DEVELOP THE DISEASE SPECIFIC BIOINFORMATICS PLATFORMS WITH INTEGRATED BIOINFORMATICS DATA

Jiannan Liu

Submitted to the faculty of the University Graduate School
in partial fulfillment of the requirements
for the degree
Doctor of Philosophy
in the School of Informatics and Computing,
Indiana University

November 2022

Accepted by the Graduate Faculty of Indiana University, in partial fulfillment of the requirements for the degree of Doctor of Philosophy.

Doctoral Committee

Jingwen Yan, PhD, Chair

Jie Zhang, PhD

August 11, 2022

Kun Huang, PhD

Chi Zhang, PhD

Timothy I. Richardson, PhD

Huanmei Wu, PhD

ACKNOWLEDGEMENT

I would like to express my deepest appreciation to my academic advisors, Dr. Jie Zhang and Dr. Huanmei Wu, without your support and guidance on the projects I have been working on, I would never have theses chievements. Your way of approaching challenges has set up a golden standard for me to follow, not only in work, but also in daily life.

I would also like to express my deepest love and appreciation for my wife, Tianhan Dong, who have helped and supported me during the past five years, without your support and unconditional love, I would never have the power to conquer the challenges I have been through in the last five years.

Finally, I want to thank my family, friends and collaborators for providing me the support during this journey, I will keep on working hard to be better and better.

Jiannan Liu

DEVELOP THE DISEASE SPECIFIC BIOINFORMATICS PLATFORMS WITH INTEGRATED BIOINFORMATICS DATA

With the advance of multiple types of omics technology and corresponding analytical methods, various type of bioinformatic data have become available. Mining and integrating these data for analysis will provide valuable insights for disease mechanism investigation, drug target identification and new drug development. However, most of the omics data are large size, heterogeneous, and complex, it is challenging for biomedical researchers to mine the data for relevant evidence, especially for those with limited computational skills. In this thesis, I aimed to develop disease specific platforms integrated with multimodal bioinformatic data types to provide researchers with strong bioinformatics support. To achieve this goal, I explored advanced transcriptomic data analytical methods and proposed a novel biomarker for the prediction of overall survival of colon cancer patients, then prototyped a user-friendly patient oriented clinical decision support system to provide accurate and intuitive colorectal cancer risk factor assessment. With the experience of the transcriptomic data analytical methods and the web-based application development, I further designed and implemented Cancer Gene and Pathway Explorer which is an integrative bioinformatics webserver that can be used for cancer publication trends investigation, gene set enrichment analysis with integrated data, and optimal cancer cell line identification. Based on the framework of CGPE, I developed another bioinformatics platform focusing on Alzheimer's disease, called Alzheimer's Disease Explorer, which is a first-of-its-kind bioinformatics server, providing rich bioinformatic support from literature, omics and chemical data to facilitate researchers in ND drug development field. By accomplishing a series of work in my thesis, I have shown that integrated disease specific bioinformatics platforms can provide great value to the research community by allowing 1.) fast and accurate investigation of currently available literature, 2.) quick hypothesis generation and validation using transcriptomic datasets, 3.) multi-dimension drug target evaluation and 4) fast querying of published bioinformatics outcomes.

Jingwen Yan, PhD, Chair

Jie Zhang, PhD

Kun Huang, PhD

Chi Zhang, PhD

Timothy I. Richardson, PhD

Huanmei Wu, PhD

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

AD: Alzheimer’s Disease

ADE: Alzheimer’s Disease Explorer

ANN: Artificial Neural Network

CCLE: Cancer Cell Line Encyclopedia

CDSS: Clinical Decisiom Support System

CGPE: Cancer Gene and Pathway Explorer

COAD: Colon Cancer

CRC: Colorectal Cancer

CTD: Comparative Toxicogenomics Database

DE: Differentially Expressed

FAP: Familial Adenomatous Polyposis

GDC: Genomic Data Commons

GEO: Gene Omnibus

GPP: Gene Publication Profile

GSEA: Gene Set Enrichment Analysis

HIPAA: Health Insurance Portability and Accountability Act

HPC: high-performance computers

HTML: Hyper Text Markup Language

HTT: High Throughput Technologies

ML: Machine Learning

MVC: Model View Controller

ND: Neurodegenerative Disease

NGS: New Generation Sequencing

NLP: Natural Language Processing

PCA: Principal Component Analysis

RF: Random Forest

RNA-seq: RNA sequencing

SID: Score Indicator Dashboard

SVM: Support Vector Machine

TCGA: The Cancer Genome Atlas

TDL: Target Development Level

TF: Transcriptional Factor

Chapter 1. Introduction

1.1 Background

The dramatic advance of sequencing technology in recent years takes the big data in biomedical field into a new level. Various types of omics data provides researchers with rich information on solving challenging biology problems. In fact, significant amount of bioinformatics discoveries have been made based on the analysis of omics data. However, the data size and heterogeneity of omics data make it difficult for biological researchers to utilize the available data without extensive training in programming and other data mining skills, thus the use of omics data is largely in the hands of bioinformatic researchers [1]. The collaboration of bioinformatic researchers and experimental researchers is still the predominant way of transferring data-inferred knowledge to experiment based studies [2]. The status of omics data utilization created two major issues. One is the process may result in repeatitive analysis that is already performed by other researchers. This issue is especially common when bioinformatic researchers perform analysis using public domain data. Another issue is for some general and easy-to-replicate data analysis, the turnaround time during collaboration usually cannot satisfy the experimental researchers' urgent needs. With these two major issues, it will be of great help if an integrated bioinformatics platform can be developed, which not only integrates the analysis results from general bioinformatics analysis (such as differentially expressed gene analysis), but also can be used to preform customizable analysis using integrated datasets. Thus, the objective of my dissertation focuses on developing an integrated bioinformatics platform that can significantly eliminate the resource waste and improve the efficiency of knowledge transfer between computational and experimental researchers [3].

To develop the proposed disease specific bioinformatics platform, four progressive steps are involved in this dissertation: 1) apply advanced analytical methods to transcriptomic data to address challenges in colon cancer research; 2) develop web-based applications to assist with patient-oriented colorectal cancer risk assessment; 3) prototype a bioinformatic data portal integrating cancer-related key bioinformatics resources and knowledge to provide bioinformatics support in cancer research; 4) reuse 3)'s disease-specific bioinformatics framework to develop a new web data portal for Alzheimer's disease and other neurodegenerative disease research.

1.2 Objectives

1.2.1. Apply advanced analytical methods to transcriptomic data to address challenges in colon cancer research

Transcriptomic data have been a valuable resource to advance biomedical research in various disease due to the fast development of sequencing technology in the past decades. With large amount of available transcriptomic data, mathematical algorithms have been utilized for uncovering insights of disease development and treatment using transcriptomic datasets. With the advance of analytical methods applied to transcriptomic data, a few mathematical analyses have gain tremendous popularity across most fields of biomedical research. Such as the differentially expressed gene analysis and gene set enrichment analysis. In the first objective of my thesis, I explored several popular mathematical methods in bioinformatics area and proposed a novel transcriptomic biomarker to address challenges of colon cancer patient overall survival prediction [4].

1.2.2 Develop web-based applications to assist with patient-oriented colorectal cancer risk assessment

With the advance of web-based application technology in recent years, biomedical researchers started to implement system that incorporate advanced algorithms to directly benefit general populations health status, the clinical decision support system (CDSS) is one many examples. In the second aim of my thesis, I have investigated, designed, and implemented a CDSS prototype which is patient-oriented and can provide risk factor assessment of colorectal cancer with patient's inputs [5]. This study incorporated many advanced web application development technologies to enhance the user experience and improve the interpretability of the risk factor assessment results. It provided valuable insights on integrating advanced prediction model with intuitive user interface to make cutting edge research outcomes more accessible and beneficial to patients.

1.2.3 Prototype a system integrating key bioinformatics resources and knowledge to provide bioinformatics support in cancer research

Cancer prognosis and treatment is still one of the most challenging problems in biomedical research [6, 7]. Plenty of databases, tools and algorithms are developed to address various challenges related to cancer research. Because of the well-established databases and data resources, cancer research is the best focus for me to prototype the disease specific bioinformatics platform. Within this objective, following steps have been implemented: Firstly, I collected and integrated heterogenous biomedical datasets, these datasets include but not limited to bulk RNA sequencing data, single cell RNA sequencing data, biomedical publication data, etc. Secondly, I connected the data sets with existing bioinformatics tools and algorithms and pipelines, such as Gene Set Enrichment Analysis (GSEA) [8], natural language processing pipelines, etc. Thirdly, I captured downloaded and processed bioinformatics information such as cell line information, DepMap data, etc. Finally, I have designed and implemented visualizations for various results integrated or generated from the platform [9].

1.2.4 Reuse disease specific bioinformatics framework to provide bioinformatics support for Alzheimer's Disease research

To fulfill this objective, I used the framework generated from aim 3 to develop a bioinformatics platform to help with the Alzheimer's disease (AD) research. Since cure or even interrupt the progression of AD remains to be a unresolved problem [10], providing an AD specific and integrated bioinformatics platform to serve the AD research community will benefit the knowledge integration and potentially promote the discovery of new AD drug targets. This new AD bioinformatics platform incorporates the generalized information from mining PubMed database using NLP pipeline, it also integrates several key multi-omics datasets and process data to help with broadcasting processed bioinformatics outcomes.

1.3 Significance

With the diverse and ever-expanding data in biomedical research, it becomes more and more challenging to grasp the useful information from multiple resources to support ongoing biomedical research. An integrated disease specific platform will help the targeted research community to easily access various data resources in bioinformatics field, these resources include but not limited to natural language processing (NLP) based text mining results, transcriptomic data analysis with various tools and algorithms, analytical results from published studies and drug target profiling information.

As new algorithms and analysis pipelines are published every day to uncover insights from omics data. With the ever-expanding bioinformatics data and algorithms, the work of applying algorithms to data are generally in the hands of people who have computational skill to do programming. The bioinformatics platform implemented in my thesis created a framework to integrate data with algorithms, so that the platform is expandable, scalable, and flexible. 1.) Expandable: the platform can easily incorporate both new omics datasets and new analysis pipelines. 2.) Scalable: depends on the needs of users, the platform can be deployed to single server for running simple analysis or it can be deployed to cloud based systems, HPC systems to handle computation intense analysis. 3.) Flexible: with the framework of the platform as the skeleton, the content of the platform can be easily replaced depending on the users' needs.

By analyzing the literature data and transcriptomic data, then develop the disease specific bioinformatics system, my thesis generated several valuable research outcomes and integrated bioinformatics platforms that have been used by thousands of researchers to guide their on-going research projects. The outcome of my thesis has a significant impact on allowing biomedical researcher with limited data processing ability to easily access data sources in biomedical field.

1.4 Innovation

My thesis project generated a set of innovative outcomes that will benefit the future development of disease specific bioinformatics platforms, the innovations of my thesis project are mainly from these two aspects:

In my thesis, I have proposed an innovative method for discovery transcriptional factor based biomarkers for colon cancer patients, which not only utilized traditional mathematical models such as cox regression but incorporated machine learning methods to enhance the prediction power of patients overall survival. I have also generated two specific bioinformatics platform prototypes that has the capability no other bioinformatic tools or platforms can provide. These capabilities include but not limited to incorporate processed bioinformatics analysis outcomes, run customizable bioinformatics analysis with integrated dataset, provide accurate publication survey guidance according to research needs. With all these capabilities, the prototype platforms addressed most of the basic needs of biomedical researchers that cannot be fulfilled by current available systems.

A generalized system framework is utilized in aim three and four to guide the development of the corresponding disease specific bioinformatics platform. To the best of our knowledge, the system framework used by CGPE and ADE is the first one to address the development structure of integrated disease specific bioinformatics systems, and it will be the first framework that can be used as the guidance of developing future bioinformatics platforms.

1.5 Organization of the Study

With the overall goal of my thesis, four progressive objectives are designed to investigate the proposed topic. The first chapter of my thesis gives general introduction to the background of my thesis including a brief introduction to each objective of the thesis, significance and innovation of the thesis work. In the Chapter 2, I studied methods of analyzing transcriptomic datasets to address biomarker discovery challenges with colon cancer patients. This step helps me to get better knowledge on how bioinformatics studies are conducted using omics datasets. The Chapter 3 introduced a patient oriented clinical decision support system for colorectal cancer patients, which provided easy-to-use colorectal cancer risk factor evaluation for general population. This chapter provides me the knowledge on developing advanced web-based application using biomedical information. The Chapter 4 described a bioinformatics platform that integrated multiple resources to help with cancer research, the platform addressed literature survey, transcriptomic dataset analysis and gene-based cell line evaluation using multiple criteria. This chapter serves as the first prototype of the disease specific bioinformatics system that tries to address the full research life cycle support of cancer research. The Chapter 5 introduces an upgraded and more advanced bioinformatics platform that focuses on Alzheimer’s Disease by utilizing the framework mentioned in Chapter 4. This chapter experimented and incorporated more advanced functions in the disease specific bioinformatics platform, it successfully validated that the generalized framework used in chapter 4 can be used as a universal guideline for developing disease specific bioinformatics platforms.

Chapter 2. Apply Advanced Analytical Methods to Transcriptomic Data to Address Challenges in Colon Cancer Research

2.1 Introduction

Colon cancer is the sixth in men and the fifth in women the most common cause of cancer-related death globally. In the United States, colon cancer is estimated to have 135,430 newly diagnosed cases and result in 50,260 deaths in 2017, accounting for 9% of cancer deaths [11]. Colon cancer is a complex disease with many risk factors, such as genetics, lifestyles, and dietary habits. Among them, inherited gene mutation, which can pass through family members, is one critical factor to increase one's colon cancer risk. A common colon cancer feature is the intra-cancer heterogeneity, which makes patients distinctive from each other in clinical presentations and responses to treatment. Colon cancer treatments should be tailored based on the individual's risk factors and genetic factors.

The inherited colon cancers can be broadly classified into two categories: familial adenomatous polyposis (FAP) and hereditary nonpolyposis colorectal cancer [12]. Molecular features in the genomics level play an essential role in treatment decision making and will continue providing more insights for pathological classification and tailored treatment for colon cancer. Proper colon cancer classification will significantly improve the survival rate, but hinders considerably by limited available prognosis assays.

