MONASH University

Machine learning approaches to identify predictive biomarkers for cell-cycle inhibitors in prostate cancer – Master thesis report

Vishal Pattabiraman - 31131441

Supervisor: Dr. Lan Nguyen,

Co-supervisors: Dr. Sungyoung Shin

FIT5126

Nov 2021

TABLE OF CONTENTS

Part I : General Literature Review

1. Introduction

Prostate cancer is one of the most diagnosed cancer among men which causes somatic mutations that have irreversible repercussions. Even with advancement in early identification when PCa is most curable using extensive screening, patients are most often overtreated or undertreated due to heterogeneity and complex nature of genomic mutations. There are a range of biomarkers that are used in clinical practice in identification of prostate cancer including, serum-based biomarkers prostate health index(PHI)[34], Kscore and other blood-based biomarkers[35], urinary based biomarkers[36], MRI images and others, for patient stratification on localised and metastatic variants. Potent and selective inhibitors of Cyclin-dependent kinases CDK4 and CDK6 have shown improved outcomes on treatment on ER+ breast cancer[20]. Given their impressive performance in breast cancer, these inhibitors are investigated on their implementation on Prostate cancer using various potential combinational therapies along with existing treatments[21]. However, there is still lack of information on the patients who may hold potential and get benefited from CDK4/6 inhibitor treatments due to resistance mechanisms that are developed over the course leading to patient relapse[20]. Patient stratification based on Single biomarkers have been shown to be unsuccessful in understanding the sensitivity and tolerance of CDK4/6 inhibitors in treating prostate cancer[23]. As a result, rather than individual biomarkers, prospective clinical trials will benefit from the validation of a multi-panel of predictive biomarkers. Several collaborative efforts on molecular profiles, drug sensitivity of cancer patient cohorts from resources like Genomics in Drug Sensitivity in Cancer(GDSE)(https://www.cancerrxgene.org/) and COSMIC cell line project (CCLP)(https://cancer.sanger.ac.uk/) have made it possible to use machine learning approaches to identify potential biomarker panels. The explosion of molecular and cellular profiling data from large numbers of samples has resulted from technological developments in genomics and imaging. Conventional research techniques are being tested by the rapid increase in biological data dimension and acquisition rate. Modern machine learning techniques, such as deep learning, promise to make accurate predictions by using very large data sets to find hidden structure[31].

This project will integrate omics (gene/protein expression) data along with drug-response data from prostate cancer cell lines and explant samples (from GDSE and CCLP) and use it as inputs into machine learning models. Further, build and train novel machine learning models to identify accurate biomarker panels that predict sensitivity (resistance) to CDK4/6 inhibitors. Finally, validate identified biomarkers against independent datasets from cell lines and prostate cancer patients and produce robust potential biomarker panel that can be used for patient stratification and personalised treatment.

2. Substantive Literature Review

2.1 Prostate cancer overview

The prostate gland is a male reproductive accessory organ that is found under the bladder and in the region surrounding the urethra. Prostate functions to regulate secretion of sperm, formulate ejaculate and maintain sperm viability[1,3]. The cells contained within prostate often cause tumorous growth in mid to late stage of life[2,3]. Prostate cancer (PCa) was once identified as a very rare disease by surgeon J. Adams in 1853. PCa is now the second leading cause of cancer death among men in the United States, after lung cancer[1,6]. Furthermore, every year, more than 1.2 million new cases of prostate cancer are diagnosed, with over 350,000 deaths worldwide, making it one of the leading causes of cancer-related death in men[3].

Prostate cancer is believed to be caused due to somatic mutation that occur in the prostate genome through the patient's lifetime. These changes arise at the tumour suppressor genes that cause translation and/or transcription functional defects which lead to deregulated cell homeostasis. These aberrations can mainly cause cell proliferation, cell growth and cell death. Most commonly the Androgen Receptors(AR) promotor regions that regulate the AR transcription are disrupted causing overexpression. This related to indolent prostate cancer[3]. Similarly, the under expression of AR leads to the rare, aggressive variant[5]. Both are broadly classified as localised disease as these are within the prostate gland. For patients with localised disease identification of specific gene alterations that can differentiate aggressive from weaker variants has been a challenge. This can be due to the heterogenous nature of the disease and variants of mutations that occur. Currently in identification of aggressiveness, molecular profiling of the tumour is not used instead genetic signatures consisting of multiple features including CNA, Gene methylation and complex mutational phenomena such as kataegis, chromothripis and chromoplexy are applied which helps to provide clinical treatment pathways and progression to Metastasis disease[3,5,8].

On the other hand, metastatic prostate cancer is more advanced stage which are longer organ confined and cause the growth of lymph nodes and/or cancer growth on bone sites. Broadly there are two variants including de novo metastatic castration-sensitive prostate cancer (mCSPC), as well as cancers that progress during or after Androgen deprivation therapy (ADT), termed castration-resistant prostate cancer (mCRPC) overexpression of AR can occur due to amplification of gene or alteration of factors that control AR expression due to somatic aberrations which leads to CSPC[7]. These somatic gain-of-function mutations result in active AR mutants which invariably activate other agonists including steroid hormones. In addition, post translational modifications enable AR even at lower levels of testosterone. Also, due to alternate splicing some tumour creates AR isoforms called AR splice variants which promotes the conversion from CSPC to CRPC. AR overexpression is always directly related with CRPC[3]. Most common regions that are directly affect from metastasis prostate cancer are internal and/or external pelvic node, prerectal node common iliac node and bones including pevis, hip and axial skeleton[7]. Modelling metastasis have been extensively done on mouse, however there is still ongoing investigation in understanding human progression. [3,5,7]

2.2 Factors influencing Prostate cancer

Prostate cancer risks are strongly correlated with age where more than 85% of the cases are of greater than 60 age group[4]. This is more evident in the highly developed countries like USA and UK in comparison with developing countries. There is a direct correlation of Prostate cancer and the (Human Development Index) and gross domestic product[3,4]. In addition, Prostate cancer risk is greater on selected ethnic group, African or Caribbean descent living in USA have twofold chance of getting aggressive tumour on comparison with the white population[3]. Also, men living in Asian countries have relatively lower risk of prostate cancer than the white population living in the USA. For other ethnic groups, the mortality and risks are unknown. Genetical inheritance is another major factor that influences the probability of prostate cancer in men. There is strong correlation with being born in a family having history of any cancer or PCa to getting prostate cancer[3,4]. Non heritable factors including exposure to cigarette smoking, obesity and predominant western diet are thought to have influence in prostate cancer however evidence for this is still lacking[3]. There are several factors that directly correlate to prostate cancer and influence the relative risk however there are lack of evidence for some of these factors.

2.3 Prostate cancer diagnosis and biomarkers

Active screening for prostate cancer is one of the major ways to detect localised prostate cancer at early stages when it is most curable. Screening methods relay on measurement of blood serum biomarker prostate specific antigen(PSA). Standard diagnostic methods include direct rectal exam procedure which involves physical examination of prostate to assess the gland enlargement, stiffness and texture along with MRI imaging. Based on suspicious results from PSA values, Digital rectal exam (DRE) and MRI, prostate biopsy is performed to evaluate the prostate for potential prostate cancer[3]. There are several other biomarkers that are used assess risk factor of prostate cancer. Firstly, serum-based markers which are derived from blood include various metrics like prostate health index(PHI), Kscore and other blood-based biomarkers[11]. Secondly, urinary markers which includes prostate cancer antigen 3(PCA3), TMPRSS2-ERG Fusion Gene, MiProstate score(MiPS), SelectMDx and ExoDx Prostate Intelliscore[11]. Thirdly, Tissue markers which includes ConfirmMDX. Finally, Imaging test based on Multiparametric Magnetic Resonance Imaging (mpMRI)[12]. Biomarkers that are used to inspect the susceptibility of prostate cancer include understanding the germline mutations in BRCA1 and BRCA2 which are shown to be directly associated with several types of cancer including breast, prostate and ovarian[9]. Molecular biomarkers in localised prostate cancer include ProMark, Prolaris, Oncotype Dx and Decipher which identifies different parts of the genome for potential mutations. Molecular biomarkers in advanced prostate cancer include DNA Repair Defects, PTEN Loss and PI3K/AKT Activation and Androgen Receptor[8]. Although with all these biomarkers there is inability to identify clinically significant Prostate Cancer has led to intense search for prognostic factors and molecular biomarkers that will help overdiagnosis and overtreatment with localised disease. Although biomarker gene panels have been efficient for patient stratification there is still growing research on using genomic data in clinical trial and need for new biomarkers and panels[8,9,10,11].

2.4 Prostate cancer treatment

Androgen deprivation therapy (ADT) has been the gold standard for treating prostate cancer as AR signalling pathways influence cell proliferation and growth[3]. In general, localised prostate cancer is treated predominantly by active surveillance and/or radical local treatment. Patients with low-risk variants undergo active surveillance followed with Radiotherapy(external beam radiotherapy(EBRT)) along with androgen deprivation therapy(ADT) or radical prostatectomy depending on patient characteristics[3]. Patients with intermediate risks undergo active surveillance followed with Radiotherapy and transient ADT in high-risk cases based on patient profiles radical prostatectomy along with or without pelvic lymph node dissection is performed. Selected patients with 10 years or lower life expectancy should receive watchful treatments based on progression. Prostate specific antigen(PSA) serum-based biomarker and DRE help understand the abnormality in these prostate glands. [3]

Even with androgen deprivation therapy (ADT), PCa will gradually build resistance known as castrate resistant prostate cancer(CRPC) and advance to the end-stage disease[7]. There are currently eight different therapy offered to treat metastatic castrate resistant prostate cancer (mCRPC) approved by FDA including cabazitaxel, sipuleucel-T, docetaxel, radium-223, abiraterone, enzalutamide, Olaparib and rucaparib. Regrettably, the average survival gain from these treatments varies from 2.4 to 4.8 months, highlighting the value of new clinical growth[1,7]. Castration-sensitive prostate cancer (CPSC) is further classified into clinically metastatic cM0 and micro metastatic disease cM1[5]. As initial treatment patients with cM0 are put through observation to understand if this variant will develop into cM1, local salvage treatments are provided along with lifelong ADT. Patients with cM1 are treated with lifelong ADT with either of EBRT or docetaxel or use of androgen receptor signalling inhibitor(ARSI)(either of abiraterone (ABI), enzalutamide (ENZ) or apalutamide (APA)) with radiotherapy depending upon the metastasis characteristics. When the metastasis develops further this stage is termed as castration-resistant prostate cancer CRPC[7]. These are treated with lifelong ADT along with ARSI inhibitors ABI, ENZ or ABI and routine MRI to understand the growth and progression. For cM1 under CRPC, cases are treated with continued lifelong ADT with either ARSI inhibitors or Radium-223 in case of Bone involvement or chemotherapy with docetaxel or cabazitaxel. Selected patients are benefited with use of Sipuleucel-T and Pembrolizumab PARPi[3,5,7]. Even with all these treatments none of these treatments are superior to other in curing Prostate cancer as genomic heterogeneity contributes to increased complexity. Understanding the correct set of treatment sequence for a given patient cohort has been the active area of research. Tailored treatment corresponding to precision medicine and possible combinational therapies are still underway.

2.5 Limitations of current clinical practice

Patient diagnosed with PCa are broadly classified based on the pathological and clinical criteria, this stratifies them into risk groups[8]. The intrinsic characteristics of the tumour (e.g., Gleason score; GS), clinical parameters such as tumour stage (TNM), and pre-treatment prostate-specific antigen (PSA) values are used in these models. While these stratifications provide vital information about tumour and their expected behaviour, often provide suboptimal results in classification of aggressive tumours[8]. Thus, leading to under or over treatment.

Furthermore, true personalised medicines led by predictive biomarkers are lacking in the clinical management of PCa[8,13]. With the added difficulty of clinical decision-making in prognosis and diagnosis, we must accurately identify patients who are most likely to react to a specific therapy, devise treatment sequences, and enrol these patients in targeted therapies[8,15].

Finally, most predictive biomarkers developed from multiple avenues including genomic data are univariate and often ineffective in generalizing malignant development. There is gap in Identification of Biomarker panels with the use of multi-omics for precision PCa oncology[11,12,13,14].

2.6 CDK4/6 Inhibitor and usage in Prostate Cancer treatment.

Multiple mitogenic and growth inhibitory signals must be integrated which informs on cell division. The cell cycle contains certain regulatory factors that control cellular growth. In metazoans, these regulations are controlled by Kinases that transitions in the G1 phase from G0 phase allowing for cellular proliferation[12]. While different cell types have collection of kinases, usually cyclin-dependent kinases CDK4 and/or CDK6 contribute towards this cellular proliferation[12]. These regulatory proteins along with cyclin D, Retinoblastoma gene product(Rb) and E2 transcription factor(E2F) transition from G1 phase to S phase (replication). The activity of cyclin dependent kinases is dependent on synthesis, accumulation and correct localisation of their cyclin partners either cyclin D1, D2 and/or D3[13]. Based on extensive mouse studies the function of CDK4 and CDK6 have been determined. For instance, if D1 is deleted will lead to small body size and a decreased mammary development. In comparison, cyclin D2 deletion causes abnormal ovarian growth and female infertility[12]. CDK4 deficiency is linked to small body size and various developmental abnormalities, whereas CDK6 deficiency is linked to mostly hematopoietic phenotypes[12]. These kinases then phosphorylate critical subunits that progress from G1 to S phase for cell replication in cell cycle. One of these subunits are the Retinoblastoma gene products(Rb) tumour suppressor proteins. There are potentially 16 binding for cyclin dependent kinases to phosphorylate these Rb proteins and to inactive state[13]. Due to these phosphorylation causes structural alteration that leads functional damage of tumour suppressor activity. In the absence of CDK4/6 activity Rb will not be hyper phosphorylated state. When the Rb is in active state it binds and inhibits the activation of E2F transcription factors. However, in inactive state Rb detaches from the E2F transcription factors, activating E2F transcription and E2F related gene products that drives DNA replication and processing to S phase. Consequently, these alteration in Rb tumour suppressor proteins elicits inhibition of E2F transcription factors.

Dysregulation of cell cycle is a common phenomenon observed in many cancers which results in uncontrolled cellular proliferation, a typical hallmark of cancer[5]. The importance of the cyclin regulation and appropriate function in cell cycle in normal tissue homeostasis is disrupted by multiple genetic events in cancer that result in hyperactivated CDK4/CDK6 including cyclin D1 amplification, CDK4 amplification, and deletion of CDKN2A are common events in multiple human cancer[5]. Hence, targeting CDKs as a potential therapeutic treatment for cancer has been promising, fuelling efforts in development of CDK inhibitors as anticancer drugs[5,12,13].

The CDK4/6 inhibitors act at the G1 to S cell transition phase checkpoint and prevents the progression and finally cell cycle arrest[16].Only recently there have been significant success with use of CDK inhibitors in clinical practice[16]. There are number of agents that can indirectly inhibit CDK4/6 activity, although currently CDK4/6 directed agents are of Palbociclib, Ribociclib and Abemaciclib. These molecules are developed to directly act towards targeting CDK4/6 and their specificity is accurate in differentiating with cyclin including D1,D2 and/or D3. CDK4/6 inhibitors first gained approval in 2015 when Palbociclib showed promising performance in treating patients with Breast cancer in clinical trials showing overall survival(OS) benefits[20]. Combinational therapy with letrozole(a selective aromatase inhibitor) in postmenopausal women with locally advanced or metastatic HER2 negative, estragon receptor positive breast cancer patients was approved. [5,12,15,17,20]. Ribociclib received its first FDA approval in 2017 for the treatment of postmenopausal hormone receptor (HR) positive HER2 negative metastatic breast cancer in combination with an aromatase inhibitor in postmenopausal hormone receptor (HR) positive HER2 negative metastatic breast cancer. [5,20]. Abemaciclib was first approved in 2017 for recurrence after endocrine therapy in women with HR positive, HER2 negative advanced breast cancer[20].

Based on the success in breast cancer patients there are currently five clinical trials involving CDK4/6 inhibitors in prostate cancer, using the "umbrella and basket" trails using Palbociclib, Ribociclib and Abemaciclib in treating prostate cancer patients[16]. A phase II study evaluating the use of ADT with or without Palbociclib is conducted on patients with mHSPC and RB positive this study has been completed and currently the safety and effectiveness is being studied[5]. A phase II study involving mCRPC patients dosed with palbociclib is under recruitment. The other study involving mCPRC patients also include the use of Ribociclib with Docetaxel is under phase II[5]. Also, Abiraterone with or without Abemaciclib is under phase II study for mCRPC patients under recruitment stage[5]. mCRPC with chemo-naive to retain RB expression is also under phase II study that uses dose of Enzalutamide with or without Ribociclib is also under recruitment stage[15]. There are other potential combinational therapies that can be used with existing cancer treatments including chemotherapy, immunotherapy, DNA repair pathway, FOXO3-FOXI Axis, PI3K/Akt Axis, FGF-FGFR Axis, Ras-Raf-MEK-ERK Axis and TP53 axis are under research[5,9,15,16]. Although there is potential use for CDK4/6 inhibitor in treating prostate cancer, there is uncertainty in understanding the resistance and sensitivity of patients who are likely to get benefited[5]. Understanding the biomarkers significantly improve the treatment for patient cohorts with CDK4/6 inhibitors[5].

2.7 Putative biomarkers of Resistance , Sensitivity

Approximately 20% of patients will not respond to CDK4/6 inhibitor treatments initially and almost all patients will develop resistance eventually[20]. A better understanding mechanism and identification of biomarkers and the intrinsic details on the acquired resistance will help guide therapy. Putative biomarkers discussed in section 2.3 including AR amplification, P53 loss of function, RAS/RAF/MAPK oncogene activation, PTEN loss of function, DNA repair, PI3K amplification, FGFR amplification SPOP oncogene activation and cell cycle activation as single biomarker identifiers cannot be fully utilised to all patients due to heterogenous nature of the cancer[14]. Even with significant information about the type of biomarkers and their corresponding function, identification of specific markers that corresponds to sensitivity remains unclear. Most of the previous research has concentrated on genetic events that are linked to CDK4/6 addiction and are widespread in cancer[12]. With diverse effects associated with each therapy and treatment identification of prescriptive pathways that best suit individual patients based on their PCa characteristics remains obscure[12,14,20].

Tumours with dysregulation of Cyclin D1 by amplification or translocation have been studied extensively[12]. Preclinical MCL models from mouse cells have shown to be sensitive to CDK4/6 inhibitors in case of Cyclin Amplification. The reason behind why CDK4/6 inhibitors are effective in these models are hard to evaluate. Direct involvement of genetic alteration of cyclin D1 gene to CDK4/6 inhibitors is still not clear which is the hallmark of cancer. Using cyclin D1 and their variants along with their amplification as predictive biomarkers to determine responsiveness of CDK4/6 treatment is obscure. Similarly, Loss of CDKN2A is also a frequent event observed in tumours[12]. CDKN2A loss has not been found to predict CDK4/6 inhibitor sensitivity in preclinical or clinical trials. There is evidence of acquired resistance to CDK4/6 inhibitors in models with p16INK4a genetic loss[12]. Loss of CDKN2A, on the other hand, has no evidence of a direct connection between response and tumour suppressor loss. Amplification of CDK4 or CDK6 is observed to be another event that is predominant in cancer[12]. Most models created for other cancer(liposarcoma) has directed towards using amplification of CDK4 or CDK6 as predictive biomarker for sensitivity of CDK4/6 inhibitors treatments. But with recent studies CDK4 has shown to drive the resistance to CDK4/6 inhibitors in preclinical models[12].

