Statistical Frameworks for Integrative Analysis of Genetic Data

A Dissertation Presented

by

Lizhen Peng

to

The Graduate School

in Partial Fulfillment of the

Requirements

for the Degree of

Doctor of Philosophy

in

Applied Mathematics and Statistics

Stony Brook University

May 2015

Stony Brook University

The Graduate School

Lizhen Peng

We, the dissertation committee for the above candidate for the

Doctor of Philosophy degree, hereby recommend

acceptance of this dissertation.

Xuefeng Wang – Dissertation Advisor
Assistant Professor, Department of Preventive Medicine

Wei Zhu – Dissertation Co-advisor
Professor, Department of Applied Mathematics and Statistics

Pei Fen Kuan - Chairperson of Defense
Assistant Professor, Department of Applied Mathematics and Statistics

Song Wu - Member
Assistant Professor, Department of Applied Mathematics and Statistics

Wadie F. Bahou – External member
Professor, Department of Medicine

This dissertation is accepted by the Graduate School

Charles Taber
Dean of the Graduate School

Abstract of the Dissertation

Statistical Frameworks for Integrative Analysis of Genetic Data

by

Lizhen Peng

Doctor of Philosophy

in

Applied Mathematics and Statistics

Stony Brook University

2015

We studied three major interconnected projects focusing on establishing frameworks for integrative analysis of genetic data, based on regularized regression models, support vector machines regressions, Cox proportional hazard models, multiple kernel learning models, and kernel Cox regressions, incorporated with dimensionality reduction and feature selections. In Project 1 we employed several machine learning algorithms for clinical predictions utilizing omics data across cancer types, to explore the potential benefits of including genetic measurements with traditional clinical information in supporting the doctors' decision making process. To predict the survival of patients with cancer, we established two predictive models. First we applied the multivariate Cox proportional hazard (Cox) models with univariate Cox screen or correlation screen, plus L1 penalized log partial likelihood (LASSO) for feature selection. Secondly, we also examined the factors that could affect prediction of dichotomized survival data by different machine learning algorithms, especially the multiple kernel learning

(MKL) algorithms for its data fusion capability. Our analysis indicates that incorporating omics data with clinical information can significantly improve prediction accuracy. Our study provides a sound framework and resources for reliable prognostic modeling and therapeutic decision making. In Project 2 we assessed comprehensively, by using genome-wide DNA methylation data as markers, the contribution of epigenetic effects on asthma and blood related quantitative traits. To evaluate the clinical utility of epigenetic markers, we constructed and compared various prediction models by including top ranked methylation loci from the genome-wide association scan, together with selected sets of known genetic markers from published genome-wide association studies. We observed a significant increase in correlation between actual and predicted IgE level when methylation markers were included. We also assessed the performance of cross platform prediction using methylation markers. Taken together, results from our assessment suggest that methylation has great potential in predicting clinical phenotypes. Finally, in Project 3, we explored the kernel Cox regression models and survival support vector machines to improve the prediction accuracy of patients with metastatic castrate resistant prostate cancer (mCRPC) treated by docetaxel. We studied the effects of utilizing clinical kernels to obtain better results than linear and Gaussian kernels, with clinical variables for prognostic modeling.

Dedicated to my beloved parents

Table of Content

List of Figures

List of Tables

Acknowledgments

Foremost, I would like to give my greatest thanks to Prof. Wei Zhu, who had offered me the precious opportunity to work in her group since the beginning of my PhD study, and later introduced me to Prof. Xuefeng Wang, who becomes my main dissertation advisor. It is my great honor to have both of them as my advisors. The projects that I have worked on are all very exciting and challenging. My deepest thanks go to my advisors, for their support and advice, for their patience, motivation, enthusiasm, immense knowledge, and for their dedication of a tremendous amount of time and guidance.

Besides my advisors, I would like to give my thanks to my other committee members Prof. Pei Fen Kuan, Prof. Song Wu, and Prof. Wadie Bahou for being so caring and supportive.

Last but not the least, I want to thank my family for always being so loving and supportive during my entire doctoral study, and my group members and friends at Stony Brook University who have cheered me on and supported me, all the way.

Vita

Lizhen Peng was born in China as the second child in her family. In summer 2009, she graduated from Shanghai Jiao Tong University with a Bachelor's degree in Biological Engineering, and then joined Purdue University Interdisciplinary Life Science Graduate Program (PULSe) in fall to start her doctoral study. Along the way during her study at Purdue, she found her greatest passion in Statistics and decided to pursue her future study and career in statistical science. In May 2012, she graduated from Purdue, with a Master's degree in PULSe/Chemistry, plus a Graduate Certificate in Applied Statistics. In the same year, she joined the doctoral program at the Department of Applied Mathematics and Statistics (AMS), Stony Brook University (SBU), and thus began her wonderful journey in statistics ever since. Lizhen Peng received her Ph.D degree in Statistics from SBU in May 2015.

Chapter 1 Introduction

High-throughput assay technologies have enabled various genomic profiles available for study, which include mRNA gene expression, DNA methylation (mDNA), reverse phase protein array (RPPA), and microRNA for over than 30 cancer types. However, the benefits of integrating genomic profiles with traditional clinical variables, for clinical management of cancer, have not yet been systematically studied.

In this work, we established three comprehensive analytic frameworks for integrative analysis of genetic data, for different research goals. This chapter will provide an overview of these three interconnected main projects (Figure 1.1) followed by a brief introduction of our general approaches in model assessment and selection.

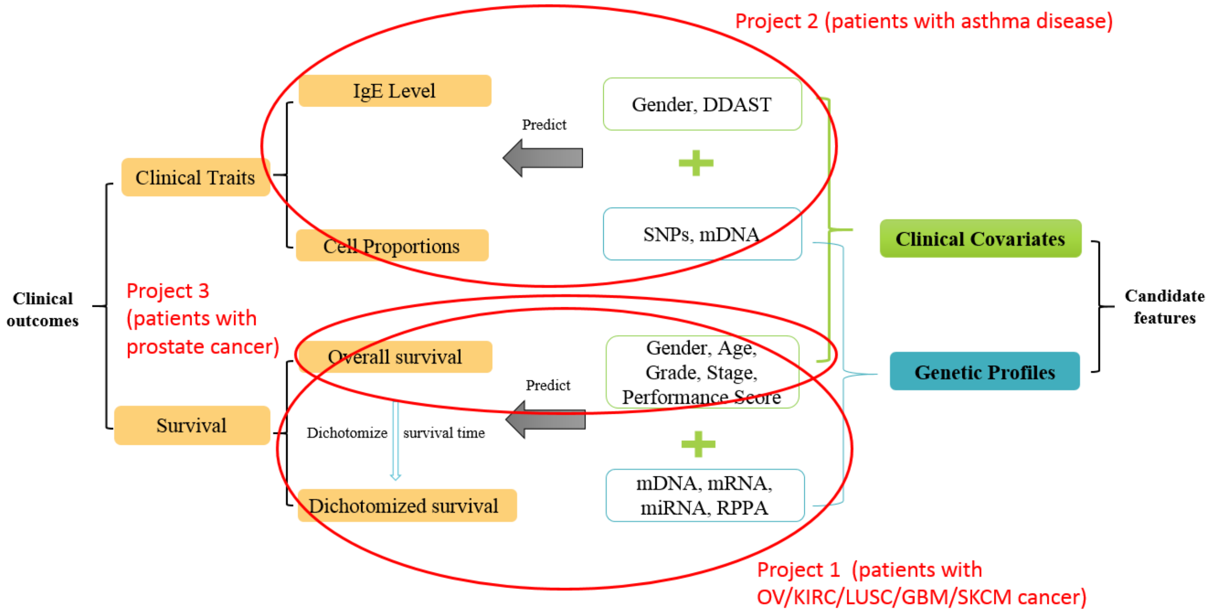


Figure 1.1 Illustration of Our Projects

1.1 Literature Review

Providing accurate patient prognosis is critical in clinical decision making for it will enable doctors to group patients into different risk groups and subsequently, choose the best treatment strategies. Commonly, prognosis is based on clinical variables such as age and cancer stage, and more recently a few gene expression-based markers have been adapted in clinical practice. Researchers have made extensive efforts to incorporate genetic measurement into prognostic modeling. For instance, important biomarkers in breast cancer, such as ER, PR, HER2 protein levels and HER2 genomic amplification have shown high clinical value [1]. However, this and other previous studies have either focused on a small number of selected genes/proteins or have used only single-platform genomic data, due to the high cost of large scale genetic profiling.

1.2 Overview of Three Interconnected Projects

Clinical predictions utilizing multi omics data across cancer types

First, for predicting survival for patients with cancer, we studied several different cancer types including ovarian serous cystadenocarcinoma (OV), kidney renal clear cell carcinoma (KIRC), glioblastoma multiforme (GBM), lung squamous cell carcinoma (LUSC), and skin cutaneous melanoma (SKCM), by applying multivariate Cox proportional hazard (Cox) models with univariate Cox screen or correlation screen, plus L1 penalized log partial likelihood (incorporated in   LASSO) for feature selection. We also explored the factors that could affect prediction of dichotomized survival data, utilizing several methods: (i) Support Vector Machines (SVMs), (ii) K-nearest neighbor (KNN), (iii) Random Forest (RF), and most importantly (iv)

MKL algorithms to optimize their prediction performance.

Recently, the availability of multi Omics data have provided scientists with complementary views on survival analysis for cancers and at the same time, highlighted a particular challenge: to integrate genetic data in different measurements and from different sources. In this study, we proposed statistical frameworks for integrative analysis of genetic data, to study the potential benefits of including genetic measurements with traditional clinical information.

In order to predict survival time of patients with cancer, we studied several cancer types including ovarian serous cystadenocarcinoma (OV), kidney renal clear cell carcinoma (KIRC), glioblastoma multiforme (GBM), and lung squamous cell carcinoma (LUSC). First we applied multivariate Cox proportional hazard (Cox) models with L1 penalized log partial likelihood (LASSO) for feature selection. We studied the Cox models with only clinical information and models incorporating both genetic data and clinical information, to identify the marginal improvement of adding genetic profiles.

In addition to analyzing censored survival data, we also examined the factors that could affect prediction of dichotomized survival data. There are many machine learning algorithms for classifying binary outcomes, and we picked three most popular and well-established methods for comparison purpose: (i) Support Vector Machines (SVM), (ii) K-nearest neighbor (KNN), (iii) Random Forest (RF). But none of these methods are capable of integrating data coming from different feature sets (different representations), that is why we are most interested in studying the multiple kernel learning (MKL) algorithms for data fusion.

Predicting serum IgE level and blood cell proportions from methylation profile

Secondly, we also studied the impact of integrating genetic and epigenetic profiles on predicting blood cell proportions of patients with asthma. To evaluate the clinical utility of epigenetic markers, we constructed and compared various prediction models by including top ranked methylation loci from the genome-wide association scan, together with selected sets of known genetic markers from published genome-wide association studies.

DNA methylation at CpG sites is an important epigenetic modification that may regulate gene expression. There is a growing interest in understanding how the methylation inheritance contributes to the development of complex diseases or traits. It has been shown that methylation modification may influence individual asthma risk and related phenotypes. The primary purpose of this study is to comprehensively assess—by using genome-wide DNA methylation data as markers—the contribution of epigenetic effects on asthma and blood related quantitative traits. To evaluate the clinical utility of epigenetic markers, we constructed and compared various prediction models by including top ranked methylation loci from the genome-wide association scan, together with selected sets of known genetic markers from published genome-wide association studies. A new prediction model based upon Best Linear Unbiased Prediction (BLUP) was further proposed where all CpG sites (on the Illumina Infinium 27K methylation array and 450K array) were simultaneously modeled. The overall prediction accuracies of the proposed methods were extensively evaluated via the cross-validation analysis. We observed a significant increase in correlation between the actual and the predicted IgE level when methylation markers were included. By using an independent sample based on Illumina 450K methylation array, we also assessed the performance of cross platform prediction using methylation markers. Taken together, results from our assessment suggest that methylation has great potential in prediction of clinical phenotypes.

Survival prediction with Kernel Methods

Our third project mainly focuses on improving the survival prediction for patients with metastatic castrate resistant prostate cancer (mCRPC). We implemented Kernel Cox regressions, especially the case with Clinical kernel, which is supposed to be better than linear and Gaussian with clinical variables. The potential benefit of this study is to establish better prognostic models, to help support clinical decisions, and to better understand the mechanism of mCRPC disease progression.

1.3 Model Assessment and Selection

To evaluate the performance of a learning method, we need to evaluate its prediction capability on test data independent to the original training set. In practice, it is extremely important to assess model performance to guide the choice of learning method or model. In this section we describe and illustrate the key methods for performance assessment, including both cross validation and information criteria (AIC, BIC), and show how they are applied in model selections. We begin this section with a discussion of the bias-variance dilemma.

Ideally, we would wish that both the bias and variance of our proposed model are as small as possible, as we prefer that with smaller expected mean square error. So in order to minimize expected MSE, we hope that we can push both bias and variance to be as small as possible. But in reality, this is not attainable, since both bias and variance are a function of model complexity, but with different trends.

For analytical approaches, popular model assessment methods are the Akaike information criterion (AIC) [2] and the Bayesian information criterion (BIC) [3]. For any statistical model, the AIC value is $AIC = -2\ln(L) + 2k$, where L denoted as the likelihood function for the model, and k is the number of parameters in the model. We prefer to choose the model with the minimum AIC value from a given set of candidate models. The BIC works in a similar as the AIC. It is also based on the likelihood function and a penalty term. Its formula is closely related to AIC, but with a slightly different penalty incorporating the sample size (N) as well: $BIC = -2\ln(L) + k\ln(N)$. Adding parameters can possibly increase the likelihood, but it may also bring in the problem of over-fitting. Introducing a penalty term for the number of parameters in the model in BIC and AIC criterion can help resolve the over-fitting issue. In general, the penalty term is larger in the BIC than in the AIC.

Except analytical methods, cross validation as an assessing approach, has also been widely used to check model error by testing on an independent data. In K-fold cross validation, first randomly partitioning the original sample into K equal size subsets, using (K – 1) subsets as training data for modeling and retaining one subset as the validation data for testing; then repeat the cross-validation process K rounds. Multiple rounds of cross-validation are performed to reduce variability of different partitions, and to average all validation results as the final result. The simplest version of K-fold cross validation is a 2-fold cross validation. And the leave-one-out cross validation is an extreme case that retains only one observation from the original sample as the validation data, while using the all but one data pints for the training.

1.4 Reference

[1] Weigel, M.T. & Dowsett, M. Current and emerging biomarkers in breast cancer: prognosis and prediction. Endocr. Relat. Cancer 17, R245–R262 (2010).

[2] Akaike, Hirotugu. "A new look at the statistical model identification." Automatic Control, IEEE Transactions on 19.6 (1974): 716-722.

[3] Schwarz, Gideon. "Estimating the dimension of a model." The annals of statistics 6.2 (1978): 461-464.

Chapter 2 Clinical Predictions Utilizing Multi Omics Data

The recent availability of multiple types of genome-wide data provides scientists with complementary views on survival analysis for cancer patients and highlights a particular challenge: to integrate genetic data in different measures and from multiple sources. The ultimate goal of this study is to establish statistical frameworks for integrative analysis of genetic data, and thereby understand the potential benefits of incorporating genetic measurements with traditional clinical information, for better patient prognostic analysis in support of clinical strategies.

For the five cancer types we have studied,Ovarian serous cystadenocarcinoma (OV), Kidney renal clear cell carcinoma (KIRC), Lung squamous cell carcinoma (LUSC), Glioblastoma multiforme (GBM), Skin Cutaneous Melanoma (SKCM),- we found that no gene expression signatures were routinely used in clinical practice for lung cancer and kidney cancer. For GBM and OV, currently findings have limited influence on clinical decision making, although MGMT promoter methylation, and the status of other few markers, are frequently used for patients with GBM and CA125 for patients with OV [46, 47].

In this chapter, we will first introduce the Cancer Genome Atlas (TCGA), where we obtained all the clinical and omics data for our study. Then we will provide an intensive and detailed workflow for TCGA data collection and preprocessing. Special attentions might be warranted for our methodologies and results, where we will show the added benefit of multi-source omics data for prediction of clinical prognostic factors (e.g. blood IgE levels and cell proportions), for survival prediction by kernel Cox regression, and for prediction of dichotomized survival data, by several machine learning algorithms, such as SVM, RF, KNN and MKL.

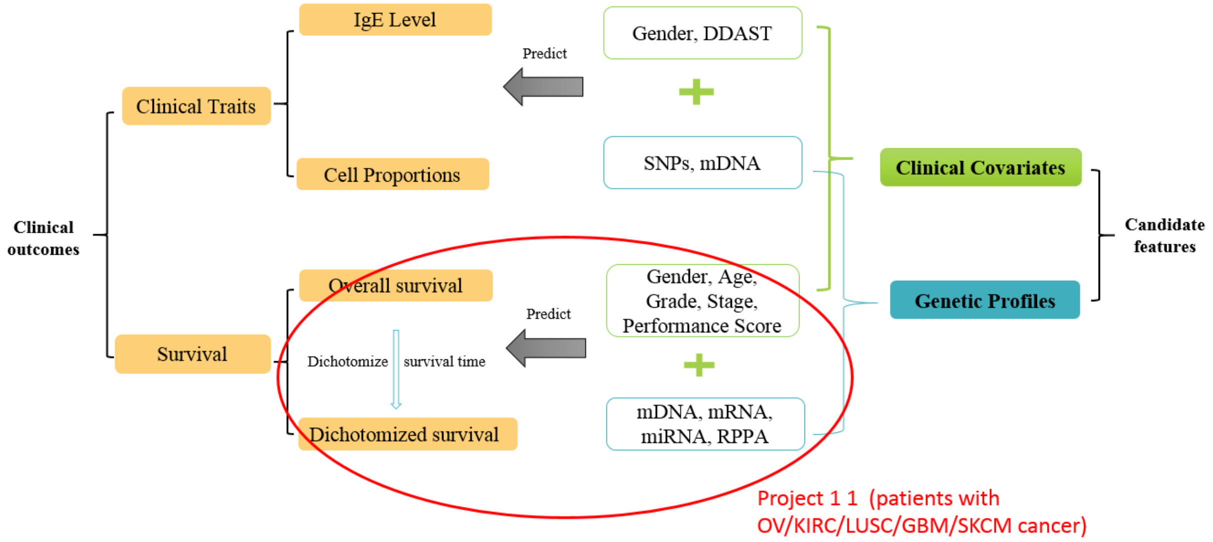


Figure 2.1 Illustration of Clinical Predictions Utilizing Multi Omics Data

2.1 Introduction

The Cancer Genome Atlas

Since its inception in 2005, the Cancer Genome Atlas (TCGA) project has made significant contributions to accelerating the interconnection study and integrative analysis of cancer genome data, using genome sequencing and bioinformatics (https://tcga-data.nci.nih.gov/tcga/) [1, 2]. As an ideal test bed for conducting and comparing different analyses, TCGA provides the public with comprehensive profiling data (e.g. Gene Expression (mRNA), DNA methylation (mDNA), microRNA expression (miRNA), and reverse phase protein array (RPPA)) for tumor samples of more than 30 cancer types [1].