Among the genetic factors, transcription factors (TFs) play a vital role in most important cellular processes, such as cell development, response to inner and outer environment change, cell cycle controls, and carcinogenesis. TFs are proteins that control the transcription of fragment DNA to messenger RNA by binding to specific DNA regions [13]. Their functions are to regulate, turn on and off genes to make sure that genes expressed in the right cell at the right time and in the right amount throughout the life of the cell and the organism [14]. For example, the NF-κB comprises a family of five TFs that form distinct protein complexes, which bind to consensus DNA sequences at promoter regions of responsive genes regulating cellular processes. NF-κB signaling and its mediated transcription play a critical role in inflammation and colorectal cancer development [15]. STAT3 is reported constitutively activated in colon-cancer-initiating cells and play a significant role in colon cancer progression [16]. FOXM1 was another TF that had been reported to be a key regulator of cell cycle progression, inflammation, tumor initiation and invasion [17].

In the past two decades, many researchers have implemented machine learning (ML) methods in the discovery and validation of cancer prognosis, especially after the population of High Throughput Technologies (HTTs) [18]. Recently, Long Nguyen Phuoc, et al. [19] developed a novel prognosis signature in colorectal cancer (CRC) by implementing several ML methods on public available CRC omics data. Their results demonstrated that the random forest method outperformed other ML methods they tried. Some researchers focused on microRNAs to find cancer prognosis signatures. Fatemeh Vafaee, et al. [20] proposed a prognostic signature of colorectal cancer comprising 11 circulating microRNAs. They also tested several different ML methods including RF and AdaBoost in their study. Their performance of the proposed prognostic signature was confirmed by an independent public dataset. Similarly, Jian Xu, et al. [21] developed a 4-microRNA expression signature for colon cancer patients by using the data from The Cancer Genome Atlas (TCGA). Their study showed that this 4-microRNA signature might play an important role in cancer cell growth after anti-cancer drug treatment. In 2016, Guangru Xu, et al. [22] discovered a 15-gene signature that could effectively predict the recurrence and prognosis of colon cancer using a Support Vector Machine (SVM) algorithm. Their study pointed out that some genes in this signature might be an indicator of new therapeutic targets. Although these previous studies implemented machine learning methods on the discovery of cancer prognosis signatures, the crucial role of TFs has not been sufficiently addressed in cancer prognosis signature development.

The goal of our study is to identify the fundamental transcript factors, which are associated with clinical outcomes of colon cancer patients, by implementing an innovative cancer prognosis signature discovery process that combines the random forest algorithm with classic Cox Proportional Hazard (Cox PH) method. Our study will emphasize on only using TFs expression data to conduct prognostic analysis and we will provide a new perspective on how we can better use gene expression profiles to conduct prognostic research. By using proposed workflow, a TFs based prediction model has been successfully developed for colon cancer prognosis. The prediction power of our model is validated on hundreds of colon cancer patient samples available in the GEO database [23]. Our TF-based colon cancer prognosis prediction model can be used for a better classification of colon cancer patients in survival. Successful findings of this study will shed lights on understanding the mechanisms of the underlying colon cancer development and metastasis.

2.2 Methods

2.2.1 Data sources

In this study, we are using the expression data of TFs from two public resources. One is TCGA colon cancer (COAD) dataset, which can be downloaded from UCSC Xena (http://xena.ucsc.edu) [24] for both the expression dataset and the clinical data of patients. There are 497 samples in the COAD dataset, including 456 primary cancer tissue samples and 41 adjacent normal tissue samples. The downloaded TCGA level 3 RNAseq data is in the log2(counts + offset) format. The TCGA COAD dataset is used as the training set in this study to build the predictive model for the colon cancer prognosis. Only patients carrying a primary tumor with the overall survival times and events were included in the training dataset. Then we further filtered the dataset by excluding patients who have missing information in cancer stage and other clinical information including sex and age. Finally, 435 patients with primary cancer tissue information were remaining in the training TCGA dataset,

The second public expression data resource is the microarray data from GEO database, which will be used to validate our prediction model. We chose four Affymetrix Human Genome U133 Plus 2.0 Array microarray study as validation datasets. The accession numbers, sequencing platform information, and sample sizes of each GEO dataset used in this study were listed in Table 1. The respective clinical data were retrieved from published literature. The GEO dataset also filtered similarly to the TCGA COAD dataset with the survival events and times. In the end, the total number of GEO samples we used for prediction model validation is 1,584. Before performing further analysis, the Affymetrix microarray data were normalized using the Robust Multi-array Average (RMA).

Table 1 Summary of the general clinicopathologic characteristics of patients in both training and testing datasets.

Characteristic	TCGA (N=435)	GSE39582 (N=563)	GSE17536 (N=177)	GSE37892 (N=130)	GSE17537 (N=55)
	N (%)	N (%)	N (%)	N (%)	N (%)
Age(years)					
Median	66	68	66	68	62
Range	31-90	22-97	26-92	22-97	23-94
<65	166 (38.2)	211 (37.5)	78 (44.1)	54 (41.5)	32 (58.2)
⩾65	269 (51.8)	351 (62.3)	99 (55.9)	76 (58.5)	23 (41.8)
Sex					
Male	202 (46.4)	309 (54.9)	96 (54.2)	69 (53.1)	26 (47.3)
Female	233 (53.6)	253 (44.9)	81 (45.8)	61 (46.9)	29 (52.7)
T Status*					
T1-2	86 (19.8)	56 (9.9)	*NA*	*NA*	*NA*
T3-4	345 (79.3)	483 (85.8)	*NA*	*NA*	*NA*
N Status*					
N0	254 (58.4)	299 (53.1)	*NA*	*NA*	*NA*
N1	100 (23.0)	133 (23.6)	*NA*	*NA*	*NA*
N2	78 (17.9)	98 (17.4)	*NA*	*NA*	*NA*
M Status*					
M0	318 (73.1)	479 (85.1)	*NA*	*NA*	*NA*
M1	60 (13.8)	61 (10.8)	*NA*	*NA*	*NA*
MX	47 (10.8)	2 (0.4)	*NA*	*NA*	*NA*
Stage					
I	73 (16.8)	32 (5.7)	24 (13.6)		4 (7.3)
II	167 (38.4)	262 (46.5)	57 (32.2)	73 (56.2)	15 (27.3)
III	124 (28.5)	204 (36.2)	57 (32.2)	57 (43.8)	19 (34.5)

IV	60 (13.8)	60 (10.7)	39 (22)	17 (30.9)

*T status: describes the size of primary tissue and whether it has invaded nearby tissue. N status: describes nearby lymph nodes that are involved. M status: describes distant metastasis.

As shown in Table 1 for the summary of the training and testing datasets, there are substantial similarities upon patient diagnosed age, gender and in the AJCC staging level. The consistency in the pathology levels renders convincing for further analysis without bias or overfitting.

2.2.2 Workflow of the study

The overall workflow of our study is demonstrated in Figure 1, which can be classified into three stages: TFs Screening, Predictive Modeling, and Model Validation. In Stage 1, we first identified a complete list of human TFs with official annotation from previous publications. Since not all the human TFs have the expression data in TCGA COAD dataset, the overlapped genes between TCGA COAD dataset and the complete list of TFs identified. Among the overlapping TFs, we further narrow down the numbers of TFs by the Cox PH Model analysis, which resulted in a limited set of TFs. Cox PH model is a widely used and performance proved statistical model in prognostic signature construction [18].

In Stage 2, since there are still too many colon prognosis TFs (more than 20 TFs), we need to decrease the final prognosis TFs to build a valid and good performance prognosis signature. The ensemble learning method, random forest method, is performed to refine further and reduce the TFs. Based on the RF training results, the most significant TFs are selected based on the top feature importance of RF. With the final TF list, we trained a predictive model for colon cancer prognosis using Cox PH regression.

Stage 3 is the validation of the predictive model. First, the prediction power is tested by accuracy analysis. Furthermore, the predictive model is validated on colon cancer datasets, collected from GEO database, including 925 samples from 4 studies. The Gene Set Enrichment Analysis (GSEA) [25] was also conducted to obtain further insights into our prediction model in the pathway level.

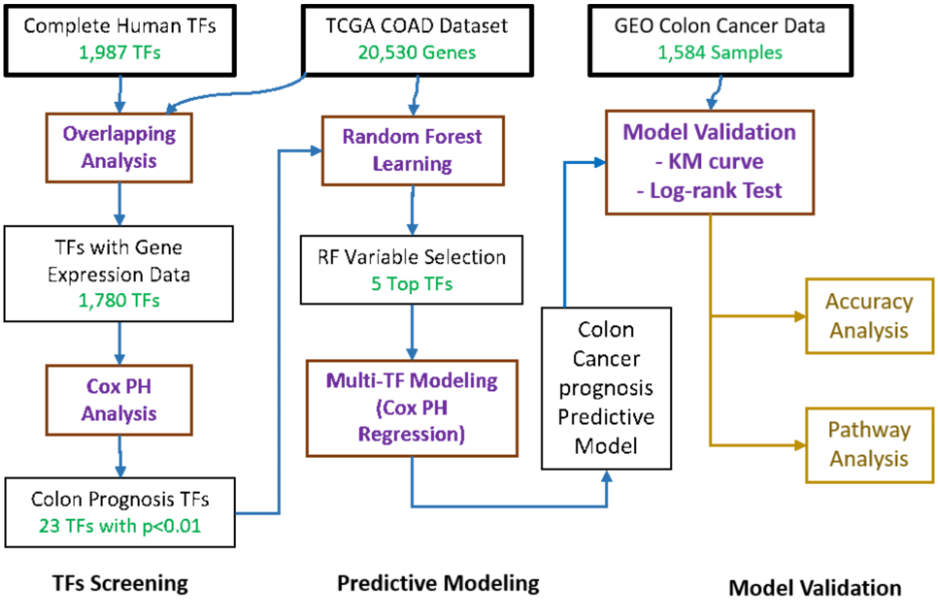


Figure 1. Workflow of the study.

2.2.3 Details on the Variable Selection and Survival Analysis Methods

In Stage 1 of the variable selection, we used the univariate Cox PH model in the statistical environment R (v3.4), the association between expression profiles of TFs and the overall survival of patients was calculated to identify the prognostic ones. Any TF with a p-value less than 0.01 was considered statistically significant and used for further investigation.

In Stage 2 of refining variable selection, we performed RF methods for variable selection given that RF can be used for both classification problems and regression problems. RF [26] is an ensemble algorithm that use a bagging method to combine the multiple decision trees. It draws a set of samples from the whole dataset with replacement to feed the decision tree. After one decision tree has been trained, another sample set will be drawn from the whole dataset to train another decision tree. The process is repeated in the RF algorithm until the desired number of decision trees are trained. The final output of the prediction RF model can be the average of each decision tree' output. In cancer prognosis signature discovery practice, RF is a performance proved method [19, 20, 27]. In our study, the *randomForestSRC* for survival package [28] was used to measure the importance of each variable's contribution to the overall survival of colon cancer patients. This package uses minimal depth variable selection. The algorithm is the termed RSF-Variable Hunting [29]. It exploits maximal subtrees for effective variable selection in survival data scenarios. In our implementation, the parameters used in the feature selection RF model were ntree = 1000 and nstep = 5.

In Stage 3, for the validation of the predictive model, the Kaplan-Meier (KM) curve [30] was used to estimate the difference in the survival between high and low risk groups in validation datasets. The log-rank test [31] was conducted to test the significance of the difference between subgroups since the log-rank test is a very robust statistical method to test important differences between two groups and is widely used in clinical trial experiments.

2.3 Results

2.3.1 The results of identifying the potential prognostic transcription factors

The complete list of 1,987 human TFs was downloaded based on the census of human TFs from the Nature Review Genetics paper by Vaquerizas, Juan M., et al. [32]. Among the listed human TFs, 1,834 of them have gene symbols annotations. After mapping to TCGA COAD dataset, only 1,780 TFs have gene expression data in TCGA COAD dataset, which were included in this study.

The univariate Cox PH regression was applied to the gene expression profiles for the overlapping 1,780 TFs and the patient clinical data in TCGA colon cohort, to identify the TFs, which are associated with the survival of the patients and have the potential using as prognostic markers. Those TFs with $p \leq 0.01$ were kept for further analysis (The selected 23 TFs are listed in Supplementary Table S1).

2.3.2 Results on building the multi-TF predictive model

To identify the minimum subset of TFs that can still achieve a good prediction of colon cancer survival, the 23 TFs from the Cox PH regression model were further evaluated with a random forest algorithm, *randomForestSRC*. In the *randomForestSRC* variable hunting mode, top P ranked variables will be selected, P is the average model size and variables are ranked by frequency of occurrence. In our study, five TFs (*i.e.*, HOXC9, ZNF556, HEYL, HOXC4, and HOXC6) were chosen for the final predictive model construction. The results of the algorithm is shown in Figure 2. The parameters for random forest are ntree = 1000 and nstep =5.

To establish a multiple molecular based regression model, the multivariate Cox PH regression was trained with gene expression data using the five TFs and clinical variables from TCGA COAD dataset. The coefficients from the Cox model were then applied to a multivariate linear regression model. The risk score was calculated with the following formula:

Risk score= 0.139*HOXC6 - 0.046*HOXC4 + 0.165*HEYL + 0.106*ZNF556 - 0.032*HOXC9

The final coefficients of the model have been modified automatically to achieve better performance and to increase accuracy overall. Thus, the coefficients of HOXC9 and HOXC3 are adjusted to slightly below zero, which are much smaller than those positive coefficients. Then we performed the KM analysis and the log-rank test result over these five selected TFs. The results and the p-value from previous Cox PH analysis, along with the hazard ratio for each of these genes are summarized in Figure 3. It can be seen that all selected 5 TFs has Cox p-value < 0.01, which indicates all these TFs are highly related to the overall survival of patients according to Cox PH analysis. For the log-rank p, only the ZNF556 has a p-value of 0.107, while all the other four have p-value < 0.05. According to the RF results, the importance of ZNF556 is ranked fourth in all 23 TFs with no significant difference with other TFs in maximum depth (Figure 2), this qualifies the ZNF556 as one of the most important prognostic TFs. The Hazard ratios of all these five TFs are more than 1.0, indicating higher risks of colon cancer prognosis.