In total review the fact that they calculate different quantities under different conditions causes disagreements between methods[25]. As a result, one approach may be suitable for a particular kinase and/or drug than another, resulting in conflicting results for certain targets. To obtain reliable selectivity profiles for drugs and understand their mechanisms, multiple profiling approaches must be combined. Also, to better inform clinical decisions and improve the effectiveness and safety of cancer treatment, researchers can look at drug selectivity profiles and adverse effects. A thorough understanding of drug selectivity landscapes could also allow for the repurposing of existing drugs to treat new indications[25].

3. Summary of State of the Art.

CDK4/6 inhibitors were developed as a significant advance in the treatment of breast cancer, and their use has increasingly expanded to other oncological indications including Prostate cancer. The exact mechanism of action of approved compounds, on the other hand, is unknown. While cytostatic cell cycle arrest is the most common result of CDK4/6 inhibition, its effect on cellular processes is highly variable and appears to be dependent on intrinsic cellular programmes and/or cell cycle dynamics[25].

The benefits of CDK4/6i's in prostate cancer can be maximised by completely comprehending cell cycle pathways, CDK4/6i's resistance patterns, and therapies that target driver mutations in mCRPC. When agents are combined based on interrelated pathways, resistance profiles, sensitivity profiles and genomics, the best chance for synergistic success exists. Precision medicine can not only help improve treatment efficacy, but also protect and recognise patients who may not benefit from therapy, reducing toxicity and morbidity[5]. The significance of studying drug selectivity profiles and adverse effects to better inform clinical decisions and increase cancer therapy effectiveness and safety. A thorough understanding of drug selectivity landscapes could also allow for the repurposing of existing drugs to treat new indications[25].

Since cancers undergo complex changes during treatment and relapse, tracking these changes using response biomarkers can be useful in predicting which patients will benefit from CDK4/6 inhibitors at different stages. Furthermore, the most effective use of CDK4/6 inhibitors necessitates the use of accurate predictive biomarkers. Although basic molecular features (other than AR status) such as loss of the RB1 gene or upregulation of the CDK4/6 binding partner cyclin D1 have been proposed as potential biomarkers for predicting CDK4/6 inhibition resistance and sensitivity, respectively. None of these alone have demonstrated clinical utility[26]. Unlike other therapeutic kinases, CDK4/6 activation has been shown to be much too complex for any of these proteins or the pathway's components to predict inhibitor response. Usage of single biomarkers to understand the sensitivity and resistance of CDK4/6 inhibitors in treating prostate cancer is found to be highly ineffective[16]. As a result, rather than individual biomarkers, the validation of a panel of predictive biomarkers would be useful in future clinical trials[16].

4. Plan for Research project

4.1 Next Generation Sequencing multi-omics

High-throughput sequencing technologies have made it possible to visualise cancer samples at various molecular levels for the first time in recent years[29]. Integrating and analysing these multi-omics datasets is a key step in gaining actionable information in a precision medicine environment. Precision medicine is focusing on the integration and study of high-throughput molecular assays to better understand patient and disease specific variations. Integrated approaches enable detailed views of genetic, biochemical, metabolic, proteomic, and epigenetic mechanisms underlying a disease that would otherwise be impossible to examine using single-omics methods. Computational multi-omics methods use machine learning techniques to identify patients into cancer subtypes, with the aim of finding biomarkers and repurposing drugs[29]. Although the dynamics of cancer continue to obstruct our understanding of how it develops and progresses, multi-omics methods have been proposed as promising tools for dissecting patient dysfunctions in multiple biological systems that may be influenced by cancer mechanisms[27]. Several collaborative efforts have been made to catalogue cancer cell line molecular profiling data and drug sensitivity data with the aim of identifying genomic biomarkers that predict anticancer drug response. The Genomics in Drug Sensitivity in Cancer (GDSC, https://www.cancerrxgene.org) database, for example, contains experimentally tested drug sensitivities of 1,001 human cancer cells to 265 anticancer compounds. Importantly, as part of the COSMIC cell line project (CCLP, https://cancer.sanger.ac.uk/cosmic), the molecular profiles of all the cancer cell lines used in GDSC were thoroughly characterised, including profiling of somatic genomic alterations. The realisation of genomic-driven precision cancer medicine is expected to benefit greatly from these tools. Despite their potential value, the high dimensionality and sophistication of such databases makes integrative analysis difficult[29].

4.2 Applications of multi-omics

To gain actionable knowledge from data produced by multi-omics sequencing, computational approaches ranging from data integration to statistical methods and artificial intelligence systems must be developed[27]. When such services are paired with clinical data on patient characteristics and treatment outcomes over the course of cancer progression and relapses, integrated approaches to improving both diagnostic and therapeutic options are possible. Unlike traditional clinical management, which treats cancers as homogeneous entities, “precision oncology” aims to find a molecularly targeted treatment for each cancer patient sub-type or individual patient (i.e., stratified or personalised medicine)[27]. Somatic aberrations, such as genetic mutations or molecular modifications, are often used to match available medication to patients, provided that therapeutically actionable markers are discovered and can be used in clinical practise. The ability to identify panels of biomarkers associated with treatment responses in each patient cohort is thus a significant pre-requisite for the precision oncology approach. We refer to such markers as "predictive biomarkers" if they generalise beyond the discovery cohort to new cancer patients[28].

4.3 Machine learning

Machine learning is a new area in computational chemistry that has a wide range of applications, including drug discovery, cheminformatics, and predictive toxicology[29]. Building a model, training the model, conducting validation, repeating training and validation until a suitable model is obtained, and finally testing the model on data not previously exposed to the model are all examples of machine learning[28,29,30,33]. A machine learning model's aim is to extract patterns from input data, or to generalise it, and then apply the results to unknown test data[29]. To step up a machine learning model the data needs to be integrated from various sources including GDSC and CCLP which contains the sensitivity profiles of anticancer drug compounds and molecular profiles of all cancer cell lines used in GDSC. After that, the data is analysed to see if it is suitable for use as model input. Missing, invalid, or unwanted data, for example, will normally be deleted. Working with unbalanced data, which is popular in machine learning, is another aspect of data processing[30]. To train high quality models, several methods were used to balance the classes in the data set. Oversampling, under sampling, or a combination of both are popular techniques used in these processes[28]. Testing and analysing raw data, as well as checking the accuracy of end-point data, are critical steps that are often ignored, resulting in the creation of weak models. Following the preparation of the data collection, the next step is to create a machine learning model. In general, there are three types of learning models: supervised learning, unsupervised learning, and semi-supervised learning[30]. Traditional analytical techniques for identifying treatment response-associated markers usually begin with unsupervised clustering of patient samples' molecular and/or genomic profiles, and then attempt to classify therapies that display therapeutic efficacy in different sample sub-clusters[29]. Alternatively, one may begin with treatment response clustering and then look for genomic or molecular correlates that explain the observed drug sensitivity or resistance clusters in patient-derived samples. The input data form, dimensionality, noise ratio, data heterogeneity, and complexity, as well as the specific prediction issue, all play a role in the development of drug response prediction models[29].

4.4 Supervised, Unsupervised and Deep Learning

A supervised machine learning model attempts to learn a function f(x) = y from a set of training pairs (x1,y1), (x2,y2), and so on. Predicting the viability of a cancer cell line when exposed to a specific drug is a popular application in biology. The input features (x) would capture cell line somatic sequence variations, drug chemical make-up, and concentration, which, along with the calculated viability (output mark y), could be used to train a support vector machine, a random forest classifier, or a similar method (functional relationship f)[31].

Without the use of output labels y, unsupervised machine learning methods seek to discover patterns from the data samples x themselves. Unsupervised models used on biological data include clustering, principal component analysis, and outlier identification, to name a few. In these models "seeing about the universe" is represented by the inputs x, which are calculated from raw data, and their selection is highly problem specific. The process of determining the most insightful features is critical for success, but it can be time-consuming and involves domain knowledge[31].

A deep neural network takes raw data from the lowest (input) layer and transforms it into increasingly abstract feature representations by data-driven merging outputs from previous layers, encapsulating extremely complicated functions in the process. Deep learning's promise in high-throughput biology is clear: in theory, it allows for better use of increasingly broad and high-dimensional data sets (for example, from DNA sequencing, RNA measurements, flow cytometry, or automated microscopy) by training complex networks with multiple layers that capture their internal structure[31]. The learned networks identify high-level features, outperform conventional models, enhance interpretability, and add to our understanding of the nature of biological data[31].

4.5 Proposed methods for Biomarker panel Identification

We will use Dr. Shin's Artificial Neural Network (ANN) to create predictive biomarkers for CDK4/6 inhibitors using a new ground-breaking ML-based pipeline. Figure 1 shows a workflow that depicts the whole project schedule.

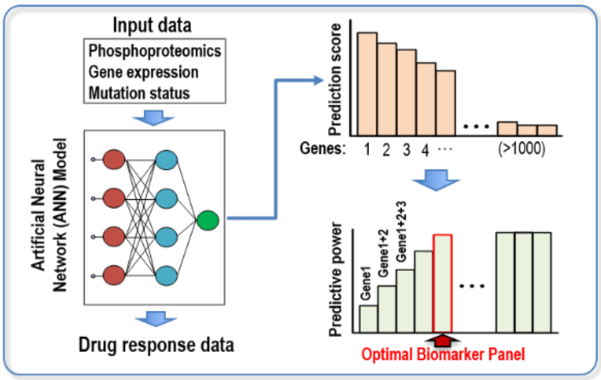


Figure 1: Workflow of proposed biomarker panel identification

Large-scale -omics data, such as phospho-proteomics (in-house and from the literature), mutation and gene expression (COSMIC/CCLE [32]) data from hundreds of cancer cell lines (including prostate cancer cell lines), and proteomic data from prostate derived explant (PDE) samples from our collaborators, will be used as input into the ML models and the ML model contribution (e.g., palbociclib/ribociclib, from the Genomics of Drug Sensitivity in Cancer project [33]). The ML models will be used to evaluate and rank the ability of individual features (e.g., single gene/protein) to predict response to CDK4/6 inhibition after splitting the data into a training and validation package. Then, using the most predictive gene as input, the 1st+2nd most predictive genes as input, and so on, until the overall model predictive score does not improve any further, we can iteratively retrain our ML model until the overall model predictive score does not improve any further, at which point the current input gene list will constitute the most predictive yet compact multi-gene biomarker for CDK4/6 inhibition sensitivity (Figure 1). Patient survival and gene expression data will be used to validate the best biomarker panel. This research is important because it will include new predictive biomarkers for anti-CDK4/6 therapy, which will greatly help the individualised care of prostate cancer patients.

5. Conclusion

Despite several excellent preclinical studies, well-designed clinical trials will be needed to determine the best use of CDK4/6 inhibition for a specific tumour[12]. Furthermore, since the clinical emphasis with CDK4/6 inhibitors has moved to combination therapy, it is highly likely that determining determinants for CDK4/6 inhibition in isolation would be a secondary concern going forward. A hybrid biomarker panel composed of multiple genes/proteins, which more accurately captures the multi-factorial processes regulating the response to CDK4/6 inhibitors, is the best way to predict sensitivity to CDK4/6 inhibitors. For multi-marker statistical regression of drug response patterns, a single cancer type and/or multiple treatment profiles are often underpowered. The biomarker discovery challenge is further complicated by the lack of available clinical treatment knowledge for the profiled patient tumours. Single markers, whether derived from somatic mutations or other omics results, are not typically reliable enough for explaining treatment responses for most drug groups, except for a few notable cases of clinical success. Instead, larger sample sizes and multivariate modelling are needed to classify accurate marker panels[28]. The solution provide as part of this project will help to identify robust predictive biomarkers that can accurately predict sensitivity and resistance to CDK4/6 inhibition for prostate cancer using machine learning (ML) approaches, which will be vital for patient stratification and personalised treatment.

6. Reference

1. Verze, P., Cai, T. & Lorenzetti, S. The role of the prostate in male fertility, health and disease. Nat. Rev.Urol. 13, 379–386 (2016).
2. Attard, G. et al. Prostate cancer. Lancet 387, 70–82
3. Rebello, Richard J, Oing, Christoph, Knudsen, Karen E, Loeb, Stacy, Johnson, David C, Reiter, Robert E, Gillessen, Silke, Van der Kwast, Theodorus, & Bristow, Robert G. (2021). Prostate cancer. Nature Reviews Disease Primers., 7(1). https://doi.org/10.1038/s41572-020-00243-0
4. Bray, F. et al. Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA Cancer J. Clin. 68, 394–424 (2018).
5. Kase, A. M., Copland, J. A., & Tan, W. (2020). Novel Therapeutic Strategies for CDK4/6 Inhibitors in Metastatic Castrate-Resistant Prostate Cancer. OncoTargets and Therapy, Volume 13, 10499-10513. doi:10.2147/ott.s266085
6. Key Statistics for Prostate Cancer. American Cancer Society. Accessed April 24, 2019., 2019,https://www.cancer.org/cancer/%20prostate-cancer/about/key-statistics.html.
7. Overview of the treatment of castration-resistant prostate cancer (CRPC). UpToDate, 2018. (Accessed April 24, 2019., 2019

https://www.uptodate.com/contents/search?search=overview-of-the-%20treatment-of-castration-resistant-prostate-cancer-crpc&sp=0&searchType=PLAIN_TEXT&source=USER_INPUT&searchControl=TOP_PULLDOWN&searchOffset=1&autoComplete=false&language=&max=0&index=&autoCompleteTerm= .

8. Couñago, Felipe, López-Campos, Fernando, Díaz-Gavela, Ana Aurora, Almagro, Elena, Fenández-Pascual, Esaú, Henríquez, Iván, Lozano, Rebeca, Linares Espinós, Estefanía, Gómez-Iturriaga, Alfonso, de Velasco, Guillermo, Quintana Franco, Luis Miguel, Rodríguez-Melcón, Ignacio, López-Torrecilla, José, Spratt, Daniel E, Guerrero, Luis Leonardo, Martínez-Salamanca, Juan Ignacio, & del Cerro, Elia. (2020). Clinical Applications of Molecular Biomarkers in Prostate Cancer. Cancers., 12(6), 1550–25. https://doi.org/10.3390/cancers12061550
9. Palmbos PL, Tomlins SA, Agarwal N, et al. Cotargeting AR signaling and cell cycle: a randomized phase II study of androgention therapy with or without palbociclib in RB-positive metastatic hormone sensitive prostate cancer (mHSPC). J Clin Oncol. 2018;36 (6_suppl):251. doi:10.1200/JCO.2018.36.6_suppl.251
10. Angeles, A.; Bauer, S.; Ratz, L.; Klauck, S.; Sültmann, H. Genome-Based Classification and Therapy of Prostate Cancer. Diagnostics 2018, 8, 62. [CrossRef]
11. Kontos, Christos K, Avgeris, Margaritis, Scorilas, Andreas, Atta-ur-Rahman, & Choudhary, M Iqbal. (2018). Biomarkers with Prognostic Potential in Prostate Cancer. In Frontiers in drug design and discovery. (Vol. 1, Issue 1, pp. 108–134). Bentham Science Pub. https://doi.org/10.2174/9781681085821118090003
12. Knudsen, E. S., & Witkiewicz, A. K. (2017). The Strange Case of CDK4/6 Inhibitors: Mechanisms, Resistance, and Combination Strategies. Trends in Cancer, 3(1), 39-55. doi:10.1016/j.trecan.2016.11.006
13. Karp G. Cell and Molecular Biology: Concepts and Experiments. 6th ed. Hoboken, NJ: John Wiley and Sons; 2010.
14. Cucchiara, Vito, Cooperberg, Matthew R, Dall'Era, Marc, Lin, Daniel W, Montorsi, Francesco, Schalken, Jack A, & Evans, Christopher P. (2018). Genomic Markers in Prostate Cancer Decision Making. European Urology : Official Journal of the European Association of Urology, 73(4), 572–582. https://doi.org/10.1016/j.eururo.2017.10.036
15. Sedlacek H, et al. Flavopiridol (L86 8275; NSC 649890), a new kinase inhibitor for tumor therapy. Int J Oncol. 1996; 9:1143–1168. [PubMed: 21541623]
16. Chong, Qing-Yun, Kok, Ze-Hui, Bui, Ngoc-Linh-Chi, Xiang, Xiaoqiang, Wong, Andrea Li-Ann, Yong, Wei-Peng, Sethi, Gautam, Lobie, Peter E, Wang, Lingzhi, & Goh, Boon-Cher. (2020). A unique CDK4/6 inhibitor: Current and future therapeutic strategies of abemaciclib. Pharmacological Research : the Official Journal of the Italian Pharmacological Society., 156. https://doi.org/10.1016/j.phrs.2020.104686
17. Meijer L. Chemical inhibitors of cyclin-dependent kinases. Prog Cell Cycle Res. 1995; 1:351–363. [PubMed: 9552377]
18. Sherr CJ. D-type cyclins. Trends Biochem Sci. 1995; 20:187–190. [PubMed: 7610482]
19. Knudsen KE, et al. Cyclin D1: polymorphism, aberrant splicing and cancer risk. Oncogene. 2006; 25:1620–1628. [PubMed: 16550162]
20. Shah, M. (2018). CDK4/6 inhibitors: Game changers in the management of hormone receptor– positive advanced breast cancer? Oncology., 32(5), 216–222. https://doi.org/info:doi/
21. Baselga J, et al. Everolimus in postmenopausal hormone-receptor-positive advanced breast cancer. N Engl J Med. 2012; 366:520–529. [PubMed: 22149876]
22. Bottcher R, Hoogland AM, Dits N, et al.Novel long non-coding RNAs are specific diagnostic and prognostic markers for prostate cancer.Oncotarget. 2015;6(6):4036–4050.
23. Shi, Jingqi, Jiang, Dongbo, Yang, Shuya, Zhang, Xiyang, Wang, Jing, Liu, Yang, Sun, Yuanjie, Lu, Yuchen, & Yang, Kun. (2020). LPAR1, Correlated With Immune Infiltrates, Is a Potential Prognostic Biomarker in Prostate Cancer. Frontiers in Oncology., 10. https://doi.org/10.3389/fonc.2020.00846
24. López-Campos, Fernando, Linares-Espinós, Estefanía, Maldonado Pijoan, Xavier, Sancho Pardo, Gemma, Morgan, Todd Mathew, Martínez-Ballesteros, Claudio, Martínez-Salamanca, Juan, & Couñago, Felipe. (2020). Genetic testing for the clinician in prostate cancer. Expert Review of Molecular Diagnostics., 20(9), 933–946.