With the rapid development of new genetic measurement methods, multi Omics data can be quantified in a high-throughput manner. Multi-dimensional genetic data can be analyzed in many different ways. Besides the initial focus on investigating each individual data type separately, there are increasing number of researches have been conducted with focus from other perspectives such as studies of interconnections between two or more types of regulations [3-10]. For instance, the correlation structure of mRNA with DNA methylation, CNV and miRNA have been studied in [4, 5, 10].

Overview of the Study

A few exploratory research projects has been conducted to access the prognostic power of genetic data for clinical utilization [11]. However, the marginal gains of predictive performance were very limited in these studies, by only incorporating one type of genetic data on top of clinical information at each time. Our overall goal is to improve the prediction performance, by examining different methods and algorithms, as well as examining several novel techniques for feature selection and integration.

We studied several different cancer types, ovarian serous cystadenocarcinoma (OV), kidney renal clear cell carcinoma (KIRC), glioblastoma multiforme (GBM), lung squamous cell carcinoma (LUSC) and Skin Cutaneous Melanoma (SKCM). First we applied multivariate Cox proportional hazard (Cox) models with L1 penalized log partial likelihood (LASSO) for feature selection. We studied the Cox models with only clinical information and models incorporating both genetic data with selected probes and clinical information, to identify the marginal improvement of adding genetic profiles.

In addition to analysis on censored survival data, we also examined the factors that could affect prediction of dichotomized survival data. There are many machine learning algorithms for classifying binary outcomes, and we picked three most popular and well-established methods for comparison purpose: (i) Support Vector Machines (SVM), (ii) K-nearest neighbor (KNN), (iii) Random Forest (RF). But none of these methods are capable of doing integrative analysis. So we are most interested in studying predictions by MKL, which showed a greater predictive power.

Table 2.1 List of available cancer types on TCGA Data Portal [6]

Available Cancer Types	# Cases Shipped by BCR*	# Cases with Data*	Date Last Updated (mm/dd/yy)
Acute Myeloid Leukemia [LAML]	200	200	04/29/15
Adrenocortical carcinoma [ACC]	80	80	04/24/15
Bladder Urothelial Carcinoma [BLCA]	412	412	04/24/15
Brain Lower Grade Glioma [LGG]	516	516	05/01/15
Breast invasive carcinoma [BRCA]	1100	1098	05/01/15
Cervical squamous cell carcinoma and endocervical adenocarcinoma [CESC]	308	308	05/01/15
Cholangiocarcinoma [CHOL]	36	36	05/01/15
Colon adenocarcinoma [COAD]	461	461	04/24/15
Esophageal carcinoma [ESCA]	185	185	04/27/15
FFPE Pilot Phase II [FPPP]	38	38	04/30/15
Glioblastoma multiforme [GBM]	529	528	05/01/15
Head and Neck squamous cell carcinoma [HNSC]	528	528	05/01/15
Kidney Chromophobe [KICH]	66	66	04/24/15
Kidney renal clear cell carcinoma [KIRC]	536	536	05/01/15
Kidney renal papillary cell carcinoma [KIRP]	291	291	05/01/15
Liver hepatocellular carcinoma [LIHC]	377	377	05/01/15
Lung adenocarcinoma [LUAD]	521	521	04/24/15

Lung squamous cell carcinoma [LUSC]	510	504	05/01/15
Lymphoid Neoplasm Diffuse Large B-cell Lymphoma[DLBC]	48	48	04/24/15
Mesothelioma [MESO]	87	87	04/24/15
Ovarian serous cystadenocarcinoma [OV]	586	586	05/01/15
Pancreatic adenocarcinoma [PAAD]	185	185	04/24/15
Pheochromocytoma and Paraganglioma [PCPG]	179	179	05/01/15
Prostate adenocarcinoma [PRAD]	498	498	04/28/15
Rectum adenocarcinoma [READ]	172	171	05/01/15
Sarcoma [SARC]	261	261	04/30/15
Skin Cutaneous Melanoma [SKCM]	470	470	04/28/15
Stomach adenocarcinoma [STAD]	445	443	04/28/15
Testicular Germ Cell Tumors [TGCT]	150	150	04/24/15
Thymoma [THYM]	124	124	05/01/15
Thyroid carcinoma [THCA]	507	507	04/24/15
Uterine Carcinosarcoma [UCS]	57	57	05/01/15
Uterine Corpus Endometrial Carcinoma [UCEC]	548	548	05/01/15
Uveal Melanoma [UVM]	80	80	05/01/15

2.2 Data Collection and Processing

2.2.1 Data Collection

In this project, we quantified the predictive power of cancer prognosis using the Concordance Index (C-index) for survival. All data were downloaded from the Broad GDAC

Firehose (http://gdac.broadinstitute.org/) as of December 2014. The obtained data include clinical information, mRNA, mDNA, RPPA and miRNA. We refer to the TCGA website and Firehose website for more detailed information.

In short, the clinical information contain survival status and survival time of patients, plus age, gender, cancer grade, etc.; for mRNA we download log2 lowess normalized (cy5/cy3) collapsed by gene symbol, or normalized reads, depending on which measurement is available for certain cancer type; for mDNA, we extract the beta values, which are scores calculated from methylated (M) and unmethylated (U) bead types and measure the percentages of methylation (HumanMethylation450 BeadChips, or HumanMethylation27 BeadChips, whichever is available); for miRNA, we download Distance Weighted Discrimination (DWD) Batch adjusted measurement, or normalized reads, whichever is available, and for RPPA, we download Z-scores, measured by reverse phase protein array.

2.2.2 Data Processing

We generated a flowchart of data processing for cross cancer types (Figure 2.2). There are three main groups of data. First we have (A) clinical outcomes (overall survival information), and later we will remove samples with overall survival time missing or equal to 0. Then there are feature candidates (B) clinical information, such as age, gender, grade and stage information of patients and (C) genomic measurements, including mDNA, mRNA, miRNA and RPPA. For clinical information, any missing values for the covariates will be imputed by Random Forest missing values imputation algorithm (randomForest R package). For omics data, the preprocessing is more complicated. First we remove duplicated measurement for a single patient, keep only the measurements associated with tumor samples, and discard ones for normal samples. And then for these probes that contain more than 10% missing value, we drop them from genomic data matrix, and impute the left missing observations by a random forest algorithm.

In principle, we can analyze the mDNA with feature size of 450,000, or 27,000, or mRNA with feature size of 20,000 directly. We conduct an un-supervised screening followed by a supervised screening, to narrow down their feature space, since that the number of genes related to cancer survival is not expected to be super large.

For a small number of probes with extremely low variations, the Cox model fitting does not converge. Such genes can either be directly removed (which is adopted in this study) or fitted under a small ridge penalization. This step is the so-called un-supervised screening.

For supervised screening, here we first need to merge omics data with clinical outcomes and extract these with overlapped samples, and then fit a Cox regression model for survival to each probe, select the top 2500 features.

Now all omics profiles, mDNA (2500 probes), mRNA (2500 probes), miRNA (several hundreds), and RPPA (about 200 probes), have relatively small feature sizes. We move on to the modeling step. We employed supervised screening (Univariate Cox regression for prognosis with only omics data; Correlation screen for prognosis with clinical covariates and Omics data), to ensure that the number of features will not exceed the number of events for downstream analysis. The very last step of the analysis is to fit a Cox regression with LASSO L1 penalty to further shrinking the covariates. Prognosis performance is accessed by the C-index.

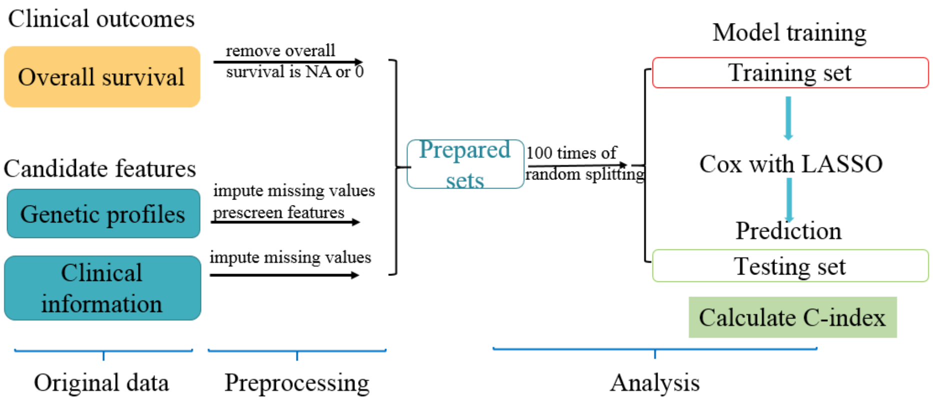


Figure 2.2 An overview of data processing for survival prediction.

In addition to prognosis with overall survival data, we also dichotomized the continuous survival data by a cutoff time and examined the power of omics data in predicting binary survival outcomes, by several machine learning algorithms. Here are the cutoffs we chose for assigning the binary labels for clinical outcomes, (A) for OV, cutoff is 3 years; (B) for KIRC, cutoff is 4 years; (C) for LUSC, cutoff is 2 years; (C) for GBM, cutoff is 1 year; and for (E) for SKCM, cutoff is 3 years. We excluded the samples with censored survival before the cutoff. The binary indicator for survival: 1 stands for samples living longer than the cutoff (strictly greater than) while 0 stands for shorter than the cutoff (less than or equal to).

2.3 Methods

Now we will introduce the methods employed for overall survival prognosis and algorithms for dichotomized survival prediction.

2.2.1 Overall Survival Modeling and Performance Comparison

Cox with LASSO

Survival analysis models the time duration until events occur; in cancer prognosis study, such event is death. The Cox proportional hazards model (Cox), introduced by Cox in 1972 [43], has achieved widespread use in the analysis of time-to-event data with censoring covariates. By saying there are right-censored data, a general case for survival analysis, it means we know the date of event or death is after a certain date observed, for example the study has ended at a certain date therefore we have no information on future events (deaths) of patients who are still alive by then. Covariates in our study can be clinical information, such as gender, age, treatments and multi-dimensional genomic profiles. The proportional hazard assumes that the covariates in the Cox model are multiplicatively related to the hazard function [44]. And remarkably, although the baseline hazard is unspecified, utilizing the method of partial likelihood, the Cox model can be estimated [43].

Let T denote the observed time, and C indicate the time corresponding to an event (if C = 1 the event has occurred and if C = 0 it is censored). Survival function $S(t)$ is defined as $S(t) = P(T > 1) = 1 - P(t)$. Lifetime distribution function is defined as $F(t) = P(T \leq 1) = 1 - S(t)$. The hazard function is defined as $\lambda(\mathrm{t}) = \lim_{dt \to 0} \frac{P(t \leq T < t+dt)}{dt \cdot S(t)} = \frac{f(t)}{S(t)} = \frac{-S(t)'}{S(t)}$. In Cox model:

$$\boldsymbol{\lambda}(\mathbf{t}|\boldsymbol{X}) = \boldsymbol{\lambda_0}(\boldsymbol{t})\boldsymbol{exp}(\boldsymbol{X\beta})$$

The partial likelihood can be constructed as $L(\beta) = \prod_{i:Ci=1} \theta_i \,/ \sum_{j:Y_j \geq Y_i} \theta_j$, where $\theta_j = exp(\beta^T X_j)$ and $X_1, \dots, X_n$ are the covariate vectors for the n individuals. Model parameters estimations are solved by maximizing partial likelihood over β. Denote the log partial likelihood by $l(\beta) = \log L(\beta)$, then estimate β via $\hat{\beta} = argmin\, l(\beta)$.

However, when predicting survival based on genomic data, we could not apply the Cox regression directly due to the issues of high dimensionality. There are several existing methods that have successfully overcome this issue [12-15], by adopting regularizations. We will apply the L1 penalized shrinkage-based Cox regression in our study for cancer prognosis [12], in which we estimate β via

$$\widehat{\boldsymbol{\beta}} = \boldsymbol{argmin\ l(\beta)}, \textbf{subject to} \sum |\boldsymbol{\beta}|_{\boldsymbol{j}} \ \leq$$

where $\boldsymbol{s} > 0$, is a user-specified parameter.

A key component to performance assessment, is to evaluate the prediction accuracy of risk algorithms in the concept of discrimination, which is known as the ability to distinguish whether a subject who will develop an event or not, often referred as the ‘C-statistic’. In the area of modern clinical medicine, it is of great interest for researchers to evaluate the predictive power of proposed biomarkers [29, 30, 31, 32].

For a censored survival outcome, the C-statistic is essentially a rank-correlation measure, calculated by some linear function of the modified Kendall’s tau [36].There are various C-statistics proposed in the literature, which cope with censored survival data [37, 38, 39]. We choose the concordance index (C-index) [40] to assess the predictive power of our clinical covariates or Omics, which is implemented in R package survcomp.

For binary outcome, area under the curve (AUC) is a popular measurement that has been well-studied and quantified [33]. The model with AUC of 0.5 indicates that it is no better than chance at determining the survival outcome of a patient; and a value close to 1.0 indicates that the model perfectly determines the prognosis of a patient. We refer to Li's work [34, 35] and others, for more relevant discussions. Overall accuracy is the alternative way to measure a model's predictive power.

2.2.2 Classification Methods for Dichotomized Survival Analysis

For ovarian cancer with dichotomized survival data, we examined four classification algorithms: Random Forest, K-nearest neighbor, Support Vector Machine and Multiple Kernel Learning (MKL). We picked the MKL as the final method for analyzing all cancer types, for it outperformed the other methods.

The performance was accessed in the same fashion as that of overall survival prediction, randomly splitting prepared data sets into training and testing sets for 100 times, and calculating AUCs for each of algorithms. After identifying MKL as the best choice, for all cancer types, we used MKL to access their performance, and narrow down the splitting time to 30 times per case, taking the computation time into consideration. Later, we will systemically introduce MKL in section 2.5.2.

2.2.3 Identification of miRNA Corresponding DMR mDNA Probes

The human pan-cancer methylation database, MethHC, provides lists of differentially methylated regions (DMR) for many cancer types, in three forms, top 250 hyper-methylated genes, top 250 hypo-methylated genes and top 250 most differentially methylated genes, respectively (http://methhc.mbc.nctu.edu.tw/) [16]. Among these DMR lists, we are particularly interested in examining miRNA corresponding methylation probes, for its robustness might give great survival predictive performance. For 3 out of 5 cancer types in our study, (KIRC, LUSC and SKCM), there are well established DMR lists from the database, which we can employ directly. However, for the other two cancer types (OV and GBM), there is no such lists existing yet, so we adapt its methods for DMR genes, and established our own lists, by t-test for testing the difference between tumor and normal samples, pick the top genes with p-value less than 0.05.

2.2.4 Identification of Hub Probes in miRNA and RPPA by remMap

The regularized multivariate regression for identifying Master predictors (remMap) method has been proposed in 2010, to fit multivariate response regression models for high dimensional data with regularization that deals with the high dimensionality and incorporates network structures [17]. It has been employed to investigate the relationships among different biological molecules based on multiple omics data. In our study, we are interested in identifying the miRNA lists sorted by the number of RPPAs regulated by its corresponding miRNA, and the RPPA lists sorted by the number of miRNAs regulating its corresponding RPPA. We then select the top probes in the lists for modeling, which are the hub regions that we later used in our prognosis for comparison purposes.

2.2.5 Multiple Kernel Learning (MKL)

When applying SVMs, it is often unclear what the best kernel for the each individual task is. In order to figure it out, researchers may want to combine several possible kernels, by giving each kernel a certain weight. One of the ultimate goals of MKL is to learn the kernel in an SVM from training data. Recently, many approaches have been proposed to combine multiple kernels via different ways. Of course, simply adding kernels is that using uniform weights is possible solution, but probably won’t be the optimal. An extreme case can be that giving a kernel positive weight, while it is not correlated with the labels at all, will only add nothing but noise [27]. MKL is a way that can optimize kernel weights while training the SVM, which can leads to good classification accuracies. Plus, MKL can also be useful for identifying relevant and meaningful features [27, 41, 42]. Until now, there are many multiple kernel learning algorithms have been proposed, and research focused on organizing and highlighting the similarities and differences between them, have been done [26]. It has been shown that overall using multiple kernels instead of a single one is useful [26].

MKL allows the scientists to optimize over linear or non-liner combinations of kernels, and also to generalize feature selection to kernel selection, by enforcing sparse confidents. These different kernels can correspond to using information from multiple sources, such as different genetic measurements or feature subsets, which is desirable for the settings of our study.

A key reason of applying MKL to classify dichotomized survival data, is its capability of combining data from different sources with different notions of similarity. Instead of creating a new kernel, MKL can combine established kernels for each individual source. It has been widely applied in many fields, such as object recognition in images [19], event recognition in video [18], and biomedical data fusion [20].

The overall MKL framework is: (i) extract features from all available sources; (ii) construct kernel matrices, based on different features, different kernel types, and different kernel parameters; (iii) find the optimal kernel combination and the kernel classifier.

Linear and Non-Linear Functional Forms

There are different ways of kernel combination and we group the existing MKL algorithms into two categories, based on the functional forms: linear MKL and non-linear MKL. Let's briefly review kernel classifier. Define $\boldsymbol{K}: \boldsymbol{X} \times \boldsymbol{X} \rightarrow \boldsymbol{R}$, called kernel, such that

$$\boldsymbol{\phi}(\boldsymbol{x_i}) \cdot \boldsymbol{\phi}(\boldsymbol{x_j}) = \boldsymbol{K}(\boldsymbol{x_i}, \boldsymbol{x_j}) \tag{1}$$

$\boldsymbol{K}$ is a similarity measure.