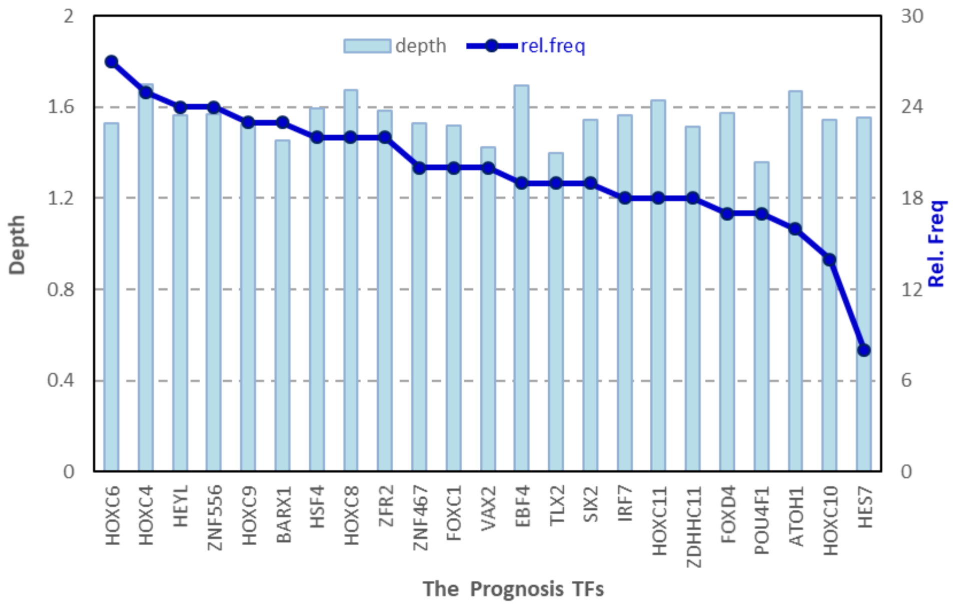


Figure 2. The RF results of the prognosis TFs for the Depth and relative frequency.

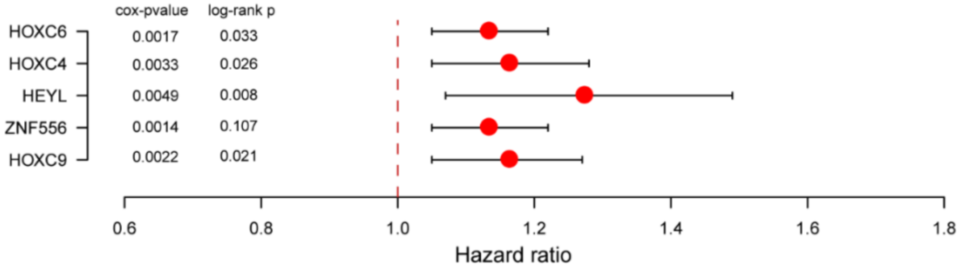


Figure 3. Information on five prognostic TFs finally selected for building the prediction model.

2.3.3 Results on validation of the five-tf based prediction model

Based on the median value of the predicted risks scores of all the patients in both the training and validation set, patients are classified into high-risk and low-risk subgroups. KM curve analysis and log-rank test were conducted to evaluate the performance of predicting power in colon cancer prognosis on TCGA COAD dataset. The results are shown in Figure 4. The scatter plot (Figure 4A(b)) shows the distribution of patients' overall survival status. The red point indicates the patient belonging to a high-risk group while a blue point indicates the patient belonging to a low-risk group. From the scatter plot, we can observe that the red points are more concentrated in the lower part of the figure. This is an indication that high-risk patients have a shorter survival time comparing to low-risk patients. The heatmap (Figure 4A(c)) shows that the five selected TFs in our predictive model were highly expressed in TCGA COAD dataset. Moreover, the KM curve (Figure 4B) shows a distinctive survival difference between the high-risk and low-risk groups in a time span of more than 10 years. All these results prove the prediction power of our predictive model on TCGA COAD dataset.

To test the five-TF based signature as colon cancer survival predictor, we further validated the predictive model on another four independent microarray datasets with a total of 1,584 samples for GEO with GSE39582 (n=563), GSE17536 (n=177), GSE37892 (n=130) and GSE17537 (n=55). The risk score of each patient in validation dataset was calculated by using the same formula established with TCGA training dataset. The same coefficients were utilized to assign weight to each of the selected TF. By using the same median cutoff strategy to divide patients to the high-risk and low-risk groups, the KM curve analysis shows the consistent patterns with the TCGA COAD dataset. Patients in the high-risk group have a significantly shorter survival time than patients in the low-risk group (Figure 5A–D), which suggests the clinical robustness among multiple centers. Therefore, our five-TF based signature is proved to be a robust predictor for colon cancer survival.

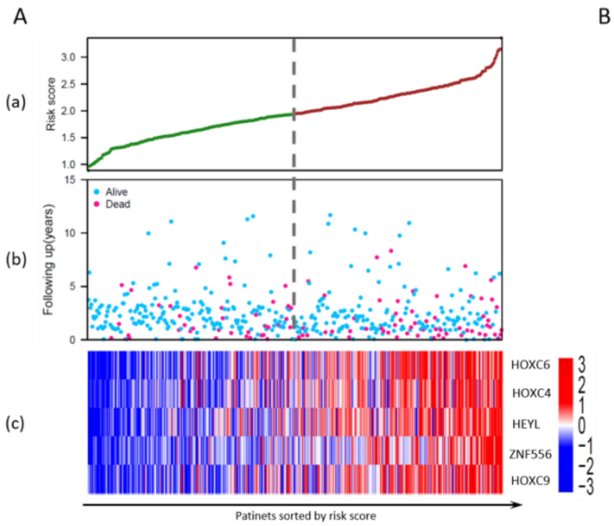


Figure 4. A multivariate linear regression model based on expression of five TFs.

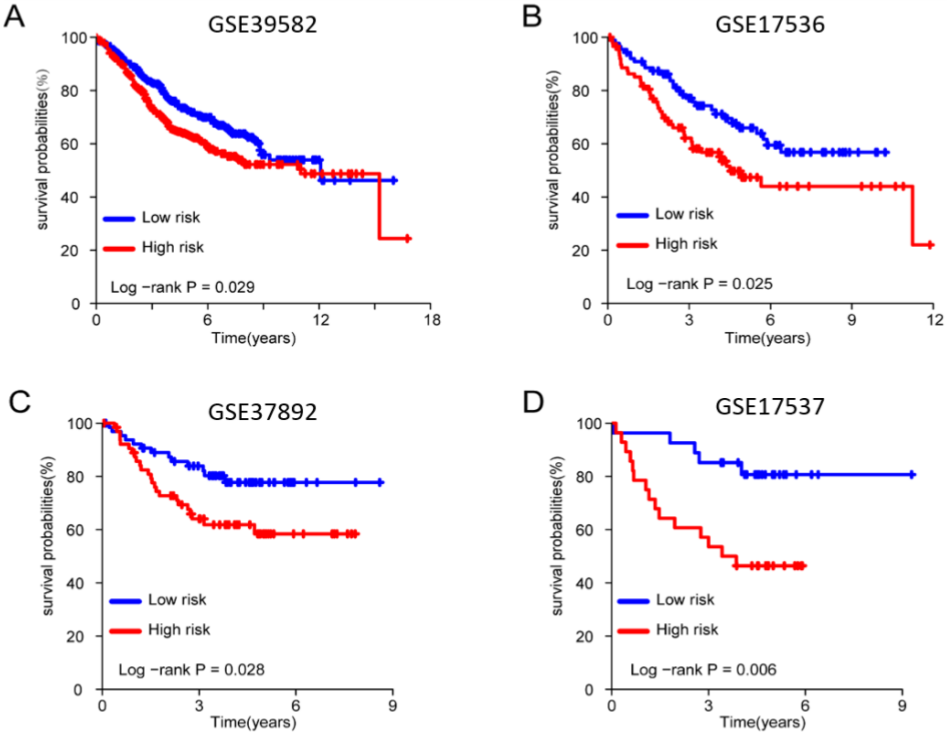


Figure 5. The KM curves of the overall survival probabilities for four independent validation datasets for predicted high-risk subgroups and low-risk subgroups.

2.3.4 Results on pathway analysis

The Gene Set Enrichment Analysis (GSEA) [25] was conducted to investigate the biological function of this five-TF based signature, including its molecular function and gene-gene network. GSEA is performed on the TCGA COAD dataset with predicted high-risk subgroup versus low-risk subgroup. In conducting the GSEA study, the reference gene pathway database is the Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway database[33]. The GSEA number of permutations is set to be 1000, and the phenotype labels are determined according to whether a patient is in the high-risk subgroup or the low-risk subgroup. As illustrated in Figure 6, the GSEA results showed that several cancer-related pathways were alternated in patients with high-risk scores, such as the pathways for the Epithelial-mesenchymal transition, the ECM receptor interaction, the cytokine-cytokine receptor interaction, and the cell adhesion molecules (Figure 5A-D). Taken together these findings, it’s indicated that the five TFs in our model may highly associate with tissue morphogenesis, intercellular regulations and cell adhesion. By affecting these cell processes, these TFs may promote the tissue malignant then result in a poor overall survival rate of colon cancer patients.

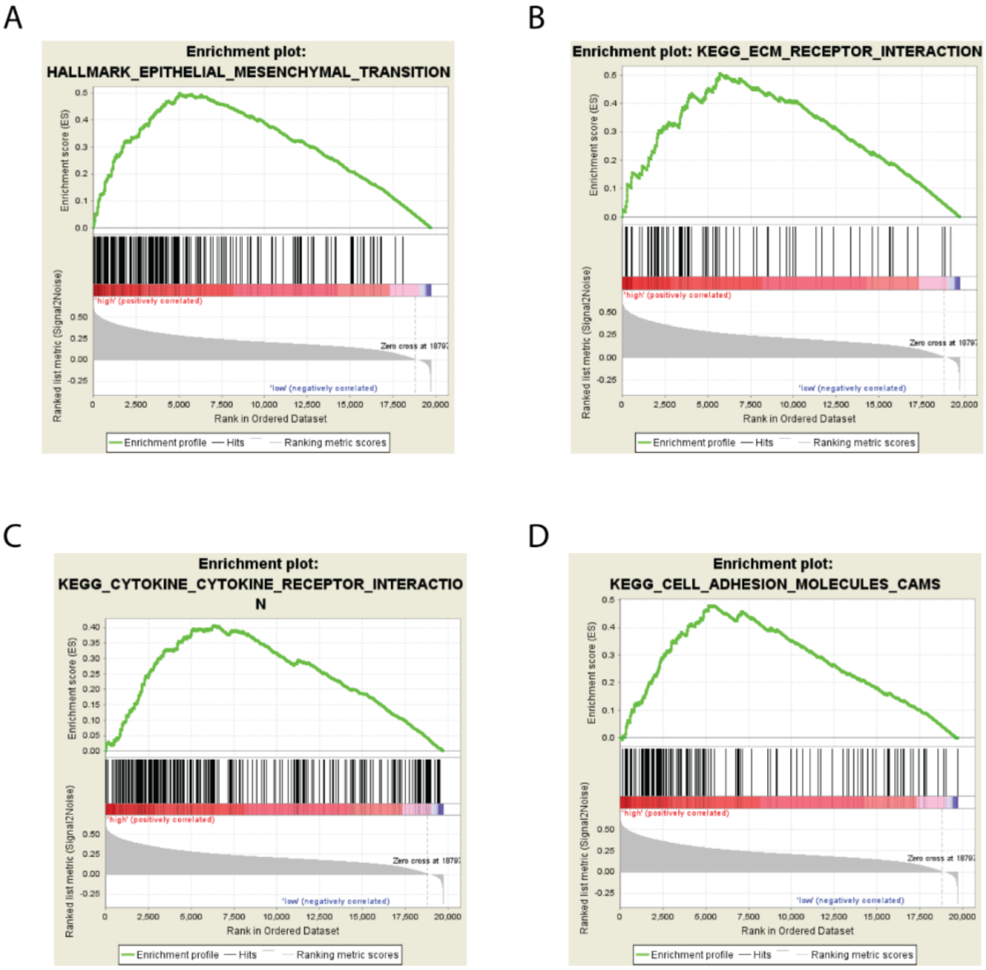


Figure 6. Enrichment plots for the top four enriched gene pathways according to the GSEA results.

2.4 Discussions

We implemented an innovative machine learning approach for signature variables, which combines the Cox PH method with the random forest algorithm. Our signature selection process can find the minimum subset of TFs to build the prognosis prediction model with satisfying performance. A five-TF predictive model was developed by training the classifiers on TCGA COAD dataset. The trained multivariable linear predictive model was validated with multiple datasets from the GEO database.

Three out of the five selected genes, *i.e.*, HOXC4, HOXC6, and HOXC9, belong to the *homeobox* family of genes. The homeobox genes are highly conserved TF family and play an essential role in morphogenesis in all multicellular organisms. Dysregulation of HOX gene expression implicated as a factor in malignancies, and up-regulation has been observed in malignant prostate cell lines and lymph node metastases [34]. HOXC6 was also reported to be overexpressed in colorectal cancer tissue, and highly correlated with poor survival outcome and acts as a significant prognostic risk factor [35].

For the other two genes selected in our predictive model, HEYL belongs to the hairy and enhancer of split-related (HESR) family of basic helix-loop-helix (bHLH)-type transcription factor. A recent study shows that HEYL may be a tumor suppressor of liver carcinogenesis through upregulation of P53 gene expression and activation of P53 ‑ mediated apoptosis [36]. ZNF556 belongs to zinc finger protein (ZNF) family. Despite the large size of ZNF gene family, the number of disease-linked genes in this family is very small [37]. To the best of our knowledge, the research on ZNF556 related to cancer is very limited. Therefore, our study provided new insight on potential relationships between overexpression of ZNF556 and the development of colon cancer.

Our study also showed that by using TFs to build a predictive signature for colon cancer prognosis is practical. The prediction power of the model is promising. Intuitively, the TFs have the overall control on the gene expressions in cells so that a TF-based predictive model should be able to indicate the different gene expression levels in some cancer types with high accuracy.

Our innovative signature discovery process can potentially be extended on other cancer types such as breast cancer or lung cancer. It will be interesting to carry out studies on whether these five TFs used by our model have tissue specific expression patterns in colon cancer. Moreover, by conducting downstream analysis such as gene regulation network analysis, we can probably identify genes that are regulated by our five TFs, these downstream genes can be potentially added to the prediction model to add more robustness to our model. Another future study is to examine the performance of combining traditional statistical methods, such as Cox PH, with other machine learning methods, such as the artificial neural network (ANN) [38], to select potential prognostic TFs or other signatures for different types of cancer.