https://doi.org/10.1080/14737159.2020.1816170

25. Hendrychová, Denisa, Jorda, Radek, & Kryštof, Vladimír. (2021). How selective are clinical CDK4/6 inhibitors? Medicinal Research Reviews., 41(3), 1578–1598. https://doi.org/10.1002/med.21769
26. Álvarez-Fernández, Mónica, & Malumbres, Marcos. (2020). Mechanisms of Sensitivity and Resistance to CDK4/6 Inhibition. Cancer Cell., 37(4), 514–529. https://doi.org/10.1016/j.ccell.2020.03.010
27. Nicora, Giovanna, Vitali, Francesca, Dagliati, Arianna, Geifman, Nophar, & Bellazzi, Riccardo. (2020). Integrated Multi-Omics Analyses in Oncology: A Review of Machine Learning Methods and Tools. Frontiers in Oncology., 10. https://doi.org/10.3389/fonc.2020.01030
28. Ali, Mehreen, & Aittokallio, Tero. (2019). Machine learning and feature selection for drug response prediction in precision oncology applications. Biophysical Reviews., 11(1), 31–39. https://doi.org/10.1007/s12551-018-0446-z
29. Chang, Yoosup, Park, Hyejin, Yang, Hyun-Jin, Lee, Seungju, Lee, Kwee-Yum, Kim, Tae Soon, Jung, Jongsun, & Shin, Jae-Min. (2018). Cancer Drug Response Profile scan (CDRscan): A Deep Learning Model That Predicts Drug Effectiveness from Cancer Genomic Signature. Scientific Reports., 8(1). https://doi.org/10.1038/s41598-018-27214-6
30. Wang, Marcus W H, Goodman, Jonathan M, & Allen, Timothy E H. (2021). Machine Learning in Predictive Toxicology: Recent Applications and Future Directions for Classification Models. Chemical Research in Toxicology., 34(2), 217–239. https://doi.org/10.1021/acs.chemrestox.0c00316
31. Angermueller, Christof, Pärnamaa, Tanel, Parts, Leopold, & Stegle, Oliver. (2016). Deep learning for computational biology. Molecular Systems Biology, 12(7). https://doi.org/10.15252/msb.20156651
32. Nusinow, David P, Szpyt, John, Ghandi, Mahmoud, Rose, Christopher M, McDonald, E Robert, Kalocsay, Marian, Jané-Valbuena, Judit, Gelfand, Ellen, Schweppe, Devin K, Jedrychowski, Mark, Golji, Javad, Porter, Dale A, Rejtar, Tomas, Wang, Y Karen, Kryukov, Gregory V, Stegmeier, Frank, Erickson, Brian K, Garraway, Levi A, Sellers, William R, & Gygi, Steven P. (2020). Quantitative Proteomics of the Cancer Cell Line Encyclopedia. Cell, 180(2), 387–402.e16. https://doi.org/10.1016/j.cell.2019.12.023
33. Yang, Wanjuan, Soares, Jorge, Greninger, Patricia, Edelman, Elena J, Lightfoot, Howard, Forbes, Simon, Bindal, Nidhi, Beare, Dave, Smith, James A, Thompson, I Richard, Ramaswamy, Sridhar, Futreal, P Andrew, Haber, Daniel A, Stratton, Michael R, Benes, Cyril, McDermott, Ultan, & Garnett, Mathew J. (2012). Genomics of Drug Sensitivity in Cancer (GDSC): a resource for therapeutic biomarker discovery in cancer cells. Nucleic Acids Research., 41(D1), D955–D961. https://doi.org/10.1093/nar/gks1111
34. Loeb, S., & Catalona, W. J. (2014). The Prostate Health Index: a new test for the detection of prostate cancer. Therapeutic advances in urology, 6(2), 74–77. https://doi.org/10.1177/1756287213513488
35. Punnen, S., Pavan, N., & Parekh, D. J. (2015). Finding the Wolf in Sheep's Clothing: The 4Kscore Is a Novel Blood Test That Can Accurately Identify the Risk of Aggressive Prostate Cancer. Reviews in urology, 17(1), 3–13.
36. Wei, John T. Urinary biomarkers for prostate cancer, Current Opinion in Urology: January 2015 - Volume 25 - Issue 1 - p 77-82 doi: 10.1097/MOU.0000000000000133

Part II : Research Paper

Abstract

Inhibitors palbociclib, ribociclib, and abemaciclib of the Cyclin-dependent kinases CDK4 and CDK6 have demonstrated to improve treatment results in ER+ breast cancer. Given their promising results, these inhibitors are being studied for use in various cancer treatments including prostate cancer. However, due to resistance mechanisms that develop over time there is currently a lack of knowledge on patients who may have potential and benefit from CDK4/6 inhibitor therapy. Patient stratification based on single biomarkers have been shown to be unsuccessful in understanding the sensitivity and tolerance of CDK4/6 inhibitors in treating prostate cancer. In this study, we incorporated a machine learning-based framework to identify robust predictive multipanel biomarkers that can effectively predict sensitivity and resistance to CDK4/6 inhibition for prostate cancer, which will be critical for patient stratification and customised treatment. We demonstrated the power of this approach by applying it on pan-cancer cell-line genomic/proteomic data with CDK4/6 inhibitor palbociclib drug sensitivity and patient-derived explant samples of prostate cancer treated using ribociclib inhibitor. To improve the model's predictive performance, a rational feature section strategy was used together with Boolean algebra approaches to construct specific expression signatures for the marker proteins.

Keywords: Predictive biomarker, machine learning, Prostate cancer, predictive medicine, Pan cancer, Expression signature, CDK4/6 inhibitor

2.1 Introduction

CDK4/6 inhibitors were initially created to treat breast cancer, but their use has since expanded to include additional oncological diseases such as prostate cancer. Understanding cell cycle pathways, CDK4/6i resistance patterns, and therapies that target driver mutations in mCRPC can maximise the benefits of CDK4/6i's in prostate cancer.

The best opportunity for synergistic success arises when drugs are coupled based on connected pathways, resistance profiles, sensitivity profiles, and genomics. Precision medicine can help increase therapeutic efficacy while protecting and identifying patients who may not benefit from therapy, by lowering toxicity and morbidity[1]. It is important to understand drug selectivity profiles and side effects to better inform clinical decisions and improve the efficacy and safety of cancer therapies.

The ability to repurpose existing medications to treat novel indications could be aided by a detailed grasp of drug selectivity landscapes [2]. As malignancies undergo complicated changes during therapy and relapse, employing response biomarkers to track these changes can help predict which patients will benefit from CDK4/6 inhibitors at various stages. Furthermore, precise predictive biomarkers are required for the most efficient usage of CDK4/6 inhibitors. Although basic biological characteristics (other than AR status) like deletion of the RB1 gene or overexpression of the CDK4/6 binding partner cyclin D1 have been proposed as potential biomarkers for predicting CDK4/6 inhibition resistance and sensitivity, none of them have been shown to be clinically useful on their own[3]. As a result, precise biomarker discovery for targeted drugs that can guide patient classification for better clinical trial design and therapeutic results is becoming increasingly relevant.

The multi-factorial regulator of cellular response to drug therapy, which is exacerbated by the vast molecular heterogeneity between patients, is a key difficulty in identifying predictive drug-response biomarkers. As a result, single markers are often insufficient for explaining treatment responses and thus unlikely to be clinically robust for most medication classes [6]. This finding has been true, apart from a few prominent cases of clinical success [4, 5]. Biomarker panels with many molecular actors, on the other hand, are more likely to capture the multi-factorial mechanisms that drive cellular drug response [7, 8].

The identification of medication response-related biomarkers has been aided by computational approaches [9-11]. Based on omics data, one typical method is to identify molecular entities (genes or proteins) that are differentially expressed between treatment-sensitive and treatment-resistant groups. Despite differentially expressed genes/proteins (DEG/Ps) are an excellent place to start [12,13], the degree of differential expression (based on fold-change and/or p-value) of a gene is frequently not a strong predictor of drug responsiveness. Furthermore, the often-lengthy lists of DEGs make prioritizing difficult in many circumstances, restricting the application of DEG-based techniques. Machine learning-based biomarker discovery techniques have recently been developed to take advantage of the growing availability of large-scale omics data. However, most machine learning studies to date have been conducted using panels of cancer cell lines [14, 15-17], which do not always reflect drug sensitivity in human tumours [18, 19]. This is due to a paucity of focused patient-relevant molecular and drug response data.

In this study, we created a novel and generalized machine learning-based methodology for identifying anti-cancer agent predictive multi-marker panels. The methodology is demonstrated by generating predictive biomarker panels and marker signatures for pan-cancer and prostate cancer datasets. Genomic of Drug Sensitivity in Cancer (GDSC) (https://www.cancerrxgene.org/) and COSMIC cell line project (CCLP) (https://cancer.sanger.ac.uk/) provide pan-cancer data. Data from GDSC provides a collection comprising of well over 1,000 human tumour cell lines of common and rare types of adult and childhood cancers of epithelial, mesenchymal and haematopoietic origin with Palbociclib inhibitor directed on them[20]. CCLE contains expanded characterizations of CCLE cancer cell lines including RNA sequencing, whole-exome sequencing, whole-genome sequencing, global histone modification profiling, and metabolomics[21]. This methodology is also used to analyse pharmacoproteomic data from prostate cancer patient-derived explants (PDEs) treated with ribociclib CDK4/6 inhibitor.

Our methodology used Boolean algebra methods to create concise and predictive expression patterns suggesting an individual patient's response to CDK4,6 inhibitor treatment, using genomics/proteomics data from pan-cancer and prostate cancer. Prior to using machine learning, we undertook a feature selection process in which features (proteins) were sensibly chosen based on how they individually affect overall drug response prediction. The marker signatures were then deduced using engineering methods such as Boolean logic operations and Boolean function reduction algorithms.

As a result, our system produced a biomarker panel of 10 genes that had a drug-response prediction accuracy of 66 percent for pan-cancer dataset. Similarly, prostate cancer data produced a biomarker panel of 10 genes with 93 percent drug-response prediction accuracy, far better than conventional approaches. Finally, we compared the gene panel between both these scenarios and found their biological functions.

After interrogating prostate cancer patient cohorts, we identified nearly half of the patients with matching expression signatures who may benefit from ribociclib treatment. Overall, this research introduces a unique machine learning-based framework for delivering accurate multi-marker panels for improved patient selection and treatment of prostate cancer, which can also be applied to other tumour types.

2.2 Background

2.2.1 Prostate Cancer , Current treatment and Limitations

Prostate cancer is one of the most diagnosed cancers in men, and it generates irreversible somatic mutations. Patients are either overtreated or undertreated due to the variety and complex nature of epigenetic modifications, even with advances in early detection when PCa is most curable with thorough screening. Patients diagnosed with PCa are divided into risk groups based on pathological and clinical criteria. While these stratifications provide important information about tumours and their predicted behaviour, they frequently yield unsatisfactory results when it comes to classifying aggressive tumors[26]. As a result, treatment may be insufficient or excessive. Moreover, in the clinical management of PCa, real customised therapies guided by prognostic biomarkers are lacking[26,27]. With the extra challenge of clinical decision-making in prognosis and diagnosis, we must reliably identify patients who are most likely to respond to a certain therapy, create treatment sequences, and enrol these patients in targeted therapies[26,28]. Finally, most predictive biomarkers derived from a variety of sources, including genetic data, are univariate and typically poor at predicting malignant progression. With the use of multi-omics for precision PCa oncology, there is a gap in the identification of biomarker panels[27,28,29,30].

2.2.2 CDK4/6 inhibitors and putative biomarkers

Multiple genetic events in cancer that result in overexpressed CDK4/6 disrupt the importance of cyclin regulation and appropriate function in cell cycle in normal tissue homeostasis, including cyclin D1 amplification, CDK4 amplification, and deletion of CDKN2A are common events in multiple human cancers[1]. As a result, targeting CDKs as a potential cancer therapeutic treatment has shown promise, spurring efforts to produce CDK inhibitors as anticancer drugs[1,27,30]. CDK4/6 inhibitors block the passage of the G1 to S cell transition phase checkpoint, resulting in cell cycle arrest[31]. The use of CDK inhibitors in clinical practise has only recently seen significant success[31]. There are several drugs that can suppress CDK4/6 activity indirectly, however the only CDK4/6 directed medications now available are Palbociclib, Ribociclib, and Abemaciclib.

Palbociclib, a CDK4/6 inhibitor, received FDA approval in 2015 after clinical trials revealed that it was effective in treating patients with breast cancer, with overall survival (OS) benefits[32]. Approximately 20% of patients will initially fail to react to CDK4/6 inhibitor medication, and almost all patients will develop resistance at some point[32]. A deeper knowledge of the mechanism, as well as the identification of biomarkers and the inherent characteristics of acquired resistance, will aid in the treatment of this resistance.

Due to the heterogeneous nature of cancer, single biomarker identifiers such as AR amplification, P53 loss of function, RAS/RAF/MAPK oncogene activation, PTEN loss of function, DNA repair, PI3K amplification, FGFR amplification, SPOP oncogene activation, and cell cycle activation cannot be fully utilised to all patients[33]. Despite having a lot of information about the different types of biomarkers and their functions, the identification of specific markers that correspond to sensitivity is still a mystery. Because each therapy and treatment have different side effects, identifying prescriptive paths that best suit specific patients based on their PCa characteristics is difficult[30,32,33].

Multiple profiling methodologies must be coupled to create credible selectivity profiles for medications and to understand their processes. Researchers can also look at medication selectivity profiles and adverse effects to better guide clinical decisions and increase the effectiveness and safety of cancer treatment. The ability to repurpose existing medications to treat novel indications could be aided by a detailed grasp of drug selectivity landscapes[2].

2.2.3 Plan for research

For the first time in recent years, high-throughput sequencing methods have enabled the visualisation of cancer samples at numerous molecular levels[34]. Integrated methodologies provide extensive perspectives of genetic, biochemical, metabolic, proteomic, and epigenetic systems that would otherwise be hard to investigate with single-omics methods. Machine learning algorithms are used in computational multi-omics methods to classify patients into cancer subgroups to locate biomarkers and repurpose drugs[34].

Several joint initiatives have been conducted to record cancer cell line molecular profiling and drug sensitivity data in the hopes of discovering genetic biomarkers that predict anticancer drug response. For example, the Genomics in Drug Sensitivity in Cancer (GDSC, https://www.cancerrxgene.org) database contains experimentally tested drug sensitivities of 1,001 human cancer cells to 265 anticancer drugs. The molecular profiles of all the cancer cell lines utilised in GDSC were thoroughly characterised as part of the COSMIC cell line project (CCLP, https://cancer.sanger.ac.uk/cosmic), which included profiling of somatic genomic changes

Computational technologies ranging from data integration to statistical methods and artificial intelligence systems must be developed to extract actionable knowledge from data generated by multi-omics sequencing[35]. Integrated approaches to improve both diagnostic and therapeutic choices are possible when such services are combined with clinical data on patient characteristics and treatment outcomes across the course of cancer progression and relapses. Unlike traditional clinical care, which treats tumours as if they were all the same, *precision oncology* tries to develop a molecularly tailored treatment for each cancer patient sub-type or individual patient (i.e., stratified or personalised medicine)[35].

To improve a machine learning (ML) model, data from many sources must be combined, including GDSC and CCLP, which contain anticancer drug sensitivity profiles and molecular profiles of all cancer cell lines used in GDSC. We can then iteratively retrain our ML model using the most predictive gene as input, the 1st+2nd most predictive genes as input, and so on, until the overall model predictive score does not improve any further. At which point the current input gene list will constitute the most predictive yet compact multi-gene biomarker for CDK4/6 inhibition sensitivity. The best biomarker panel will be validated using patient survival and gene expression data.

This study is significant because it will uncover novel predictive biomarkers for anti-CDK4/6 therapy, which will considerably aid in the treatment of prostate cancer patients who require individualised care. The solution developed as part of this project will aid in the identification of robust predictive biomarkers for prostate cancer that can accurately predict sensitivity and resistance to CDK4/6 inhibition using ML approaches, which will be critical for patient stratification and individualised treatment.

2.3 Materials and Methodology

2.3.1 Dataset of ribociclib drug response from prostate PDE tissue

We previously developed an *ex vivo* culturing model of prostate cancer tissue that retains the structure and stromal-epithelial interactions of the tumour microenvironment and provides the level of disease heterogeneity seen in patients to better recapitulate the in vivo response of prostate cancer to therapies[3]. In a previous work, we used this technique to create 30 prostate cancer patient-derived explants (PDEs) and treated them after 48 hours with either vehicle (DMSO) or ribociclib (500 nM) [28]. The relative expression of the proliferative marker Ki67, assessed by immunohistochemical analysis after medication treatment, was used to quantify treatment response [28]. On the matching 30 PDEs, we also did mass spectrometry-based proteome profiling and HRM-DIA data processing, which revealed the expression of 4675 measurable proteins prior to ribociclib therapy [28].These unique datasets will be used in this study to create novel multi-protein biomarker panels that properly predict ribociclib therapy response.

To label the data, the PDE samples were divided into two groups based on Ki67 positive after ribociclib therapy [28]. These are shown in Figure 1B:(i) the RD (responders) group, which included PDEs with a two-fold decrease in Ki67 positivity; and (ii) the NR (non-responders) group, which included PDEs with Ki67 positivity levels that were in the middle of the two cut-offs. As a result, 15 PDEs were labelled RD and 15 were labelled NR. The PDE datasets are made up of 4675 PDE-specific protein expression levels that act as 'inputs' and Ki67-based response classification that serve as 'labelled outputs' for predictive model construction.

2.3.2 Description of GDSC and CCLE dataset for Pan-cancer

The CCLE and GDSC are two large panels of comprehensively characterised human cancer models that have provided a rigorous framework for studying genetic variants, candidate targets, small-molecule and biological therapeutics, and identifying new marker-driven cancer dependencies.

Characterizations of cancer cell lines to include; genetic, RNA splicing, DNA methylation, histone H3 modification, microRNA expression, and reverse-phase protein array data for 1,072 cell lines from individuals of various lineages and ethnicities to improve our understanding of the molecular features that contribute to cancer phenotypes, including drug responses [36].

Palbociclib, a CDK4/6 inhibitor that targets the cell cycle pathway that contains 901 different cell lines, was employed in this study's GDSC data. This collection contains genomic profiles of cell lines as well as pharmacological perturbation results in terms of IC50 values, AUC, and cytotoxic activities of compounds [37]. The tissue and tissue subtype for the matching cell line are also included in the data, which comprises of 13 unique tissues and 53 unique tissue subtypes. The CCLE dataset provides gene RNA sequencing based on differential expression of 56203 genes across 1019 cell lines as rows and columns. The tissue type and cell line names are separated by an underscore in the cell line names. As shown in Figure 1C, new cell lines for the GDSC dataset are formed by combining cell line names with tissue type names.

For each cell line, a category column is constructed based on the IC50 value median; if IC50>4.66107, then *Sensitive* otherwise, *Insensitive*. With 558 cell lines and 53059 genes, both the GDSC and CCLE datasets are ensembled on cell line names. The Pvalues is then calculated using a two-sample T-test comparing the sensitive and insensitive groups, and Pearson's correlation between gene expression and IC50 values is derived (Rho). The distribution between -log10 of Pvalues and absolute of Rho is shown in Figure 2F. Pvalue of 0.01 is utilised with a correlation coefficient of +0.2 or -0.2 shown by red lines. The most significant observations were obtained using these thresholds with 668 genes and 559 cell lines. Following that, the dataset is transposed to have genes as columns and cell lines as rows, which will be utilised to train various models in the following steps.

2.3.3 A novel machine learning framework maximises prediction accuracy through rational selection of input features

We used a systematic 'feature drop-out' analysis to choose the most relevant and informative input features as part of the feature space. Rather of using all differentially expressed proteins (DEPs) as input features, we systematically deleted one protein from the feature space at a time to see how the dropped-out protein affected the overall model's drug-response prediction. We compared the performance of machine learning models using the remaining feature space to the performance of the original model for each eliminated protein (a workflow is given in Figure. 1A).

The Caret package in R was used to create machine learning models such as Support vector machines (SVM) with multiple kernels such as linear, radial, and polynomial, Random Forest (Rf), Extreme gradient boosting (XGBoost), Gradient boosting (GBM), and Logistic regression. Based on their performance, the best model is picked and used for future study. If removing a protein lowers (or raises) model prediction accuracy, the protein is said to have a positive (or negative) impact on drug response prediction shown in Figure 2E for PDEs and Figure 2G for Pan-cancer dataset. To quantify these impacts, an *influence score* (IS) for each protein was calculated, which was defined as the difference in prediction accuracy between the drop-out and original machine learning models. Positive-effect proteins have an IS > 0, negative-impact proteins have an IS < 0, and proteins with no impact on response prediction have an IS = 0. (Figure. 3A, also see Appendices for more detail).