Linear MKL can either be unweighted sum or weighted sum. There is an example of linearly parameterized combination function:

$$\boldsymbol{K}_{\mu} = \sum_{\boldsymbol{m}=\boldsymbol{1}}^{\boldsymbol{p}} \boldsymbol{\mu_m} \boldsymbol{K_m} (\boldsymbol{x_i}, \boldsymbol{x_j}) \tag{2}$$

Hence $\boldsymbol{K}_{\mu}$ is a new kernel, a linear combination of a set of $\boldsymbol{p}$ kernels, where $\boldsymbol{\mu_m}$ is a vector of coefficients for each kernel. This new function $\boldsymbol{K}_{\mu}$ is still a kernel, since kernels are additive, according to the properties of reproducing kernel Hilbert spaces [21].

Although linear MKL is quite popular, but it may also be restrictive. Nonlinear MKL can possibly use nonlinear functions of kernels, such as, multiplication, power, and exponentiation, polynomial combination, kernel ridge regression, and so on [22,23,24,25]. An example of kernel ridge regression and polynomial combination, can be formulates as [22]

$$\boldsymbol{K}_{\mu} = \sum_{\boldsymbol{0} \leq \boldsymbol{k_1} + \cdots + \boldsymbol{k_p} \leq \boldsymbol{d},\ \ \boldsymbol{k_p} \geq \boldsymbol{0},\ \ \boldsymbol{i} \in [\boldsymbol{0}, \boldsymbol{p}]} \boldsymbol{\mu_{k_1 \cdots k_p}} \boldsymbol{K_1^{k_1}} \cdots \boldsymbol{K_p^{k_p}} \tag{3}$$

where $\boldsymbol{\mu_{k_1 \cdots k_p}} \geq \boldsymbol{0}$. However, the number of coefficients $\boldsymbol{\mu_{k_1 \cdots k_p}}$ is in $\boldsymbol{O}(\boldsymbol{p^d})$, which may be too large to learn. To reduce the learning complexity, the combined kernel can be simplified:

$$\boldsymbol{K}_{\mu} = \sum_{\boldsymbol{k_1}+\cdots+\boldsymbol{k_p}=\boldsymbol{d}} \boldsymbol{\mu_1^{k_1}} \cdots \boldsymbol{\mu_p^{k_p}} \boldsymbol{K_1^{k_1}} \cdots \boldsymbol{K_p^{k_p}} \tag{4}$$

where $\boldsymbol{\mu} = (\boldsymbol{\mu_1} \cdots \boldsymbol{\mu_p})^{\boldsymbol{T}} \in \boldsymbol{R^p}$.

Major Learning Algorithms for MKL

In this section, we will introduce the following categorization proposed by Gonen and Alpaydın, where they categorize the existing MKL algorithms into 5 major groups [26]. These five major categories are: A) fixed rules, which are functions without any parameters without any training; B) heuristic approaches that use a parameterized combination function and based on some measure from each kernel to learn the parameters; C) optimization approaches, which also use a parametrized combination function and by solving an optimization problem to learn the parameters; D) bayesian approaches that treat the kernel combination parameters as random variables and assume priors on these parameters, and perform inference to learn both kernel combination parameters and base kernel parameters; and E) boosting approaches, which will iteratively add a new kernel until obtain the a cutoff of performance improvement. Now we go over some examples of fixed rules briefly.

Multiple kernels can be written as a combination function of existing kernels:

$$\boldsymbol{K}_{\mu} = \boldsymbol{f}_{\mu}(\boldsymbol{K_1}, \dots, \boldsymbol{K_p}) \tag{5}$$

where $\boldsymbol{f}_{\mu}\colon \boldsymbol{R^p} \to \boldsymbol{R}$, as discussed in the functional forms section, can be linear or nonlinear.

Fixed rules obtain $\boldsymbol{K}_{\mu}$ using function $\boldsymbol{f}_{\mu}$. Since each kernel itself is positive semidefinite, by definition, $\boldsymbol{v^T K_i v} \geq \boldsymbol{0}$, for all $\boldsymbol{v} \in \boldsymbol{R^N}$, $\boldsymbol{N}$ is the sample size. Easily, we can derive that both

$\boldsymbol{K}_{\mu}$ of summation and multiplication are also positive semidefinite.

$$\begin{cases} \boldsymbol{v}^{\boldsymbol{T}}\boldsymbol{K}_{\boldsymbol{\mu}}\boldsymbol{v} \ = \boldsymbol{v}^{\boldsymbol{T}}\boldsymbol{K}_{\boldsymbol{1}}\boldsymbol{v} + \boldsymbol{v}^{\boldsymbol{T}}\boldsymbol{K}_{\boldsymbol{2}}\boldsymbol{v} \geq \boldsymbol{0} \\ \boldsymbol{v}^{\boldsymbol{T}}\boldsymbol{K}_{\boldsymbol{\mu}}\boldsymbol{v} \ = \boldsymbol{v}^{\boldsymbol{T}}\boldsymbol{K}_{\boldsymbol{1}}\boldsymbol{K}_{\boldsymbol{2}}\boldsymbol{v} \geq \boldsymbol{0} \end{cases} \tag{6}$$

In the same fashion, we can derive that both summation and multiplication of $\boldsymbol{p}$ kernels are also valid kernels

$$\begin{cases} \boldsymbol{K}_{\mu}(\boldsymbol{x_i}, \boldsymbol{x_j}) = \sum_{\boldsymbol{m=1}}^{\boldsymbol{p}} \boldsymbol{K_m}\,(\boldsymbol{x_i^m}, \boldsymbol{x_j^m}) \\ \boldsymbol{K}_{\mu}(\boldsymbol{x_i}, \boldsymbol{x_j}) = \prod_{\boldsymbol{m=1}}^{\boldsymbol{p}} \boldsymbol{K_m}(\boldsymbol{x_i^m}, \boldsymbol{x_j^m}) \end{cases} \tag{71}$$

2.4 Results

2.4.1 Prediction of Overall Survivals from Clinical Variables and Omics data

We examined five cancer types: OV, KIRC, LUSC, GBM, and SKCM, for their TCGA data sets included survival data with long enough follow-up time and big enough sample sizes of multiple genetic profiles, such as mDNA, mRNA, miRNA and RPPA. Table 2.2 summarized all TCGA samples for our study, number in each cell is the sample size for its associated cancer and feature candidate.

Table 2.2 Summary of TCGA Samples

Cancer	**Survival**	**Clinical**	**mDNA**	**mRNA**	**miRNA**	**RPPA**
OV	576	576	592	581	453	412
KIRC	526	526	319	533	516	454
LUSC	473	473	370	533	478	195
GBM	591	591	420	511	575	214
SKCM	428	428	470	468	448	204

In the following several sub-section, we will present survival prediction comparison of clinical covariates, genetic data and their combinations, plus several feature selections, crossing these five cancer types listed here. For each cancer, we chose clinical information only (patient age, patient gender, tumor stage and grade, and Karnofsky performance score, varies cross cancer types, upon its availability) Cox models' performance as baseline for comparison and focused on the potential marginal gains of Cox models incorporating genetic data (mDNA, mRNA, miRNA, and RPPA), upon existing clinical variables. We modified the multivariate Cox method to fit both clinical and molecular features, in two ways. (A).We performed a feature-selection step against the residuals of the clinical-variable-only models, and combine the selected features that have better goodness of fit, to build a new multivariate integrative Cox model. (B) Or, when a relatively small size of subset of features were selected by other feature selection methods, such as univariate Cox regression, or remMap selection, we chose to throw both clinical variables and subset of genetic features into Cox model, simultaneously, and directly allow LASSO to pick its features. We used the R package "survival" to build Cox models and the R package "glmnet" to perform LASSO with penalty parameter λ chosen by fivefold cross-validation within the training data set. We repeated each procedure 100 times to generate 100 C-indexes and compared their predictive power by its coordinating median C-index.

Overall survival prediction of OV

First, we generated prognosis models by clinical variables, and each type of genetic data alone, and then integrated clinical variables with one type of genetic data at time as feature combined models. Figure 2.3 shows our preliminary results for OV case. As we can see, there is no case for which models without clinical variables can outperform the model with only the clinical variables.

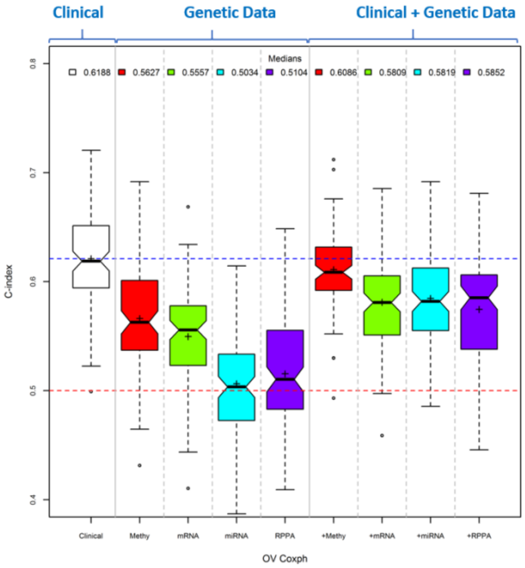


Figure 2.3 Preliminary comparison of survival prediction of OV

In order to obtain better prognosis models utilizing genetic data, we further explored models with the top miRNA corresponding DMR mDNA probes. Since there is no well-established DMR gene lists available from MethHC database, we performed t-test to select ideal probes. By t-test, we selected 18 mDNA probes of miRNA genes. Here comes the performance of miRNA corresponding DMR mDNA probes (denoted as methyG in plot) Figure 2.4(a). We can see the predictive power improvements from methyG, compared to that from mDNA profile. So we can now update our comparison result with adding predictive performance from methyG, as shown in Figure 2.4(b).

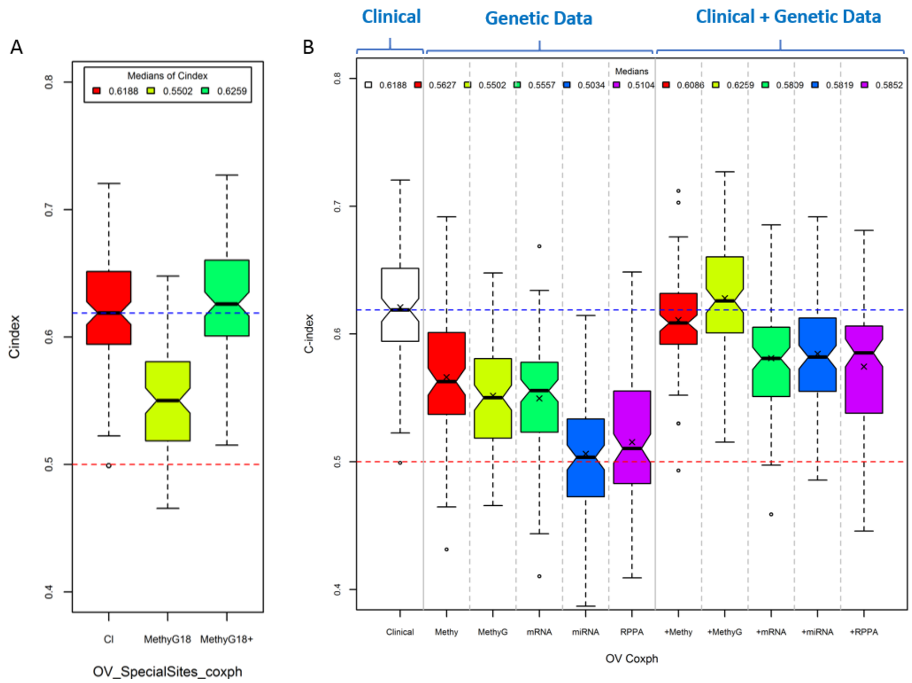


Figure 2.4 Updated comparison of survival prediction of OV. (a) Comparison of perform of clinical variables and methyG, (b) Secondary comparison, after taking methyG into consideration.

We then updated miRNA and RPPA in our prognosis models with remMap selected miRNA and RPPA. We find out that for miRNA, models with top30, top20,top10, top5 probes in the hub range, are all better than clinical only, while top30<top20<top10<top5. And the predictive power is also improved for RPPA, but still can’t outperform the model of clinical information only. It does not hurt to try other top probes selection methods and examine its effects. So we also used simple univariate Cox selection approach to identify the top most influential probes and its performance is quite similar to that from remMap selected probes.

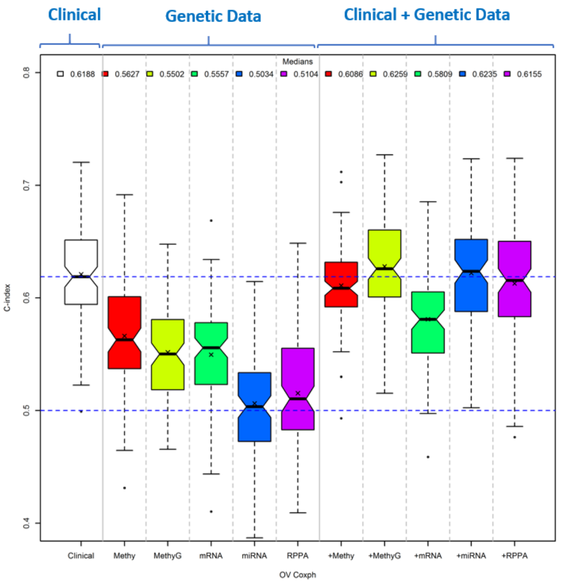


Figure 2.5 Final comparison of survival prediction of OV

We summarize our final comparison for OV in Figure 2.5. We find out the for OV cancer, clinical variables alone can give a good prognosis power, with C-index of 0.6188. And also 5 out of 5 cases, the models built from genetic data sets alone showed significant predictive power (range: 0.5502-0.5627). However, incorporating genetic information can't add a significant marginal gains, even with multiple feature selection techniques applied. It does show that miRNA corresponding DMR mDNA probes, and remMap selection and univariate Cox selection, are promising approaches for achieving better prognostic power. It is worth exploring for other cancer types.

Overall survival prediction of KIRC

Similar as we examined OV cancer, we first built prognosis models by clinical variables, and each type of genetic data alone, and then built our integrative models. Figure 2.6 shows a comparison of survival predictive performance for KIRC. Overall, prognosis power for KIRC are better than patients with OV. As we can see from it, the clinical-variable-only model, showed a pretty good predictive power, with C-index of 0.754, which is significantly higher than 0.5. And also 5 out of 5 of our models built from genetic data sets alone showed significant predictive power (range: 0.6518-0.7086). And 5 out of 5 of our integrative models give us better prediction than the clinical-variable-only model, with the highest predictive power of 0.7922 from methylation profile and the second highest predictive power of 0.7919 from miRNA corresponding top miRNA corresponding DMR mDNA probes (methyG), both are marginally gained of 5%.

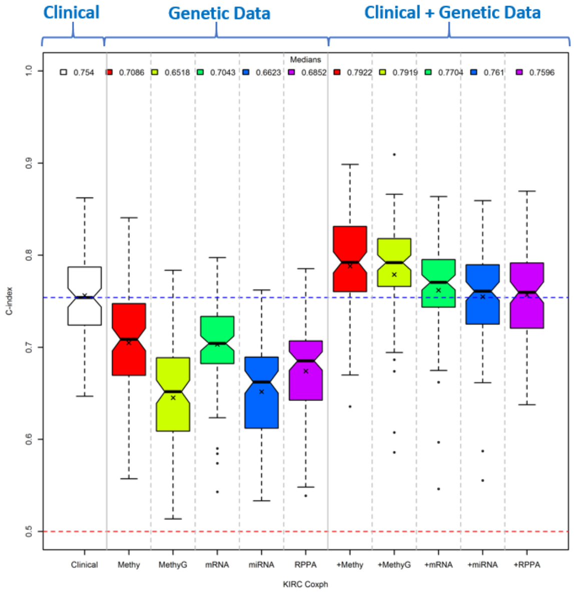


Figure 2.6 Preliminary comparison of survival prediction of KIRC

Since among all integrative Cox models, that from miRNA and RPPA, were not as good as others, in terms of marginal gains, so it is worth to try feature selections for miRNA and RPPA. We then updated our preliminary comparison with model performance from univariate

Cox selected miRNA and RPPA, and summarized in Figure 2.7. Updated model performance showed that models from both miRNA only and miRNA incorporating clinical information, are better than before. And also for RPPA: models from both miRNA only and miRNA incorporating clinical information, are better than before.

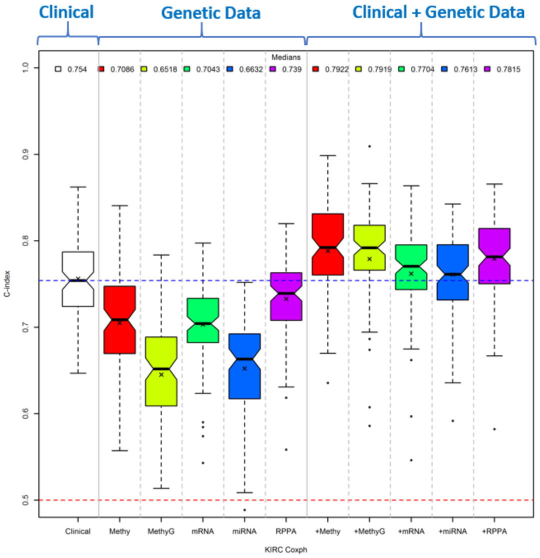


Figure 2.7 Secondary comparison of survival prediction of KIRC

We checked how good the prognosis can be from remMap identified miRNA and RPPA probes, with results presented in Figure 2.8. None of them outperformed the above models with univariate Cox selected miRNA and RPPA probes. By examining at the remMap identified lists in details, we realized that for KIRC, it is difficult to outperform whole RPPA, by top remMap identified probes, except the case of top 130 out of 155 RPPA, which employed the number of RPPAs regulated by the corresponding miRNA= 20, as a cutoff. Comparing to the list for OV, the top RPPAs with the largest number of RPPAs regulated by the corresponding miRNA, is only 16, this may indicate that for KIRC, remMap can't serve as a selection tool for narrow down the number of RPPA probes, since all of them are claimed to be important. So we realized that our previous univariate Cox selection based on the whole RPPA profile, is equivalent to that we add a univariate Cox selection, after remMap identification, which serves as a tool to future narrow down the number of probes by only keep the high related ones. It perfectly explains why the predictive powers are better than remMap only. Similarly idea explains the results for model performance from miRNA

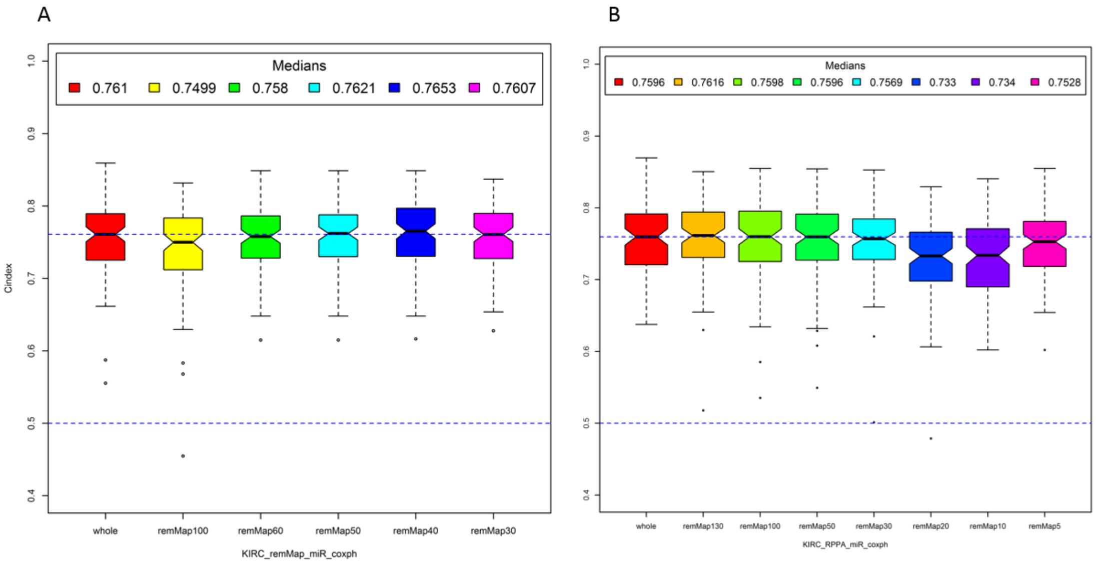


Figure 2.8 Exploring predictive power of models from remMap selected miRNA and RPPA for KIRC. (A) Predictive powers from top 100, top60, top 50, top 40 and top30 miRNA probes identified by remMap, sorted by number of RPPAs regulated by the corresponding miRNA, and top 40 give the best and the only result that is slightly better than that from the whole miRNA profiles. (B) Predictive powers from top 130, top100, top 50, top 30, top20, top10 and top5 RPPA probes identified by remMap, sorted by number of miRNAs regulating the corresponding RPPA, and top 130 give the best and the only result that is slightly better than that from the whole miRNA profiles.

