, et al. ArrayExpress—a public repository for microarray gene expression data at the EBI. Nucleic acids research. 2003;31(1):68-71.

234. Osborne JW, Overbay A. The power of outliers (and why researchers should always check for them). Practical Assessment, Research, and Evaluation. 2004;9(1):6.

235. Du Z, Sun T, Hacisuleyman E, Fei T, Wang X, Brown M, et al. Integrative analyses reveal a long noncoding RNA-mediated sponge regulatory network in prostate cancer. Nature communications. 2016;7(1):1-10.

doi:10.6342/NTU202300922

236. Xu J, Li Y, Lu J, Pan T, Ding N, Wang Z, et al. The mRNA related ceRNA–ceRNA landscape and significance across 20 major cancer types. Nucleic acids research. 2015;43(17):8169-82.

237. Zhang J, Le TD, Liu L, Li J. Inferring miRNA sponge co-regulation of protein-protein interactions in human breast cancer. BMC bioinformatics. 2017;18(1):1-12.

238. List M, Dehghani Amirabad A, Kostka D, Schulz MH. Large-scale inference of competing endogenous RNA networks with sparse partial correlation. Bioinformatics. 2019;35(14):i596-i604.

239. Sardina DS, Alaimo S, Ferro A, Pulvirenti A, Giugno R. A novel computational method for inferring competing endogenous interactions. Briefings in Bioinformatics. 2017;18(6):1071-81.

240. Li R, Qu H, Wang S, Wei J, Zhang L, Ma R, et al. GDCRNATools: an R/Bioconductor package for integrative analysis of lncRNA, miRNA and mRNA data in GDC. Bioinformatics. 2018;34(14):2515-7.

241. Hornakova A, List M, Vreeken J, Schulz MH. JAMI: fast computation of conditional mutual information for ceRNA network analysis. Bioinformatics. 2018;34(17):3050-1.

doi:10.6342/NTU202300922

242. Chiu H-S, Llobet-Navas D, Yang X, Chung W-J, Ambesi-Impiombato A, Iyer A, et al. Cupid: simultaneous reconstruction of microRNA-target and ceRNA networks. Genome research. 2015;25(2):257-67.

243. Furió-Tarí P, Tarazona S, Gabaldón T, Enright AJ, Conesa A. spongeScan: A web for detecting microRNA binding elements in lncRNA sequences. Nucleic acids research. 2016;44(W1):W176-W80.

244. Navada S, Lai P, Schwartz A, Kalemkerian G. Temporal trends in small cell lung cancer: analysis of the national Surveillance, Epidemiology, and End-Results (SEER) database. Journal of Clinical Oncology. 2006;24(18_suppl):7082-.

245. Ala U, Karreth FA, Bosia C, Pagnani A, Taulli R, Léopold V, et al. Integrated transcriptional and competitive endogenous RNA networks are cross-regulated in permissive molecular environments. Proceedings of the National Academy of Sciences. 2013;110(18):7154-9.

246. Bosson AD, Zamudio JR, Sharp PA. Endogenous miRNA and target concentrations determine susceptibility to potential ceRNA competition. Molecular cell. 2014;56(3):347-59.

247. Fiannaca A, Paglia LL, Rosa ML, Rizzo R, Urso A. miRTissue ce: extending miRTissue web service with the analysis of ceRNA-ceRNA interactions. BMC bioinformatics. 2020;21(8):1-21.

doi:10.6342/NTU202300922

248. Wang Y, Xiong X, Hua X, Liu W. Expression and Gene Regulation Network of Metabolic Enzyme Phosphoglycerate Mutase Enzyme 1 in Breast Cancer Based on Data Mining. BioMed Research International. 2021;2021.

249. Garon EB, Hellmann MD, Rizvi NA, Carcereny E, Leighl NB, Ahn M-J, et al. Five-year overall survival for patients with advanced non–small-cell lung cancer treated with pembrolizumab: results from the phase I KEYNOTE-001 study. Journal of Clinical Oncology. 2019;37(28):2518.

250. Zhu C, Deng X, Wu J, Zhang J, Yang H, Fu S, et al. MicroRNA-183 promotes migration and invasion of CD133+/CD326+ lung adenocarcinoma initiating cells via PTPN4 inhibition. Tumor Biology. 2016;37(8):11289-97.