We reasoned that the positive-impact DEPs would be suitable candidate features for maximising the response prediction to ribociclib treatment because of their influence as shown in Figure 3B. A new technique for picking the optimum combination of features from the positive-impact DEPs to improve model prediction was proposed. This process was repeated by gradually adding the next most important DEP to the feature space. If the new DEP improves prediction accuracy, it is retained as a feature; if it decreases (or has no effect), the protein is skipped, and we move on to the next positive-impact DEP. This was done until all the positive-impact DEPs had been considered and the model's predicted performance had not improved any further (Figure. 3C)

2.3.4 The development of a compact biomarker panel with significant prediction power

In general, the size of a predictive biomarker panel and its practical application are trade-offs. Including more relevant proteins in a panel improves prediction accuracy, it comes at the expense of having to test a larger number of proteins from patient samples, which isn't always easy for novel or poorly described markers. As a result, to make possible predictive biomarker panels more clinically applicable needed to be used. We've narrowed down the DEPs described above to create a more compact panel with reasonable predictive value that might serve as a useful companion biomarker panel. To do this, we examined all conceivable combinations of proteins from the DEP pool into panels of increasing size. We then used each of these panels as input features to test the machine learning model's prediction performance.

2.3.5 Derivation of specific expression signatures of biomarkers for patient stratification

So far, our research has developed a protein biomarker panel that, when combined, provides the most information for predicting drug response sensitivity. The next step is to create precise expression profiles for these markers that can be used to stratify patients in the future.

Traditional statistical procedures like the T-test and boxplot analysis are frequently used to compare the expression patterns of individual marker proteins across response groups. While beneficial, this method has limitations because it ignores variation in the expression of individual markers among samples within each response group and does not account for possible hidden inter-relationships between markers.

We reasoned that our protein panel's strong predictive performance was attributable in part to possible hidden functional linkages between the markers, rather than their sheer inclusion in the panel. Instead of analysing the indicators separately, signatures encapsulating the group-specific heterogeneity and combinatorial expression patterns of the biomarkers (e.g., protein 1/2/3/4/5 are high/high/low/high/low; see Figure 5A) should be generated. Based on Boolean algebra, we offer a systematic process for identifying these combinatorial expression patterns for each response group.

The procedure is broken down into four phases, as shown in Figure-5A for three hypothetical proteins A-C. In step 1, the continuous expression data is discretized into binary values, with 1 and 0 indicating high and low expression, respectively. This is accomplished by normalising the protein expression data to the sample's median value: normalised values > 1 or 0 will be transformed to 1 or 0, accordingly. Step 2 identifies and summarises all combinatorial binary expression patterns of proteins A-C in a 'truth' table, which is subsequently transformed into logical expressions of the proteins (Figure. 5A). The separate patterns logical expressions are then added together in a Sum-of-Products (SOP) form using the Boolean operator (+) [52]. In Step 3, the summed logical expression is further simplified using the Quine-McCluskey algorithm [52], a Boolean function minimization algorithm that turns it into a simpler and more compact form without sacrificing information. Finally, in Step 4 the resulting simplified logical expression is converted back into binary expression patterns of the biomarkers.

In our example, we started with 4 expression patterns involving 3 proteins (A′B′C + A′BC + ABC′ + ABC) which were simplified into 2 patterns involving 2 proteins (A′C + AB) having only two proteins by applying Boolean algebra [52] (see Figure. 5A). Here, the prime (′) sign indicates the respective protein should be low, and high otherwise. Finally, the simplified logical expression is transformed back into binary biomarker expression patterns in Step 4. By using Boolean algebra, we were able to simplify four expression patterns involving three proteins (A'B'C + A'BC + ABC' + ABC) into two patterns involving two proteins (A'C + AB) with only two proteins [52]. (see Figure. 5A). The prime (') sign indicates that the associated protein should be low expressed, and high otherwise.

2.4 Result and Discussion

2.4.1 Machine learning model selection

We initially evaluated whether unsupervised hierarchical clustering could correctly predict the response groups for ribociclib therapy using all 4675 proteins as inputs. However, there was a lack of consistency between the three detected clusters and the answer groups that were labelled (Figure. 1D). Both Cluster1 and Cluster2 have a heterogeneous mix of RD and NR samples. We can easily observe that the data in both clusters is not equally partitioned into response groups. These findings reveal that in our prostate cancer PDE cohort, unsupervised clustering did not reliably predict response to ribociclib treatment.

As a next step, The dataset was separated randomly into two parts. A training set (70%) and a test set (30%) for PDEs due to smaller sample size(n=30) ; Training set (80%) and a test set (20%) for Pan-cancer, with the machine learning model being trained on the former and validated on the latter. We used 20 cross-validation tests to prevent potential biases related with data splitting (Figure 2A). We can see that out of all the machine learning models used to predict the response groups of PDEs (see Figure 2C), the SVM model with Linear kernel and polynomial kernel, as well as the XGBoost machine learning model appear to perform better, with an average prediction accuracy of 73 percent. SVM Linear is used to train PDEs because of the complexity of the implementation and the time it takes to compute the model.

Similarly, the SVM model with Radial kernel outperformed other models in the Pan-cancer dataset, with an average prediction accuracy of 58 percent. Following model identification, the cross-validation size must be assessed; for this SVM Linear model, several cross-validation partition sizes ranging from 2 to 22 must be trained. Figure 2B demonstrates that the model performance for CV size 6 was 88 percent, which is higher than the other sizes. Due to the increased size of the training set in the Pan-cancer dataset, repeated cross validation of size 5 with 3 repeats was used.

These findings imply that, while supervised machine learning algorithms outperform unsupervised hierarchical clustering methods, employing all DEPs as input features for labelled groups with a small sample size may not be ideal. One reason for this could be the noise produced by proteins in the DEPs that had no relevance in predicting reaction to response groups, which would have disguised the predictive signals from those that do, decreasing the models overall predictive performance [32].

2.4.2 Optimal predictive marker genes selection

We discovered that the models had an average prediction accuracy of 58 percent. Because the amount of input variables (4765 proteins, also known as input features) is significantly more than the sample size (n=30), a phenomenon known as the curse of dimensionality in machine learning [49]. This is an expected result. When the number of input features is increased, the prediction accuracy generally rises; however, when the number reaches a (ideal) threshold value, adding more input features considerably reduces the machine's performance [49]. This is because the input data's high dimensionality renders each observation appear indistinguishable from the others, making meaningful clustering impossible [49]. Another explanation is that input features that are irrelevant or only partially relevant can have a negative impact on model performance[34]. As a result, we applied a feature selection strategy to reduce the number of input features, avoiding the dimensionality curse [50, 51]. Because genes/proteins that are differentially expressed between response groups are frequently effective beginning points for discovering possible biomarkers [28, 31], we used the varImp function in R [38] to predict the response groups for all characteristics (4765 proteins). The threshold of 50% importance was used to decrease the number of characteristics used to pick the most important proteins. Thus, 559 differentially expressed proteins (DEPs) for PDEs were discovered.

We gradually increased the number of input features by adding the DEPs of PDEs one by one to the training dataset and re-evaluated the model prediction accuracy after investigating how modification of the input feature space may affect model performance for the SVM Linear model. Figure 2E demonstrates that prediction accuracy increases as the size of the feature space grows, and the model's performance with all 559 DEPs was not far behind the best-performing model. However, we discovered that adding specific DEPs to the feature space significantly reduced the model's predictive power, as indicated by a decline in the prediction accuracy trajectory. Specifically, we discovered that 297 DEPs contributed positively to model performance, while 262 negatively impacted it, and some had no discernible impact (Fig. 2D). These findings back up the premise that irrelevant input features can hurt a model's capacity to predict medication response, and that rational feature selection is key to improve predictive performance.

Figure 3B shows a ranked list of the 559 DEPs, which were ordered by their individual IS scores. Surprisingly, a large percentage of the DEPs (49%) had a negative impact on drug response prediction (Fig. 3B), implying that included them in the feature space could lower the model's overall predictive performance.

This also explains the original model's unsatisfactory prediction, which contains all 559 DEPs as inputs. Moreover, half of the DEPs, had a positive impact on drug response prediction (Fig. 3B), with the top 20 positive-impact DEPs given as examples in Figure 3D.

Figure 3E illustrates the ROC curve using these top 20 DEPs as input features. We reasoned that the positive-impact DEPs would be suitable candidate features for maximising the response prediction to ribociclib treatment because of their influence. We trained the model with the highest IS score, STAT3, and found that the model accuracy was greater than 50%, therefore used it as the baseline threshold. Then the next highest IS score DEP is chosen and evaluated for a 1% improvement in model accuracy. The threshold is changed when model performance improves, more DEPs are added, and model accuracy is assessed iteratively. Finally, we had a panel of ten DEPs with a 93 percent overall accuracy, including STAT3, PRDX3, PCP4, TPM4, TMEM192, CPSF2, GPX4, SRPX, DAZAP1, ALG11. Figure 4A depicts the gradual rise in model accuracy with the inclusion of DEPs, whereas Figure 4B depicts the ROC curve. The sensitivity of 82 percent for the NR and 91 percent for the RP answer groups is shown in the confusion matrix (Figure 4D) of training data with n=22 for PDEs of most significant DEP. The precision is 90 percent for the NR and 83 percent for the RP response groups. Similarly, when assessing testing data with a n=8 sample size, the precision for NR is 80% and 100% for RP, while the sensitivity is 80% for NR and 75% for RP response groups(Figure 4E).

By applying the same logic to Pan-cancer dataset we identified 8 genes that increased the model performance with their entrez id shown in Figure 4F. Identified genes are RP11-338N10.1 ,RP11-338N10.2, SLC2A3P2, CASZ1, HIST2H2BC, AGBL4, CDKN2A, PDXDC1. Confusion matrix(Figure 4G) of the testing data of Pan-cancer dataset with n=166 of most significant DEP shows a sensitivity of 60% for insensitive and 27% for sensitive response groups also the precision is 70% for insensitive and 36% for sensitive response groups. All these results indicate significant improvement to prediction of drug response within each individual labelled response groups for both PDE and Pan-cancer datasets. This finding emphasises the importance of expanding biomarker studies beyond the single gene/protein paradigm, which is currently prominent in biomarker research.

2.4.3 Expression signature of biomarker panel

We found a total of ten expression signatures of markers for prostate cancer PDEs that highlight the RD group (Figure. 5B). Similarly, 10 signatures were found for the NR group (Figure. 5B). Signature 1 (Figure. 5B, right table), which has low expression of STAT3, PRDX3, PCP4, TPM4, TMEM192, CPSF2, GPX4, SRPX, DAZAP1, ALG11 and low expression of STAT3, PRDX3, PCP4, TPM4, TPM4, TMEM192, CPSF2, GPX4, SRPX, DAZAP1, ALG11, is the most common pattern (22 percent) in the RD group. The next most common signature found in NR and RD is high PRDX3, PCP4, TPM4, GPX4 expression and low CPSF2 expression (Figure. 5B) suggesting that there is no difference in protein expression between these DEPs. In the NR group, the most abundant signature (Figure. 5C) is high PRDX3 plus low expression of STAT3, PCP4, TPM4, TMEM192, CPSF2, GPX4, SRPX, DAZAP1, ALG11, implying that PDEs with this signature are likely to have poor response to ribociclib treatment.

In the Pan-cancer dataset, a total of 9 expression profiles of markers that highlight the sensitive group (Figure. 5D) was found. In the same way, 11 signatures were discovered in the insensitive group. Signature 1 (Fig. 5D), which has low expression of RP11-338N10.1, RP11-338N10.2, CASZ1, HIST2H2BC, AGBL4 and high expression of SLC2A3P2, CDKN2A, PDXDC1, is the most common pattern (14 percent) in the sensitive group, implying that this signature most likely predicts sensitive response to Palbociclib treatment. High expression of RP11-338N10.2, HIST2H2BC, AGBL4 plus low expression of RP11-338N10.1, CASZ1, SLC2A3P2, CDKN2A, PDXDC1 is the most abundant signature in the insensitive group, indicating that PDEs exhibiting this signature are likely to have poor response to Palbociclib treatment.

2.4.4 The molecular functions of biomarkers and their links to cancer

The 10-protein panel includes proteins such as STAT3, PRDX3, TPM4 and SRPX which have been linked to prostate cancer and ribociclib treatment resistance. Signal transducer and activator of transcription 3 (STAT3) signalling, for example, is thought to have key carcinogenic activities, and targeting STAT3 as a treatment option in patients with metastatic CRPC in prostate cancer is being investigated [39]. STAT3 is involved in cell cycle control and is particularly critical during the transition from the G1 to the S cell cycle [39]. Cyclin D1, a direct STAT3 target gene whose overexpression has been linked to androgen independent metastatic PCa, is a crucial regulator of G1 phase progression [39]. peroxiredoxin-3 (PRDX-3) as a cell-surface protein that is androgen regulated in the prostate cancer (PCa) cell line. In antiandrogen resistant PCa cell lines, the PRDX-3 protein is over expressed, resulting in greater resistance to oxidative stress and failure to activate pro-apoptotic pathways. PRDX-3 has an important role in regulating oxidation-induced apoptosis and could be used to treat castrate-independent PCa [40]. TPM4 is a member of the tropomyosin family of actin-binding proteins that control the contractile mechanism in muscle and non-muscle cells. In past studies, TPM4 has been found to be downregulated in metastatic lung cancer but elevated in ovarian cancer. Hypermethylation of TPM4 has been associated to transcriptional downregulation in PCa, implying a cell type-specific effect. Sushi protein with a repeating pattern (SRPX) had lower expression and more methylation in its promoter, which is a hallmark of tumour suppressors [41]. SRPX expression in tumours was also significantly reduced in high-risk tumours compared to low- and intermediate-risk tumours, suggesting that it could be a valuable indication of PCa development [41].

Of the 8 gene panel identified for pan-cancer gene such as CASZ1, HIST2H2BC, AGBL4, CDKN2A, PDXDC1 are associated with various human cancer and palbociclib treatment. Castor zinc finger 1 (CASZ1) was discovered to be a neural fate-determination gene that plays an important role in cell differentiation, as well as brain and cardiac development [42]. Colorectal cancer and bladder cancer have both been linked to an abnormal fusion transcript of CASZ1 [42]. In neuroblastoma, CASZ1 is downregulated and acts as a tumour suppressor. In epithelial ovarian cancer, on the other hand, CASZ1 is highly expressed and responsible for cell migration and invasion, indicating that CASZ1 has diverse expression patterns and roles in different human malignancies [42]. Silencing of the cyclin-dependent kinase inhibitor 2A (CDKN2A) tumour suppressor gene, which produces the p16INK4a protein, has been linked to prostate, skin, lung, pancreatic, oesophageal, ovarian, renal, gastric, and head and neck cancers [43]. Through the control of the cyclin-dependent kinase (CDK) 4/6 and cyclin D complexes, the p16INK4a protein plays an executional role in cell cycle and senescence[43]. Several CDKN2A genetic and epigenetic abnormalities result in increased tumorigenesis and metastasis, as well as cancer recurrence and poor prognosis. In these instances, restoring CDKN2A's genetic and epigenetic reactivation is a viable strategy for cancer prevention and treatment [43].

These findings reinforce the validity of the simplified biomarker panel by demonstrating a relationship between the discovered biomarker panel of proteins/genes for prostate cancer and pan-cancer.

2.5 Limitations and Future Work

The current research has certain possible drawbacks. First, all DNA methylation analyses were limited to tissue samples based on ribociclib treatment. In the future, we want to create a blood-based DNA methylation test for PCa . Although blood-based analyses are outside the focus of this paper, our top candidate PCa hypermethylation markers should be tested in the analysis of Circulating tumour DNA (ctDNA) in plasma samples as a following step. Only nucleosome-associated DNA fragments are protected from DNase activity, therefore abnormally methylated tumour DNA may not be released in significant numbers or maintained in plasma. Furthermore, we must use a very stringent biomarker discovery approach to ensure the high specificity of future ctDNA-based methylation tests, which should eliminate (or at least minimise the risk of) false positive signals in the plasma from leukocyte-derived genomic DNA, particularly in patients with early-stage cancer and low tumour burden [44].

Second, the efficacy and superiority of the proposed algorithm are not guaranteed by a comparison of algorithms and statistical approaches employed to produce biomarkers [45]. To acquire consistent results, the possibility of attaining good results must be validated regularly. Furthermore, the current investigation was based on available RP specimens and data collected retrospectively, which could lead to selection bias. Because not all RP patients will acquire metastatic disease or die from PCa throughout the course of their lives.

We didn't have enough occurrences in our RP patient population to investigate these additional clinical outcomes [44]. As a result, more research is needed to evaluate the true clinical utility of methylation levels of STAT3, PRDX3, PCP4, TPM4, TMEM192, CPSF2, GPX4, SRPX, DAZAP1, ALG11 for predicting metastatic development and PCa-specific survival [44].

Additionally, the utilised biomarkers panel should be clinically verified so that they may be employed in prostate cancer patient screening, detection, diagnosis, and prognosis and survival outcomes. The application of biomarkers could be improved by focusing on quantifying the identified biomarkers panel for their signalling pathways or medications[46]. However, due to the slow progression of early stage PCa, such future studies would necessitate a clinical follow-up period of more than ten years [44].

2.6 Conclusion

In conclusion, we have created a new machine learning-based computational framework that facilitates in the development of multi-gene predictive biomarkers for targeted cancer treatments. To derive the specific expression signature of the biomarker proteins, we used Boolean algebra and logic minimization methods. While these techniques are routinely used in engineering areas to create digital logics, their application to biomarker discovery is completely novel [33, 34]. The goal of the Boolean function minimization algorithm is to find the underlying phenomenon's essential logics. We used the method to extract numerous expression signatures of biomarker proteins, which are responsible for each of the two responder groups. We validated its effectiveness by focusing on prostate and pan-cancers, and we demonstrated that the framework has broad applicability and can be used to other medications and cancer types in future research.