Overall survival prediction of LUSC

Similar as in our examination of the OV and KIRC cancers, we first built prognosis models by clinical variables, or genetic data individually, and then the integrated variables. Figure 2.9 shows a comparison of survival predictive performance for LUSC. As we can see, the clinical-variable-only model, showed a significant predictive power, with C-index of 0.5506. And also 5 out of 5 cases, the models built from genetic data sets alone showed significant predictive power (range: 0.545-0.62). And 5 out of 5 of our integrative models give us better prediction than the clinical-variable-only model, with the highest predictive power of 0.61415 from mRNA profile and the second highest predictive power of 0.6121 from RPPA, both are marginally gained of 11.5% and 11.2%, respectively.

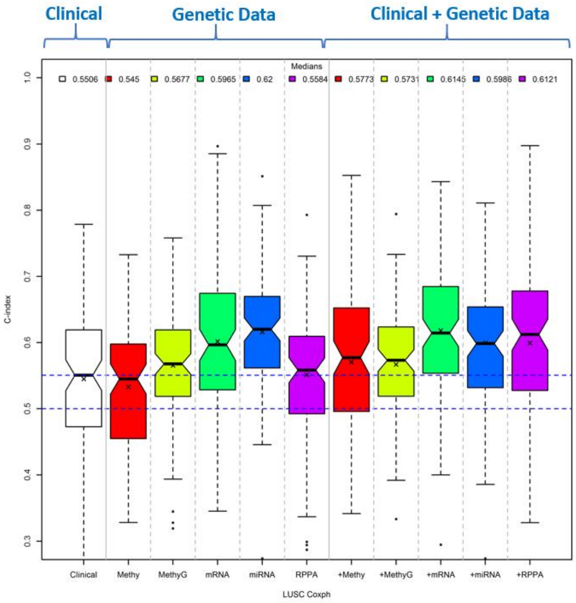


Figure 2.9 Preliminary comparison of survival prediction of LUSC

We can of course also update our preliminary comparison with model performance from univariate Cox selected miRNA and RPPA, and further investigate with remMap identified miRNA and RPPA lists.

GBM overall survival prediction

Similar as we examined OV, KIRC and LUSC cancer, we first built prognosis models by clinical variables, genetic data individually, and then built our integrative models. Figure 2.10 shows a comparison of survival predictive performance for GBM. As we can see from it, the clinical-variable-only model, showed a significant predictive power, with C-index of 0.6636. Also 5 out of 5 cases, the models built from genetic data sets alone showed significant predictive power (range: 0.5285-0.6012). And 3 out of 5 of our integrative models provide better prediction than the clinical-variable-only model, with the highest predictive power 0.6777 from miRNA corresponding DMR mDNA probes.

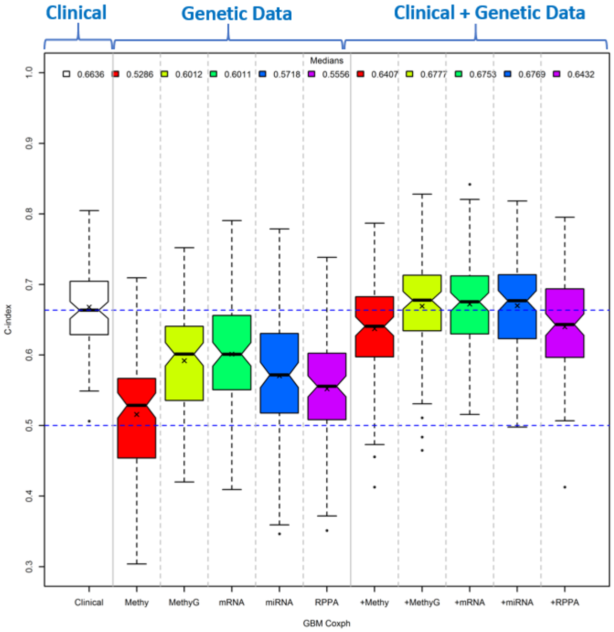


Figure 2.10 Preliminary comparison of survival prediction of GBM

We can also further examine and update our preliminary comparison with predictions from univariate Cox selected miRNA and RPPA probes, and models with remMap identified miRNA and RPPA lists.

Overall survival prediction of SKCM

Similar as we examined OV, KIRC, LUSC, and GBM cancer, we first built prognosis models from clinical variables, and each type of genetic data individually, and then built our integrative models. Figure 2.11 shows a comparison of survival predictive performance for GBM. As we can see from it, the clinical-variable-only model, showed a significant predictive power, with C-index of 0.6279. Also 5 out of 5 our models built from genetic data sets alone showed significant predictive power (range: 0.5739-0.6155). And 4 out of 5 of our integrative models give us better predictions than the clinical-variable-only model, with the highest predictive power 0.6716 from methylation profile. Among 5 of the integrative prognosis models, model from clinical information incorporating with entire RPPA, couldn't adding any marginal gains in predictive power.

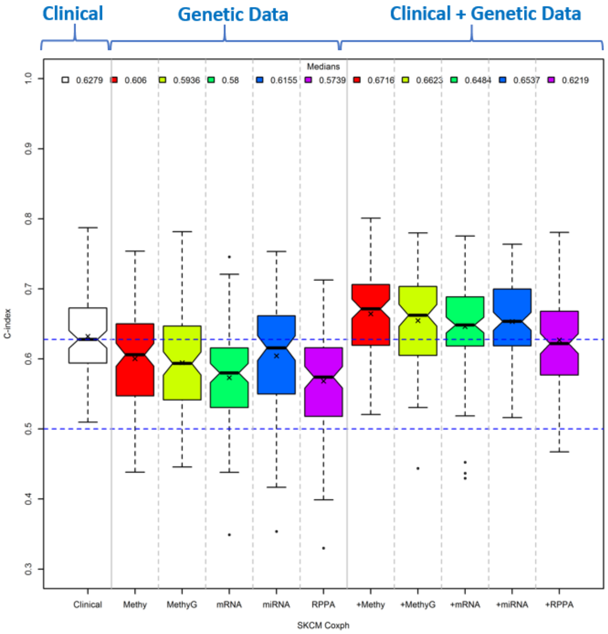


Figure 2.11 Preliminary comparison of survival prediction of SKCM

Later, we can also update our preliminary comparison by examining model performance from univariate Cox selected miRNA and RPPA probes, and further investigating models with remMap identified miRNA and RPPA lists.

Comparison of Cox model and kernel Cox regression

We also explored another computational method Kernel Cox regression to compare their performance with Cox models, from clinical variable only, cross four cancer types. Kernel Cox models were built using the R package “survpack”. It shows quite consistent performance between Cox models and Kernel Cox regression for clinical variables only models in Figure 2.12.

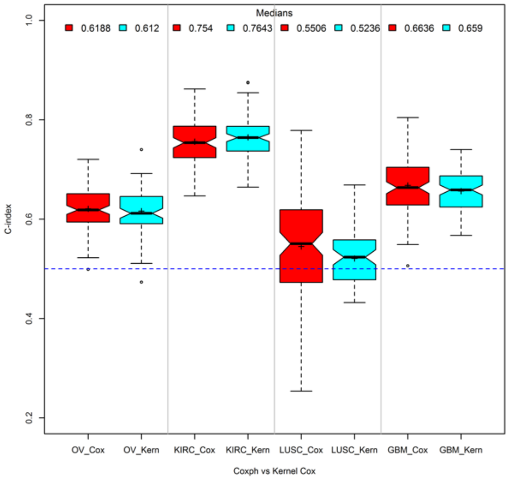


Figure 2.12 Predictive Power Comparison of Cox Model and kernel Cox Regression Cross Cancer types.

Summary for overall survival predictions

In this project, we extensively evaluated the predictive powers of our single sourced models and integrative models were via different ways of comparison. We summarize our all the models and results in Table 2.3, for OV, KIRC, LUSC, GBM and SKCM cancer types. Out of 25 integrative of multivariate Cox models, 20 out of 25 of them outperformed their baseline models from clinical variables only (color labeled in green). We conclude that among five cancer types, KIRC is the best case, with respect to the predictive power by both the baseline model from clinical information only (median C-index = 0.754) and the integrative model from clinical information incorporated with genetic measurements (median C-index range: 0.7610 ~ 0.7922), with its highest marginal gain of ~5%. While for LUSC, by Cox model, we discovered the highest marginal gain cross five cancer types, which is ~11.6%.

Table 2.3 Summary of predictive performance of Cox models for overall survival time cross cancer types. Each cell in the table is the median C-index of 100 repeated training and testing for its corresponding model and cancer type. Green color labeled cells are the cases that integrative

models outperformed its baseline models with only clinical variables.

Models / Cancers	~ Clinical Variables	~ Genetic Measurements					~ Clinical Variables + Genetic Measurements				
		mDNA	DMR	mRNA	miNRA	RPPA	Clinical + mDNA	Clinical + DMR	Clinical + mRNA	Clinical + miNRA	Clinical + RPPA
OV	0.6188	0.5627	0.5502	0.5557	0.5034	0.5104	0.6086	0.6259	0.5809	0.6235	0.6155
KIRC	0.754	0.7086	0.6518	0.7043	0.6632	0.739	0.7922	0.7919	0.7704	0.7613	0.7815
LUSC	0.5506	0.545	0.5677	0.5965	0.62	0.5584	0.5773	0.5731	0.6145	0.5986	0.6121
GBM	0.6636	0.5286	0.6012	0.6011	0.5718	0.5556	0.6407	0.6777	0.6753	0.6769	0.6432
SKCM	0.6279	0.606	0.5936	0.58	0.6155	0.5739	0.6716	0.6623	0.6484	0.6537	0.6219

Feature selections for miRNA and RPPA to further improve predictive power, have been extensively explored by both univariate Cox regression and remMap identification. As we identified that Cox models from corresponding DMR mDNA probes, are better than that from mDNA profile, which indicates that a few mDNA probes can be super robust in predicting overall survival time, while incorporating a large size of probes may bring more noise than useful information. We suspected that it could be the similar cases for other genetic profiles too. Let's review the predictive powers from genetic profiles and its integrative cases of OV, in Figure 2.3, before exploring feature selection for miRNA and RPPA data. It shows that the performance of Cox model from miRNA, is no better than a pure guess, with median C-index of 0.5034, and that from RPPA is slightly better than a pure guess, with median C-index of 0.5104, meaning that it does contain a small amount of using information. However, in there corresponding integrative models with clinical variables, the predictive powers dropped dramatically, comparing to the baseline, with median C-index of 0.5819 and 0.5852, respectively. That indicates adding these two types of genetic data on top of clinical variables, only brings noise, which is contradict to their individual performance. So we conducted feature selection to narrow down the size of probes adding to the models and evaluate their performance again, by trying both univariate Cox regression and remMap identification. Luckily, both improved model performance, by a larger median C-index, close to, or higher than its baseline, meaning that feature selections are essential to our integrative models, even for relatively small feature sized miRNA and RPPA, comparing to mDNA and miRNA. Since both univariate Cox regression and remMap identification improved the predictive performance to a quite similar level, it is hard to conclude which feature selection method is better, in terms of performance gained. But, biologically, remMap identified probes are more meaningful.

For the tumor type of KIRC, we also explored these two types of feature selection technique, trying to further improve its predictive powers from integrative models, which were already very well (median C-index range: 0.7596 ~ 0.7922, refer to Figure 2.5). Because among all integrative Cox models, that from miRNA and RPPA, were not as good as others, in terms of marginal gains, so it is worth to try feature selections for miRNA and RPPA. After univariate Cox regression for selection top important probes of miRNA and RPPA, the predictive power of their corresponding models (both genetic only model and integrative model) were all improved, as in Figure 2.7. In order to get a biologically more meaningful selected feature and also to compare two feature selection techniques, we then checked how good the prognosis it can be from remMap identified miRNA and RPPA probes. It is interesting to find out that remMap identified all 155 RPPA probes with large number of RPPAs regulated by the corresponding miRNA (45 out of 155 with number greater than 30; 134 out of 155 with number greater than 20; all probes with number great than 10), while for OV, there are only 26 out 152 RPPA probes identified with numbers of RPPAs regulated by the corresponding miRNA, greater than 10 and no probe with numbers greater than 20. It may indicate that since all are biologically important probes, remMap for KIRC, did not truly serve as a feature selection, which explained the possible reason why it did not outperform models with univariate Cox selected RPPA probes, (Figure 2.8).

So our univariate Cox selection based on the whole RPPA profile, is equivalent to that we add an additional univariate Cox selection step, after remMap identification, which serves as a tool to future narrow down the number of probes by only keep the high related ones. Similarly idea explains the results for model performance from miRNA too.

We can conclude that feature selections for miRNA and RPPA with relatively smaller feature space, can certainly further improve survival predictive performance. We identified that top important probes of mDNA, miRNA and RPPA were super robust, comparing to whole profile.

2.4.2 Prediction of Dichotomized Survival Data

Adapting our workflow for overall survival prediction, we modified it for dichotomized survival data, as shown in Figure 2.13. We first examined three popular classification algorithms: Random Forest, K-nearest neighbor and Support Vector Machine (by R packages, “randomForest”, “class”, and “kernlab”, respectively) for dichotomized survival data of ovarian cancer. A preliminary comparison was done for clinical variables, RPPA data, and their combination (Figure 2.14). The performance evaluation was accessed in the same fashion as that of overall survival prediction, randomly splitting prepared data sets into training and testing sets for 100 times, and calculating AUCs for each of algorithms. It is hard to conclude which method was the best, since three were quite consistent, except that SVM was slightly better than the other two.

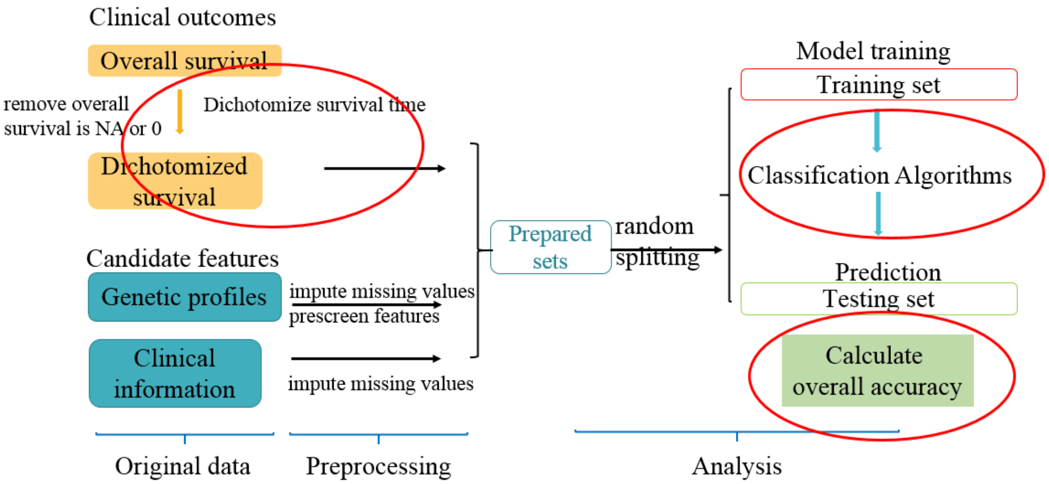


Figure 2.13 An overview of data processing for dichotomized survival prediction

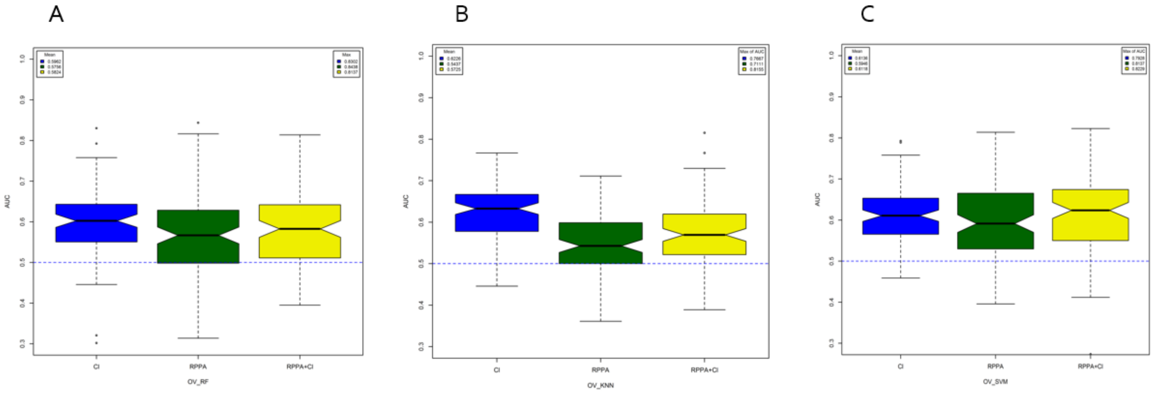


Figure 2.14 Classification Algorithm Comparison for Dichotomized Survival Prediction. (A) RF, (B) KNN and (C) SVM

As discussed before, Multiple Kernel Learning (MKL) has substantial advantages, comparing to Support Vector Machines, when dealing with data from different sources. We decided to employ MKL for classification of dichotomized survival data prediction and access our model performance via overall accuracy. For this type of prediction, we narrowed down the splitting time to 30 per model per cancer type, considering the cost of computational time.