2.5 Conclusion

We have successfully identified a five-TF signature and built a predictive model for colon cancer prognosis signature with the selected five TFs by using a machine learning approach. Our five-TFs based linear model was validated on hundreds of publicly available patient data from the GEO database. The results showed that our model has a good predicting power in predicting colon cancer overall survival. Our predictive model and its biological functions would provide more insights in the precision treatment of colon cancer, which leads to further investigation on these five TF genes and their roles during the development of colon cancer at the molecular level.

Chapter 3. Develop Web-Based Applications to Assist with Patient-Oriented Colorectal Cancer Risk Prediction

3.1 Introduction

Colorectal Cancer (CRC) affects caecum, colon, and rectum, which is the third leading cause of cancer death among men and women in the United States [39]. The lifetime risk of developing CRC is about 1 in 21 (4.6%) for men and 1 in 24 (4.2%) for women. It is estimated to have 135,430 new diagnosed cases and result in 50,260 deaths in 2017, accounting for the 9% of cancer deaths. The mortality rates have been decreasing for several decades because of changes in risk factors such as the introduction and dissemination of screening tests, and improvements in treatments [40-42]. Statistics showed that between 66% and 75% of CRC cases could be avoided with a healthy lifestyle [43] and appropriate dietary changes. Regular physical activities and maintenance of healthy weight could substantially reduce the morbidity and mortality associated with colorectal cancer [44]. There are many researchers worked on CRC risk factors and CRC risk scores calculations [45, 46]. One is the absolute risk score calculation model by Andrew et al.'s [46] to be discussed further, which is adopted in our work.

However, the public knowledge on the significance of CRC is limited. Many do not recognize the significant impact of lifestyle on the development of CRC. It is the essential motivation for this project on constructing the patient-oriented CDSS. Currently, CDSS is serving an important role in patient visits, it was reported that 30% of annual US patient visits will use Electric Health Report (EHR) systems and 57% of EHR involved patient visits will use CDSS [47]. Several CDSS features such as automated decision support as part of workflow, provision of recommendations, have been proved to improve patient care significantly [48]. Previous study has also shown that 92% of existing CDSS enrolled physicians as primary users [49], the number of patient-oriented CDSS is very limited and there is no research on patient-oriented CDSS specialized in CRC. Therefore, a CDSS for CRC risk assessment, education, and preventative care, that are not only open to the public access but also connected to EHR system will play a critical role in the preventative care of CRC patients.

3.2 Methods

3.2.1 The underlying algorithms of the CDSS

The group first conducted literature reviews of the potential CRC risk factors, CRC risk score calculation, and CRC screening approaches. We adopt the absolute risk score calculation model [46] for the CRC risk score calculation in our CDSS. It used population-based case-control studies as source data to train a prediction model for estimating the risk of developing CRC in a certain period (e.g., 10 or 20 years). In this model, the projected probability will be the absolute risk score with a confidence interval of 95%. Eq. 1 summarizes the primary components of their model:

$$\begin{aligned} absolute\, risk &= f_1(relative\ risk\ parameters) + f_2(age\ specific\ cancer\ hazards) \\ &\quad + f_3(attributable\ risks) \end{aligned}$$

(Eq. 1)

The detailed mathematic model and risk factor coefficients have been explained in Freedman’s report [46]. The relative risk parameters are estimated from population-based case-control data. Sample risk factors include the numbers of relatives with CRC, the patient physical activity, smoking habit, diet preference, body mass index, and others.

The $f_1()$ function will calculate the relative risk based on tumor sites, including the proximal (cecum through transverse colon), distal (splenic flexure, descending, and sigmoid colon), and rectal (rectosigmoid junction and rectum) tumor sites. The $f_2()$ is a function to predict the CRC risk based on different ages and risk factor profiles. The $f_3()$ function will assess the attributable risks from the case-control data., The baseline age-specific cancer hazards and attributable risks are all estimated from the case-control data. The final CRC absolute risk predicted by this model combines the three absolute risks (proximal, distal, rectal) and risks of competing causes of death other than CRC. A SAS Macro program which implemented the proposed model is publicly available online. This program eased our effort on integrating the absolute risk score calculation model into our CDSS.

In our CDSS implementation, we adopted the 20-year absolute risk score as the projected risk score. We then rescaled the absolute risk reported by Andrew's model to a range of [0, 10] based on the maximum and minimum risk scores. Based on the risk scores, the CRC risks are classified into three levels, according to a previous study [9] by Jane et al. The low-risk level, medium risk level, and high-risk level, reported from our CDSS system, are corresponding to the scaled risk score ranges of [0, 3], (3, 7], (7, 10], respectively. For example, if the rescaled risk score is higher than 7, our CDSS will report high-risk score.

Our developed CDSS also provides the recommended CRC screening methods. Information on multiple screening methods that are suitable to the identified CRC risk factors are gathered. The screening method details, such as the performance complexity and test time intervals, are stored in the backend database and used for giving screening recommendations to patients based on their risk factors. The recommendation algorithm is a simple structured decision tree [50]. For instance, if a patient reports that he/she has inflammatory bowel disease, the decision tree will report Fecal Immunochemical Test (FIT) as one of the recommended screening methods because of its low complexity, low side effect, and low cost.

3.2.2 The application framework

Figure 7 illustrates the system infrastructure for the prototype of the patient-oriented CDSS. It is developed using Django, a Python-based Model-View-Controller (MVC) web application development framework. An MVC framework separates application functionalities into three domains. The *models* describe the data structures of the backend database. The *views* display application outputs and collect inputs, which can consist of several files, such as HTML, CSS, JavaScript, and others. The *controllers* define the internal logic of the application. It is also responsible for data processing [51]. In our CDSS, we use the D3 data visualization package to visualize the risk score data as bullet chart and create the interactive dashboard [52].

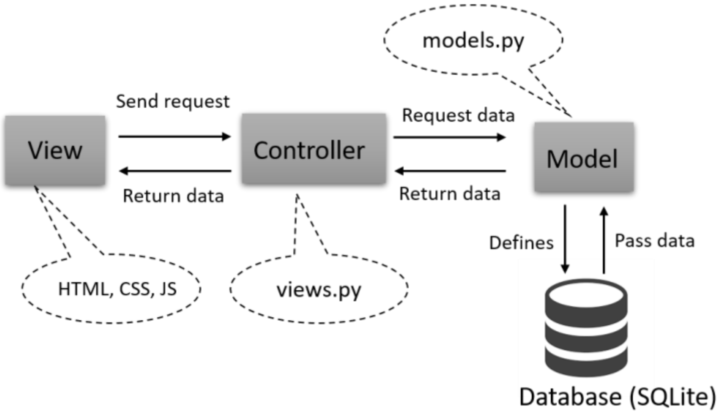


Figure 7. The Django MVC framework.

3.2.3 The backend database

We use MySQL [53] as the backend database. Figure 8 shows the primary data structure. The *User* table stores the information of the CDSS users. When the CDSS is connected to an EHR system, the user information can be transformed to a patient table in the EHR system. The *asmt_results* table is the main component that stores the assessments patients have done. Since our CDSS will give recommendations on CRC screening methods, the *result_scrn_test* table serves as a relation table, which represents a many-to-many relationship between assessment results and screening methods. All the detailed information (e.g., test name, test time interval, test performance, etc.) on screening methods will be kept in the *scrn_tests* table. The *asmt_questionnaire* table defines questionnaire title and theme. In each questionnaire, there will be several sections, which contain some similar type of questions. The *asmt_sections* table describes the section title and its preferred style. All the questions will be stored in the *asmt_questions* table. Each question has a status attribute, which has a potential value of active or disabled. This attribute will help CDSS administrator add or delete questions for each questionnaire easily, making the database design more flexible and extensible. Options for each question will be kept in the *asmt_options* table, with a type attribute to indicate the type of input (such as a radio button or a text input) and a value indicating the risk score for each risk factor. The full list of attribute descriptions can be found in Appendix 1.

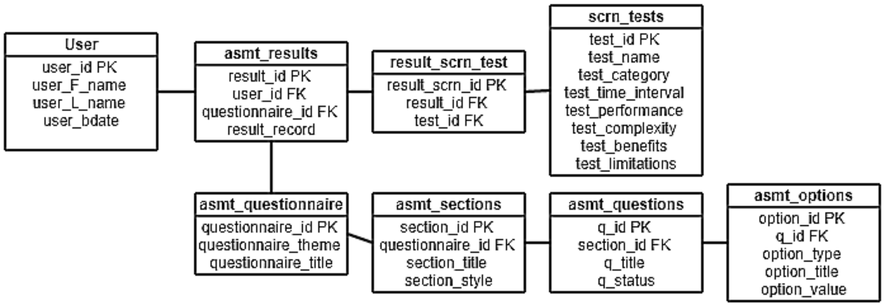


Figure 8. The entity relationship diagram for the backend database.

The data structure of our database keeps all the relative information used by our CDSS. It also keeps the flexibility of changing questions and options in the questionnaire. By designing such a database structure, we have also maintained the flexibility for CDSS upgrades in the future.

3.2.4 The website design

Figure 9 demonstrates the design and workflow of the CRC CDSS website. The green-colored textboxes indicate a webpage in the CDSS. Other boxes describe the content of the pages. Our CDSS prototype has the following components and primary CDSS functions. The first is an interactive website with an anonymous scientific questionnaire to obtain the information about essential CRC risk factors. These questions are designed according to previous studies on CRC risk factors [46]. The second is a user-friendly display module with the risk scores calculated based on the input risk information. The third innovative component is an interactive visualization dashboard to show how changing lifestyle habits and diet preferences will affect their CRC risk level. The visualization is personalized based on the user input to the survey questions. Fourth, we incorporate a CDSS module to provide individualized recommendations on screening methods based on survey results and risk scores. The fifth is an appointment scheduling system with CRC providers based on user preferences on doctor characteristics and geographical locations. Last, the CDSS provide educational information on CRC preventative care.

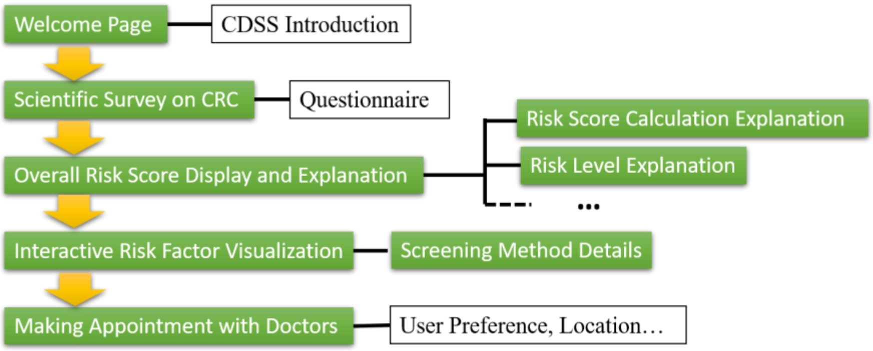
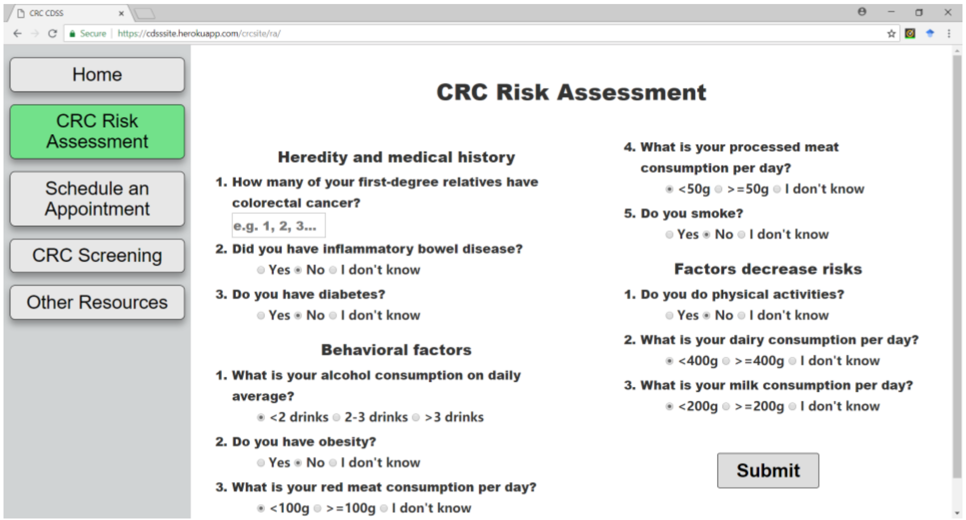

Welcome Page
CDSS Introduction
Scientific Survey on CRC
Questionnaire
Risk Score Calculation Explanation
Risk Level Explanation
Overall Risk Score Display and Explanation
...
Interactive Risk Factor Visualization
Screening Method Details
Making Appointment with Doctors
User Preference, Location…

Figure 9. The design and workflow of the CRC CDSS website.

3.3 Results

We have designed an interactive website to provide an easy-to-use questionnaire for potential risk factors of the users, as illustrated in Figure 10. After completing the survey, the CDSS will first report the overall risk score with a bullet chart to visually display the users their risk levels, as shown in Figure 11. The CRC risk level has different highlighted colors according to different risk levels. Several useful links are provided to help the users to understand the risk scores and risk levels.



CRC Risk Assessment

Heredity and medical history

1. How many of your first-degree relatives have colorectal cancer?

   e.g. 1, 2, 3...