FIGURES

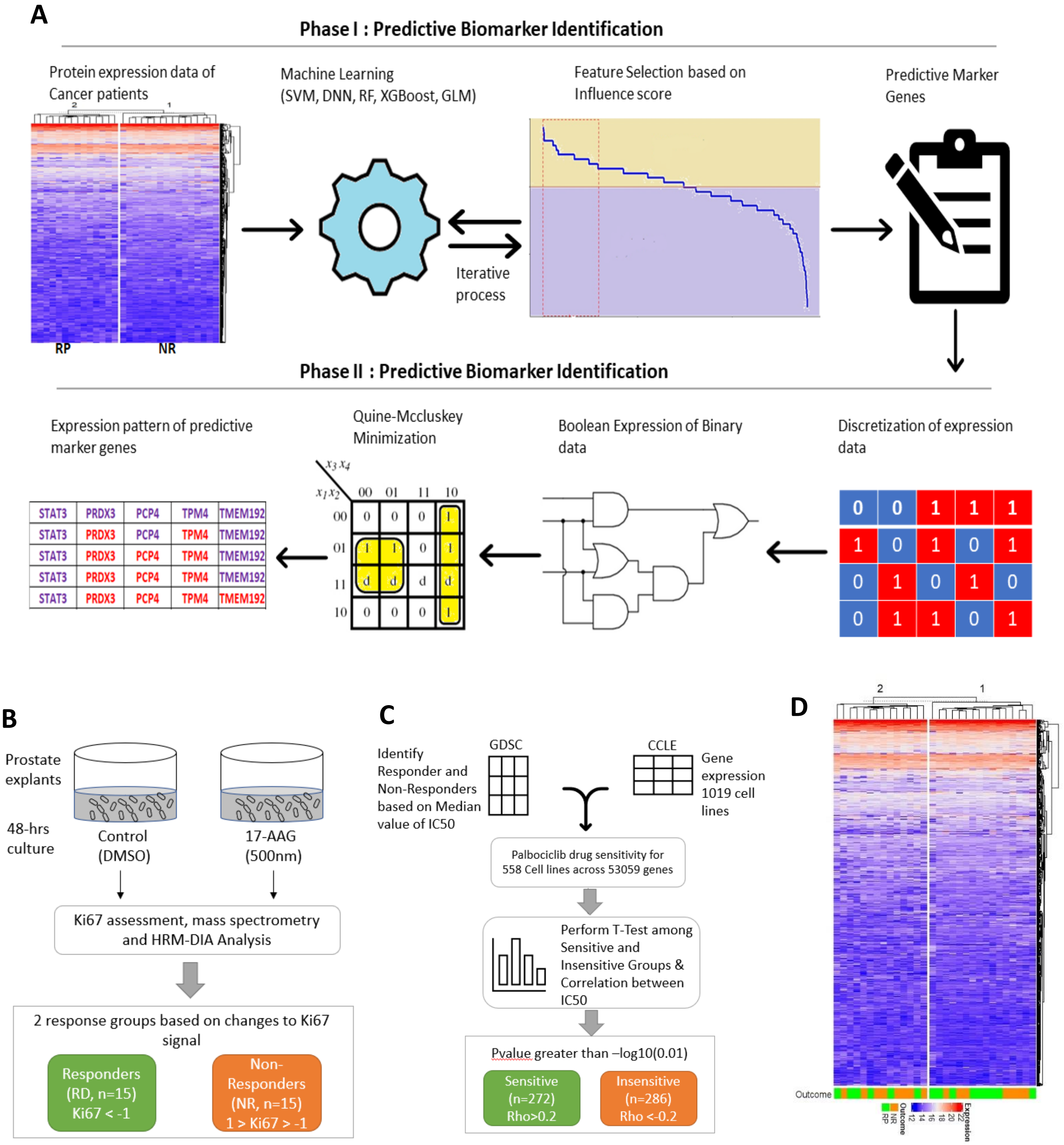

A
Phase I : Predictive Biomarker Identification
Protein expression data of Cancer patients
2
1
RP
NR
Machine Learning (SVM, DNN, RF, XGBoost, GLM)
Iterative process
Feature Selection based on Influence score
Predictive Marker Genes
Phase II : Predictive Biomarker Identification
Expression pattern of predictive marker genes
STAT3 PRDX3 PCP4 TPM4 TMEM192
STAT3 PRDX3 PCP4 TPM4 TMEM192
STAT3 PRDX3 PCP4 TPM4 TMEM192
STAT3 PRDX3 PCP4 TPM4 TMEM192
STAT3 PRDX3 PCP4 TPM4 TMEM192
Quine-Mccluskey Minimization
x_3x_4
x_1x_2
00 01 11 10
00 0 0 0 1
01 1 1 0 1
11 d d d d
10 0 0 0 1
Boolean Expression of Binary data
Discretization of expression data
0 0 1 1 1
1 0 1 0 1
0 1 0 1 0
0 1 1 0 1
B
Prostate explants
48-hrs culture
Control (DMSO)
17-AAG (500nm)
Ki67 assessment, mass spectrometry and HRM-DIA Analysis
2 response groups based on changes to Ki67 signal
Responders (RD, n=15) Ki67 < -1
Non-Responders (NR, n=15) 1 > Ki67 > -1
C
Identify Responder and Non-Responders based on Median value of IC50
GDSC
CCLE
Gene expression 1019 cell lines
Palbociclib drug sensitivity for 558 Cell lines across 53059 genes
Perform T-Test among Sensitive and Insensitive Groups & Correlation between IC50
Pvalue greater than –log10(0.01)
Sensitive (n=272) Rho>0.2
Insensitive (n=286) Rho <-0.2
D
2
1
Outcome
Outcome
NR
RP
Expression
12
14
16
18
20
22

Figure 1 : Pipeline and Drug response datasets (A) Workflow of our multi-step machine learning based framework for the identification of predictive biomarkers. **(B)** Experimental design for 30 prostate cancer patient-derived explants (PDEs), which was subjected to either vehicle (DMSO) or ribociclib (500 nM) treatment for 48 hours with classification of the PDE samples into two distinct response groups based on the corresponding Ki67 positivity upon treatment with ribociclib: the RD group (responders), a fold decrease in Ki67 positivity (-1 in log2 scale); the NR group (non-responders), the fold change in between +1 and -1 in log2 scale. **(C)** Pan cancer dataset derived from GDSC and CCLE dataset for Palbociclib drug sensitivity. Ttest and Pearson's correlation is performed for across genes to obtain 558 observations having Pvalue greater than –log10(0.01) ; Sensitive group with Rho>0.2 and Insensitive group with Rho<-0.2. **(D)** Unsupervised hierarchical clustering of response groups of PDEs which failed to predict response to Hsp90 inhibitor ribociclib.

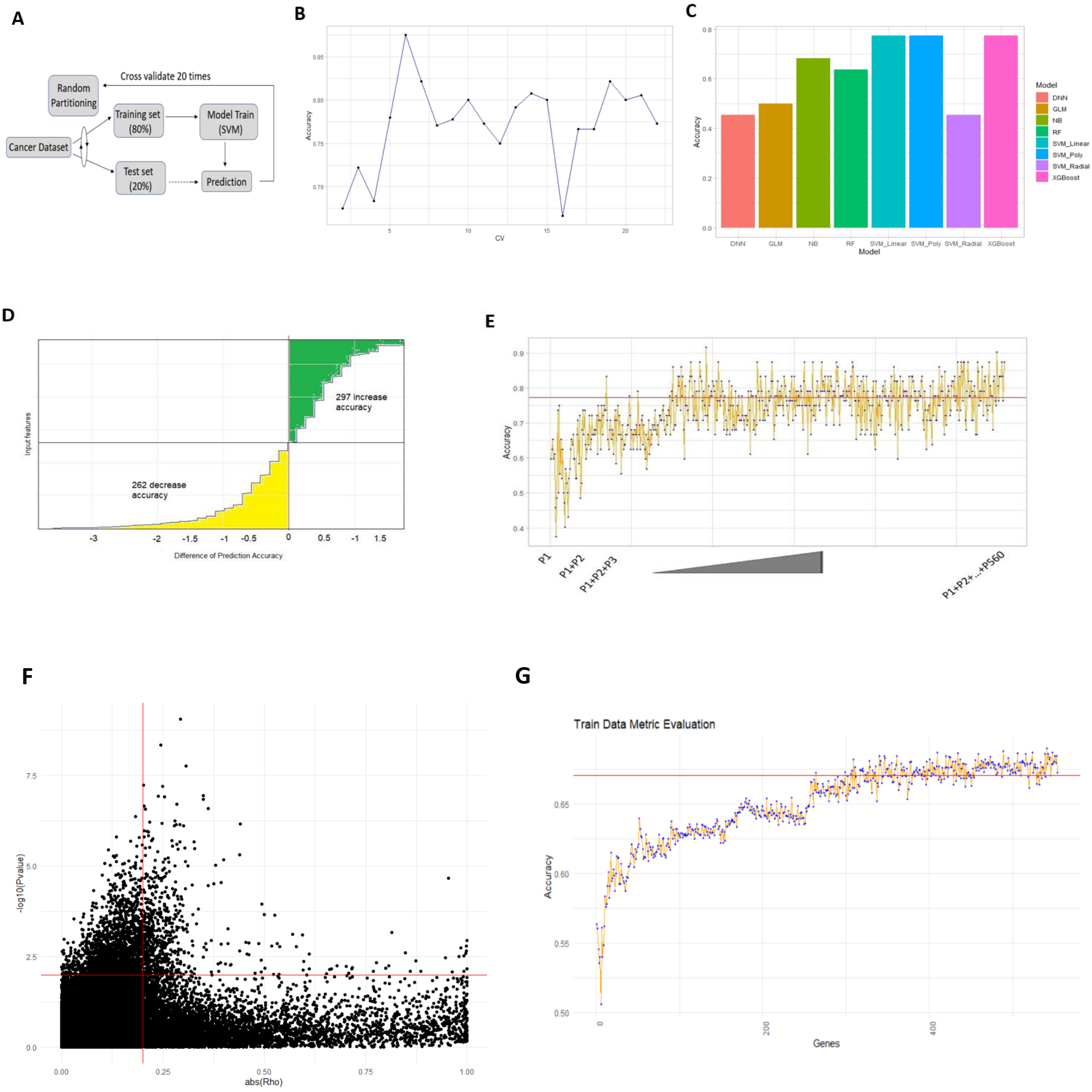


Figure 2 : Model training and feature selection (A) A workflow of the machine learning process. The dataset was randomly divided into a training (80%) and a test set (20%). The prediction accuracy was cross validated 20 times .**(B)** Distribution of the prediction accuracy for different cross validation number from 2 to 22 **(C)** Distribution different model performance for prostate cancer Training dataset. **(D)** Difference in the prediction accuracy compared to the original model when a new input feature is added to the training data. **(E)** Prediction accuracy depending on modulation of the input feature space for prostate cancer training data. Adding specific input features worsened the predictive performance of the models. Red line shows model performance with all proteins.**(F)** Scatterplot depicting the distribution of -log10 Pvalue of genes on Y axis and with Pearson's correlation (Rho) on X axis for Sensitive and Insensitive groups. Red lines show the threshold of 0.01 for Pvalue on Y axis and 0.2 Rho value on X axis. **(G)** Prediction accuracy depending on modulation of the input feature space for Pan cancer training data. Red line shows the SVM Radial kernel model prediction accuracy with all the genes as input parameters.

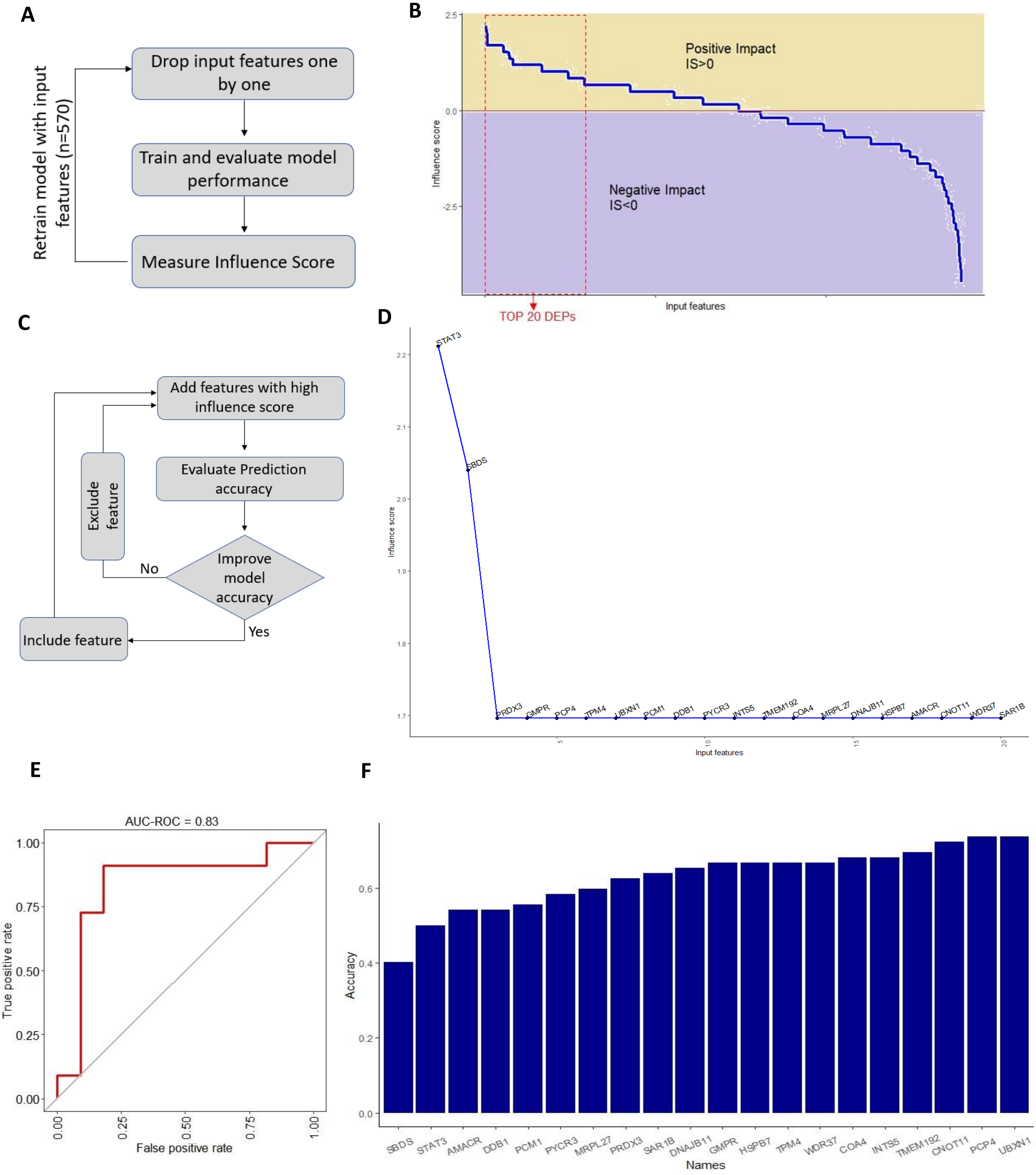


Figure 3 : Optimal feature selection to maximize the prediction accuracy for prostate cancer dataset (A) A workflow of evaluation of protein influence.**(B)** Influence score of input features (sorted in a descending order). Input feature having IS > 0 indicates positive-impact and IS<0 indicates negative impact. **(C)** A workflow to identify optimal input features that maximize the prediction score. **(D)** Top 20 DEPs having high influence score. **(E)** ROC curve of model with top 20 DEP with high influence score **(F))** Top 20 proteins with high influence score added one by one to check SVM radial kernel model performance.

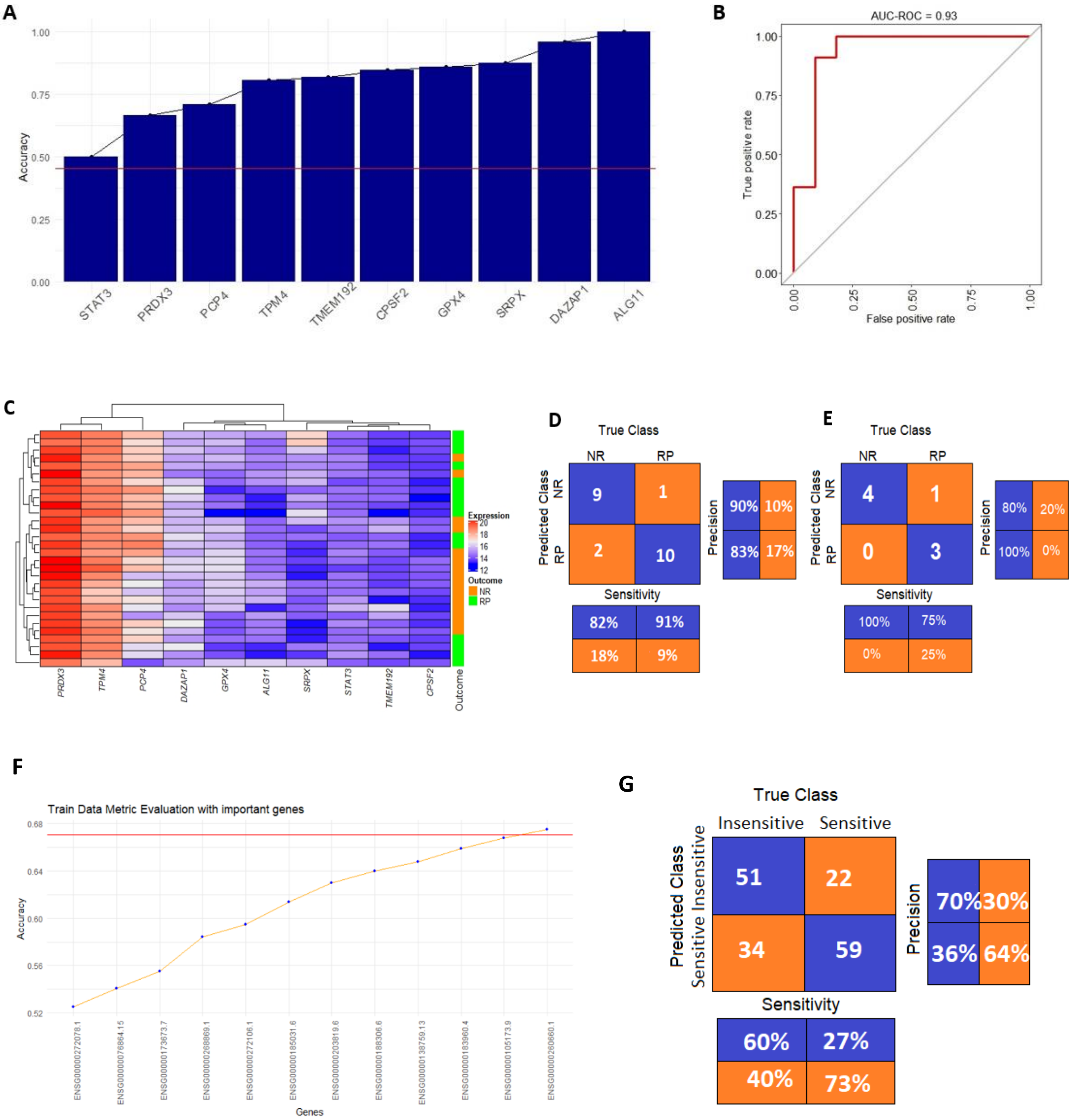
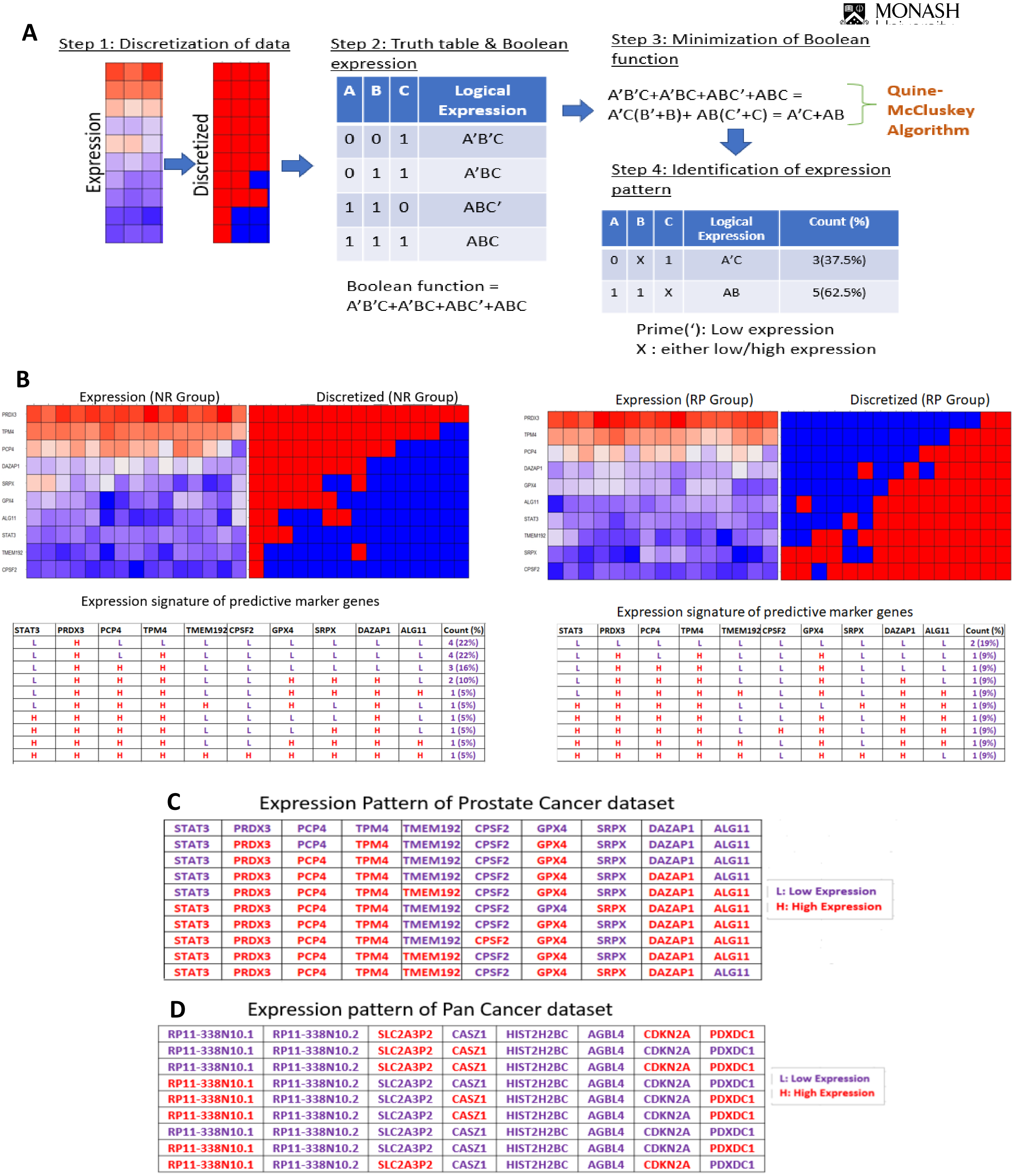


Figure 4 : Optimal feature set, Expression signature and confusion Matrix (A) Distribution of the prediction accuracy of the optimal features for prostate cancer dataset. The averaged prediction accuracy is 93% **(B)** Roc curve of SVM Radial kernel performance with optimal feature set for prostate cancer data **(C)** Expression pattern of optimal 10 proteins. Outcome shows the sensitivity distribution of Respondent and Not respondent groups. **(D)** Confusion matrix of Prostate cancer Training dataset (80%) with sensitivity and precision. **(E)** Confusion matrix of Prostate cancer Validation dataset (20%) with sensitivity and precision. **(F)** Distribution of the prediction accuracy of the optimal features for Pan cancer dataset. The averaged prediction accuracy is 67% **(G)** Confusion matrix of testing dataset of Pan cancer dataset with sensitivity and precision.