Prediction of Dichotomized Survival Data by SimpleMKL

We build the MKL models via SimpleMKL, using the matlab package “SimpleMKL toolbox” [45]. We conducted feature selection for genetic data to choose about 20 most important variables for each data set, by applying random forest variable importance list, using R package “randomForest”. Predictive performance crossing cancer types has been summarized in the following five figure, Figure 2.15~ Figure 2.19, for OV, KIRC, LUSC, GBM, and SKCM, respectively. It showed a great performance from all integrative models of OV, significantly higher than the baseline, in terms of median overall accuracy of 30 times repeats, with the highest power from (Clinical + RPPA + miRNA), and (Clinical + RPPA), marginal gains of ~10.5% and 6.8%. And for KIRC, we also identified that all integrative models are significantly better than clinical-variable-only model, with the best performance with marginal gains of ~ 11%, from (Clinical + RRPA) and (Clinical + RPPA + DMR) models. While for LUSC, there was only one integrative model from (Clinical + RPPA + DMR) outperformed its baseline, with a marginal gain of ~ 9%. The predictive powers of integrative models for GBM and SKCM, were not promising.

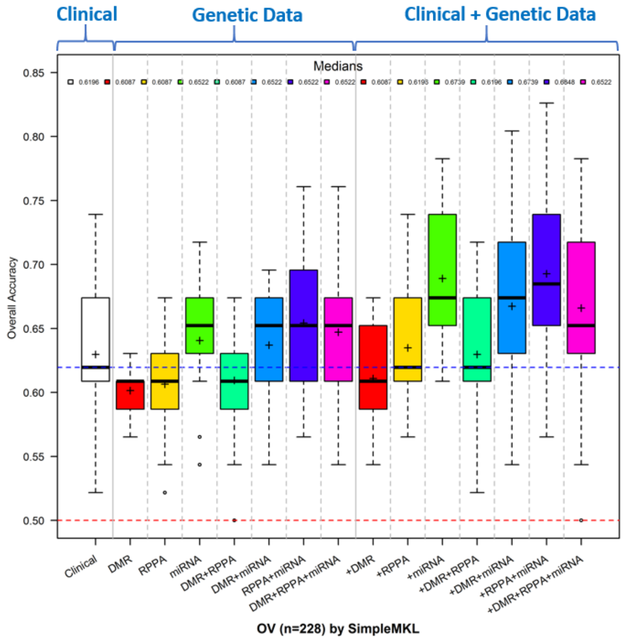


Figure 2.15 Predictive performance comparison of dichotomized survival data for OV

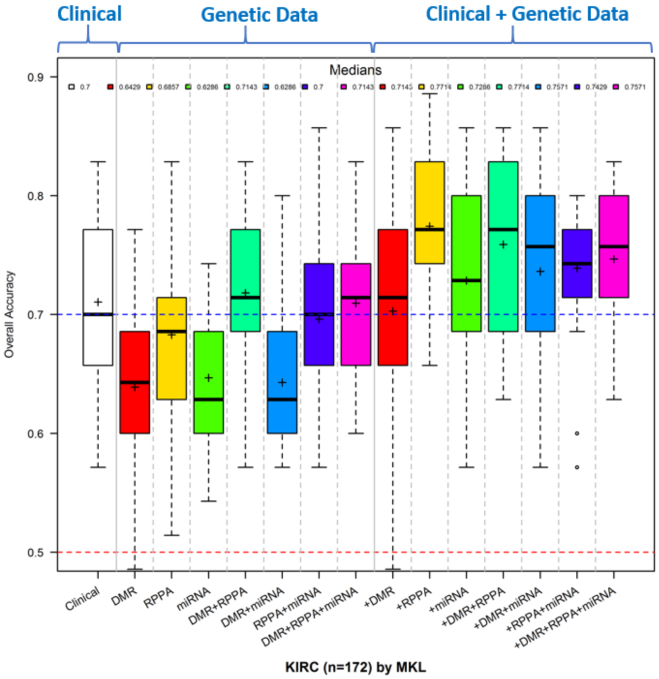


Figure 2.16 Predictive performance comparison of dichotomized survival data for KIRC

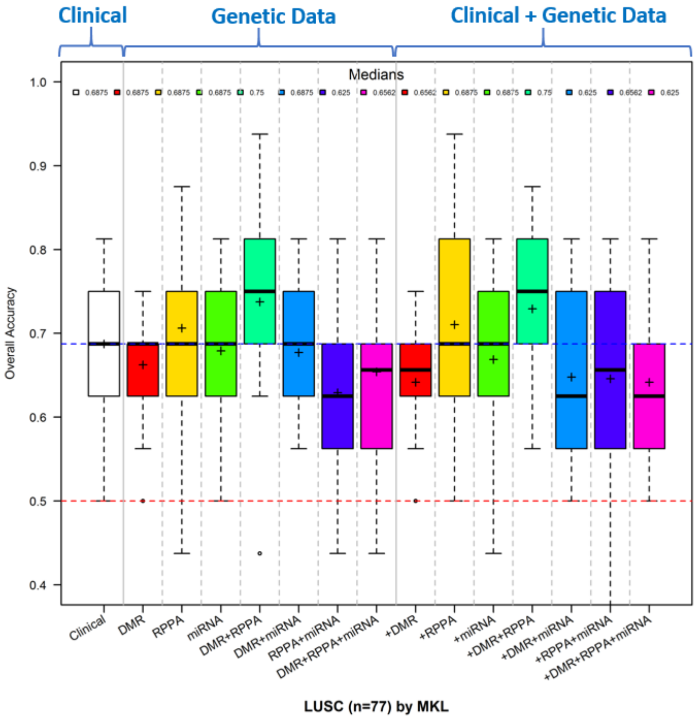


Figure 2.17 Predictive performance comparison of dichotomized survival data for LUSC

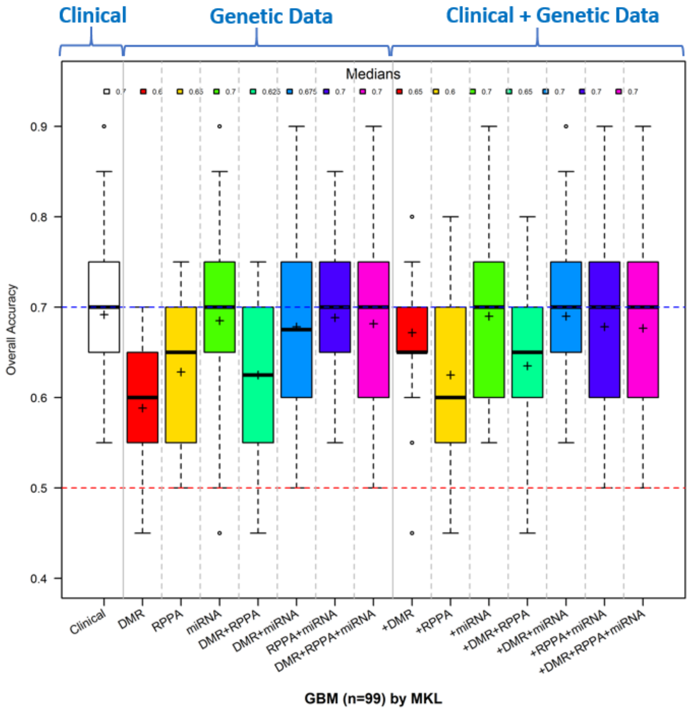


Figure 2.18 Predictive performance comparison of dichotomized survival data for GBM

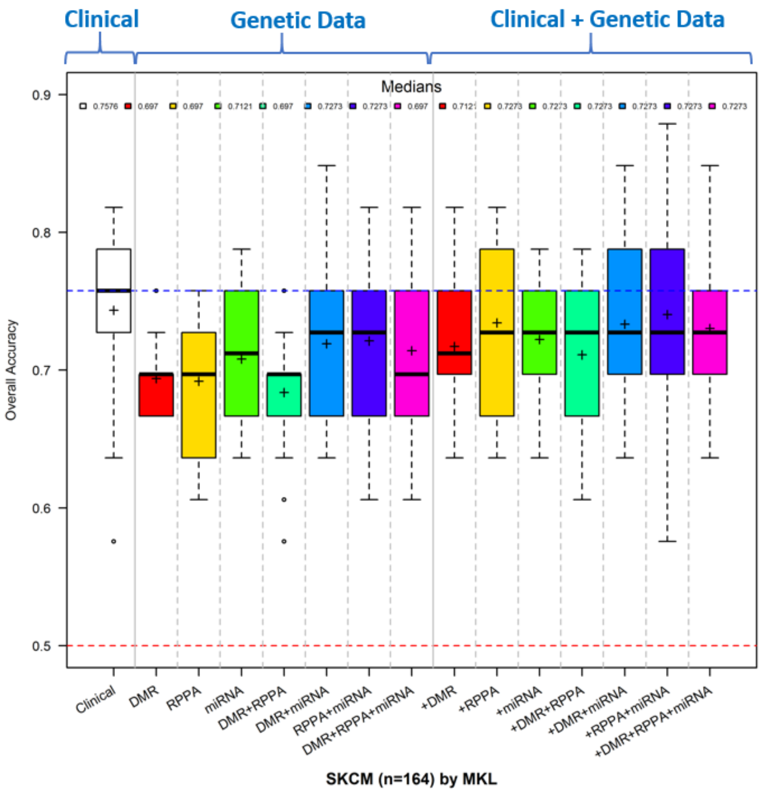


Figure 2.19 Predictive performance comparison of dichotomized survival data for SKCM

Summary for dichotomized survival predictions

SimpleMKL algorithm worked very well as a powerful classification approach to predictive dichotomized survival data, for some cancer types, such as OV, KIRC and LUSC, but might not be a good choice for prediction of GBM and SKCK. Since we only examined a MKL algorithm, which is SimpleMKL, it is also worth exploring other MKL algorithms in the future.

2.5 Discussion

We employed several different methods for overall survival prognosis and algorithms for dichotomized survival prediction, knowing some of them have never been applied in this way before. Our findings are quite interesting and promising for future investigations in this field and further benefit clinical decision makings. Except multivariate Cox regressions for integrative analysis of overall survival time, Kernel Cox also appears very promising for survival predictions. When dealing with clinical variables only, different kernels, such as linear, Gaussian, and especially clinical kernels are worth exploring, with focus in improving its predictive power. For dichotomized survival time prediction, as a classification problem, there are already plenty of existing and well established algorithms available. However, most of them can not handle features from very different representations, except the MKLs that are powerful enough to train models and to choose the best kernels or kernel combinations for each data resource, and hence providing a more accurate predictive performance. We explored the SimpleMKL algorithm in our study, however there are still quite a lot of other MKL algorithms worth trying in the future, such as SpicyMKL, OnlineMKL and so on, which might further improve the prediction performance.

2.6 Reference

1. Cancer Genome Atlas Research Network, Weinstein JN, Collisson EA, Mills GB, Shaw KR, Ozenberger BA, et al. The Cancer Genome Atlas Pan-Cancer analysis project. Nat Genet 2013;45:1113–20.
2. NIH Launches Cancer Genome Project Washington Post Dec 14, 2005
3. Bussey, Kimberly J., et al. "Integrating data on DNA copy number with gene expression levels and drug sensitivities in the NCI-60 cell line panel." Molecular cancer therapeutics 5.4 (2006): 853-867.
4. Andrews, Joseph, et al. "Multi-platform whole-genome microarray analyses refine the epigenetic signature of breast cancer metastasis with gene expression and copy number." PLoS One 5.1 (2010): e8665.
5. Soneson, Charlotte, et al. "Integrative analysis of gene expression and copy number alterations using canonical correlation analysis." Bmc Bioinformatics 11.1 (2010): 191.
6. Xu C, Liu Y, Wang P, et al. Integrative analysis of DNA copy number and gene expression in metastatic oral squamous cell carcinoma identifies genes associated with poor survival. Mol Cancer 2010;9:142.
7. Lu TP, Lai LC, Tsai MH, et al. Integrated analyses of copy number variations and gene expression in lung adenocarcinoma. PLoS One 2011;6(9):e24829
8. van Wieringen WN, Unger K, Leday GG, et al. Matching of array CGH and gene expression microarray features for the purpose of integrative genomic analyses. BMCBioinform 2012;13:80.

9. de Cubas AA, Leandro-Garcia LJ, Schiavi F, etal. Integrative analysis of miRNA and mRNA expression profiles in pheochromocytoma and paraganglioma identifies genotypespecific markers and potentially regulated pathways. Endocr Relat Cancer 2013;20(4):477–92.

10. van Iterson M, Bervoets S, de Meijer EJ, et al. Integrated analysis of microRNA and mRNA expression: adding biological significance to microRNA target predictions. Nucleic Acids Res 2013;41(15):e146.

11. Yuan, Yuan, et al. "Assessing the clinical utility of cancer genomic and proteomic data across tumor types." *Nature biotechnology* 32.7 (2014): 644-652.

12. Tibshirani, Robert. "The lasso method for variable selection in the Cox model." Statistics in medicine 16.4 (1997): 385-395.

13. Tibshirani, Robert J. "Univariate shrinkage in the Cox model for high dimensional data." Statistical applications in genetics and molecular biology 8.1 (2009): 1-18.

14. Witten, Daniela M., and Robert Tibshirani. "Survival analysis with high-dimensional covariates." Statistical methods in medical research 19.1 (2010): 29-51.

15. Bradic, Jelena, Jianqing Fan, and Jiancheng Jiang. "Regularization for Cox’s proportional hazards model with NP-dimensionality." Annals of statistics 39.6 (2011): 3092.

16. Huang, Wei-Yun et al. “MethHC: A Database of DNA Methylation and Gene Expression in Human Cancer.” Nucleic Acids Research 42.Database issue (2015): D856–D861. PMC. Web. 4 May 2015.

17. Peng, Jie, et al. "Regularized multivariate regression for identifying master predictors with application to integrative genomics study of breast cancer." The annals of applied statistics 4.1 (2010): 52.

18. Lin Chen, Lixin Duan, and Dong Xu, "Event Recognition in Videos by Learning From Heterogeneous Web Sources," in IEEE International Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 2666-2673
19. Serhat S. Bucak, Rong Jin, and Anil K. Jain, Multiple Kernel Learning for Visual Object Recognition: A Review. T-PAMI, 2012.
20. Yu et al. L2-norm multiple kernel learning and its application to biomedical data fusion. BMC Bioinformatics 2010, 11:309
21. Aronszajn, Nachman (1950). "Theory of Reproducing Kernels". Transactions of the American Mathematical Society 68 (3): 337–404
22. Cortes, Corinna, Mehryar Mohri, and Afshin Rostamizadeh. "Learning non-linear combinations of kernels." Advances in neural information processing systems. 2009.
23. Gönen, Mehmet, and Ethem Alpaydin. "Localized multiple kernel learning." Proceedings of the 25th international conference on Machine learning. ACM, 2008.
24. Lewis, Darrin P., Tony Jebara, and William Stafford Noble. "Nonstationary kernel combination." Proceedings of the 23rd international conference on Machine learning. ACM, 2006.
25. Varma, Manik, and Bodla Rakesh Babu. "More generality in efficient multiple kernel learning." Proceedings of the 26th Annual International Conference on Machine Learning. ACM, 2009.
26. Mehmet Gönen, Ethem Alpaydın. Multiple Kernel Learning Algorithms Jour. Mach. Learn. Res. 12(Jul):2211−2268, 2011
27. G. Lanckriet, T. De Bie, N. Cristianini, M. I. Jordan, and W. Stafford Noble. A statistical framework for genomic data fusion. Bioinfomatics, 20(16):2626–2635, 2004.

28. Gert R. G. Lanckriet, Nello Cristianini, Peter Bartlett, Laurent El Ghaoui, and Michael I. Jordan. Learning the kernel matrix with semidefinite programming. In Proceedings of the 19th International Conference on Machine Learning, 2002

29. Anderson KM, Odell PM, Wilson PW, Kannel WB. Cardiovascular risk profiles. American Heart Journal. 1991; 121:293–8. [PubMed: 1985385]

30. D'Agostino RB, Vasan RS, Pencina MJ, Wolf PA, Cobain MR, Massaro JM, Kannel WB. General cardiovascular risk profile for use in primary care: The Framingham Heart Study. Circulation. 2008;117(6):743–52.10.1161/CIRCULATIONAHA.107.699579 [PubMed: 18212285]

31. Shariat SF, Karakiewicz PI, Roehrborn CG, Kattan MW. An updated catalog of prostate cancer predictive tools. Cancer. 2008; 113(11):3075–99.10.1002/cncr.23908 [PubMed: 18823041]

32. Parikh NI, Pencina MJ, Wang TJ, Benjamin EJ, Lanier KJ, Levy D, D'Agostino RB Sr, Kannel WB,Vasan RS. A risk score for predicting near-term incidence of hypertension: the Framingham Heart Study. Annals of Internal Medicine. 2008; 148(2):102–10. [PubMed: 18195335]

33. Hanley JA, McNeil BJ. The meaning and use of the area under a receiver operating characteristic (ROC) curve. Radiology. 1982; 143:29–36. [PubMed: 7063747]

34. Li JL, Fine JP. ROC analysis with multiple classes and multiple tests: methodology and its application in microarray studies. Biostatistics 2008;9(3):566–76.

35. Li JL, Fine JP. Weighted area under the receiver operating characteristic curve and its application to gene selection. JR Stat Soc C-Appl 2010;59:673–92.

36. Pencina MJ, D’Agostino RB. Overall C as a measure of discrimination in survival analysis: model specific population value and confidence interval estimation. Stat Med 2004; 23(13):2109–22.