251. Zhao B, Han H, Chen J, Zhang Z, Li S, Fang F, et al. MicroRNA let-7c inhibits migration and invasion of human non-small cell lung cancer by targeting ITGB3 and MAP4K3. Cancer letters. 2014;342(1):43-51.

252. Wang K, Saito M, Bisikirska BC, Alvarez MJ, Lim WK, Rajbhandari P, et al. Genome-wide identification of post-translational modulators of transcription factor activity in human B cells. Nature biotechnology. 2009;27(9):829-37.

253. Hsu S-D, Lin F-M, Wu W-Y, Liang C, Huang W-C, Chan W-L, et al. miRTarBase: a database curates experimentally validated microRNA–target interactions. Nucleic acids research. 2011;39(suppl_1):D163-D9.

doi:10.6342/NTU202300922

254. Xiao F, Zuo Z, Cai G, Kang S, Gao X, Li T. miRecords: an integrated resource for microRNA–target interactions. Nucleic acids research. 2009;37(suppl_1):D105-D10.

255. Paraskevopoulou MD, Georgakilas G, Kostoulas N, Vlachos IS, Vergoulis T, Reczko M, et al. DIANA-microT web server v5. 0: service integration into miRNA functional analysis workflows. Nucleic acids research. 2013;41(W1):W169-W73.

256. Gaidatzis D, van Nimwegen E, Hausser J, Zavolan M. Inference of miRNA targets using evolutionary conservation and pathway analysis. BMC bioinformatics. 2007;8(1):1-22.

257. Betel D, Koppal A, Agius P, Sander C, Leslie C. Comprehensive modeling of microRNA targets predicts functional non-conserved and non-canonical sites. Genome biology. 2010;11(8):1-14.

258. Wong N, Wang X. miRDB: an online resource for microRNA target prediction and functional annotations. Nucleic acids research. 2015;43(D1):D146-D52.

259. Kertesz M, Iovino N, Unnerstall U, Gaul U, Segal E. The role of site accessibility in microRNA target recognition. Nature genetics. 2007;39(10):1278-84.

260. Loher P, Rigoutsos I. Interactive exploration of RNA22 microRNA target predictions. Bioinformatics. 2012;28(24):3322-3.

doi:10.6342/NTU202300922

261. Agarwal V, Bell GW, Nam J-W, Bartel DP. Predicting effective microRNA target sites in mammalian mRNAs. elife. 2015;4:e05005.

262. Goldman M, Craft B, Zhu J, Haussler D. The UCSC Xena system for cancer genomics data visualization and interpretation. Cancer Research. 2017;77(13_Supplement):2584-.

263. Datar M, Gionis A, Indyk P, Motwani R. Maintaining stream statistics over sliding windows. SIAM journal on computing. 2002;31(6):1794-813.

264. da Silva MAF, de Carvalho RL, da Silva Almeida T. Evaluation of a Sliding Window mechanism as DataAugmentation over Emotion Detection on Speech. Academic Journal on Computing, Engineering and Applied Mathematics. 2021;2(1):11-8.

265. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proceedings of the National Academy of Sciences. 2005;102(43):15545-50.

266. Olshen AB, Venkatraman E, Lucito R, Wigler M. Circular binary segmentation for the analysis of array-based DNA copy number data. Biostatistics. 2004;5(4):557-72.

doi:10.6342/NTU202300922

267. Tickle T, Tirosh I, Georgescu C, Brown M, Haas B. inferCNV of the Trinity CTAT Project. Klarman Cell Observatory, Broad Institute of MIT and Harvard. 2019.

268. Wang K, Li M, Hadley D, Liu R, Glessner J, Grant SF, et al. PennCNV: an integrated hidden Markov model designed for high-resolution copy number variation detection in whole-genome SNP genotyping data. Genome research. 2007;17(11):1665-74.

269. Sen A, Srivastava MS. On tests for detecting change in mean. The Annals of statistics. 1975:98-108.

270. Linn SC, West RB, Pollack JR, Zhu S, Hernandez-Boussard T, Nielsen TO, et al. Gene expression patterns and gene copy number changes in dermatofibrosarcoma protuberans. The American journal of pathology. 2003;163(6):2383-95.