A	B	C	Logical Expression
0	0	1	A'B'C
0	1	1	A'BC
1	1	0	ABC'
1	1	1	ABC

Boolean function = A'B'C+A'BC+ABC'+ABC

A	B	C	Logical Expression	Count (%)
0	X	1	A'C	3(37.5%)
1	1	X	AB	5(62.5%)

Prime('): Low expression
X : either low/high expression

Expression signature of predictive marker genes

STAT3	PRDX3	PCP4	TPM4	TMEM192	CPSF2	GPX4	SRPX	DAZAP1	ALG11	Count (%)
L	H	L	L	L	L	L	L	L	L	4 (22%)
L	H	L	H	L	L	L	L	L	L	4 (22%)
L	H	H	H	L	L	L	L	L	L	3 (16%)
L	H	H	H	L	L	H	H	H	L	2 (10%)
L	H	H	H	L	L	H	H	H	H	1 (5%)
L	H	H	H	H	L	H	L	H	L	1 (5%)
H	H	H	H	L	L	L	L	H	L	1 (5%)
H	H	H	H	L	L	L	H	H	L	1 (5%)
H	H	H	H	L	L	H	H	H	H	1 (5%)
H	H	H	H	H	H	H	H	H	H	1 (5%)

Expression signature of predictive marker genes

STAT3	PRDX3	PCP4	TPM4	TMEM192	CPSF2	GPX4	SRPX	DAZAP1	ALG11	Count (%)
L	L	L	L	L	L	L	L	L	L	2 (19%)
L	H	L	H	L	L	H	L	L	L	1 (9%)
L	H	H	H	L	L	H	L	L	L	1 (9%)
L	H	H	H	L	L	H	L	H	L	1 (9%)
L	H	H	H	H	L	H	L	H	H	1 (9%)
H	H	H	H	L	L	L	H	H	H	1 (9%)
H	H	H	H	L	L	H	L	H	H	1 (9%)
H	H	H	H	L	H	H	L	H	H	1 (9%)
H	H	H	H	H	L	H	L	H	H	1 (9%)
H	H	H	H	H	L	H	H	H	L	1 (9%)

Expression Pattern of Prostate Cancer dataset

STAT3	PRDX3	PCP4	TPM4	TMEM192	CPSF2	GPX4	SRPX	DAZAP1	ALG11
STAT3	PRDX3	PCP4	TPM4	TMEM192	CPSF2	GPX4	SRPX	DAZAP1	ALG11
STAT3	PRDX3	PCP4	TPM4	TMEM192	CPSF2	GPX4	SRPX	DAZAP1	ALG11
STAT3	PRDX3	PCP4	TPM4	TMEM192	CPSF2	GPX4	SRPX	DAZAP1	ALG11
STAT3	PRDX3	PCP4	TPM4	TMEM192	CPSF2	GPX4	SRPX	DAZAP1	ALG11
STAT3	PRDX3	PCP4	TPM4	TMEM192	CPSF2	GPX4	SRPX	DAZAP1	ALG11
STAT3	PRDX3	PCP4	TPM4	TMEM192	CPSF2	GPX4	SRPX	DAZAP1	ALG11
STAT3	PRDX3	PCP4	TPM4	TMEM192	CPSF2	GPX4	SRPX	DAZAP1	ALG11
STAT3	PRDX3	PCP4	TPM4	TMEM192	CPSF2	GPX4	SRPX	DAZAP1	ALG11
STAT3	PRDX3	PCP4	TPM4	TMEM192	CPSF2	GPX4	SRPX	DAZAP1	ALG11

L: Low Expression
H: High Expression

Expression pattern of Pan Cancer dataset

RP11-338N10.1	RP11-338N10.2	SLC2A3P2	CASZ1	HIST2H2BC	AGBL4	CDKN2A	PDXDC1
RP11-338N10.1	RP11-338N10.2	SLC2A3P2	CASZ1	HIST2H2BC	AGBL4	CDKN2A	PDXDC1
RP11-338N10.1	RP11-338N10.2	SLC2A3P2	CASZ1	HIST2H2BC	AGBL4	CDKN2A	PDXDC1
RP11-338N10.1	RP11-338N10.2	SLC2A3P2	CASZ1	HIST2H2BC	AGBL4	CDKN2A	PDXDC1
RP11-338N10.1	RP11-338N10.2	SLC2A3P2	CASZ1	HIST2H2BC	AGBL4	CDKN2A	PDXDC1
RP11-338N10.1	RP11-338N10.2	SLC2A3P2	CASZ1	HIST2H2BC	AGBL4	CDKN2A	PDXDC1
RP11-338N10.1	RP11-338N10.2	SLC2A3P2	CASZ1	HIST2H2BC	AGBL4	CDKN2A	PDXDC1
RP11-338N10.1	RP11-338N10.2	SLC2A3P2	CASZ1	HIST2H2BC	AGBL4	CDKN2A	PDXDC1
RP11-338N10.1	RP11-338N10.2	SLC2A3P2	CASZ1	HIST2H2BC	AGBL4	CDKN2A	PDXDC1

L: Low Expression
H: High Expression

Figure 5 : Expression signatures derivation of biomarkers for patient stratification. (A) A four-step workflow that identifies the biomarkers combinatorial expression fingerprints for each response group. Step 1: Convert expression data to binary data (1 and 0 for high and low expression, respectively). Step 2: A truth table is created, which is subsequently converted into logical (Boolean) expressions. Step 3: Converting the Boolean function into simpler and more compact forms using the Quine-McCluskey algorithm, for example. Step 4: Biomarker expression signatures are identified. **(B)** In the RD and NR groups of PDE samples, the expression, and discretized data of the prediction 10 marker genes of prostate cancer dataset. Using the Boolean function minimization approach, the expression signatures of the marker proteins were discovered. Low and high protein expression are indicated by the letters L and H, respectively. **(C)** Expression pattern of Respondent group for prostate cancer dataset **(D)** Expression pattern of Respondent group for Pan cancer dataset

2.7 Reference

1. Kase, A. M., Copland, J. A., & Tan, W. (2020). Novel Therapeutic Strategies for CDK4/6 Inhibitors in Metastatic Castrate-Resistant Prostate Cancer. OncoTargets and Therapy, Volume 13, 10499-10513. doi:10.2147/ott.s266085
2. Hendrychová, Denisa, Jorda, Radek, & Kryštof, Vladimír. (2021). How selective are clinical CDK4/6 inhibitors? Medicinal Research Reviews., 41(3), 1578–1598. https://doi.org/10.1002/med.21769
3. Álvarez-Fernández, Mónica, & Malumbres, Marcos. (2020). Mechanisms of Sensitivity and Resistance to CDK4/6 Inhibition. Cancer Cell., 37(4), 514–529. https://doi.org/10.1016/j.ccell.2020.03.010
4. Quintás-Cardama, A. and J. Cortes, Molecular biology of bcr-abl1-positive chronic myeloid leukemia. Blood, 2009. 113(8): p. 1619-1630.
5. Dieci, M.V., et al., Biomarkers for HER2-positive metastatic breast cancer: Beyond hormone receptors. Cancer Treat Rev, 2020. 88: p. 102064.
6. Nguyen, L., C.C. Dang, and P.J. Ballester, Systematic assessment of multi-gene predictors of pan-cancer cell line sensitivity to drugs exploiting gene expression data. F1000Res, 2016. 5.
7. Lima, A.R., et al., Identification of a biomarker panel for improvement of prostate cancer diagnosis by volatile metabolic profiling of urine. British Journal of Cancer, 2019. 121(10): p. 857-868.
8. Zhu, C.S., et al., A framework for evaluating biomarkers for early detection: validation of biomarker panels for ovarian cancer. Cancer prevention research (Philadelphia, Pa.), 2011. 4(3): p. 375-383.
9. Fortino, V., Wisgrill, L., Werner, P., Suomela, S., Linder, N., Jalonen, E., Suomalainen, A., Marwah, V., Kero, M., Pesonen, M., Lundin, J., Lauerma, A., Aalto-Korte, K., Greco, D., Alenius, H., & Fyhrquist, N. (2020). Machine-learning–driven biomarker discovery for the discrimination between allergic and irritant contact dermatitis. Proceedings of the National Academy of Sciences of the United States of America., 117(52), 33474–33485. https://doi.org/10.1073/PNAS.2009192117
10. Tabl, A.A., et al., A Machine Learning Approach for Identifying Gene Biomarkers Guiding the Treatment of Breast Cancer. Frontiers in Genetics, 2019. 10(256).
11. Menden, M.P., et al., Machine learning prediction of cancer cell sensitivity to drugs based on genomic and chemical properties. PLoS One, 2013. 8(4): p. e61318.
12. Nguyen, E.V., et al., Identification of Novel Response and Predictive Biomarkers to Hsp90 Inhibitors Through Proteomic Profiling of Patient-derived Prostate Tumor Explants. Mol Cell Proteomics, 2018. 17(8): p. 1470-1486.
13. Chen, C., et al., Bioinformatics analysis of differentially expressed proteins in prostate cancer based on proteomics data. Onco Targets Ther, 2016. 9: p. 1545-57.
14. Barretina, J., Caponigro, G., Stransky, N., Venkatesan, K., Margolin, A., Kim, S., . . . Garraway, L. (2012). 22 The Cancer Cell Line Encyclopedia - Using Preclinical Models to Predict Anticancer Drug Sensitivity. European Journal of Cancer, 48. doi:10.1016/s0959-8049(12)70726-8
15. Garnett, M.J., et al., Systematic identification of genomic markers of drug sensitivity in cancer cells. Nature, 2012. 483(7391): p. 570-5.
16. Seashore-Ludlow, B., et al., Harnessing Connectivity in a Large-Scale Small-Molecule Sensitivity Dataset. Cancer Discov, 2015. 5(11): p. 1210-23.
17. Iorio, F., et al., A Landscape of Pharmacogenomic Interactions in Cancer. Cell, 2016. 166(3): p. 740-754.
18. Borst, P. and L. Wessels, Do predictive signatures really predict response to cancer chemotherapy? Cell Cycle, 2010. 9(24): p. 4836-40.
19. Gillet, J.P., S. Varma, and M.M. Gottesman, The clinical relevance of cancer cell lines. J Natl Cancer Inst, 2013. 105(7): p. 452-8.
20. Yang, W., Lightfoot, H., Bignell, G., Behan, F., Cokelear, T., Haber, D., . . . Garnett, M. (2016). Genomics of Drug Sensitivity in Cancer (GDSC): A resource for biomarker discovery in cancer cells. European Journal of Cancer, 69. doi:10.1016/s0959-8049(16)32839-8
21. Barretina, J., Caponigro, G., Stransky, N., Venkatesan, K., Margolin, A., Kim, S., . . . Garraway, L. (2012). 22 The Cancer Cell Line Encyclopedia - Using Preclinical Models to Predict Anticancer Drug Sensitivity. European

Journal of Cancer, 48. doi:10.1016/s0959-8049(12)70726-8

22. Loeb, S., & Catalona, W. J. (2014). The Prostate Health Index: a new test for the detection of prostate cancer. Therapeutic advances in urology, 6(2), 74–77. https://doi.org/10.1177/1756287213513488
23. Punnen, S., Pavan, N., & Parekh, D. J. (2015). Finding the Wolf in Sheep's Clothing: The 4Kscore Is a Novel Blood Test That Can Accurately Identify the Risk of Aggressive Prostate Cancer. Reviews in urology, 17(1), 3–13.
24. Wei, John T. Urinary biomarkers for prostate cancer, Current Opinion in Urology: January 2015 - Volume 25 - Issue 1 - p 77-82 doi: 10.1097/MOU.0000000000000133
25. Rebello, Richard J, Oing, Christoph, Knudsen, Karen E, Loeb, Stacy, Johnson, David C, Reiter, Robert E, Gillessen, Silke, Van der Kwast, Theodorus, & Bristow, Robert G. (2021). Prostate cancer. Nature Reviews Disease Primers., 7(1). https://doi.org/10.1038/s41572-020-00243-0
26. Couñago, Felipe, López-Campos, Fernando, Díaz-Gavela, Ana Aurora, Almagro, Elena, Fenández-Pascual, Esaú, Henríquez, Iván, Lozano, Rebeca, Linares Espinós, Estefanía, Gómez-Iturriaga, Alfonso, de Velasco, Guillermo, Quintana Franco, Luis Miguel, Rodríguez-Melcón, Ignacio, López-Torrecilla, José, Spratt, Daniel E, Guerrero, Luis Leonardo, Martínez-Salamanca, Juan Ignacio, & del Cerro, Elia. (2020). Clinical Applications of Molecular Biomarkers in Prostate Cancer. Cancers., 12(6), 1550–25. https://doi.org/10.3390/cancers12061550
27. Karp G. Cell and Molecular Biology: Concepts and Experiments. 6th ed. Hoboken, NJ: John Wiley and Sons; 2010.
28. Sedlacek H, et al. Flavopiridol (L86 8275; NSC 649890), a new kinase inhibitor for tumor therapy. Int J Oncol. 1996; 9:1143–1168. [PubMed: 21541623]
29. Kontos, Christos K, Avgeris, Margaritis, Scorilas, Andreas, Atta-ur-Rahman, & Choudhary, M Iqbal. (2018). Biomarkers with Prognostic Potential in Prostate Cancer. In Frontiers in drug design and discovery. (Vol. 1, Issue 1, pp. 108–134). Bentham Science Pub. https://doi.org/10.2174/9781681085821118090003
30. Knudsen, E. S., & Witkiewicz, A. K. (2017). The Strange Case of CDK4/6 Inhibitors: Mechanisms, Resistance, and Combination Strategies. Trends in Cancer, 3(1), 39-55. doi:10.1016/j.trecan.2016.11.006
31. Chong, Qing-Yun, Kok, Ze-Hui, Bui, Ngoc-Linh-Chi, Xiang, Xiaoqiang, Wong, Andrea Li-Ann, Yong, Wei-Peng, Sethi, Gautam, Lobie, Peter E, Wang, Lingzhi, & Goh, Boon-Cher. (2020). A unique CDK4/6 inhibitor: Current and future therapeutic strategies of abemaciclib. Pharmacological Research : the Official Journal of the Italian Pharmacological Society., 156. https://doi.org/10.1016/j.phrs.2020.104686
32. Shah, M. (2018). CDK4/6 inhibitors: Game changers in the management of hormone receptor– positive advanced breast cancer? Oncology., 32(5), 216–222. https://doi.org/info:doi/
33. Cucchiara, Vito, Cooperberg, Matthew R, Dall'Era, Marc, Lin, Daniel W, Montorsi, Francesco, Schalken, Jack A, & Evans, Christopher P. (2018). Genomic Markers in Prostate Cancer Decision Making. European Urology : Official Journal of the European Association of Urology, 73(4), 572–582. https://doi.org/10.1016/j.eururo.2017.10.036
34. Chang, Yoosup, Park, Hyejin, Yang, Hyun-Jin, Lee, Seungju, Lee, Kwee-Yum, Kim, Tae Soon, Jung, Jongsun, & Shin, Jae-Min. (2018). Cancer Drug Response Profile scan (CDRscan): A Deep Learning Model That Predicts Drug Effectiveness from Cancer Genomic Signature. Scientific Reports., 8(1). https://doi.org/10.1038/s41598-018-27214-6
35. Nicora, Giovanna, Vitali, Francesca, Dagliati, Arianna, Geifman, Nophar, & Bellazzi, Riccardo. (2020). Integrated Multi-Omics Analyses in Oncology: A Review of Machine Learning Methods and Tools. Frontiers in Oncology., 10. https://doi.org/10.3389/fonc.2020.01030
36. Ghandi, M., Huang, F. W., Jané-Valbuena, J., Kryukov, G. V., Lo, C. C., McDonald, E. R., Barretina, J., Gelfand, E. T., Bielski, C. M., Li, H., Hu, K., Andreev-Drakhlin, A. Y., Kim, J., Hess, J. M., Haas, B. J., Aguet, F., Weir, B. A., Rothberg, M. V., Paolella, B. R., & Lawrence, M. S. (2019). Next-generation characterization of the Cancer Cell Line Encyclopedia. Nature., 569(7757), 503–508. https://doi.org/10.1038/s41586-019-1186-3
37. Sharma, A., & Rani, R. (2020). Drug sensitivity prediction framework using ensemble and

multi-task learning. International Journal of Machine Learning and Cybernetics., 11(6), 1231–1240. https://doi.org/10.1007/s13042-019-01034-0