37. Heagerty PJ, Zheng Y. Survival model predictive accuracy and ROC curves. Biometrics 2005;61(1):92–105.

38. Uno, Hajime, et al. "On the C-statistics for evaluating overall adequacy of risk prediction procedures with censored survival data." Statistics in medicine 30.10 (2011): 1105-1117.

39. Ma S, Huang J. Combining clinical and genomic covariates via Cov-TGDR. Cancer Inform 2007;3:371–78.

40. Harrel Jr, F. E. and Lee, K. L. and Mark, D. B. (1996) "Tutorial in biostatistics: multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing error", Statistics in Medicine, 15, pages 361–387.

41. K. M. Borgwardt, C. S. Ong, S. Sch ¨onauer, S. V. N. Vishwanathan, A. J. Smola, and H.-P. Kriegel. Protein function prediction via graph kernels. In Proceedings of the International Conference on Intelligent Systems for Molecular Biology, 2005.

42. S. Sonnenburg, G. R ¨atsch, and C. Sch ¨afer. A general and efficient multiple kernel learning algorithm. In Neural Information Processings Systems, 2005.

43. Cox DR. 1972. Regression models and life tables. J. Royal Stat. Soc. Ser. B 34:187-220

44. Breslow, N. E. (1975). "Analysis of Survival Data under the Proportional Hazards Model". International Statistical Review / Revue Internationale de Statistique 43 (1): 45-57.

45. Alain Rakotomamonjy, Francis Bach, Stephane Canu, Yves Grandvalet. SimpleMKL. Journal of Machine Learning Research, Microtome Publishing, 2008, 9, pp.2491-2521.

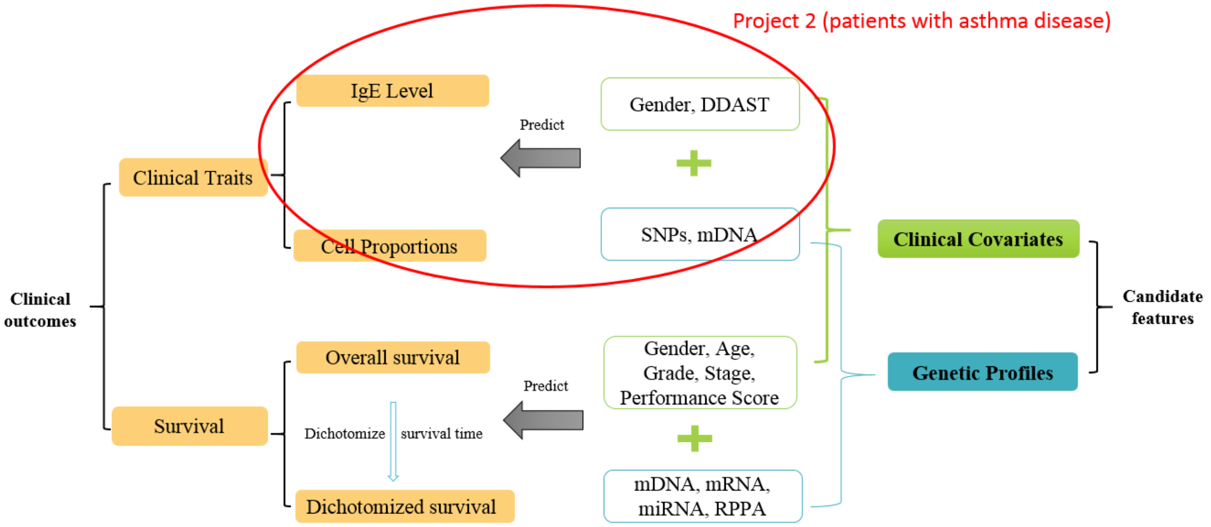
46. Holdhoff, M. et al. Use of personalized molecular biomarkers in the clinical care of adults with glioblastomas. J. Neurooncol. 110, 279–285 (2012).

47. Sturgeon, C. et al. National Academy of Clinical Biochemistry laboratory medicine practice guidelines for use of tumor markers in testicular, prostate, colorectal, breast, and ovarian cancers. Clin. Chem. 54, e11–e79 (2008).

Chapter 3 Predicting Serum IgE level and Blood Cell Proportions from Methylation

DNA methylation at CpG sites is an important epigenetic modification that may regulate gene expression. There is growing interest in understanding how the methylation inheritance contributes to the development of complex diseases or traits. It has been shown that methylation modification may influence individual asthma risk and related phenotypes. The primary purpose of this study is to comprehensively assess—by using genome-wide DNA methylation data as markers—the contribution of epigenetic effects on asthma and blood related quantitative traits. To evaluate the clinical utility of epigenetic markers, we constructed and compared various prediction models by including top ranked methylation loci from the genome-wide association scan, together with selected sets of known genetic markers from published genome-wide association studies. A new prediction model based upon Best Linear Unbiased Prediction (BLUP) was further proposed where all CpG sites (on the Illumina Infinium 27K methylation array and 450K array) were simultaneously modeled. The overall prediction accuracies of the proposed methods were extensively evaluated via the cross-validation analysis. We observed a significant increase of correlation coefficient between actual and predicted IgE level when methylation markers were included. By using an independent sample based on Illumina 450K methylation array, we also assessed the performance of cross platform prediction using methylation markers. Taken together, results from our assessment suggest that methylation has great potential in prediction of clinical phenotypes.

Figure 3.1 Illustration of Serum IgE level and Blood Cell Proportions Predictions



3.1 Introduction

DNA methylation is a crucial factor in regulating gene expression. It changes the structure of DNA by attaching methyl groups to target DNA region (generally in CpG islands), and therefore controls gene function without changing DNA sequence. It becomes increasingly important for us to understand mechanism of DNA methylation, which may associated with the development of human diseases or traits. With the advent of emerging high-throughput assay technologies, a methylation profiling on the entire genome can be collected. Most efforts have focused on a large-scale searching for methylation variations that are associated with a phenotype and gene expression data [1], also known as epigenome-wide association studies (EWAS). However, the ability to predict trait phenotypes by leveraging information on genome-wide methylation data has not been well investigated to date. Assessing the performance of predicting unobserved trait values using current epigenetic and genetic data will not only drive progress in personalized medicine but also help us to better understand the mechanisms of complex diseases.

In genetic analysis, it is now a routine step to predict phenotypic values based on top single nucleotide polymorphism (SNPs) or other genetic markers selected from genome-wide association studies (GWAS). However, the prediction accuracy is often low such that for many complex traits only a small proportion of phenotypic variance can be explained by the top associated SNPs. Based on mixed effect model and calculation of genetic similarity matrix, several methods have recently been proposed for simultaneously using all SNPs in whole genome [2, 3]. Although performance improvement of the new methods has been reported based on both simulated and real data, they are still limited by the heritability estimated based on the SNP panel and total sample size [4] . The availability of methylation data raises two interesting questions in terms of prediction: first, whether the addition of epigenetic information can significantly improve the prediction performance; and second, how to best determine the prediction models utilizing the current methylation data.

Therefore, instead of identifying specific differently methylated regions (DMRs), the main goal of this study is to investigate the performance of predicting various phenotypes based on epigenome-wide methylation patterns. The first phenotype we investigate is the serum IgE level. The total serum IgE level is an important indicator in allergic inflammatory diseases in human such as Asthma. An elevated serum IgE level is often seen in the asthmatic patient, and those with low IgE level have low prevalence of Asthma. To our knowledge, no study has examined the association between serum IgE level and whole-genome DNA methylation markers.

In this study, we also extended the predictive model based on genome-wide methylation profiles to predict blood cell proportions. The purpose of deconvoluting cell type proportions is two-fold. First, cell type is an important immune response, and quantifying the cell mixture is essential to identify the disease subtypes, status and the underlying immune-biology [5] . Second, the estimated cell mixture can be incorporated into EWAS analyses to delineate DMRs from the methylation change due to the case-control differences in blood composition. We construct several prediction models that can incorporate top methylation sites as well as whole genome methylation markers. We test these models on the phenotypes and methylation data (Illumina Infinium HumanMethylation27 and HumanMethylation450 BeadChips) collected for our asthma study, and compare their performance using cross-validation. We find that methylation markers have better ability in predicting asthma related phenotypes and blood related traits than available genetic data. The results also indicate that the model that considers genome-wide methylation sites achieve the best accuracy. Unlike gene mutations, the process of DNA methylation is reversible, which thus can be utilized in therapy development. This study shows the importance of epigenetic alternations in the development of asthma, and demonstrates that the prediction method can serve as an important tool for understanding the nature of diseases.

3.2 Methods

Methylation assays

DNA samples were quantified and bisulfite converted according to manufacturer's instructions and assayed using the HumanMethylation27 BeadChips or HumanMethylation450 BeadChip (Illumina Inc, San Diego, CA, USA). Data were visualized using the BeadStudio software, and examined using both sample-dependent and sample-independent quality control criteria. Individual data points with detection pvalue>0.01 were set as missing data.

Methylation normalization and correction for batch effects

Raw methylation data was exported from the GenomeStudio software. For the Illumina HumanMethylation27 BeadChip data, quantile normalization of intensity was applied to all methylated and unmethylated probes for all samples together. The methylation β values were recalculated as the ratio of methylated probe signal/(total signal + 100). For HumanMethylation450 BeadChip, we used the pipeline developed by Touleimat and Tost [7]. Individual data points with detection *P*>0.01 or number of beads <3 were treated as missing data, as were samples with more than 20% missing probes. Probe overlaps with any frequent SNP (MAF >5% in 1000 Genomes Project phase 1 EUR population) in the probe sequence or in position +1 or +2 of the query site (depending on Infinium I or Infinium II status) were removed. The lumi package [8] was used for background and colour bias correction. Quantile normalization across samples was applied to probes within each functional category (CpG island, shelf, shore, etc.) separately to correct the shift of methylation beta value between Infinium I and Infinium II probes by aligning the distribution of Infinium II probes to the reference distribution built upon Infinium probes.

3.3.1 Variance component model

The prediction of phenotypes can be carried out by considering the following linear mixed model (in matrix notation):

$$\boldsymbol{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{m} + \boldsymbol{\varepsilon} \tag{1}$$

where $\boldsymbol{y}$ is the phenotypic vector being analyzed, β is a vector of fixed effects, m is a vector of random additive effects attributable to methylation and it follows $N(0, \boldsymbol{K}_m \sigma_m^2)$, ε is the vector of random residual effects and follows $N(0, \boldsymbol{I}\sigma_\varepsilon^2)$. X is the incidence matrix for the fixed effects. $\boldsymbol{I}$ is an identify matrix and $\boldsymbol{K}_m$ can be interpreted as the relationship matrix between individuals attributable to DNA methylation. Therefore, the variance of phenotypic observation can be decomposed into $\boldsymbol{V}_y = \boldsymbol{K}_m \sigma_m^2 + \boldsymbol{I}\sigma_\varepsilon^2$. Here we use correlation coefficient as the realized relationship for each pair of individuals. The variance components σ_m^2 and σ_ε^2 can be estimated using restricted maximum likelihood (REML). Similarly, the above mixed model can be extended to incorporate genetic data using the following model:

$$\boldsymbol{y} = \mathbf{X\beta} + \mathbf{m} + \mathbf{g} + \boldsymbol{\varepsilon} \qquad (2)$$

where g is a vector of random genetic effect which can be based on whole genome SNP information and $\boldsymbol{g} \sim N\left(0, \boldsymbol{K}_g \sigma_g^2\right)$. $\boldsymbol{K}_g$ is the realized relationship matrix of SNP data and can be estimated by $\boldsymbol{K}_g = \boldsymbol{G}\boldsymbol{G}^T/N$, where $\boldsymbol{G}$ is a standardized genotype matrix[9]. However, we noticed that most functions implemented in current R packages can only estimate one variance component at each time. Estimating multiple components iteratively is very time consuming and may have convergence issue. Therefore, we developed our own efficient function which can estimate multiple variance components in a model simultaneously based on REML, in R. If only a small number of candidate SNPs are available (as in our data analysis), we can easily incorporate them into covariates.

3.3.2 Mixed model based prediction

In order to perform prediction on phenotypic values, the above described random effect model needs to be converted into the following fixed effect model:

$$\boldsymbol{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{Z}_{\boldsymbol{m}}\boldsymbol{u}_{\boldsymbol{m}} + \boldsymbol{\varepsilon} \tag{3}$$

where Z_m is the incidence matrix of methylation, i.e., methylation score matrix (Beta values). $\boldsymbol{u}_m$ is a vector of fixed effects of CpG sites included. In REML analysis, the best linear unbiased predictors (BLUP) of the random effect can be obtained as $\widehat{\boldsymbol{m}} = \boldsymbol{V}_{\boldsymbol{m}}\boldsymbol{V}_{\boldsymbol{y}}^{-\boldsymbol{1}}\left(\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}\right),$ where $\widehat{\boldsymbol{\beta}}$ is the best linear unbiased estimators (BLUE): $\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}^{\boldsymbol{T}}\widehat{\boldsymbol{V}}^{-\boldsymbol{1}}\boldsymbol{X})^{-\boldsymbol{1}}\boldsymbol{X}^{\boldsymbol{T}}\widehat{\boldsymbol{V}}^{-\boldsymbol{1}}\boldsymbol{y}$. By equating the random and fixed effects models, we can get $\widehat{\boldsymbol{u}_m} = \boldsymbol{Z}^{\boldsymbol{T}}\boldsymbol{K}_{\boldsymbol{m}}^{-\boldsymbol{1}}\widehat{\boldsymbol{m}}/N$, where N is the total number of CpG sites. Therefore, we can predict the effect of methylation in a test data set by plugging in the BLUP of $\boldsymbol{u}$, i.e., $\widetilde{\boldsymbol{m}} = \boldsymbol{Z}_{test}\widehat{\boldsymbol{u}_m}$. The prediction of unknown phenotypes in the test dataset can be performed through:

$$\widetilde{\boldsymbol{y}}_{\boldsymbol{test}} = \boldsymbol{X}_{\boldsymbol{test}}\widehat{\boldsymbol{\beta}} + \widetilde{\boldsymbol{m}}_{\boldsymbol{new}} = \boldsymbol{X}_{\boldsymbol{test}}\widehat{\boldsymbol{\beta}} + \boldsymbol{Z}_{\boldsymbol{test}}\boldsymbol{Z}^{\boldsymbol{T}}\boldsymbol{K}_{\boldsymbol{m}}^{-\boldsymbol{1}}\widehat{\boldsymbol{m}}/\boldsymbol{N} \tag{4}$$

3.3.3 Cross-validation of predictive models

We construct six models in predicting IgE (Table 2.1). Gender, DDAST and candidate SNPs that are associated with IgE were adjusted as covariates separately or jointly in these models. To evaluate the accuracy of prediction, we employ a leave-one-family-out cross-validation (CV) by iteratively excluding samples from one family from the training set. This strategy will help eliminate spurious prediction improvement due to familial correlation. Instead of using formula (4), the CV prediction can be carried out more efficiently by employing an equivalent form: $\tilde{y}_{\boldsymbol{test}} = X_{\boldsymbol{test}}\widehat{\beta} + \mathbf{K}_{21}\mathbf{K}_{11}^{-1}\widehat{\mathbf{m}}$, where $\mathbf{K}_{21}$and $\mathbf{K}_{11}$ are the corresponding submatrices of similarity matrix of whole samples. In model (E), epigenome-wide association analyses (EWAS) are first preformed to select top methylation sites. The EWAS scan and top methylation site selection is repeated for each training set generated in the cross-validation process. The prediction in model (E) is then based on

$$\widetilde{\boldsymbol{y}}_{\boldsymbol{test}} = \boldsymbol{X}_{\boldsymbol{test}}\widehat{\boldsymbol{\beta}} + \boldsymbol{W}_{\boldsymbol{test}}\widehat{\boldsymbol{\alpha}} \tag{5}$$

$\boldsymbol{W_{test}}$ and $\widehat{\alpha}$ are the incidence matrix and effect vector for the selected top methylation sites, respectively.

3.3.4 LASSO, Ridge and Elastics Net regression models

For ordinary least square regression, we obtain the estimates, by objective function of minimize the least squares of error. When are assumptions for least square regression hold, it generates good models for data with relative small number of feature. Cancer genome data fall into the category of high dimension, while in this situation the number of features (p) is extremely large, comparing to sample size (n). For example, DNA methylation pattern (mDNA), characterized based on Illumina Infinium HumanMethylation450 BeadChip panel, interrogates about 450k CpG sties in total, while sample size is several hundreds.

Regression of high dimensional data, (A) there might be redundant features, which contributes no additional information, other than noises; (B) it brings trouble for interpretation and visualization; (C) from computational perspective, it increases difficulty to store and process data; (D) last but not the least, complexity of decision rule tends to grow with number of feature. Therefore, when dealing with high dimensional data, we are facing unique challenges. Problems include how to select feature and test variable relationships and how to build a model based on high-dimensional data while in the meantime protecting against over-fitting.

The LASSO, least absolute shrinkage and selection operator[10], and Elastic Net[11] are shrinkage and selection methods that are applied frequently for high dimensional data linear regression. Statistically, the phrase high dimensional data describe situations in which the number of measurements (p) is large, especially relative to the number of experimental units (n). In the fitting of linear regression models, the elastic net combines both L1 and L2 [12]. $y = X\beta + \varepsilon$, where y is the phenotypic vector being analyzed, $\boldsymbol{X}$ is an input measurements matrix, and ε is the vector of random residual effects, which follows $N(0, \boldsymbol{I}\sigma_{\varepsilon}^{2})$. By minimizing the residual sum of squares subject to the constraints, the LASSO and Elastic Net naturally tend to set some of the coefficients to be zero, which helps interpret the model. The estimate $\hat{\beta}$ from the elastic net method are defined as

$$\begin{aligned} \hat{\boldsymbol{\beta}} &= \mathbf{arg\,min}\left\{\sum_{i=1}^{n}\left(y_i - \sum_j \boldsymbol{\beta}_j \mathbf{x}_{ij}\right)^2 + \lambda_2 \left|\left|\boldsymbol{\beta}\right|\right|^2 + \lambda_1 \left|\left|\boldsymbol{\beta}\right|\right|_1\right\} \\ &= \mathbf{arg\,min}\left\{(y - X\boldsymbol{\beta})^T(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) + \lambda_2 \left|\left|\boldsymbol{\beta}\right|\right|^2 + \lambda_1 \left|\left|\boldsymbol{\beta}\right|\right|_1\right\} \end{aligned} \qquad (6)$$

Where $\left|\left|\beta\right|\right|^2 = \sum_j \boldsymbol{\beta}_j^2$ and $\left|\left|\beta\right|\right|_1 = \sum_j |\boldsymbol{\beta}_j|$. When $\boldsymbol{\lambda_2} = 0$, the model only contains L1 penalty and becomes LASSO. The Elastic Net method overcomes the limitations of the LASSO method by the additional quadratic penalty $\left|\left|\beta\right|\right|^2 = \sum_j \boldsymbol{\beta}_j^2$. When $\boldsymbol{\lambda_1} = 0$, the model only contains L2 penalty and becomes ridge regression. While there are two steps involved in the naive version of elastic net method to find an estimator: first for each fixed $\boldsymbol{\lambda_2}$ it solves the coefficients for the ridge regression, and then does a LASSO type shrinkage. Unnecessary extra bias is introduced by this type of estimation with doubled amount of shrinkage and hence incurs bad predictive performance. To improve the predictive performance, rescaled coefficients are introduced into the naive version by multiplying the estimated coefficients by $(1 + \boldsymbol{\lambda_2})$. Adaptive LASSO is also introduced by adding weights to L1 penalty[13].