2. Did you have inflammatory bowel disease?

   ○ Yes ● No ○ I don't know

3. Do you have diabetes?

   ○ Yes ● No ○ I don't know

Behavioral factors

1. What is your alcohol consumption on daily average?

   ● <2 drinks ○ 2-3 drinks ○ >3 drinks

2. Do you have obesity?

   ○ Yes ● No ○ I don't know

3. What is your red meat consumption per day?

   ● <100g ○ >=100g ○ I don't know

4. What is your processed meat consumption per day?

   ● <50g ○ >=50g ○ I don't know

5. Do you smoke?

   ○ Yes ● No ○ I don't know

Factors decrease risks

1. Do you do physical activities?

   ○ Yes ● No ○ I don't know

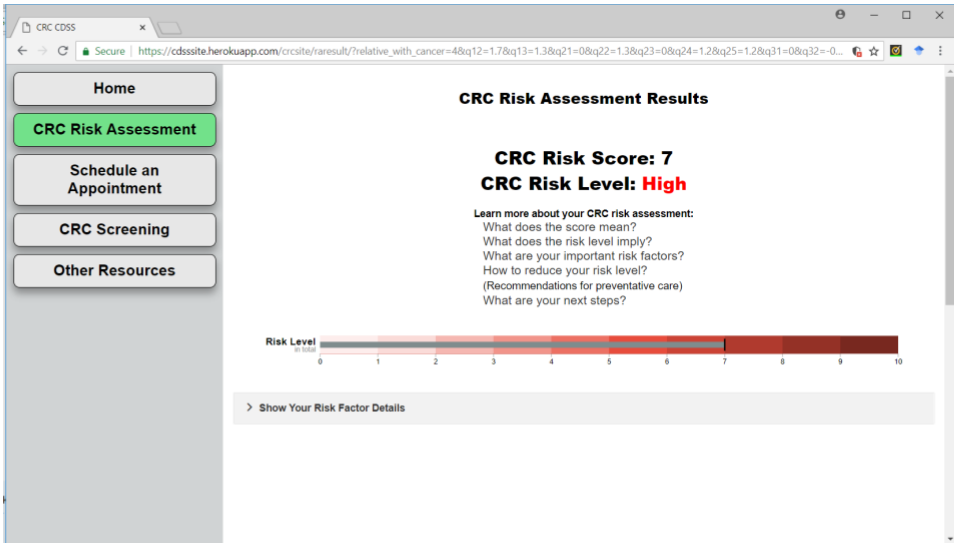
2. What is your dairy consumption per day?

   ● <400g ○ >=400g ○ I don't know

3. What is your milk consumption per day?

   ● <200g ○ >=200g ○ I don't know

Figure 10. The interface for survey CRC risk factors.


CRC CDSS
Secure
https://cdsssite.herokuapp.com/crcsite/raresult/?relative_with_cancer=4&q12=1.7&q13=1.3&q21=0&q22=1.3&q23=0&q24=1.2&q25=1.2&q31=0&q32=-0...
Home
CRC Risk Assessment
Schedule an Appointment
CRC Screening
Other Resources
Risk Level
in total
0
1
2
3
4
5
6
7
8
9
10
Show Your Risk Factor Details

CRC Risk Assessment Results

CRC Risk Score: 7

CRC Risk Level: High

Learn more about your CRC risk assessment:

What does the score mean?

What does the risk level imply?

What are your important risk factors?

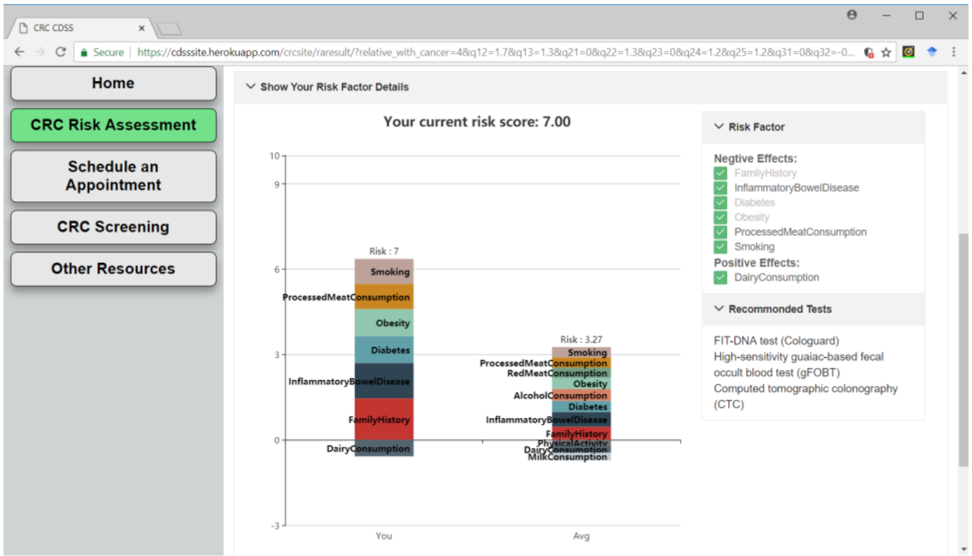
How to reduce your risk level?

(Recommendations for preventative care)

What are your next steps?

Figure 11. CRC risk assessment results.

The salient feature is the interactive stacked bar chart, generated according to user inputs. The stacked bar chart shows the total score on the top of each bar, with different risk factors. A user would be able to understand the impact of each risk factor interactively and visually. As illustrated in Figure 12, the interactive visualization interface will allow a user to modify their preferences on the side buttons to observe the dynamic changes of risk scores. It helps them to decide what healthy lifestyle (such as smoking vs. non-smoking, drinking milk or not) will reduce their CRC risks. On the other hand, risk factors such as family history cannot be modified since a user cannot change this kind of risk factors. Thus, these factors are not clickable.



Your current risk score: 7.00

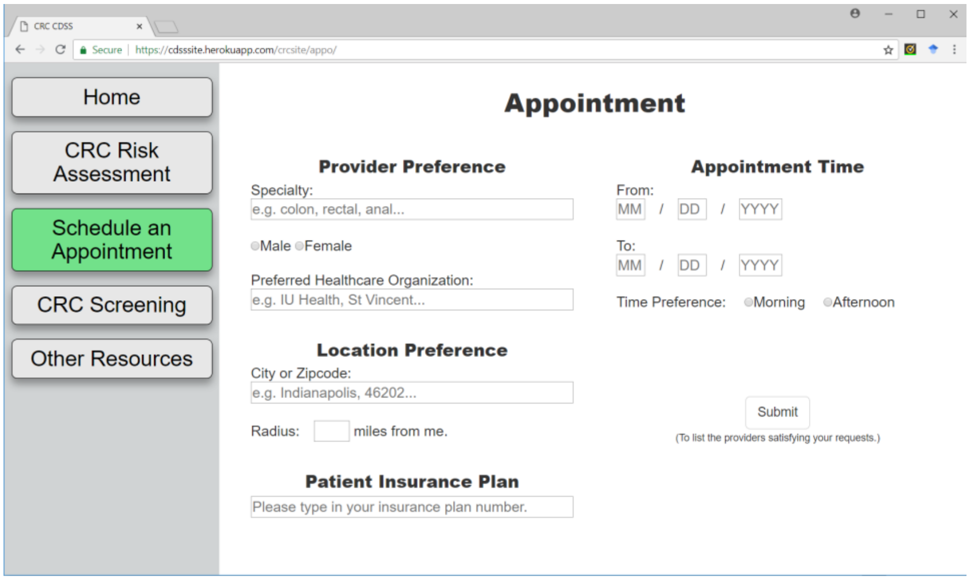
Negtive Effects:
☑ FamilyHistory
☑ InflammatoryBowelDisease
☑ Diabetes
☑ Obesity
☑ ProcessedMeatConsumption
☑ Smoking
Positive Effects:
☑ DairyConsumption

FIT-DNA test (Cologuard)
High-sensitivity guaiac-based fecal occult blood test (gFOBT)
Computed tomographic colonography (CTC)

Figure 12. The interactive visualization of risk factors.

On the side of the stacked bar chart, we provide recommendations for CRC screening methods, which are ranked using the simple decision tree method based on the questionnaire input and total risk score. Every recommendation method is a clickable link, which will lead users to an information page with detailed description on recommended screening method.

To schedule an appointment with healthcare provider, the CDSS provides options for user inputs, such as location and doctor preferences, as manifested in Figure 13. Based on the information, the CDSS offers recommendations on which hospitals or community health centers a user can visit. It also lists the providers available for appointment. This page is designed to connect to hospitals and community health centers' scheduling systems to get information about available doctors. For the testing purposes, we populated the system with sample hospitals in the Indianapolis area and synthesized doctors to simulate and test the system function. The Google map API is applied to achieve the hospital search function based on zip code, as illustrated in Figure 14.



Appointment

Provider Preference

Specialty: e.g. colon, rectal, anal...

○Male ○Female

Preferred Healthcare Organization: e.g. IU Health, St Vincent...

Location Preference

City or Zipcode: e.g. Indianapolis, 46202...

Radius: miles from me.

Patient Insurance Plan

Please type in your insurance plan number.

Appointment Time

From: MM / DD / YYYY

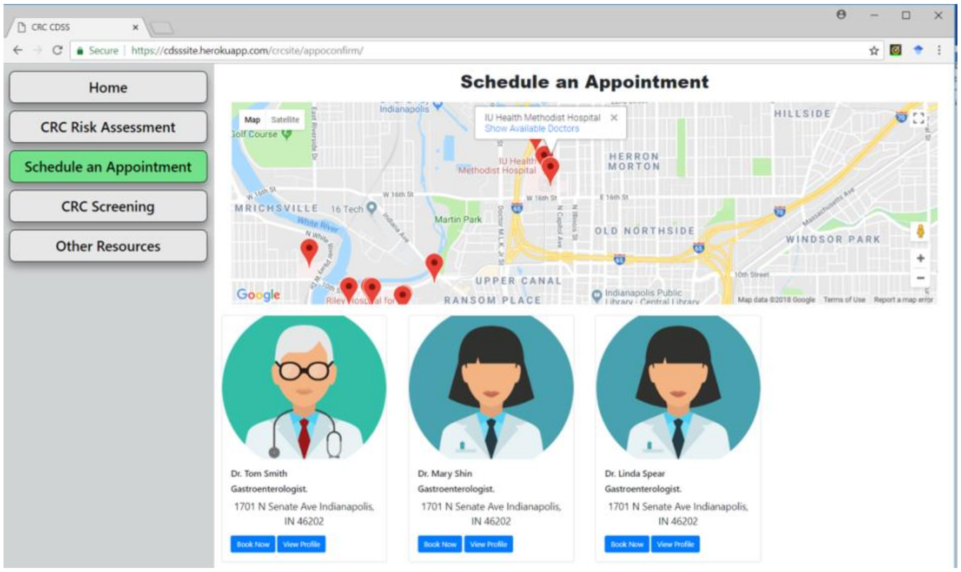
To: MM / DD / YYYY

Time Preference: ○Morning ○Afternoon

Submit

(To list the providers satisfying your requests.)

Figure 13. The dashboard for making an appointment.



Schedule an Appointment

Dr. Tom Smith
Gastroenterologist.
1701 N Senate Ave Indianapolis, IN 46202
Book Now | View Profile

Dr. Mary Shin
Gastroenterologist.
1701 N Senate Ave Indianapolis, IN 46202
Book Now | View Profile

Dr. Linda Spear
Gastroenterologist.
1701 N Senate Ave Indianapolis, IN 46202
Book Now | View Profile

Figure 14. The results of the nearby hospital search and doctor availability.

3.4 Discussions

Currently, there are several available online tools for CRC risk score prediction. Colorectal Cancer Risk Assessment Tool (CCRAT), sponsored by the National Cancer Institute (NCI), provides an interactive tool to help estimate a person's risk of developing CRC. It has a well-designed questionnaire to collect related information. The CRC risk calculation also follows the Freedman algorithm [46]. However, the CCRAT only displays the risk calculation results in an absolute percentage, which is hard for users to understand. Second, the simple bar chart for result presentation is the overall risk, lacking the detailed information on various individual risk factors and their impacts on the overall CRC risk. Another CRC risk calculation tool is the Colorectal Cancer Predicted Risk Online (CRC-PRO) [54], which can be used to calculate 10-year CRC risk score. It has an easy-to-use interface. However, the CRC-PRO only presents the calculated probability without any interpretation of the result. It is difficult to interpret the risk calculation results, especially for those with a low literacy level.

By using the Django MVC web application framework, together with the backend MySQL database, our CDSS has the flexibilities and the extensibilities of updating the content and modifying the questionnaire. It also makes the system transformable to other applications. For example, by changing the questionnaire contents and the risk score calculation, we can modify and reposition our CDSS for different cancer types, such as breast cancer and stomach cancer.

With the easy-to-follow design of the CRC risk assessment steps, our CDSS embeds scientific CRC risk score calculations into a user-friendly interface. This feature ensures the accessibility of our CDSS to the low literacy population. In our CDSS, a user does not need to have any prior knowledge of CRC risk factors and screening methods for CRC. The system provides all the information on CRC risks and screening in an intuitive way. The innovative risk factor dashboard with customizable stacked bar chart further facilitates the readability and interpretation of the CRC risk level prediction results.

The capability of online appointment scheduling in our CDSS makes it easier to create a link between our CDSS and any EHR in hospitals. After the user fills the risk factors and receives the CRC risk assessments, they can directly make appointments with proper providers with their specific requirements of locations and preferred provider characteristics. The recommendations of appropriate screening methods will be available to the EHR system with original scientific questionnaire data. This feature could assist care providers preparing better before seeing a patient and making more precise care decisions based on patient-specific health conditions.

Given the current system is only a prototype of the CRC CDSS, there are several future directions can be carried out based on the current system.

- For our system, one crucial ongoing task is to perform a systematic evaluation of the CDSS [55], before implementing into a production version. We will work with a working group of patients, providers, health care organizations, and HIT professionals. Multiple-step evaluation processes will be carried out. For instance, we will follow the Software Development Life Cycle for the development and evaluation of the CDSS [56]. A system-wide review of its performance and stabilities will be assessed by IT professionals. On the other hand, the different influential factors and risk calculation algorithms will be validated and evaluated by CRC experts. Also, the interactive website design, the dashboard visualization, and system usabilities will be evaluated with potential users (including patients and providers) for better user experience.

- For healthcare providers, one potential future improvement of the CDSS is to connect the CDSS with different EHR systems. In this way, our CDSS can support effective adoption and achieve health IT interoperability goals. It is also possible to allow each patient to create the patient account so that the system can provide individualized preliminary CRC risk reports based on our interactive dashboard.
- With the development of omics technologies and genomic data analysis, we can integrate the genetic factors or biological factors into our CRC CDSS to expand the assessment function. For instance, based on the gene expression profile, a seven-gene signature has been discovered to predict the overall survival (OS) of CRC patients [57]. We can adopt or modify the survival risk score system to CRC risk score calculation, which could potentially be integrated into our CDSS to improve the accuracy of CRC risk score prediction.