38. Kuhn, M. (2008). Package. Journal of Statistical Software., 28(5), 1–26. https://doi.org/10.18637/jss.v028.i05
39. Culig, Z., Pencik, J., Merkel, O., & Kenner, L. (2016). Breaking a paradigm: IL-6/STAT3 signaling suppresses metastatic prostate cancer upon ARF expression. Molecular & Cellular Oncology, 3(2). doi:10.1080/23723556.2015.1090048
40. Whitaker, H. C., Patel, D., Howat, W. J., Warren, A. Y., Kay, J. D., Sangan, T., Marioni, J. C., Mitchell, J., Aldridge, S., Luxton, H. J., Massie, C., Lynch, A. G., & Neal, D. E. (2013). Peroxiredoxin-3 is overexpressed in prostate cancer and promotes cancer cell survival by protecting cells from oxidative stress. British Journal of Cancer., 109(4), 983–993. https://doi.org/10.1038/bjc.2013.396
41. Kamdar, S., Isserlin, R., Van der Kwast, T., Zlotta, A. R., Bader, G. D., Fleshner, N. E., & Bapat, B. (2019). Exploring targets of TET2-mediated methylation reprogramming as potential discriminators of prostate cancer progression. Clinical Epigenetics., 11(1). https://doi.org/10.1186/s13148-019-0651-z
42. Wang, Ji-Long, Yang, Meng-yuan, Xiao, Shuai, Sun, Bo, Li, Yi-Ming, & Yang, Lian-Yue. (2018). Downregulation of castor zinc finger 1 predicts poor prognosis and facilitates hepatocellular carcinoma progression via MAPK/ERK signaling. Journal of Experimental & Clinical Cancer Research., 37(1). https://doi.org/10.1186/s13046-018-0720-8
43. Zhao, R., Choi, B. Y., Lee, M.-H., Bode, A. M., & Dong, Z. (2016). Implications of Genetic and Epigenetic Alterations of CDKN2A (p16 INK4a) in Cancer. EBioMedicine., 8, 30–39. https://doi.org/10.1016/j.ebiom.2016.04.017
44. Bjerre, M., Strand, S., Nørgaard, M., Kristensen, H., Rasmussen, A., Mortensen, M., . . . Sørensen, K. (2019). Aberrant DOCK2, GRASP, HIF3A and PKFP Hypermethylation has Potential as a Prognostic Biomarker for Prostate Cancer. International Journal of Molecular Sciences, 20(5), 1173. doi:10.3390/ijms20051173
45. Sharma, A., & Rani, R. (2019). Drug sensitivity prediction framework using ensemble and multi-task learning. International Journal of Machine Learning and Cybernetics, 11(6), 1231-1240. doi:10.1007/s13042-019-01034-0
46. Verma, Mukesh, Patel, Payal, & Verma, Mudit. (2011). Biomarkers in Prostate Cancer Epidemiology. Cancers., 3(4), 3773–3798. https://doi.org/10.3390/cancers3043773
47. Bottcher R, Hoogland AM, Dits N, et al.Novel long non-coding RNAs are specific diagnostic and prognostic markers for prostate cancer.Oncotarget. 2015;6(6):4036–4050.
48. Shi, Jingqi, Jiang, Dongbo, Yang, Shuya, Zhang, Xiyang, Wang, Jing, Liu, Yang, Sun, Yuanjie, Lu, Yuchen, & Yang, Kun. (2020). LPAR1, Correlated With Immune Infiltrates, Is a Potential Prognostic Biomarker in Prostate Cancer. Frontiers in Oncology., 10. https://doi.org/10.3389/fonc.2020.00846
49. Ali, Mehreen, & Aittokallio, Tero. (2019). Machine learning and feature selection for drug response prediction in precision oncology applications. Biophysical Reviews., 11(1), 31–39. https://doi.org/10.1007/s12551-018-0446-z
50. Wang, Marcus W H, Goodman, Jonathan M, & Allen, Timothy E H. (2021). Machine Learning in Predictive Toxicology: Recent Applications and Future Directions for Classification Models. Chemical Research in Toxicology., 34(2), 217–239. https://doi.org/10.1021/acs.chemrestox.0c00316
51. Angermueller, Christof, Pärnamaa, Tanel, Parts, Leopold, & Stegle, Oliver. (2016). Deep learning for computational biology. Molecular Systems Biology, 12(7). https://doi.org/10.15252/msb.20156651
52. Duşa, A., & Thiem, A. (2015). Enhancing the Minimization of Boolean and Multivalue Output Functions WitheQMC. The Journal of Mathematical Sociology, 39(2), 92-108. doi:10.1080/0022250x.2014.897949

Part III : Appendices

Influence score (IS)

Instead of using all 157 DEPs as input features, we performed a systematic 'feature drop-out' analysis in which we removed one protein from the feature space at a time and assessed the influence of the dropped-out protein on the overall model's drug-response prediction, which we called 'influence score' (IS), which was measured as the difference in prediction accuracy between the 'drop-out' model and the original SVM model as follows:

$$\mathrm{IS(i)} = -\frac{PA_i - PA_O}{PA_O}$$

where PA_0 and PA_i are the original SVM prediction accuracy and a 'drop-out' model in which an input feature I is removed from the training dataset, respectively.

IS > 0 denotes positive-impact features, IS 0 denotes negative-impact features, and IS = 0 denotes input features that have no effect on response prediction.

Boolean function minimization algorithm

A Boolean function is described by an algebraic expression consisting of n-binary variables represented by f(x1, x2, …, xn). There are different ways of representing a Boolean function; for instance, Sum of Product (SOP) or Product of Sum (POS). SOP is a form of expression in Boolean algebra in which different product terms of inputs are being summed together. This product is not arithmetical multiply, but it is Boolean logical 'AND' and the Sum is Boolean logical 'OR'. For example, x′+xy + yz′ where x, y and z are binary variables and prime (′) represent complement of a variable. E.g., if x = 0 then x′ = 1. POS is a form in which products of different sum terms of inputs are taken. For example, (x′)·(x+y)·(y+z′). By using Boolean laws and theorems (e.g., De Morgan's laws), the Boolean functions can be simplified [23, 24]. However, when the number of literals in such an expression grows higher, the algebraic expression's complexity skyrockets. As a result, the algebraic equation must be written in the simplest yet mathematically comparable form possible. Minimization is the process of simplifying a Boolean function's algebraic formulation. To minimise the Boolean function, we utilised the Quinine-MacCluskey algorithm [52], which was implemented in R using Alrik Thiem's eQMC function (http://www.alrik-thiem.net/).

Source Code for Pan-cancer dataset

```
library(tidyverse)
library(CePa)
library(data.table)
library(matrixStats)
library(ggplot2)
library(caret)
library(rminer)
library(EnsDb.Hsapiens.v79)
library(doSNOW)
library(deepnet)
library(QCApro)

# register parallel
getDoParWorkers()
getDoParName()
# register parallel
registerDoSNOW(makeCluster(4, type = "SOCK"))
getDoParVersion()


#BiocManager::install('EnsDb.Hsapiens.v79')

set.seed(67)
############ change the data file
read.gct('CCLE_RNAseq_genes_rpkm_20180929.gct')->CCLE
as.data.frame(CCLE)->CCLE
write.csv(CCLE,"ccle-full.csv")


read.csv("GDSC.csv")->GDSC

########### Read files ###########

read.csv("GDSC.csv")->GDSC
fread("ccle-full.csv")->CCLE

########### Check for sensitive and insensitive #############


gdsc.median<- median(GDSC$IC50)
GDSC$Category<-if_else(GDSC$IC50<gdsc.median,'Sensitive','Insensitive')

############ Extract Cell lines ############

names(CCLE)->cellcols
str_extract(cellcols,"(\\w+?)_")->cellcols
str_extract(cellcols,"([0-9a-zA-Z]+)")->cellcols
cellcols[1]<-"ID"
cellcols[2]<-"Gene"

names(CCLE)<-cellcols


#############################################################################
###

########### Prune Columns ###########

# Extract common cell lines from GDSC and CCLE
keep<-intersect(GDSC$Cell_line,names(CCLE))

# Obtain
CCLE[,..keep]->CCLE.prune

# Add the Gene ID
cbind(CCLE[,"Gene"],CCLE.prune)->gene.cell

########### T-TEST ###########
```

```
# Store only IC50 and celline from GDSC
rho.gdsc<-GDSC %>% filter(Cell_line %in% keep) %>% select(Cell_line,IC50)

# Store only Cellline and Category from GDSC
cat.gdsc<-GDSC %>% filter(Cell_line %in% keep) %>% select(Cell_line,Category)

#############################################################################
###

# create function for t-test pvalue and Correlation values

ttest.rho<-function(x){

  # transpose the row
  conv.tab<-melt(x,id.vars = "Gene")

  # convert into data frame
  as.data.frame(conv.tab)->conv.tab

  # filter 0 values
  conv.tab %>% filter(value!=0)->conv.tab

  # merge and fetch category information for T-test
  conv.tab.T<-merge(conv.tab,cat.gdsc,by.x="variable",by.y="Cell_line")

  # sensitive group
  sens<-conv.tab.T %>% filter(Category == 'Sensitive') %>% pull(value)

  # insensitive group
  insens<-conv.tab.T %>% filter(Category == 'Insensitive') %>% pull(value)

  # result
  result<-c()

  # T-Test
  if(length(sens)==0|| length(insens)==0){

    result[1]<-NA

  }else{

    # Catch the ttest error of low columns.
    tryCatch(
      {
        # obtain the pvalue
        value<-t.test(sens,insens,var.equal=T)
        # return pvalue
        result[1]<-value$p.value
      },
      error=function(cond) {
        result[1]<-NA
        }

    )

  }
  # Correlation coefficient for gene
  # Merge two dataframes
  merge.tab<-merge(conv.tab,rho.gdsc,by.x="variable",by.y="Cell_line")

  # filter the 0 values from correlation
  merge.tab %>% filter(value!=0)->merge.tab

  # Spearman correlation or RHO
  coeff<-cor(merge.tab$value, merge.tab$IC50, method = "pearson")

  result[2]<-coeff

  return(result)

}
##############################################################################
###

########### Evaluate the ttest and correlation values ###########

# create temporary datatable for pvalues.
temp.p<-c()
temp.r<-c()

# calculate the Pvalue using ttest
for(i in 1:dim(gene.cell)[1]){
  temp<-gene.cell[i,ttest.rho(.SD)]
  temp.p[i]<-temp[1]
  temp.r[i]<-temp[2]
}

# assign the p-values back to the gene.cell
gene.cell[,Pvalue:=temp.p]

# assign the RHO back to the gene.cell
gene.cell[,Rho:=temp.r]

# Remove Na values and store to new Data.table
# Set Threshold for Pvalue as 0.001
GCT<-gene.cell[!is.na(Pvalue)]

# Store the file for future reference.
fwrite(GCT,"Gene_Cell_Ttest_Pvalue.csv")


##############################################################################
###

###########  Volcano plot. ###########

as.data.frame(GCT)->GCT

ggplot(data=GCT, aes(y=-log10(Pvalue), x= abs(Rho))) +
  geom_point() +
  theme_minimal()+
  geom_vline(xintercept=0.2, col="red") +
  geom_hline(yintercept=-log10(0.01), col="red")
```

```
#############################################################################
###

# Prune the data based on the threshold for correlation and -log10(Pvalue)

GCT %>%  filter(abs(Rho)>0.2 & Pvalue<0.01) %>% select(-Pvalue,-Rho) -> final
rownames(final)<-final$Gene
final %>% select(-Gene)->final

# check dimension of final data.
dim(final)

write.csv(final,"GDSC_CCLE_FINAL.csv")

#############################################################################
###

############ Transpose the dataframe ############

t_final <- transpose(final)
colnames(t_final) <- rownames(final)
rownames(t_final) <- colnames(final)

# store Gene as a column to merge and obtain category
t_final$Gene<-rownames(t_final)
# perform merge
merge(t_final,cat.gdsc,by.x="Gene",by.y="Cell_line",all.x = TRUE)->mt_final

# Remove duplicated Cell Lines
mt_final[!duplicated(mt_final$Gene),]->mt_final

# Set Rownames as gene and remove gene column
rownames(mt_final)<-mt_final$Gene

# drop gene column
mt_final %>% select(-Gene)->mt_final
mt_final$Category<-as.factor(mt_final$Category)

write.csv(mt_final,"Train.csv")

#############################################################################
###
#=======================>   START HERE FOR MODELS
<===========================

# Load data just for model.
mt_final<-read.csv("Train.csv")
rownames(mt_final)<-mt_final$X
mt_final<-mt_final[,c(-1,-2)]
mt_final$Category<-as.factor(mt_final$Category)

# For Prostate cancer.
#mt_final<- read.csv("ProstateCancer.csv")
#mt_final<-mt_final[,c(-1)]
#mt_final$Category<-as.factor(mt_final$Category)


#remove zero variance columns and correlated variables
```

```
# Skip this for prostate cancer
mt_final_T<-mt_final[,-nearZeroVar(mt_final)]
mt_final_c<-mt_final_T[,-findCorrelation(cor(mt_final_T[,-
length(mt_final_T)]), .8)]
mt_final<- mt_final_c

##############################################################################
###


########### Build SVM model ###########

# Create Train validation

TrainIndex<-createDataPartition(mt_final$Category, p = 0.8, list = FALSE,
times = 1)
Train<-mt_final[TrainIndex,]
Validation<-mt_final[-TrainIndex,]

# Traincontrol
train_control_5 <- trainControl(
  method = "repeatedcv",
  number = 5,
  repeats = 3
)
# Train control 10
train_control_10 <- trainControl(
  method = "repeatedcv",
  number = 10,
  repeats = 3
)

# SVM model
svm1 <- train(Category ~., data = Train, method = "svmLinear", trControl =
train_control_5,metric="Accuracy")
svm1
svm2 <- train(Category ~., data = Train, method = "svmPoly", trControl =
train_control_5,metric="Accuracy")
svm2
svm3 <- train(Category ~., data = Train, method = "svmRadial", trControl =
train_control_5,metric="Accuracy")
svm3

svmt<- train(Category ~., data = Train, method = "svmRadial", trControl =
train_control_5,metric="Accuracy",
             tuneGrid = expand.grid(C =
c(0.25,0.5,1),sigma=1/seq(20,50,length=20)^2))
svmt
besttune<-svmt$bestTune
# Random Forest model

rf <- train(Category ~., data = Train, method = "rf", trControl =
train_control_5,metric='Accuracy')
rf
# variable importance
#varImp(rf)

# XGBTree model
```

```
xgb <- train(Category ~., data = Train, method = "xgbTree", trControl =
train_control_5,
             metric='Accuracy')
xgb

pred<-predict(svm3, Validation[,1:554])
confusionMatrix(pred,Validation$Category)

# Naive Bayes
nb <- train(Category ~., data = Train, method = "naive_bayes", trControl =
train_control_5,metric='Accuracy')
nb

# neural net

nnet <- train(Category ~.-Category, data = Train, method = "nnet", trControl
= train_control_5,metric='Accuracy')
nnet

# logistic regression
glm <- train(Category ~., data = Train, method = "glm", trControl =
train_control_5,metric='Accuracy')
glm

# Gradient booseted Tree
gbm<- train(Category ~., data = Train, method = "gbm", trControl =
train_control_5,metric='Accuracy')
gbm

# deepnet
dnn<- train(Category ~., data = Train, method = "dnn", trControl =
train_control_5,metric='Accuracy')


.####################### Gene Importance
#############################################

## Identify Gene Importance using Looping ##
#nb_gene <- train(Category ~., data = Train, method = "naive_bayes",
trControl = train_control_5,metric='Accuracy')
#mean(nb_gene$results$Accuracy)

############################# using loop to identify inference score
#############################
j<-1
res<-vector("list", length(Train))
for(i in 1:length(Train)){

  nb_gene <- train(Category ~., data = Train[-i], method = "svmRadial",
trControl = train_control_5,metric='Accuracy')
  res[[j]]<-max(nb_gene$results$Accuracy)
  print(j)
  j<-j+1
}

#unlist the result and store to data frame
unlist(res)->res
```

```
data.frame(Gene=names(Train[-length(Train)]),Result=res)->Gene.Importance

# plot the importance
Gene.Importance %>%
  ggplot(aes(x=Gene,y=Result))+
  geom_point()+
  geom_hline(yintercept=max(nb$results$Accuracy), col="red")

#percentage contribution

per_cont<-function(x) ((x-
mean(nb$results$Accuracy))/mean(nb$results$Accuracy))*100

unlist(lapply(Gene.Importance$Result, per_cont))->Gene.Importance$Percentage

# filter the genes that don't contribute

Gene.Importance %>% filter(round(Result,3)> 0.64 | round(Result,3)< 0.63)-
>Gene.percent

# using variable importance
varImp(nb)$importance %>%
  as.data.frame() %>%
  rownames_to_column() %>%
  arrange(desc(Sensitive)) %>%
  mutate(Gene=forcats::fct_inorder(rowname ),Importance=Sensitive) %>%
  select(Gene,Importance)->Gene.var.imp

# check the correlation
merge(Gene.percent,Gene.var.imp,by.x = "Gene",by.y = "Gene")->imp.vs.inf

# plot graph

imp.vs.inf %>% ggplot(aes(x=Percentage,y=Importance))+ geom_point()

# filter important genes
Train %>% select(Gene.percent$Gene,Category) ->Train.cut


################################# Step 1 ##########################
# need to add genes one by one and see if they increase or decrease . store
the accuracy and then plot.

j<-1
#res<-vector("list", 667)
res<-vector("list", length(Train))
#for(i in 1:667){
for(i in 1:length(Train)-1){
  res[[j]] <- max(train(Category ~., data = Train[c(1:i,length(Train))],
method = "svmRadial",
                       trControl =
train_control_5,metric='Accuracy')$results$Accuracy)
  print(j)
  j<-j+1
}

# transform the accuracy data.
unlist(res)->res
```

```
#data.frame(Gene=names(Train[-668]),Accuracy=res)->Gene.Importance
data.frame(Gene=names(Train[-length(Train)]),Accuracy=res[1:length(res)-1])-
>Gene.Importance
#percentage contribution

per_cont<-function(x) ((x-
max(svm3$results$Accuracy))/max(svm3$results$Accuracy))*100

unlist(lapply(Gene.Importance$Accuracy, per_cont))-
>Gene.Importance$Percentage


# plot without sorting.
ggplot(Gene.Importance,aes(x=1:nrow(Gene.Importance),y=Accuracy))+
geom_line(color="orange")+
  geom_point(size=0.75,color="blue")+
  geom_hline(yintercept=max(svm3$results$Accuracy), col="red")+
  xlab("Genes")+
  labs(title = "Train Data Metric
Evaluation",x="Genes",y="Accuracy",color="Metrics\n")+
  theme_minimal()+
  theme(axis.text.x = element_text(angle = 90, hjust = 1))


Gene.Importance %>% arrange(Accuracy)->reorder.gene

# plot the accuracy and show increase and decrease with addition of genes
  ggplot(reorder.gene %>%
arrange(Accuracy),aes(x=1:nrow(Gene.Importance),y=Accuracy))+
geom_line(color="blue")+
  geom_point(size=0.5)+
  geom_hline(yintercept=max(svm1$results$Accuracy), col="red")+
  labs(title = "Train Data Metric
Evaluation",x="Genes",y="Accuracy",color="Metrics\n")+
  theme_minimal()+
  theme(axis.text.x = element_text(angle = 90, hjust = 1))


write.csv(reorder.gene,"Gene_Importance_Step1.csv")

############################ Step 2 ###########################

reorder.gene<-read.csv("Gene_Importance_Step1.csv")
reorder.gene %>% arrange(desc(Accuracy)) %>% select(-X)->reorder.gene

set.seed(56)

# add gene from the ascending order. add one by one and see if
# there is a 2% increase to accuracy only then include that gene into the
model.
# if there is no increase remove the gene and move on to the next one.

thershold<-0
cols<-c()
acc<-c()
#for(i in 1:667){
for(i in 1:(length(Train)-1)){
  cols<-c(cols,reorder.gene$Gene[i])
  temp <- max(train(Category ~., data = Train[c(cols,"Category")], method = 
"svmRadial",
                       trControl = 
train_control_5,metric='Accuracy')$results$Accuracy)

  # check for 2% increase in the accuracy
  if(temp>((0.01*thershold)+thershold))
  {
    thershold<-temp
    acc<-c(acc,thershold)
    print(thershold)

  }else{
    cols<-cols[1:length(cols)-1]
  }

}

########################### Retrain with important genes 
###########################

write(cols,"Important Genes.txt")

# Important Genes
cols

# Plot the data
plot.data<-data.frame(Genes=cols,Accuracy=acc)
plot.data$Genes<-with(plot.data, reorder(Genes, Accuracy))

# plot the graph
ggplot(plot.data,aes(x=Genes,y=Accuracy,group=1))+ geom_line(color="orange")+
  geom_point(size=1,color="blue")+
  geom_hline(yintercept=max(svm3$results$Accuracy), col="red")+
  labs(title = "Train Data Metric Evaluation with important 
genes",x="Genes",y="Accuracy")+
  theme_minimal()+
  theme(axis.text.x = element_text(angle = 90, hjust = 1))

# prune the train
Train.cut<-Train[c(cols,"Category")]
Validation.cut<-Validation[c(cols,"Category")]


# retrain the model with important cols.
svm.imp<-train(Category ~., data = Train.cut, method = "svmRadial",
      trControl = train_control_5,metric='Accuracy')
svm.imp



# Test data prediction
pred<-predict(svm.imp, Validation[,1:552])
confusionMatrix(pred,Validation$Category)

######### Gene names #########

col_changed<-str_extract(cols,"([0-9a-zA-Z]+)")
```

```
data.frame(ensembldb::select(EnsDb.Hsapiens.v79, keys= cols,
                              keytype = "GENEID", columns =
c("SYMBOL","GENEID")))->GTemp

GTemp

#############################################################################
#######

Truth<-mt_final[c(cols,"Category")]
Truth %>% dplyr::filter(Category=="Sensitive") %>% dplyr::select(-Category)-
>Truth.Sen
Truth %>% dplyr::filter(Category=="Insensitive") %>% dplyr::select(-
Category)->Truth.Insen

med.s<-reshape::melt(lapply(Truth.Sen,median))$value
med.is<-reshape::melt(lapply(Truth.Insen,median))$value


# create truth table
kcalc<-function(r,flag){
  if(flag=="Sen"){
    temp<-if_else(Truth.Sen[r]<med.s,0,1)
  }else if(flag=="Insen"){
    temp<-if_else(Truth.Insen[r]<med.is,0,1)
  }

  return(temp)
}

# define truth table
Truth.Table.sensitive<-data.frame(C=NA)
Truth.Table.insensitive<-data.frame(C=NA)

# Calculate Sentive truth table
for(i in 1:length(Truth.Sen)){
  temp<-data.frame(kcalc(i,flag = "Sen"))
  Truth.Table.sensitive<-cbind(Truth.Table.sensitive,temp)
}
# update colnames
Truth.Table.sensitive<-Truth.Table.sensitive[-1]
colnames(Truth.Table.sensitive)<-names(Truth.Sen)
Truth.Table.sensitive$outcome<-1


# Calculate InSentive truth table
for(i in 1:length(Truth.Insen)){
  temp<-data.frame(kcalc(i,flag = "Insen"))
  Truth.Table.insensitive<-cbind(Truth.Table.insensitive,temp)
}

Truth.Table.insensitive<-Truth.Table.insensitive[-1]
colnames(Truth.Table.insensitive)<-names(Truth.Insen)
Truth.Table.insensitive$outcome<-0

# final table with sensitive 1 and insenstive 0.
Truth_table<-rbind(Truth.Table.sensitive,Truth.Table.insensitive)
```

```
# Use Enchanced Quine-McCluskey Algorithm
QMO<-eQMC(Truth_table,outcome = 'outcome')
# display the truth table
print(QMO$tt)

# display equation
QMO

# write to file
capture.output(QMO,file = "truthtable_equation.txt")
capture.output(QMO$tt,file = "truthtable_output.txt")


#############################################################################
###
```