3.3.6 Support Vector Machine Regression Prediction

Support Vector Machines (SVMs) [18], originally introduced by Vapnik and co-workers [19-22], have been successively extended by other researchers for a variety of classification and regression purposes. The detailed derivation of SVMs for classification problems can be find from Vapnik's work [19-22]. Vandewalle reviewed the basic work on SVM [23], we will now give a brief summary based on his reviews.

As a powerful machine learning technique, SVM has been shown to perform well in multiple areas, including biological data analysis, such as gene microarray expression data [15]. SVM utilizes "Kernels" to achieve a general mapping to a feature space, automatically, which can be linear or non-linear [24]. The operations in the feature space of kernel functions, can be done easily by computing the inner products between all pairs of data, with no worry about computing the coordinates of the data (e.g. [25]). A SVM constructs a hyperplane in a high- or infinite-dimensional space, and separate a given set of training data by maximize the distance to the nearest training data point of any class. Support vector regression (SVR) is one type of support vector machine (SVM), when it is employed for regression problems [26].

Let $\{\boldsymbol{y_k}, \boldsymbol{x_k}\}_{\boldsymbol{k=1}}^{\boldsymbol{N}}$ denote N data points from a training set, where $\boldsymbol{x_k} \in \boldsymbol{R^n}$ is the k-th input and $\boldsymbol{y_k} \in \boldsymbol{R}$ is the k-th output. SVM aims at constructing a classifier:

$$\boldsymbol{y_k = sign}[\sum_{\boldsymbol{k=1}}^{\boldsymbol{N}} \boldsymbol{\alpha_k y_k}\, \boldsymbol{\psi}(\boldsymbol{x}, \boldsymbol{x_k}) + \boldsymbol{b}] \qquad (2)$$

where $\boldsymbol{\alpha_k}$ are positive constants and $\boldsymbol{b}$ is a constant. For $\boldsymbol{\psi}(.,.)$ one can choose following optionss: $\boldsymbol{\psi}(\boldsymbol{x}, \boldsymbol{x_k}) = \boldsymbol{x_k^T x}$ (linear SVM); $\boldsymbol{\psi}(\boldsymbol{x}, \boldsymbol{x_k}) = (\boldsymbol{x_k^T x} + \boldsymbol{1})^{\boldsymbol{d}}$ (polynomial SVM of degree d); $\boldsymbol{\psi}(\boldsymbol{x}, \boldsymbol{x_k}) = \mathbf{exp}\{-\left\|\boldsymbol{x_k} - \boldsymbol{x}\right\|_2^2/\boldsymbol{\sigma}^2\}$ (RBF SVM), where $\boldsymbol{\sigma}$ is constant.

3.3 Results

3.2.1 Cross validation for the prediction of IgE

We collected our data from 195 siblings and their parents in 95 nuclear pedigrees identified by a proband of asthma. The sample has 355 subjects (183 male) with a mean age in children of 12.2 years (ranging from 2 to 39) and adults of 42 (27 to 61). DNA extracted from peripheral blood leukocytes (PBL) and lymphoblastoid cell line (LCL) was used in the assay. The methylation pattern was characterized based on Illumina Infinium HumanMethylation27 BeadChip panel, which interrogates 27,578 CpG sties in total. All the data used in the prediction analysis has been processed through proper pre-processing steps including background correction and normalization.

From the catalogue of published GWAS provided by the National Human Genome Research Institute (www.genome.gov/gwastudies), we identified two top SNPs (rs2251746 and rs2571391) that are significantly associated with IgE and also available in our SNP dataset. These two SNPs were included in our prediction model together with two other covariates, gender and DDAST (doctor diagnosed asthma). As shown in Table 2.1, using two SNPs slightly improves the prediction accuracy (by comparing model C to model A, R^2 was improved from 0.083 to 0.106). When all methylation sites were included in the analysis using our proposed method (model D), we observe a significant increase of cross-validation correlation coefficient between actual and predicted IgE level (R^2 was improved from 0.106 before to 0.327 after including methylation data). In model E, we considered a more classical prediction model by including different number of top methylation markers (identified from a genome wide scan). It is, however, always difficult to decide an optimal number of predictors because too few will limit predication capability while too many can cause over-fitting.

Figure 3.2 shows the predictive performance of model E, by increasing the number of methylation markers from 1 to 20, with R^2 between 0.20 and 0.342. The highest CV correlation 0.342 was reached when 16 markers were included, and it is very close to the proposed model (D), which uses the whole methylation profile while avoids a grid search. Results from this study suggest that DNA methylation explains much larger variability in IgE level than known genetic variants (2% due to top genetic markers in large GWAS vs. 15% due to top 3 CpG sites), suggesting that methylation has important influence in asthma and has great potential in prediction of clinical phenotypes. We also explored the prediction model with the interaction term between level of eosinophils (EOS), which is known to be associated with IgE and top methylation sites. The CV correlation from model F (Figure 3.1) is highest (0.419) when around five methylation makers and their interactions are considered but rapidly decreases when more markers are added into the model.

Table 3.1 Prediction models for predicting IgE and performance summary

Model	**Cross validation r^2**
(A) Gender+DDAST	0.083
(B) Gender+DDAST+rs2251746	0.099
(C) Gender+DDAST+rs2251746+rs2571391	0.106
(D) Gender+DDAST+rs2251746+rs2571391+ Methylation profile	0.327
(E) Gender+DDAST+rs2251746+rs2571391+top methylation sites	0.20 ~ 0.342
(F) Gender+DDAST+ rs2251746+rs2571391+EOS + top methylation sites + (METHY×EOS)	Maximum r^2 : 0.419

Figure 3.2 Predictive performances of IgE obtained from model E and model F.

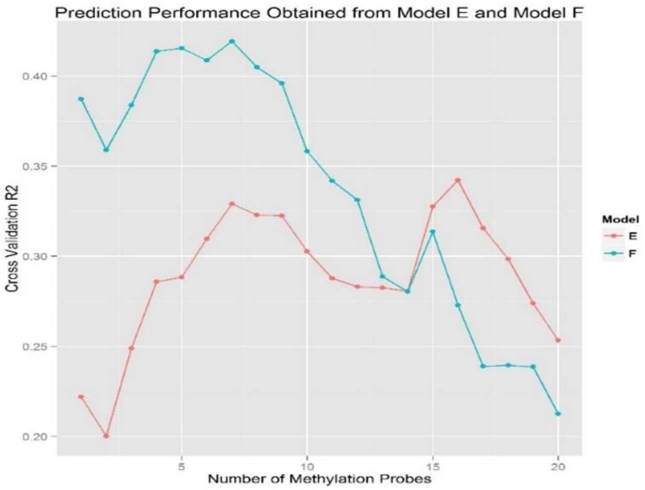


Table 3.2 summarizes the predictive performance of the proposed prediction model (D) in predicting cell proportions using cross validation. The model is similar to model D for predicting IgE but without adjusting for DDAST and top IgE SNPs. For both data sets (with and without batch effect correction), it is found that the model with methylation profile only can provide highest accuracy in predicting Neutrophils (NEU), followed by Eosinophils (EOS) and Lymphocytes (LYM), and have poor perdition Monocytes (MON).

Table 3.2 Cross validation (LOFO) predictive performance of cell proportions using methylation profile where 27K are used as training data

Cell type	**Cross validation r^2**				
	Top 1	Top 5	Top 10	Top 50	27K_VC
NEU	0.588	0.685	0.656	0.649	0.709
EOS	0.505	0.607	0.598	0.593	0.678
LYM	0.529	0.632	0.627	0.618	0.621
MON	0.07	0.165	0.202	0.183	0.001

3.2.2 Cross-platform validation for the prediction of cell proportions

We also evaluated the predictive performance of cell proportions across different methylation assay platforms. Two additional data sets used for this purpose are the 149 Caucasian subjects selected equally from the top and bottom deciles of IgE distribution in 1614 unselected volunteers (students and staff from Swansea University); and in 160 (80 male) subjects in an asthmatic family panel from the Saguenay–Lac-Saint-Jean region (SLSJ) of Quebec[6] with a mean age in children of 16 years (ranging from 5 to 50; 40 DDAST) and adults of 44 years (31 to 79). Table 3.3 presents the results from the predictions model where the previous 27K data were used as training dataset and each of the two 450K data sets were used as a testing data set. For each dataset we applied six prediction models include models with top 1, 5, 10, 50 probes and ~25K overlapping probes between 27K and 450K array, respectively. Overall, the prediction on Lymphocytes achieved best performance, followed by Eosinophils and Neutrophils. It is shown that the variance component based model which includes all 25K probes provided optimal prediction accuracy for all cell types except Monocytes. The variance component prediction model was further applied to include all probes in 450K arrays.

Table 3.3 Cross data validation predictive performance of cell proportions using 27K methylation profile as training data and testing on both Swansea and SLSJ data.

Cell type	Testing data set	Prediction probes/models r2					
		Top 1	Top 5	Top 10	Top 50	25K_VC	Top5+25K VC
LYM	Swansea	0.078	0.078	0.000	0.014	0.873	0.873
	SLSJ	0.003	0.003	0.014	0.010	0.846	0.846
EOS	Swansea	0.573	0.573	0.535	0.541	0.873	0.774
	SLSJ	0.581	0.581	0.537	0.707	0.699	0.699
NEU	Swansea	0.179	0.179	0.043	0.058	0.750	0.750
	SLSJ	0.110	0.110	0.018	0.025	0.837	0.837

MON	Swansea	0.119	0.119	0.121	0.116	0.119	0.119
	SLSJ	0.033	0.033	0.034	0.034	0.033	0.033

Table 3.4 presents the results from the predictions model where Swansea 450K and SLSJ 450K were alternatively used as training dataset to predict each other. For each dataset we applied five prediction models include models with top 1, 5, 10, 50 probes and ~25K overlapping probes between 27K and 450K array, and all 450K probes, respectively. Overall, the prediction on Lymphocytes achieved best performance, followed by Eosinophils and Neutrophils. Figure 3.3 shows the plots of the predicted cell proportions in Swansea data set (model trained using the SLSJ data) versus the observed cell proportions, and the predicted cell proportions in SLSJ data set (model trained using the Swansea data) versus the observed cell proportions.

Table 3.4 Cross data validation predictive performance of cell proportions where Swansea and SLSJ are used to predict each other.

Cell type	**Testing data set**	**Prediction probes/models r^2**					
		Top 1	Top 5	Top 10	Top 50	25K_VC	450K_VC
LYM	Swansea	0.007	0.007	0.001	0.000	0.902	0.885
	SLSJ	0.007	0.007	0.000	0.013	0.805	0.801
NEU	Swansea	0.023	0.023	0.037	0.016	0.809	0.768
	SLSJ	0.003	0.003	0.050	0.000	0.735	0.768
EOS	Swansea	0.008	0.008	0.009	0.010	0.606	0.813
	SLSJ	0.021	0.021	0.037	0.023	0.468	0.760
MON	Swansea	0.118	0.118	0.125	0.470	0.014	0.000
	SLSJ	0.067	0.067	0.053	0.263	0.022	0.000

3.2.3 Machine learning methods for the prediction of cell proportions

For the same two data sets, Swansea 450K and SLSJ 450K data, we also evaluate the predictive performance of cell proportions using methylation profile by applying several machine learning methods. Table 3.5 presents the correlation determination results from the

LASSO and Elastic Net predictions where Swansea 450K and SLSJ 450K were used as training dataset to predict each other. For each dataset we applied five prediction models include LASSO, Ridge, and Elastic Net (alpha=0.2), respectively. Overall, the prediction on Eosinophils achieved best performance (R^2=89.1%), followed by Neutrophils and Lymphocytes. It is also shown that the Elastic Net outperforms LASSO and Ridge.

Table 3.6 presents the results from Support Vector Regression predictions where Swansea 450K and SLSJ 450K were used as training dataset to predict each other. For each dataset we applied seven prediction models include models with top 10, 100, 1K, 10K probes and 25K overlapping probes between two data sets, and the utilized the whole 450K profiles, respectively. It also shows correlation determination between predicted cell proportions and observed cell proportions, by applying SVR. In predicting Eosinophils, the best performance has been achieved that 91.5% of variance can be explained by SVR with top 10K selected feature, followed by Neutrophils and Lymphocytes.

Table 3.5 LASSO, Ridge and Elastic Net predictive performance of cell proportions where Swansea and SLSJ are used to predict each other.

Cell type	**Testing data set**	**Prediction probes/models**		
		LASSO	Ridge	Elastic Net (alpha=0.2)
EOS	Swansea	0.882	0.740	0.867
	SLSJ	0.888	0.609	0.891
NEU	Swansea	0.692	0.825	0.833
	SLSJ	0.823	0.780	0.841
LYM	Swansea	0.800	0.900	0.809
	SLSJ	0.838	NA	0.849
MON	Swansea	0.457	0.196	0.432
	SLSJ	0.387	0.110	0.484

Table 3.6 SVR Predictive performance of cell proportions based on methylation profile and gender information.

Cell type	**Testing data set**	**Prediction probes/models**					
		Top10	Top100	Top1K	Top10K	25K	450K
EOS	Swansea	0.905	0.771	0.907	0.915	0.888	0.734
	SLSJ	0.867	0.730	0.899	0.899	0.890	0.587
NEU	Swansea	0.909	0.673	0.815	0.872	0.863	0.816
	SLSJ	0.835	0.765	0.838	0.844	0.831	0.771
LYM	Swansea	0.891	0.792	0.910	0.903	0.910	0.898
	SLSJ	0.835	0.766	0.858	0.859	0.856	0.802
MON	Swansea	0.491	0.092	0.596	0.700	0.605	0.187
	SLSJ	0.376	0.410	0.520	0.680	0.731	0.104

By now, we have compared the prediction accuracy of mixed linear models, variance component models, LASSO, Ridge, Elastic Net and SVR, and found out that regularized least square regressions and SVR outperforms others, in which SVR achieved the highest R^2 of 0.915, slightly better than that of the best performance of Elastic Net with R^2 of 0.891.

3.2.4 Prediction based on Houseman's top DMR CpGs

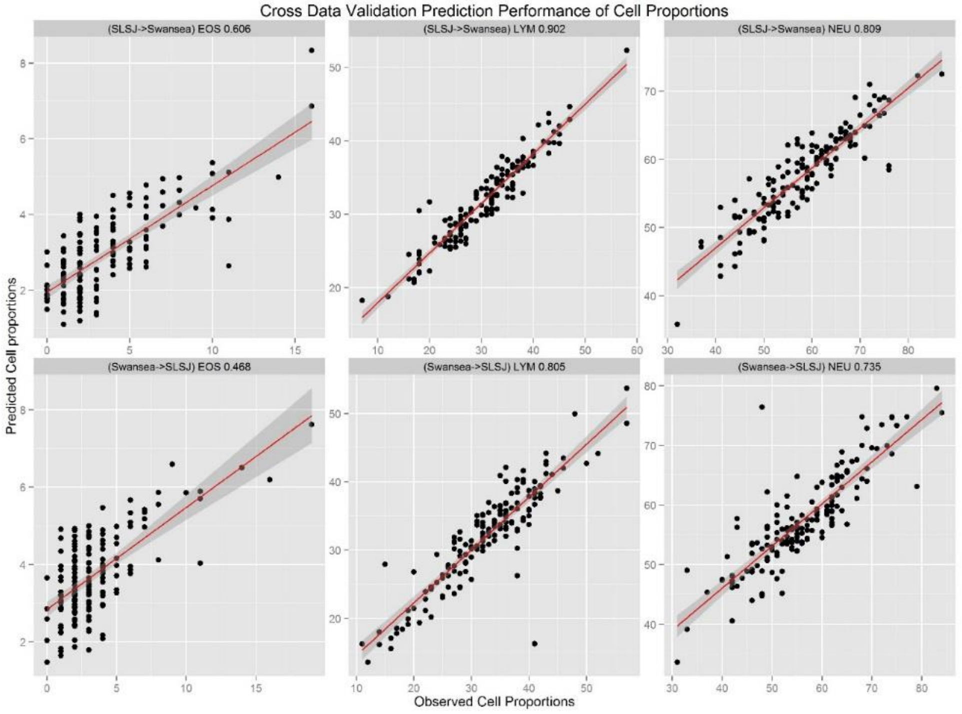
We also applied the above methods to build predicate model based on the top 500 DMR CpGs identified through the algorithm proposed in Houseman et al. (2012). We evaluated the performance of prediction through cross-data and cross-platform validation. The corresponding results from applying various models introduced in previous sections are summarized in Table 3.7. Overall, similar performance was achieved for predicting cell type proportions. It indicates that using the whole methylation sites in the training model will not affect the predictive performance due to noise and outliers in high-dimensional data, which is one major concern in the prediction using whole genome-wide markers. Among different methods tested including

LMM, elastic LASSO and other machine learning methods, there is no universally best method for all scenarios. LASSO and SVR showed very promising results in predicting lymphocytes and neutrophils, respectively. However, these two methods yielded poor reproductive performance when predicting neutrophils and eosinophils, respectively. LMM models (either trained on 27K or 450K data) and ridge regression achieved relatively robust prediction. This is as expected because both LMM and ridge regression are based on L2 penalty, where the information from all markers will be incorporated in the testing model. When number of predictors is limited (e.g. 500 CpGs here), applying a model that further induces sparsity such as L1 may be risky, where important predictors with small or medium effects may not be picked out and included in the final testing model. This results further demonstrate the advantage of the whole genome-wide prediction models which does not need to be constrained on a preselected panel, and their results tend to provide more reproducible results across different cell types – and more importantly, across different sources of data.