With all these existing features of our CDSS and potential upgrades, we believe our CRC CDSS would be a valuable patient-oriented tool in CRC preventative care field.

3.5 Conclusion

In this study, we have developed a CRC CDSS prototype which gives risk assessment and interactive interpretation of the risk outcomes using innovative data visualizations for personalized CRC screening. The demonstration project is deployed online with Heroku web application deployment platform [58]. The patient-oriented design of our CDSS will help more people to assess their CRC risk and learn more about the significant impact of lifestyle on the development of CRC. Moreover, with the easy-to-follow steps of our CDSS, patients can conveniently build a connection with hospitals and physicians and book their screening test appointments. This feature will make a significant difference in the preventative care of CRC.

Chapter 4. Prototype a System Integrating Key Bioinformatics Resources and Knowledge to Provide Bioinformatics Support in Cancer Research

4.1 Introduction

High-throughput technologies empower researchers to investigate the transcriptional expression data upon hundreds of samples at once. The massive genomic data provides advantages for phenotype marker identification [59], gene pattern discovery [60] and pathway analysis [61]. Researchers have been tried to use microarray and next generation sequencing technology to reveal the alteration of cancer genomics since the technologies just been developed [62]. More and more projects aimed to uncover the trigger of tumor initiation, development and metastasis have already been done or being undergoing, as the Cancer Genome Atlas (TCGA) include more than 11,000 multiple omics data for 33 cancer types. Genomics Evidence Neoplasia Information Exchange (GENIE) [63] includes data for over 80 major cancer types, including data from more than 7,500 patients with lung cancer, nearly 5,500 patients with breast cancer, and more than 5,100 patients with colorectal cancer. There are more and more cancer genomics data been added to the NCI Genomic Data Commons' next generation cancer knowledge network (GDC) [64]. Meanwhile, there are thousands separated cancer study with transcriptional expression data been released through Gene Expression Omnibus (GEO). The Cancer Cell Line Encyclopedia (CCLE) [65] gives a compilation of gene expression and parallel sequencing data from 1019 human cancer cell lines. The accumulation of tens of thousands of cancer and non-cancer samples providing an unprecedented opportunity for many biomedical related fields including cancer biology. However, this large amount of data cannot be easily used by biomedical researcher to extract useful information for guiding their studies due to the lack of professional computer skills, such as large data management, programming skills, supercomputer usage, etc. [66].

There are numerous databases and webservers already developed for downloading and analyzing public gene expression data, such as cBioportal [67], UCSC Xena [68] and GEPIA [69]. The cBioportal integrated more than 5000 tumor samples from more than 20 cancer studies and it provides functions including but not limited to survival analysis, network analysis, correlation analysis and source data download. With the help of cBioportal, biomedical researchers are able to rapidly and intuitively translate large-scale genomics data into biological insights. UCSC Xena is a web-based visual integration and exploration tool for multi-omic data and associated clinical and phenotypic annotations. UCSC Xena helps users to explore functional genomic data sets for correlations between genomic and/or phenotypic variables. GEPIA is another webserver for TCGA data analysis, it provides customizable functions such as tumor/normal differential expression analysis, profiling according to cancer types or pathological stages, patient survival analysis, similar gene detection, correlation analysis and dimensionality reduction analysis. However, there are still many unsatisfied needs from experimental biologists that cannot be fulfilled easily by theses existing tools, these needs include 1.) gene specific research trend inferred from publications, which could help biological researchers to quickly build the foundation of their next step of research, 2.) pathway alterations caused by certain gene or gene signature, which serves as an important step in studying gene functions and mechanism, 3.) choose the optimal cell lines to conduct biological experiments, since more than one thousand cancer cell lines are currently available, traditional way of choosing cell lines may cause resource waste and weak experiment results due to the inter-cell-line heterogeneity [70].

In the present study, we developed a user-friendly cancer gene and pathway investigation tool Cancer Gene and Pathway Explorer (CGPE), to provide a highly integrated webserver for investigating the TCGA and GEO gene expression data. CGPE provides an interactive and customizable analyze portal to address the challenges mentioned above. The CGPE mainly includes three functions, they are gene specific PubMed research trend analysis, gene (or gene signature) associated pathway alteration analysis and cancer cell line selection based on patient genomic data. CGPE aims to deliver the most concerned information to biomedical researchers to help unveiling the potential association from big data cohort view.

4.2 Methods

4.2.1 Data sources

Cancer gene expression datasets and the corresponding clinical data were downloaded from TCGA and the GEO database. Integrated gene expression and clinical data of TCGA were downloaded from the GDC data portal (https://portal.gdc.cancer.gov). In our study, more than ten thousand patients (about 10652 patient) were included. Datasets from the GEO database [71] were another importance source of cancer transcriptional data. Clinical information was extracted from the original publications as well. We totally collected 48 datasets from different platforms (TCGA Hi-seq 2000, Affymetrix HG U133 plus 2, HG U133 A and HG U133B etc.).

The publication data used in CGPE is extracted from PubMed database using Python scripts, the Entrez module of Bio package is used to communicate with PubMed database and parse retrieved data. The extracted PubMed data is processed with 6 steps using natural language processing (NLP) then stored in structured SQLite database. The cell line expression and annotation data are downloaded from CCLE database (https://portals.broadinstitute.org/ccle).

4.2.2 System implementation

The CGPE is implemented with Django web application development framework. Django follows the Model-View-Controller (MVC) model [51], which uses 1) Model that defines the database structure and handles the data flow from database backend, 2)View as the front end to collect user request and display system outputs, 3) Controller that process the user inputs, do analysis and interact with Model. Due to the integration of Gene Set Enrichment Analysis (GSEA) [8] into our system, the single GSEA analysis may take several minutes to be finished. To address the problem of running time takes too long, we implemented an asynchronous task queue by using Celery (http://www.celeryproject.org/) to handle the submitted GSEA tasks, the RabbitMQ (https://www.rabbitmq.com/) is used as a message broker between our main application and Celery task queues. Celery can create as many workers as needed based on CPU availability to handle the submitted GSEA tasks. The system infrastructure is shown in Figure 15. The database backend is implemented with SQLite database (https://www.sqlite.org/index.html), most of the processed data are stored in database including PubMed data, general gene information, cell line annotation data, etc. Large expression data files are store on server as .txt static files.

The font-end of the CGPE is implemented with HTML5, the layout and styling of the webpage is mainly achieved by Boostrap V4 (https://getbootstrap.com). Several

Javascript libraries are used in CGPE project. The jQuery (https://jquery.com) is used to create dynamic functions of the webpage, the ajax method is used to create asynchronous autocompletion in all search boxes. All interactive visualizations in CGPE project is created with D3 library (https://d3js.org). A website structure design is show in Figure 16.

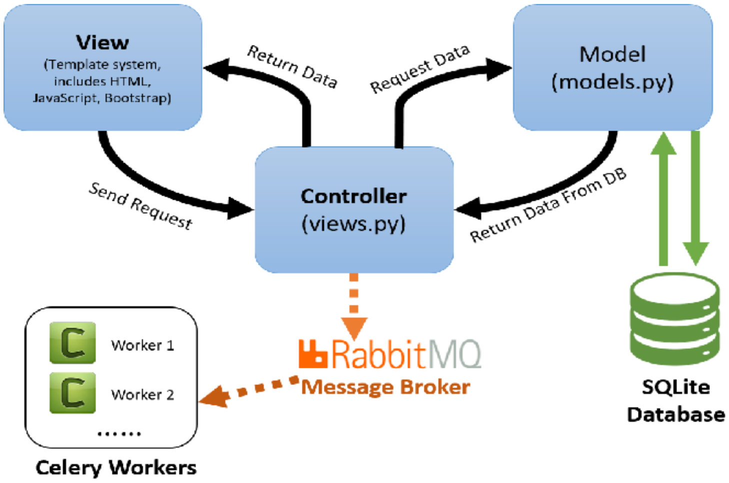


Figure 15. The infrastructure of CGPE project.

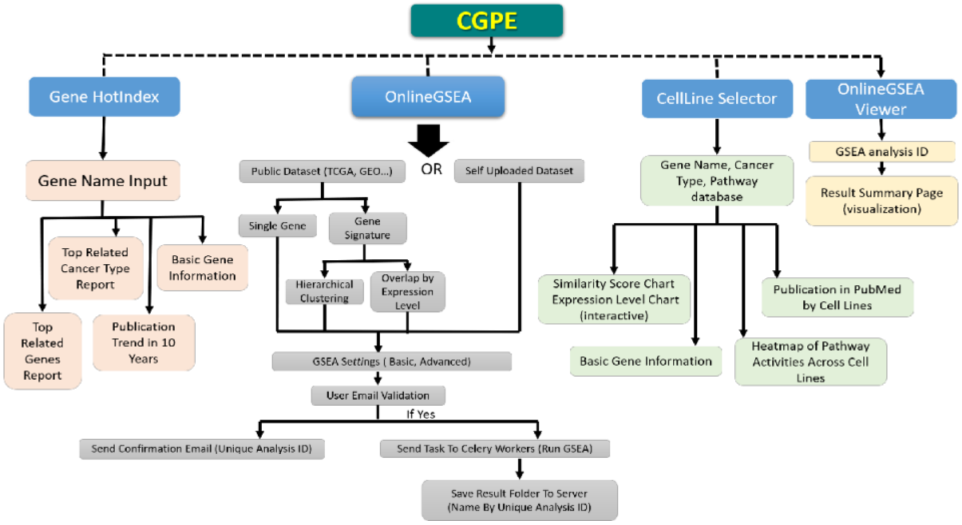

CGPE
Gene HotIndex
OnlineGSEA
CellLine Selector
OnlineGSEA Viewer
Gene Name Input
Top Related Cancer Type Report
Basic Gene Information
Top Related Genes Report
Publication Trend in 10 Years
Public Dataset (TCGA, GEO...)
OR
Self Uploaded Dataset
Single Gene
Gene Signature
Hierarchical Clustering
Overlap by Expression Level
GSEA Settings (Basic, Advanced)
User Email Validation
If Yes
Send Confirmation Email (Unique Analysis ID)
Send Task To Celery Workers (Run GSEA)
Save Result Folder To Server
(Name By Unique Analysis ID)
Gene Name, Cancer Type, Pathway database
Similarity Score Chart
Expression Level Chart
(interactive)
Basic Gene Information
Heatmap of Pathway Activities Across Cell Lines
Publication in PubMed by Cell Lines
GSEA analysis ID
Result Summary Page
(visualization)

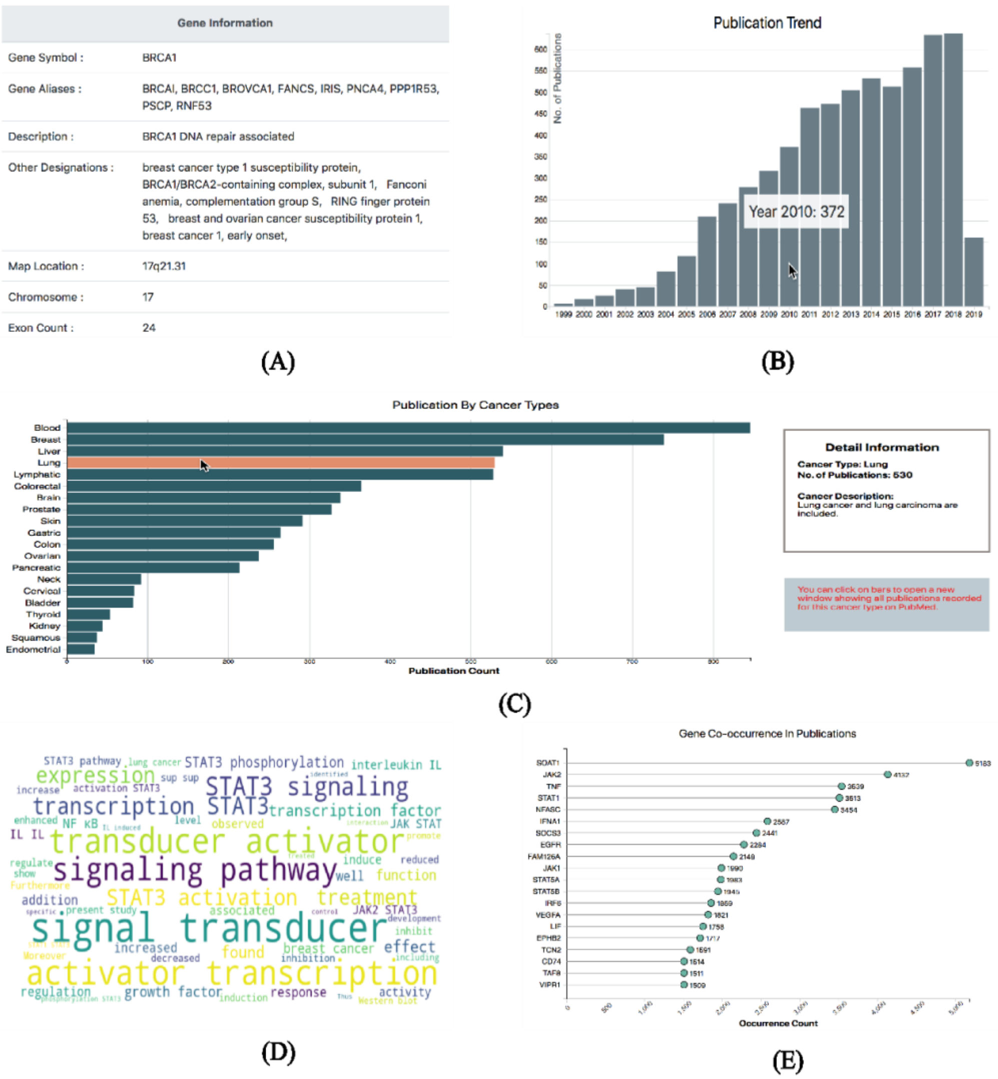
Figure 16. The website structure of CGPE.