271. Consortium GO. The Gene Ontology (GO) database and informatics resource. Nucleic acids research. 2004;32(suppl_1):D258-D61.

272. Kanehisa M, Goto S. KEGG: kyoto encyclopedia of genes and genomes. Nucleic acids research. 2000;28(1):27-30.

273. Anderson TW, Darling DA. Asymptotic theory of certain" goodness of fit" criteria based on stochastic processes. The annals of mathematical statistics. 1952:193-212.

doi:10.6342/NTU202300922

274. Weinstein JN, Collisson EA, Mills GB, Shaw KR, Ozenberger BA, Ellrott K, et al. The cancer genome atlas pan-cancer analysis project. Nature genetics. 2013;45(10):1113-20.

275. Wang P, Zhi H, Zhang Y, Liu Y, Zhang J, Gao Y, et al. miRSponge: a manually curated database for experimentally supported miRNA sponges and ceRNAs. Database. 2015;2015.

276. Wang Z, Jensen MA, Zenklusen JC. A practical guide to the cancer genome atlas (TCGA). Statistical Genomics: Springer; 2016. p. 111-41.

277. Chen C-H, Lu T-P. Utilizing gene expression profiles to characterize tumor infiltrating lymphocytes in cancers. Annals of translational medicine. 2019;7(Suppl 8).

278. Kim SJ, Sota Y, Naoi Y, Honma K, Kagara N, Miyake T, et al. Determining homologous recombination deficiency scores with whole exome sequencing and their association with responses to neoadjuvant chemotherapy in breast cancer. Translational oncology. 2021;14(2):100986.

279. Quick C, Wen X, Abecasis G, Boehnke M, Kang HM. Integrating comprehensive functional annotations to boost power and accuracy in gene-based association analysis. PLoS Genetics. 2020;16(12):e1009060.

doi:10.6342/NTU202300922

280. Troyanskaya OG, Garber ME, Brown PO, Botstein D, Altman RB. Nonparametric methods for identifying differentially expressed genes in microarray data. Bioinformatics. 2002;18(11):1454-61.

281. Morris AP, Lindgren CM, Zeggini E, Timpson NJ, Frayling TM, Hattersley AT, et al. A powerful approach to sub-phenotype analysis in population-based genetic association studies. Genetic epidemiology. 2010;34(4):335-43.

282. Kourou K, Exarchos TP, Exarchos KP, Karamouzis MV, Fotiadis DI. Machine learning applications in cancer prognosis and prediction. Computational and structural biotechnology journal. 2015;13:8-17.

283. Krämer A, Green J, Pollard Jr J, Tugendreich S. Causal analysis approaches in ingenuity pathway analysis. Bioinformatics. 2014;30(4):523-30.

284. Anastasiadou E, Jacob LS, Slack FJ. Non-coding RNA networks in cancer. Nature Reviews Cancer. 2018;18(1):5-18.

285. Fu X-D. Non-coding RNA: a new frontier in regulatory biology. National science review. 2014;1(2):190-204.

286. Nyamundanda G, Poudel P, Patil Y, Sadanandam A. A novel statistical method to diagnose, quantify and correct batch effects in genomic studies. Scientific reports. 2017;7(1):1-10.

doi:10.6342/NTU202300922

287. Jew B, Alvarez M, Rahmani E, Miao Z, Ko A, Garske KM, et al. Accurate estimation of cell composition in bulk expression through robust integration of single-cell information. Nature communications. 2020;11(1):1-11.

288. Menyhárt O, Győrffy B. Multi-omics approaches in cancer research with applications in tumor subtyping, prognosis, and diagnosis. Computational and Structural Biotechnology Journal. 2021;19:949.

289. Zhu W, Xie L, Han J, Guo X. The application of deep learning in cancer prognosis prediction. Cancers. 2020;12(3):603.

290. Outeiral C, Strahm M, Shi J, Morris GM, Benjamin SC, Deane CM. The prospects of quantum computing in computational molecular biology. Wiley Interdisciplinary Reviews: Computational Molecular Science. 2021;11(1):e1481.

291. Taylor-Weiner A, Aguet F, Haradhvala NJ, Gosai S, Anand S, Kim J, et al. Scaling computational genomics to millions of individuals with GPUs. Genome biology. 2019;20(1):1-5.

doi:10.6342/NTU202300922