Table 3.7 Predictive performance of cell proportions based on 500 probes and gender information, by various of models.

Cell type	**Testing data set**	**Prediction models with 500 probes**					
		LMM Train on 27K	LMM Train on 450K	LASSO	Ridge	Elastic Net (alpha = 0.2)	SVR
LYM	Swansea	0.803	0.819	0.896	0.767	0.705	0.794
	SLSJ	0.811	0.685	0.851	0.662	0.616	0.843
NEU	Swansea	0.791	0.806	0.378	0.904	0.923	0.757
	SLSJ	0.793	0.728	0.811	0.846	0.857	0.828
EOS	Swansea	0.745	0.794	0.608	0.836	0.481	0.754
	SLSJ	0.590	0.662	0.611	0.821	0.826	0.728
MON	Swansea	0.488	0.547	0.393	0.492	0.371	0.427
	SLSJ	0.392	0.302	0.242	0.519	0.370	0.362

Figure 3.3 Cross data validation predictive performance of cell proportions where Swansea and SLSJ are used to predict each other, based on predictive models with ~25K overlapping probes between 27K and 450K array



3.4 Discussion

We have presented several prediction methods to analyze DNA methylation data from peripheral blood leukocytes (PBL) and lymphoblastoid cell line (LCL), to predict asthma related phenotypes, serum IgE level and cell type proportions, both within the same platform and cross platforms. For IgE level prediction, we find that methylation markers have better predictive performance than genetic markers identified from genome-wide association studies. For cell proportions study, we explored mixed model based prediction, LASSO and other machine learning based methods. Among all the non-LMM based methods, elastic net outperforms

LASSO in general, across all of the different cell types. Due to the potential significant non-linear interactions between methylation sites, we want to perform SVR with not only the liner kernel, but also the Gaussian kernel. Our results show that in SVR predictive performance, the linear kernel outperforms the Gaussian kernel, except that for top 100 selected features, Gaussian kernel generated better correlation coefficients than linear kernel, consistently. Our results showed that the model based on genome-wide methylation panel could predict three cell types (EOS, NEU and LYM) with high accuracy and reliability. The predictive performance is less accurate for MON prediction and much lower for basophils (results not listed)—due to the scarcity in cell counts. Among the methods examined, we found that, if based on all 27K or 450K methylation markers, Elastic Net and SVR often achieve slightly better predictive performance on both cross-validation and independent test data. If based on a pre-select marker set such as the 500 CpGs from Houseman's, it is found that LMM-based method can achieve better performance. Therefore, for the prediction with full panel of methylation markers, we recommend to use elastic net and SVR with linear kernel. We also strongly recommend applying LMM as a benchmark method in all analysis due to its reliability. While for the prediction with a small subset of markers, the algorithms based on non-spare solutions such as LMM and ridge regression should be preferred.

Over-fitting can be especially important for prediction based on high-dimensional data. This issue is controlled by regularization in the predictive models. In the linear mixed model, regularization is incorporated implicitly and the penalty parameter is estimated automatically. In fact, the likelihood for ridge regression penalized model is mathematically equivalent to the likelihood for LMM. One advantage of the LMM formulation is that, as a byproduct, we can estimate the individual variance components, and thus the proportion of phenotypic variance explained by variation from each variant attributes. In this sense, this estimate can be treated as an epigenetic counterpart of the hereditability in GWAS and will be useful in decomposing the contribution from epigenetic and genetic factors-if both the genetic and epigenetic panels are modeled. Furthermore, LMM provides a natural way to extend the predictive model to integrate multiple types of omic data (e.g. SNP, copy-number varation, gene expression and DNA methylation), by treating each of them as a variance component. In model evaluation, we used both cross-validation and performed validation on independent test datasets. The predictive performance demonstrated that our methods are applicable to methylation data generated from different platforms and even from different sources. While we focused on epigenome-wide prediction in this study, we do see the value in using a panel of pre-select methylation markers for prediction, such as the 500 DMR CpGs supplied in Houseman et al. (2012). Using small number of markers is computationally efficient, less susceptible to the over-fitting issue, and thus tend to be more reliable, especially when the sample size is limited.

In summary, our analysis shows the importance of epigenetic alternations in the level of blood related complex traits, and demonstrates that the prediction method can serve as an important tool for understanding the status and stages of diseases. We demonstrated that methylation profile could be used to reliably predict proportions of three blood cell types, e.g. EOS, NEU and LYM. In our study, we have datasets with sample size around 150 each. With larger samples in ongoing epidemiology studies, we expect that our methods could be future improved by leveraging information from more samples and finer tuning of model parameters.

3.5 References

1. Rakyan VK, Down TA, Balding DJ, Beck S. Epigenome-wide association studies for common human diseases. Nature Reviews Genetics 12, 529-541 (2011).

2. Lee SH, van der Werf JHJ, Hayes BJ, Goddard ME, Visscher PM. Predicting unobserved phenotypes for complex traits from whole-genome SNP data. *PLoS Genetics* **4**, e1000231 (2008).

2. Ober U*, et al.* Using whole-genome sequence data to predict quantitative trait phenotypes in Drosophila melanogaster. *PLoS Genetics* **8**, e1002685 (2012).

4. Daetwyler HD, Villanueva B, Woolliams JA. Accuracy of predicting the genetic risk of disease using a genome-wide approach. *PloS one* **3**, e3395 (2008).

5. Houseman EA*, et al.* DNA methylation arrays as surrogate measures of cell mixture distribution. *BMC bioinformatics* **13**, 86 (2012).

6. Begin P*, et al.* Association of urokinase-type plasminogen activator with asthma and atopy. *American journal of respiratory and critical care medicine* **175**, 1109-1116 (2007).

7. Touleimat N, Tost J. Complete pipeline for Infinium((R)) Human Methylation 450K BeadChip data processing using subset quantile normalization for accurate DNA methylation estimation. *Epigenomics* **4**, 325-341 (2012).

8. Du P, Kibbe WA, Lin SM. lumi: a pipeline for processing Illumina microarray. *Bioinformatics* **24**, 1547-1548 (2008).

9. Yang J, Lee SH, Goddard ME, Visscher PM. GCTA: a tool for genome-wide complex trait analysis. *American Journal of Human Genetics* **88**, 76-82 (2011).

10. Tibshirani R. Regression shrinkage and selection via the LASSO: a retrospective. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **73**, 273-282 (2011).

11. Zou H, Hastie T. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **67**, 301-320 (2005).

12. Zou H. The Adaptive LASSO and Its Oracle Properties. *Journal of the American Statistical Association* **101**, 1418-1429 (2006).

12. Huang J, Ma S, Zhang C-H. Adaptive LASSO for sparse high-dimensional regression models. *Statistica Sinica* **18**, 1603 (2008).

14. Drucker H, Burges CJ, Kaufman L, Smola A, Vapnik V. Support vector regression machines. *Advances in neural information processing systems*, 155-161 (1997).

15. Brereton RG, Lloyd GR. Support Vector Machines for classification and regression. *Analyst* **135**, 230-267 (2010).

16. Cristianini N, Shawe-Taylor J. An introduction to support vector machines and other kernel-based learning methods. Cambridge university press (2000).

17. Keerthi SS, Lin C-J. Asymptotic behaviors of support vector machines with Gaussian kernel. *Neural computation* **15**, 1667-1689 (2003).

18. Cristianini, Nello, and John Shawe-Taylor. An introduction to support vector machines and other kernel-based learning methods. Cambridge university press, 2000.

19. Schölkopf B., Sung K.-K., Burges C., Girosi F., Niyogi P., Poggio T., Vapnik V., “Comparing support vector machines with Gaussian kernels to radial basis function classifiers,” IEEE Transactions on Signal Processing, Vol.45, No.11, pp.2758-2765, 1997.

20. Vapnik V., “The nature of statistical learning theory,” Springer-Verlag, New-York, 1995.

21. Vapnik V., “Statistical learning theory,” John Wiley, New-York, 1998.

22. Vapnik V., “The support vector method of function estimation,” In Nonlinear Modeling: advanced black-box techniques, Suykens J.A.K., Vandewalle J. (Eds.), Kluwer Academic Publishers, Boston, pp.55-85, 1998.

22. Suykens, Johan AK, and Joos Vandewalle. "Least squares support vector machine classifiers." Neural processing letters 9.3 (1999): 293-300.

24. Furey, T.S. et al, "Support vector machine classification and validation of cancer tissue samples using microarray expression data," Bioinformatics, 16(10):906-914 (2000).

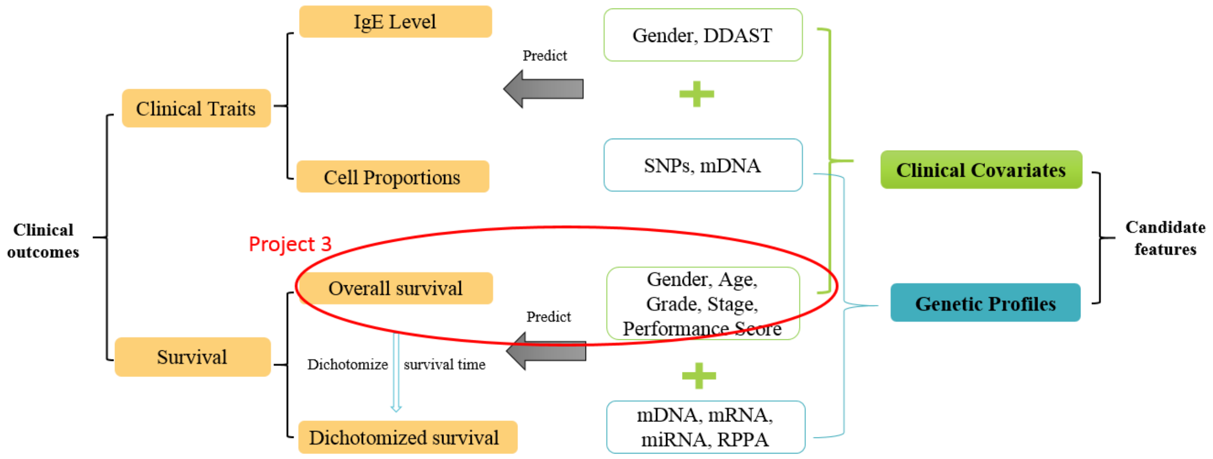
25. B. Sch ölkopf and A. J. Smola. Learning with Kernels. MIT Press, Cambridge, MA, 2002.

26. V. Vapnik, S. Golowich and A. Smola, (1997), “Support Vector Method for Function Approximation, Regression Estimation, and Signal Processing”, in M. Mozer, M. Jordan, and T. Petsche (eds.), Neural Information Processing Systems, Vol. 9. MIT Press, Cambridge, MA.

Chapter 4 Current and Future Work

In this chapter, we introduce a current project with focus on predicting overall survival time of patients with metastatic castrate resistant prostate cancer (mCRPC), utilizing clinical information alone and also discuss our future directions for integrative analysis.

Figure 4.1 Illustrate of Survival Prediction with Kernel Methods



In order to improve the predictive performance, we explored kernel Cox regression from clinical variables, exploring the efficiencies of linear, Gaussian and clinical kernels. Our preliminary findings are included in the following sections. The potential benefit of this study is to establish better prognostic models in support of clinical decisions, and better understanding the mechanism of mCRPC disease progression.

4.1 Introduction to Survival Prediction with Kernel Methods

As the most common cancer among men, prostate cancer ranks third in terms of mortality after lung cancer and colorectal cancer [1]. The mainstay of treatment for metastatic disease has been androgen deprivation therapy (ADT), though inevitably patients develop resistance. This condition is called metastatic castrate-resistant prostate cancer (mCRPC) and accounts for one third of all patients with metastatic disease. Although a number of options exist for treatment of mCRPC including chemotherapies and supportive care, the most effective single therapy or sequence of therapies remains unclear. Innovative research approaches using the phase III trials made available through Project Data Sphere may hold meaningful promise for improving the treatment of patients with mCRPC.

Factors that are highly related to the risk of prostate cancer include: age, race, a family history of the disease, and genetic background. The goal of this project is to develop models for predicting the overall survival for patients with mCRPC, based on just clinical variables. Prognostic models are then served as tools for doctors to use and to determine the best treatment options for patients. It has been showed that better prognostic models can be built, by leveraging the most up-to-date and well-sampled mCRPC data [2], which highlights the importance of prognostic research to integrate the current clinical context and patient health status into treatment choices and clinical management decision making. In this project, we leverage 4 clinical trials compiled by Project Data Sphere to develop models for predicting overall survival of chemotherapy-naive mCRPC patients receiving docetaxel. Models will be evaluated according to the C-index.

4.2 Methods

4.2.1 Kernel Cox regression

Kernel Cox regression models were originally proposed for linking gene expression profile to overall survival, with linear kernels, by Li and Luan [3], who extended the SVM for categorical data to censored survival data. In their study, they also claimed that kernel Cox regression models are more efficient and powerful than Cox regression models, with respect to: (A) It can automatically perform feature selection to identify genes highly related to survival and further identify the optimal combination of the features in predicting the risk of cancer; (B) there is no limitation for the number of genes used in the prediction of patient's overall survival, in terms of computational or methodological cost.

The following are the formulations and derivations of kernel Cox regression models, originally proposed for relating gene expression to overall survival, in the framework of SVM, utilizing kernels. Following the definition of Cox regression, and assuming the generalized Cox model with the hazard function for the *i*th patient being

$$\boldsymbol{\lambda_i(\mathrm{t}|x_i) = \lambda_0(t) exp\big(f(x_i)\big)} \qquad (1)$$

Same as in the usual Cox model, $\lambda_0(t)$ is the unspecified baseline hazard function, while $f(x_i)$ is a function of gene expression data x_i.

For gene expression data, the standard Cox model is not feasible for estimating unspecified function, due to that the dimension of x_i vector being much larger than the sample size. They then proposed to solve this problem, through kernel tricks, by defining the function $f(x_i)$ as the following

$$\boldsymbol{f(x_i) = b + \sum_{i=1}^{n} a_i K(x, x_i)} \qquad (2)$$

where b can be absorbed into the baseline function, and choosing the simplest case of natural inner product kernel, $K(x, x_i) =< x_i, x_j >$, so we then have

$$\boldsymbol{f(x_i) = \sum_{i=1}^{n} a_i K(x, x_i) = \sum_{j=1}^{p} (\sum_{i=1}^{n} a_i x_{ij}) x^{(j)} = \sum_{j=1}^{p} \beta_j \, x^{(j)}} \tag{3}$$

It is easy to replace the kernel with other kernel options that we are familiar with, such as polynomial or Gaussian kernels.

4.2.2 Clinical kernel for Cox regression

In order to apply kernel Cox regression models on clinical data to obtain better predictive performance, it will make more sense for us to utilize a new kernel, instead of the linear kernel, since clinical datasets are quite different from gene expression profiles, in terms of containing continuous, ordinal, and categorical variables. Linear kernel in this situation has some disadvantages. Researchers have proposed clinical kernel and clinical polynomial kernel for overall survival analysis, which were shown to have improved performance, comparing to survival SVM [4]. The clinical kernel is an additive summation of p kernels $K^{(p)}$, with each calculated differently based on the types of clinical variables.

$$\boldsymbol{K_{clin}(x_i, x_j) = \sum_{p=1}^{d} K_{clin}^{(p)}(x_i^p, x_j^p)} \tag{4}$$

For continuous and ordinal clinical variables, the kernel is defined as [5]

$$\boldsymbol{K_{clin,1}^{(p)}(x_i^p, x_j^p) = \frac{\left(max^{(p)} - min^{(p)}\right) - |x_i^p - x_j^p|}{max^{(p)} - min^{(p)}}} \tag{5}$$

Here $max^{(p)} and\ min^{(p)}$are the maximal and minimal values of clinical variable $\boldsymbol{p}$, of the training data.

For nominal variables, the kernel is defined as

$$\boldsymbol{K}_{\boldsymbol{clin},\mathbf{2}}^{(\boldsymbol{p})}(\boldsymbol{x}_{\boldsymbol{i}}^{\boldsymbol{p}}, \boldsymbol{x}_{\boldsymbol{j}}^{\boldsymbol{p}}) = \begin{cases} \mathbf{1}, \boldsymbol{if}\ \boldsymbol{x}_{\boldsymbol{i}}^{\boldsymbol{p}} = \boldsymbol{x}_{\boldsymbol{j}}^{\boldsymbol{p}} \\ \mathbf{0}, \boldsymbol{if}\ \boldsymbol{x}_{\boldsymbol{i}}^{\boldsymbol{p}} \neq \boldsymbol{x}_{\boldsymbol{j}}^{\boldsymbol{p}} \end{cases} \tag{5}$$

They also proposed clinical polynomial kernel, based on clinical kernel, in the same ways generating polynomial kernel from linear kernel [4].

We applied functions from the R package “survpack” for kernel Cox regression to our clinical data, with linear and Gaussian kernels implemented in the package. We hope to extend this R package to including clinical kernel, which is supposed to be more suitable for incorporating clinical variables in the Cox regression.

4.3 Future work

Currently, efficiently combining multiple types of molecular data remains a technical challenge due to the over-fitting issue and the co-linearity issue of large-scale biological data. Therefore, one future direction is to build prognostic models that better integrating clinical variables and multi Omics data, carrying potential key complementary information, for not only dichotomized survival data, but also censored overall survival data. In that regard, our future work is to develop more effective feature selection strategies to overcome over-fitting and co-linearity issues. For dichotomized survival prediction, we should extend our study to other MKL methods, such as SpicyMKL, GeneralizedMKL, and onlineMKL, and also developing our own customized MKL algorithms. For overall survival prediction, based on clinical variables only, we can further explore kernel Cox regression with Clinical kernels.

4.4 Reference

[1] IARC. International Agency for Research on Cancer (IARC). Accessed December 20, 2007b.

[2] Halabi S, et al. Updated prognostic model for predicting overall survival in first-line chemotherapy for patients with metastatic castration-resistant prostate cancer. JCO 2014: 23(7):671-7.

[3] Li, Hongzhe, and Yihui Luan. "Kernel Cox regression models for linking gene expression profiles to censored survival data." *Pacific Symposium on Biocomputing*. Vol. 8. 2002.

[4] Van Belle, Vanya, et al. "On the use of a clinical kernel in survival analysis."*ESANN*. 2010.

[5] Daemen A. and De Moor B. Development of a kernel function for clinical data. In the 31th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), pages 5913–5917, Minneapolis, Minnesota, September 2009.