4.2.3 CGPE functional modules

Gene HotIndex. Before conducting any gene related cancer research, a common practice for biological researchers would be doing a research survey on related genes of their research topic. PubMed serves as a great resource to find literatures for guiding future research. However, the literature review process is often time consuming and literature search result from PubMed sometimes can be biased and unfocused. Currently there is no tool available to help biological researchers get a public literature overview related to certain gene. The Gene HotIndex, as first functionality of CGPE, is trying to address this issue using natural language processing technology to mine information and categorize gene related publications based on PubMed database. The CGPE provides a simple search box for users to input gene names they are interested in, the acceptable gene names include HUGO gene symbol (e.g. STAT3) and ESEMBL IDs (e.g. ENSG00000168610). During the user input, autocomplete can be triggered to give a recommendation list of genes based on user input. The recommendation list is generated based on pre-processed gene aliases information, all gene's aliases are linked with their HUGO gene symbols and ESEMBL IDs. In the current stage of our application, we included 17813 genes occurred in TCGA datasets.

To prepare the data in Gene HotIndex, we firstly build gene publication profiles (GPP) for each gene. The GPP is defines as a set of publications whose title or abstract mentioned this gene at least once. During the PubMed data collecting process, all genes' aliases are used for searching PubMed database. The dataset of GPPs serves as the raw data in our Gene HotIndex function. After the GPPs are collected and built, we further processed the GPPs in 6 steps, including 1.) GPPs filtering, to filter out publications not mentioning the target gene 2.) WordCloud plots for each GPP 3.) identifying cancer types mentioned in publications of each GPP 4.) identifying gene-gene co-occurrence in each GPP 5.) summarizing year of publications in each GPP. 6.) identifying cancer cell lines mentioned in publications of each GPP.

Based on the processed PubMed data, we created several visualizations using D3 library to help biological researchers to better interpret the research trend of certain gene. The first part of Gene HotIndex search result is the basic information about the searched gene, including official HUGO symbol, aliases, description of the gene, genome map location, etc (Figure 17A). Then a bar chart is shown on the right to show the publication trend by year for the searched gene (Figure 17B). Based on the publication by cancer type data we get from data processing, we created a visualization panel with a horizontal bar char and an information box on the side (Figure 17C). The horizontal bar chart indicates the number of publications related to certain cancer types in this gene's GPP. The horizontal bar chart also has some interactive features. If the mouse is placed on the bar, counts of publications and the description of this cancer type will be shown in the Detail Information box on right side. The bars of the horizontal bar char are clickable, click on the bar will open a new window in browser to show all publications in this category on PubMed website. Then we will display the preprocessed WordCloud plot (Figure 17D) for the searched gene's GPP to show users the most frequent words in GPP, the WordCloud plot can show a general information that inferred from the GPP. Lastly, we include a lollipop plot (Figure 17E) to visualize the gene-gene co-occurrence in GPP, the figure shows the occurrence of related genes based the searched gene's GPP. Generally, this lollipop plot can be used to infer the related genes with searched gene from a perspective of public literatures.



Gene Information	
Gene Symbol :	BRCA1
Gene Aliases :	BRCAI, BRCC1, BROVCA1, FANCS, IRIS, PNCA4, PPP1R53, PSCP, RNF53
Description :	BRCA1 DNA repair associated
Other Designations :	breast cancer type 1 susceptibility protein, BRCA1/BRCA2-containing complex, subunit 1, Fanconi anemia, complementation group S, RING finger protein 53, breast and ovarian cancer susceptibility protein 1, breast cancer 1, early onset,
Map Location :	17q21.31
Chromosome :	17
Exon Count :	24

Detail Information

Cancer Type: Lung
No. of Publications: 530

Cancer Description:
Lung cancer and lung carcinoma are included.

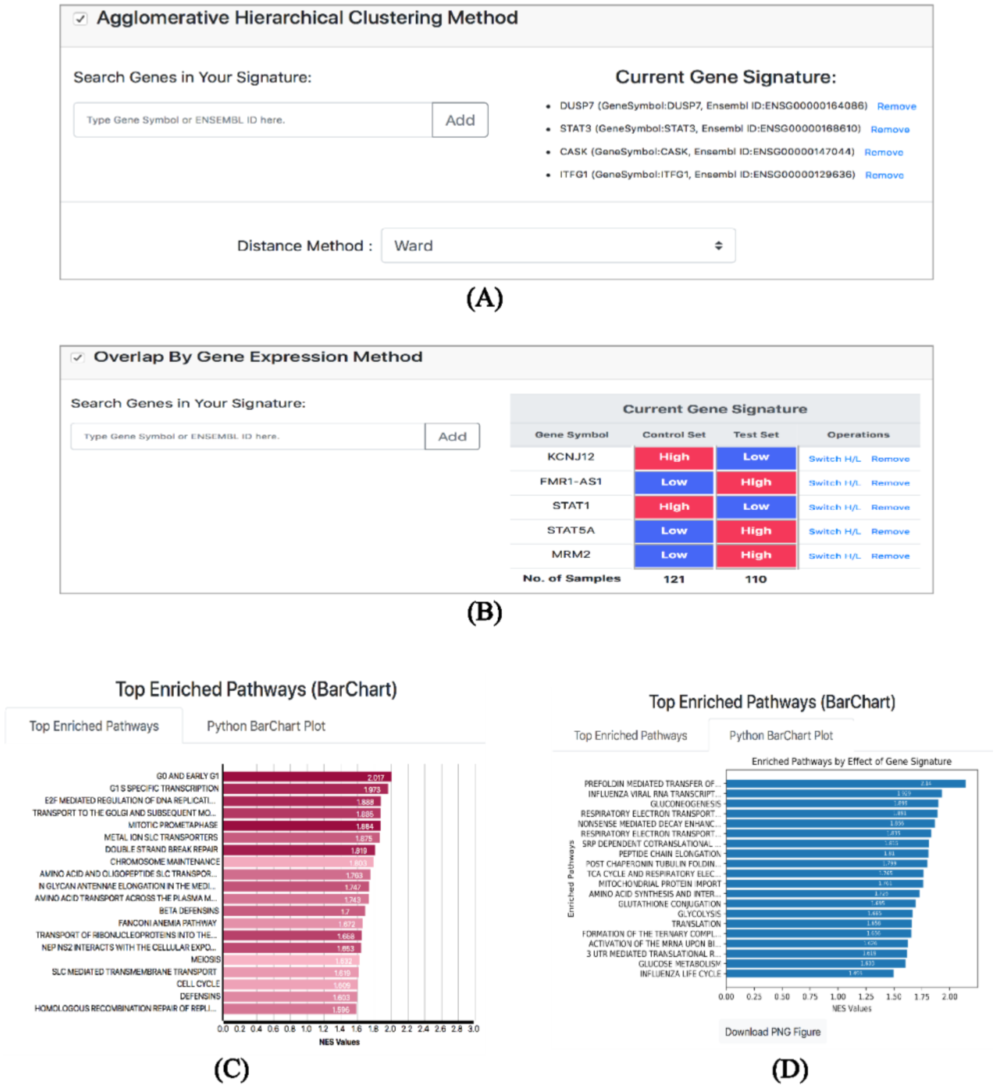
You can click on bars to open a new window showing all publications recorded for this cancer type on PubMed.

Figure 17. Gene information table, publication trend, publications by cancer type visualization, Wordcloud plot and gene-gene co-occurrence lollipop plot.

OnlineGSEA. It is a common goal for the biological research to elucidating the mechanism of gene(s) may have related to the cancer cell development. Typically, the gene sets enrichment analysis will give this kind of clues. In our application we deployed a web-based Gene Set Enrichment Analysis (GSEA) tool which integrated with thousands of publicly available patient samples, the algorithm is a classic GSEA approach, and it is an over-representation analysis method (fisher's hypergeometric test) based pathway enriched approach. This part of CGPE is to provide more convincing evidence to researchers in guiding their gene(s) driven studies. The OnlineGSEA of CGPE allow users to investigate gene or gene signature caused pathway alterations based on large amount of publicly available genomic data. Our OnlineGSEA eased the effort of biological researchers on downloading and processing gene expression data by themselves, it will also be able to significantly accelerate the preliminary gene screening process when conducting gene related cancer research.

The CGPE webserver provides two data source options, one is user self-uploaded data, the other one is publicly available genomic data we collected and process from TCGA and GEO databases. For self-uploaded data, users need to upload the expression data together with the phenotype label file to run the GSEA algorithm, then they can view the analysis result online. For the public available datasets, we downloaded the expression profiles from TCGA or GEO database. The criteria we used for filtering studies in the GEO database is that the number of patient samples in the study should be big enough (normally > 200 samples). Moreover, the study needs to provide clinical information such as survival time and survival status, this is because we will use the survival information to calculate the Hazard Ratio for patient groups to determine the control group and test group in GSEA analysis. In the current CGPE, we included three cancer types (Breast cancer, colorectal cancer and), datasets for different cancer types are displayed on different panels of our public data page, the short descriptions of studies is also provided for users

For investigating single gene caused pathway alterations with publicly available datasets, we divide all patients in a dataset into to two groups by using the median expression value of this gene, then we use Cox Proportional Hazards (Cox PH) regression [72] to estimate the Hazard Ratio (HR) for these two groups of patients. The group of patients with higher HR will be defined as test group when conducting GSEA analysis, the other group with lower HR will be served as control group. For investigating gene signatures (a set of genes), we implemented two methods for dividing patient sample into two groups for GSEA, the first method uses Agglomerative Hierarchical Clustering (AHC) algorithm to separate patient samples (Figure 18A), the second method called Overlap By Gene Expression (OBGE), which allows users to define the control group and test group based on gene expression levels, high and low expression patient sample groups of each gene are defined with median expression value as cutoff (Figure 18B). If user choose to use AHC method, a gene search box is provided on left side of the panel for searching genes in current selected dataset, after clicking the Add button, the gene will be added to the gene signature. Once the Signature is defined and the GSEA parameters are set by the user, CGPE will extract all expression values of genes in the gene signature, then use them as input of AHC algorithm to cluster patients into two groups, the Cox PH regression will also be used to determine control group and test group. Finally, an automated process will generate cls phenotype file for GSEA analysis based on previous steps and run the GSEA on server. If users choose the OBGE method to define the gene signature, a search box is also provided to search genes in current selected dataset. After clicking the Add button, the gene will be added to the gene signature table on right. The first column of gene signature table shows the official symbol of the gene, the second column shows user-defined expression level (high or low) of a gene in control group of GSEA analysis, the third column shows user-defined expression level (high or low) of a gene in test group, the last column allows users to remove the gene from the gene signature or switch the high/low expression level between control group and test group. The last row of the table will always be updated with number of samples left in current control group and test group, if the number of samples in control group or test group falls below 10, operations such as adding more genes and switching high/low will be restricted. After the gene signature is defined and GSEA parameters are set, the CGPE will generate phenotype file based on user-defined gene signature and run the GSEA on server.



☑ Agglomerative Hierarchical Clustering Method

Search Genes in Your Signature:

Current Gene Signature:

- DUSP7 (GeneSymbol:DUSP7, Ensembl ID:ENSG00000164086) Remove
- STAT3 (GeneSymbol:STAT3, Ensembl ID:ENSG00000168610) Remove
- CASK (GeneSymbol:CASK, Ensembl ID:ENSG00000147044) Remove
- ITFG1 (GeneSymbol:ITFG1, Ensembl ID:ENSG00000129636) Remove

Distance Method : Ward

☑ Overlap By Gene Expression Method

Search Genes in Your Signature:

Current Gene Signature

Gene Symbol	Control Set	Test Set	Operations
KCNJ12	High	Low	Switch H/L Remove
FMR1-AS1	Low	High	Switch H/L Remove
STAT1	High	Low	Switch H/L Remove
STAT5A	Low	High	Switch H/L Remove
MRM2	Low	High	Switch H/L Remove
No. of Samples	121	110	