Source code for Prostate Cancer

```
library(tidyverse)
#library(CePa)
library(data.table)
library(matrixStats)
library(ggplot2)
library(caret)
library(rminer)
#library(EnsDb.Hsapiens.v79)
#library(doSNOW)
library(deepnet)
#library(QCApro)
library(PRROC)

# register parallel
getDoParWorkers()
getDoParName()
# register parallel
registerDoSNOW(makeCluster(4, type = "SOCK"))
getDoParVersion()


#BiocManager::install('EnsDb.Hsapiens.v79')

set.seed(67)


#############################################################################
# START HERE FOR MODELS

# Load data just for model.
mt_final<-read.csv("ProstateCancer.csv")
rownames(mt_final)<-mt_final$X
mt_final<-mt_final[,c(-1)]
mt_final$Category<-as.factor(mt_final$Category)

# For Prostate cancer.
mt_final<- read.csv("ProstateCancer.csv")
mt_final<-mt_final[,c(-1)]
```

```
mt_final$Category<-as.factor(mt_final$Category)


#remove zero variance columns and correlated variables
# Skip this for prostate cancer
#mt_final_T<-mt_final[,-nearZeroVar(mt_final)]
mt_final_c<-mt_final[,-findCorrelation(cor(mt_final[,-length(mt_final)]),
.8)]
mt_final<- mt_final_c

##############################################################################
###


########### Build SVM model ###########

# Create Train validation

TrainIndex<-createDataPartition(mt_final$Category, p = 0.7, list = FALSE,
times = 1)
Train<-mt_final[TrainIndex,]
Validation<-mt_final[-TrainIndex,]

# Traincontrol
train_control_5 <- trainControl(
  method = "cv",
  number = 6
)
# Train control 10
train_control_10 <- trainControl(
  method = "cv",
  number = 6,
  classProbs = TRUE,
  savePredictions = T,
  summaryFunction = twoClassSummary
)

####




####
# SVM model
svm1 <- train(Category ~., data = Train, method = "svmLinear", trControl =
trainControl(
  method = "cv",
  number = i
),metric="Accuracy")
svm1
svm2 <- train(Category ~., data = Train, method = "svmPoly", trControl =
train_control_5,metric="Accuracy")
svm2
svm3 <- train(Category ~., data = Train, method = "svmRadial", trControl =
train_control_5,metric="Accuracy")
svm3

svmt<- train(Category ~., data = Train, method = "svmPoly", trControl =
train_control_5,metric="Accuracy",
             tuneGrid = expand.grid(
                                    degree= c(1,2),
                                    scale=c(0.1,0.01,0.001),
                                    C=seq(0, 1, by = 0.1)
                                    ))
svmt
besttune<-svmt$bestTune
# Random Forest model

rf <- train(Category ~., data = Train, method = "rf", trControl =
train_control_5,metric='Accuracy')
rf
# variable importance
#varImp(rf)

# XGBTree model

xgb <- train(Category ~., data = Train, method = "xgbTree", trControl =
train_control_5,
             metric='Accuracy',preProcess = c('center', 'scale'))
xgb

#pred<-predict(xgb, Validation[,1:554])
#confusionMatrix(pred,Validation$Category)


# Naive Bayes
nb <- train(Category ~., data = Train, method = "naive_bayes", trControl =
train_control_5,metric='Accuracy')
nb

# neural net

nnet <- train(Category ~., data = Train, method = "nnet", trControl =
train_control_5,preProcess = c('center', 'scale'))
nnet

# logistic regression
glm <- train(as.factor(Category) ~., data = Train, method = "glm", trControl
= train_control_5,metric='Accuracy')
glm

# Gradient booseted Tree
gbm<- train(Category ~., data = Train, method = "gbm", trControl =
train_control_5,metric='Accuracy')
gbm

# deepnet
dnn<- train(Category ~., data = Train, method = "dnn", trControl =
train_control_5,metric='Accuracy')
dnn

# Show results of all models
acc<-c()
model<-list(svm1,svm2,svm3,xgb,rf,nb,glm,dnn)
m_names<-
c('SVM_Linear','SVM_Poly','SVM_Radial','XGBoost','RF','NB','GLM','DNN')
for(i in model){
  acc<-c(acc,max(na.rm = T,i$results$Accuracy))
}

# create the plot with accuracy
pl<-data.frame(Model=m_names,Accuracy=acc)
pl %>% ggplot(aes(x=Model,y=Accuracy,fill=Model))+geom_bar(stat =
'identity')+theme_light()


######################## Gene Importance
#############################################
##Variable Importance to filter important genes from 3400

varImp(svm1)->vimp
as.data.frame(vimp$importance)->vimp
vimp$gene<-rownames(vimp)
vimp %>% dplyr::filter(NR>50) %>% dplyr::select(gene)->gene_cols
rownames(gene_cols)<-NULL
as.vector(gene_cols)->gene_cols

Train %>% dplyr::select(gene_cols$gene,Category)->Train
Validation %>% dplyr::select(gene_cols$gene,Category)->Validation

# Identify the CV number
cv_num<-c()
cv_acc<-c()
for(i in 2:nrow(Train)){
  cv_acc <- c(cv_acc,max(train(Category ~., data = Train, method =
"svmLinear", trControl = trainControl(
    method = "cv",
    number = i
  ),metric="Accuracy")$results$Accuracy))
  cv_num<-c(cv_num,i)

}

data.frame(CV=cv_num,Accuracy=cv_acc)->pp
pp %>%
ggplot(aes(x=CV,y=Accuracy))+geom_line(color='blue')+geom_point(color='black'
)+theme_light()

svm1<-train(Category ~., data = Train, method = "svmLinear", trControl =
train_control_5,metric="Accuracy")

################################# Step 1 ###########################
# need to add genes one by one and see if they increase or decrease . store
the accuracy and then plot.

j<-1
#res<-vector("list", 667)
res<-vector("list", length(Train))
#for(i in 1:667){
for(i in 1:length(Train)-1){
  res[[j]] <- max(train(Category ~., data = Train[c(1:i,length(Train))],
method = "svmLinear",
                        trControl =
train_control_5,metric='Accuracy')$results$Accuracy)
  print(j)
  j<-j+1
}

# transform the accuracy data.
unlist(res)->res
#data.frame(Gene=names(Train[-668]),Accuracy=res)->Gene.Importance
data.frame(Gene=names(Train[-length(Train)]),Accuracy=res[1:length(res)-1])-
>Gene.Importance
#percentage contribution

per_cont<-function(x) ((x-
max(svm3$results$Accuracy))/max(svm3$results$Accuracy))
diff_cont<-function(x) ((x-
mean(Gene.Importance$Accuracy))/sd(Gene.Importance$Accuracy))

unlist(lapply(Gene.Importance$Accuracy, per_cont))-
>Gene.Importance$Percentage
unlist(lapply(Gene.Importance$Accuracy, diff_cont))-
>Gene.Importance$Difference

# plot without sorting.
ggplot(Gene.Importance,aes(x=1:nrow(Gene.Importance),y=Percentage))+
geom_line(color="orange")+
  geom_point(size=0.75,color="blue")+
  geom_hline(yintercept=0, col="red")+
  xlab("Genes")+
  labs(title = "Train Data Metric
Evaluation",x="Genes",y="Accuracy",color="Metrics\n")+
  theme_light()+
  theme(axis.text.x = element_text(angle = 90, hjust = 1))



Gene.Importance %>% arrange(desc(Difference))->reorder.gene

# plot the accuracy and show increase and decrease with addition of genes
ggplot(reorder.gene %>% arrange(desc(Difference)) %>%
head(20),aes(x=1:20,y=Difference))+
  geom_point(aes(label=Gene),size=2)+
  geom_line(color="blue",size=1)+
  geom_text(aes(label=Gene),hjust=0, vjust=0,angle = 25)+
  labs(x="Input features",
       y="Influence score",color="Metrics\n")+
  theme_classic()+
  theme(axis.text.x = element_text(angle = 90, hjust = 1))


top_20<-reorder.gene %>% arrange(desc(Difference)) %>% head(20)
res<-c()
for(i in 1:20){
  Train %>% dplyr::select(top_20$Gene[1:i],Category)->temp
  res<-c(res,max(train(Category ~., data = temp, method = "svmLinear",
            trControl = train_control_5,metric='Accuracy')$results$Accuracy))
}

data.frame(Accuracy=res,Names=top_20$Gene)->pp
```

```
pp %>% mutate(Names = fct_reorder(Names, Accuracy, .fun='median')) ->pp
ggplot(pp,aes(x=Names,y=Accuracy))+geom_bar(stat='identity',fill='darkblue')+
theme_classic()+
  theme(axis.text.x = element_text(angle = 25,vjust = 0.45))

Train %>% dplyr::select(top_20$Gene[1:20],Category)->temp
Validation %>% dplyr::select(top_20$Gene[1:20],Category)->tv
topSvm<-train(Category ~., data = temp, method = "svmLinear",trControl =
train_control_10,metric='Accuracy')
evalm(topSvm,gnames = c('RP','NR'))

pred<-predict(topSvm, tv[,1:20])
confusionMatrix(pred,tv$Category)

# ROC curve
evalm(svm1,gnames = c('RP','NR'))

write.csv(reorder.gene,"Gene_Importance_Step1.csv")

############################ Step 2 ###########################

reorder.gene<-read.csv("Gene_Importance_Step1.csv")
reorder.gene %>% arrange(desc(Accuracy))->reorder.gene

set.seed(56)

# add gene from the ascending order. add one by one and see if
# there is a 2% increase to accuracy only then include that gene into the
model.
# if there is no increase remove the gene and move on to the next one.

thershold<-0
cols<-c()
acc<-c()
#for(i in 1:667){
for(i in 1:(length(Train)-1)){
  cols<-c(cols,reorder.gene$Gene[i])
  temp <- max(train(Category ~., data = Train[c(cols,"Category")], method =
"svmLinear",
                  trControl =
train_control_5,metric='Accuracy')$results$Accuracy)

  # check for 2% increase in the accuracy
  if(temp>((0.01*thershold)+thershold))
  {
    thershold<-temp
    acc<-c(acc,thershold)
    print(thershold)

  }else{
    cols<-cols[1:length(cols)-1]
  }

}

########################### Retrain with important genes
###########################
```

```
write(cols,"Important Genes_PC.txt")

# Important Genes
cols

# Plot the data
plot.data<-data.frame(Genes=cols,Accuracy=acc)
plot.data$Genes<-with(plot.data, reorder(Genes, Accuracy))

# plot the graph
ggplot(plot.data,aes(x=Genes,y=Accuracy,group=1))+ geom_line(color="black")+
  geom_point()+
  geom_bar(stat='identity',fill="darkblue")+
  geom_hline(yintercept=max(svm3$results$Accuracy), col="red")+
  labs(title = "Train Data Metric Evaluation with important 
genes",x="Genes",y="Accuracy")+
  theme_minimal()+
  theme(axis.text.x = element_text(angle = 45, hjust = 0.5,size=12))

# plot to show inference score.
res<-c()
for(i in 1:10){
  Train %>% dplyr::select(plot.data$Gene[1:i],Category)->temp
  res<-c(res,max(train(Category ~., data = temp, method = "svmLinear",
                     trControl = 
train_control_5,metric='Accuracy')$results$Accuracy))
}

data.frame(Accuracy=res,Names=plot.data$Gene)->pp
pp$Names<-with(pp, reorder(Names,typical))
#pp$typical<-acc

ggplot(pp,aes(x=Names,y=typical,group=1))+geom_bar(stat='identity',fill='dark
blue',width = 0.5)+
  geom_line(aes(x=Names,y=))+
  geom_point(aes(y=Accuracy),shape=2)
  theme_classic()+
  theme(axis.text.x = element_text(angle = 25,vjust = 0.45))

# Heatmaps
t(Train[c(cols,'Category')])->heat
colnames(heat)<-heat['Category',]
heat[-8,]->heat
as.data.frame(heat)->heat

heatmap(as.matrix(heat))

# prune the train
Train.cut<-Train[c(cols,"Category")]
Validation.cut<-Validation[c(cols,"Category")]


# retrain the model with important cols.
svm.imp<-train(Category ~., data = Train.cut, method = "svmLinear",
               trControl = train_control_10,metric='Accuracy')
svm.imp

svm.imp1<-train(Category ~., data = Train.cut, method = "svmLinear",
               trControl = train_control_10,metric='Accuracy')
svm.imp1

evalm(svm.imp1,gnames = c('NR','RP'))

# Test data prediction
pred<-predict(svm.imp, Validation.cut[,-11])
confusionMatrix(pred,Validation.cut$Category)


# plot to identify best number

Gene.Importance %>% dplyr::filter(Gene %in% plot.data$Genes)->plot.data
plot.data$Gene<-with(plot.data, reorder(Gene, Accuracy))

ggplot(plot.data,aes(x=Gene,y=Accuracy,group=1))+ geom_line(color="black")+
  geom_point()+
  geom_bar(stat='identity',fill="darkblue")+
  geom_hline(yintercept=max(svm3$results$Accuracy), col="red")+
  labs(title = "Train Data Metric Evaluation with important
genes",x="Genes",y="Accuracy")+
  theme_minimal()+
  theme(axis.text.x = element_text(angle = 45, hjust = 0.5,size=12))


######### Gene names #########

col_changed<-str_extract(cols,"([0-9a-zA-Z]+)")

data.frame(ensembldb::select(EnsDb.Hsapiens.v79, keys= cols,
                             keytype = "GENEID", columns =
c("SYMBOL","GENEID")))->GTemp

GTemp

##############################################################################
#######

write.csv(plot.data,'PC_plot_data.csv')

Truth<-mt_final[c(cols,"Category")]
Truth %>% dplyr::filter(Category=="RP") %>% dplyr::select(-Category)-
>Truth.Sen
Truth %>% dplyr::filter(Category=="NR") %>% dplyr::select(-Category)-
>Truth.Insen

med.s<-reshape::melt(lapply(Truth.Sen,median))$value
med.is<-reshape::melt(lapply(Truth.Insen,median))$value


# create truth table
kcalc<-function(r,flag){
  if(flag=="RP"){
    temp<-if_else(Truth.Sen[r]<med.s,0,1)
  }else if(flag=="NR"){
    temp<-if_else(Truth.Insen[r]<med.is,0,1)
  }
  return(temp)
}

# define truth table
Truth.Table.sensitive<-data.frame(C=NA)
Truth.Table.insensitive<-data.frame(C=NA)

# Calculate Sentive truth table
for(i in 1:length(Truth.Sen)){
  temp<-data.frame(kcalc(i,flag = "RP"))
  Truth.Table.sensitive<-cbind(Truth.Table.sensitive,temp)
}
# update colnames
Truth.Table.sensitive<-Truth.Table.sensitive[-1]
colnames(Truth.Table.sensitive)<-names(Truth.Sen)
Truth.Table.sensitive$outcome<-'RP'


# Calculate InSentive truth table
for(i in 1:length(Truth.Insen)){
  temp<-data.frame(kcalc(i,flag = "NR"))
  Truth.Table.insensitive<-cbind(Truth.Table.insensitive,temp)
}

Truth.Table.insensitive<-Truth.Table.insensitive[-1]
colnames(Truth.Table.insensitive)<-names(Truth.Insen)
Truth.Table.insensitive$outcome<-'NR'

# final table with sensitive 1 and insenstive 0.
Truth_table<-rbind(Truth.Table.sensitive,Truth.Table.insensitive)


# Use Enchanced Quine-McCluskey Algorithm
QMO<-eQMC(Truth_table,outcome = 'outcome')
# display the truth table
print(QMO$tt)

# display equation
QMO

# write to file
capture.output(QMO,file = "truthtable_equation.txt")
capture.output(QMO$tt,file = "truthtable_output.txt")


library(ComplexHeatmap)

ha = rowAnnotation(
  df = data.frame(Outcome=Truth_table$outcome),
  annotation_height = unit(4, "mm"),
  show_annotation_name = TRUE,
  col= list(Outcome = c('RP' = "Green", 'NR' = "darkorange"))
)

Heatmap(Truth_table[cols],
        border = T,rect_gp = gpar(col='Black'),
        name = "Expression",
        right_annotation=ha,
        row_names_gp = gpar(fontsize = 10),
        column_names_gp = gpar(fontsize = 10,font=3)
        )

Heatmap(Truth[cols],
        border = T,rect_gp = gpar(col='Black'),
        name = "Expression",
        right_annotation=ha,
        row_names_gp = gpar(fontsize = 10),
        column_names_gp = gpar(fontsize = 10,font=3)
)



# all the data.
Heatmap(mt_final[-3793],km=2,
        show_column_names = FALSE,
        name = "Expression",
        right_annotation=ha,
        row_names_gp = gpar(fontsize = 10),
        column_names_gp = gpar(fontsize = 10,font=3)
)




##############################################################################
###
```