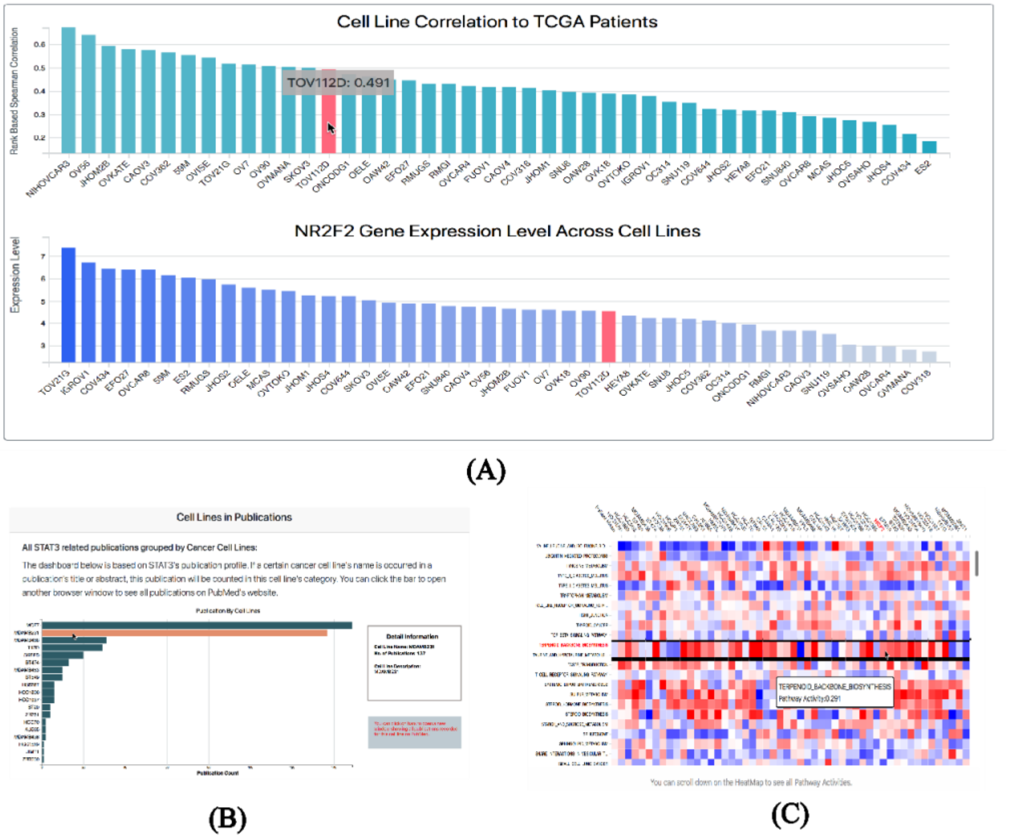
(A)
Type Gene Symbol or ENSEMBL ID here.
Add
(B)
Type Gene Symbol or ENSEMBL ID here.
Add
Top Enriched Pathways (BarChart)
Top Enriched Pathways
Python BarChart Plot
G0 AND EARLY G1 2.017
G1 S SPECIFIC TRANSCRIPTION 1.973
E2F MEDIATED REGULATION OF DNA REPLICATI... 1.888
TRANSPORT TO THE GOLGI AND SUBSEQUENT MO... 1.886
MITOTIC PROMETAPHASE 1.884
METAL ION SLC TRANSPORTERS 1.875
DOUBLE STRAND BREAK REPAIR 1.819
CHROMOSOME MAINTENANCE 1.803
AMINO ACID AND OLIGOPEPTIDE SLC TRANSPOR... 1.763
N GLYCAN ANTENNAE ELONGATION IN THE MEDI... 1.747
AMINO ACID TRANSPORT ACROSS THE PLASMA M... 1.743
BETA DEFENSINS 1.7
FANCONI ANEMIA PATHWAY 1.672
TRANSPORT OF RIBONUCLEOPROTEINS INTO THE... 1.668
NEP NS2 INTERACTS WITH THE CELLULAR EXPO... 1.653
MEIOSIS 1.632
SLC MEDIATED TRANSMEMBRANE TRANSPORT 1.619
CELL CYCLE 1.609
DEFENSINS 1.603
HOMOLOGOUS RECOMBINATION REPAIR OF REPLI... 1.596
0.0 0.2 0.4 0.6 0.8 1.0 1.2 1.4 1.6 1.8 2.0 2.2 2.4 2.6 2.8 3.0
NES Values
(C)
Top Enriched Pathways (BarChart)
Top Enriched Pathways
Python BarChart Plot
Enriched Pathways by Effect of Gene Signature
Enriched Pathways
PREFOLDIN MEDIATED TRANSFER OF... 2.14
INFLUENZA VIRAL RNA TRANSCRIPT... 1.929
GLUCONEOGENESIS 1.895
RESPIRATORY ELECTRON TRANSPORT... 1.891
NONSENSE MEDIATED DECAY ENHANC... 1.866
RESPIRATORY ELECTRON TRANSPORT... 1.835
SRP DEPENDENT COTRANSLATIONAL ... 1.815
PEPTIDE CHAIN ELONGATION 1.81
POST CHAPERONIN TUBULIN FOLDIN... 1.799
TCA CYCLE AND RESPIRATORY ELEC... 1.765
MITOCHONDRIAL PROTEIN IMPORT 1.761
AMINO ACID SYNTHESIS AND INTER... 1.725
GLUTATHIONE CONJUGATION 1.695
GLYCOLYSIS 1.665
TRANSLATION 1.658
FORMATION OF THE TERNARY COMPL... 1.656
ACTIVATION OF THE MRNA UPON BI... 1.626
3 UTR MEDIATED TRANSLATIONAL R... 1.619
GLUCOSE METABOLISM 1.603
INFLUENZA LIFE CYCLE 1.495
0.00 0.25 0.50 0.75 1.00 1.25 1.50 1.75 2.00
NES Values
Download PNG Figure
(D)

Figure 18. (A) User-defined gene signature by using AHC. (B) User-defined gene signature by overlapping high/low expression groups (C) Visualization of NES and p value (D) PNG download of the NES score bar chart.

After the GSEA analysis is submitted, an email with a unique analysis ID will be sent to the user. Once the GSEA is finished, users will be able to use the unique ID to extract the analysis result from the server, we created an individual page called OnlineGSEA Viewer for viewing the GSEA results. The OnlineGSEA Viewer page summarized the most important information provided by GSEA. A horizontal bar chart

(Figure 18C) is generated to visualize the Normalized Enrichment Score (ENS) together with p-values, the length of bars indicate NES and the color of bars indicate p-values, we also provide a PNG figure (Figure 18D) of the horizontal bar chart for users to download. On the OnlineGSEA Viewer page, we also included top 8 enriched pathways' enrichment plots to allow users quickly check the GSEA results.

In the CGPE webserver, the third main function is call CellLine Selector. The CellLine Selector requires three inputs 1.) cancer type, in the current stage of CGPE, we implemented the algorithm on three cancer types in our system 2.) Gene name, the gene's name you are interested in 3.) pathway database, when conducting the GSVA, we are able to use different pathway databases for the analysis, in current stage, we provide two of the most popular pathway databases, KEGG and REACTOME [73, 74]. On the result page of CellLine Selector, the first part shows the searching criteria, the second part shows some basic information about the searched gene, such as gene aliases, map location, exon count, etc. The next part is a visualization panel (Figure 19A) that displays two bar charts, the first bar chart shows the dependency score of the query gene, the second bar chart shows the mRNA expression level of the searched gene across cell lines. When mouse is on one of the bars in the dependency bar chart, a dependency score will be displayed, current bar is highlighted, the same cell line will also be highlighted on the second bar chart which shows the mRNA expression value. The same cell line will be highlighted on the first bar chart if the mouse is on bars of second bar chart. By implementing this visualization, we try to guide users to select cell lines not only using the dependency socre, but also using the mRNA expression level of the target gene. The best cell lines we recommended would be the ones which have low dependency score and at the same time have high mRNA expression level. In the next part of the result page, we created a PubMed visualization panel (Figure 19B) similar with the one on Gene HotIndex result page, but rather than categorize publications in searched gene's GPP by cancer types, the visualization panel on CellLine Selectors result page categorize GPP publications by cell lines. Moreover, in the last part of the CellLine Selector page, we create a heatmap (Figure 19C) to visualize pathway activities across all selected cell lines, each cell of the heatmap corresponds to a pathway's activity of the certain cell line. This heatmap will help researchers, who want to focus on some certain pathways, to choose the cell lines according to their requirements (either high or low pathway activity).



All STAT3 related publications grouped by Cancer Cell Lines:

The dashboard below is based on STAT3's publication profile. If a certain cancer cell line's name is occured in a publication's title or abstract, this publication will be counted in this cell line's category. You can click the bar to open another browser window to see all publications on PubMed's website.

You can scroll down on the HeatMap to see all Pathway Activities.

Figure 19. (A) Double bar charts. (B) Publications categorized by cell line. (C) Heatmap of pathway activities across all cell lines.

4.3 Discussion

By focusing on the preliminary research stage of biomedical field, CGPE integrates multiple data sources such as PubMed, GEO, TCGA and CCLE. The CGPE web server provides experimental biologists with a user-friendly exploratory tool to help their preliminary research. The three-step workflow of CGPE covers the publication survey, patient-based gene (or gene set) function inferring and cell line selection with patient-based evidence. To the best of our knowledge, there is no such bioinformatic tool available to biomedical researchers which mainly focusing on the guidance of preliminary research. The development of CGPE could build another bridge between bioinformatics field and biological research field to convey insights hidden in large amount of public available data to experimental biologists.

In an overall view of CGPE, three main functions provided by CGPE are logically related and cover the three of most important steps before conducting experiments. The first part, Gene HotIndex, summarizes the gene specific publication data and use intuitive visualization to give user a general view of the current research status of the searched gene. Since we used an automated process to mine the information from millions of publications in PubMed, minor errors and noise information still exist in the processed data, further manual checks are still needed. The integration of curated PubTator [75] data into Gene HotIndex might be a future work for the next version of CGPE to provide more accurate gene specific publication summary. The processed gene-gene co-occurrence data could be potentially used with protein-protein interaction (PPI) network to assist with PPI inferrming [76]. The second part, OnlineGSEA, integrates the GSEA algorithm with large amount of public-available gene expression data. Experimental biologists can use single gene or self-defined gene set to investigate gene or gene set functions based on patient data. More cancer types and datasets could be added to CGPE continuously, more gene signature defining options could be provided to allow users conduct more customizable experiments. The third part, CellLine Selector, implemented an innovative algorithm which tries to give gene specific cell line recommendations based on gene expression profiles' similarity between cell lines and TCGA samples. Our algorithm is the first of this kind which gives gene specific cell line recommendations using gene expression profiles, but other type of genomic data, such as copy number variations, mutation, etc. could potentially be added to our algorithm as co-factors of the similarity. It is also worth mentioning that our CellLine Selector aims to provide a new insightful perspective for biomedical researchers to choose cell lines but should not be considered as the only factor when guiding the cell line selection, other factors such as mRNA expression level of interested gene, preliminary research results should also be considered when selecting cell lines.

4.4 Conclusion

The CGPE webserver is a user-friendly, intuitive, and informative bioinformatic tool which allows biomedical researchers to explore large amount of public available bioinformatics data. The CGEP eases the effort of biomedical researcher's effort of collecting, processing, and analyzing the data during the preliminary research phase, it can serve as complements with other powerful bioinformatic tools like cBioPortal and GEPIA.

Chapter 5. Reuse Disease Specific Bioinformatic Framework to Provide Bioinformatics Support for Alzheimer's Disease Research

5.1 Introduction

Neurodegenerative diseases (ND), which include Alzheimer's Disease (AD), Parkinson's Disease (PD), Huntington's Disease (HD), and many others, put a major health threat to the currently aging society, especially to the life quality of elderlies [77, 78]. These NDs share some common features such as aggregation and deposition of abnormal proteins in the brain, which helped researchers to investigate the pathology of ND development and identify potential drug targets[79, 80]. As the advance and prevalence of next-generation sequencing (NGS) technology, multiple omic data in ND fields are quickly accumulated. The single-cell RNA (scRNA) Sequencing technology has further advanced our understanding of NDs development and allowed us to investigate disease pathology with cellular level information[81].

Many ND transcriptomic datasets are available in the GEO database. In addition, a few specially designed databases have been developed to aggregate publicly available scRNA data from AD patients [82-84]. However, the disease etiologies of most NDs are still unclear. Currently, there are not many effective drugs that can cure or even slow down the disease progression in general patients population [85]. To address challenges and promote ND research, a few web servers have been developed recently by leveraging various bioinformatic resources. The Agora, which aggregated AD-related information and experimental resources, provides a platform for the AD research community to propose and identify novel AD targets[86]. The AlzCode integrates rich AD functional genomic datasets and offers a web server for conducting multiview analysis of genes

[87]. The AD Atlas combines over 20 large studies to provide a multi-omics global view of user interested AD research results[88].

These web portals focus on providing a general view of current AD/ND research by integrating and leveraging various resources. However, they also lack some important features for the ND research and drug development field: 1) most of them are impractical to generate an overall picture of a specific gene's role in the ND research field; 2) they did not address a few important aspects in AD/ND research such as literature survey, novel drug target profiling, hypothesis generation and validation using publicly available large cohort data. For the former, those gene-related results are usually scattered in multiple places which makes users difficult to summarize. For the latter, the size and complex structure of public available ND omic data also posed a challenge on biomedical researchers with limited programming skills to fully utilized the data for their hypothesis generation and testing. For example, with thousands of publications in AD/ND research published each year, and their highly diversified and specialized research focuses, it becomes increasingly difficult for researchers to accurately pinpoint the publications related to their research topic without well-annotated publication database. Similar situation happens to the utilization of AD/ND transcriptomic data, the transcriptomic data related to AD/ND research with decent number of samples usually come from a few large studies such as ROSMAP, MSBB, etc., the rest are often deposited in GEO database, these datasets have various NGS pipelines so that only researchers with sufficient NGS knowledge are able to parse these datasets and build the desirable subcohort for their hypothesis generation and testing.

To address these limitations and challenges of full research life cycle bioinformatic support for AD/ND researchers, in this study, we designed and implemented this open web portal Alzheimer's Disease Explorer (ADE), which aims to fulfill the missing links mentioned above to generate a comprehensive picture of a specific gene/gene sets in the related ND field (with a focus on AD), while at the same time provides comprehensive bioinformatic data and analytical tools and druggability information for researchers in the ND research and drug development.

5.2 Methods

5.2.1 System architecture

To accommodate the functional design of the ADE, we firstly designed a specialized web application architecture to support all designed functions in ADE (Figure 1). ADE system mainly consists of three sub-systems, 1) the database system, which oversees the management of all data integrated from various resources and the storage of user generated data such as analysis jobs, it serves as the cornerstone of the ADE system. 2) the job handling system, since several tools are offered for conducting customizable bioinformatics analysis, some of them will take minutes to finish due to the large file size and computational intense steps involved in the analysis, a job handling system is necessary to handle the asynchronous jobs submitted by users. 3) the visualization system, which generates all illustrations in various functional modules and provides streamlined user interface to guide users through diverse functions provided in ADE.

5.2.2 Functional modules

In the Alzheimer's Disease Explorer (ADE) web server, we designed and developed five functional modules to address different aspects of AD/ND research, namely PubAD, GeneAD, ToolboxAD, TargetAD, and DataAD. The detailed modular functions and resources are illustrated in Figure 20.

- The PubAD is designed to support literature survey of AD researchers by doing text mining on AD related publications in the PubMed database. The information extracted from AD related publications will be presented to the users with various visualizations. As shown in Figure 20, this module includes the NLP mining of the PubMed publications regarding AD/ND research and the visualization of AD/ND related information in CTD database.
- The GeneAD focuses on providing gene based bioinformatic information, such as differentially expressed (DE) gene results derived from both bulk RNA sequencing data and single cell RNA (scRNA) sequencing data, functional gene modules reported from publications, clinical trait-related gene modules, and potential cell makers information from published studies. The GeneAD serves as an integrated portal for all gene related bioinformatic outcomes in AD/ND field.
- The TooboxAD integrates transcriptomic data with popular bioinformatic tools to provide users with an easy-to-use experience for conducting customizable bioinformatic analysis. The ToolboxAD provides users with a list of well-developed tools that focus on the AD/ND research. Such as the ID Converter allows users to convert commonly used gene IDs easily; the OnlineGSEA allows users to perform gene set enrichment analysis with integrated transcriptomic data; the PCA Plot allows users to perform principle component analysis to investigate transcriptomic data variance.

- The TargetAD addresses the challenge of AD/ND drug target profiling, it incorporates multiple important databases within which the data can be used for measurements of drug target potency. Interactive comparison tools enable users to efficiently evaluate drug targets using various critical criteria, the multi-gene query capability allows users to compare multiple potential drug targets at once.
- The DataAD provides the transcriptomic data preparation and download functions, which allow users to filter out the samples by clinical traits and download the mRNA gene expression profiles for downstream analysis. The streamlined user interface provides intuitive and efficient transcriptomic data manipulation and processing capability.

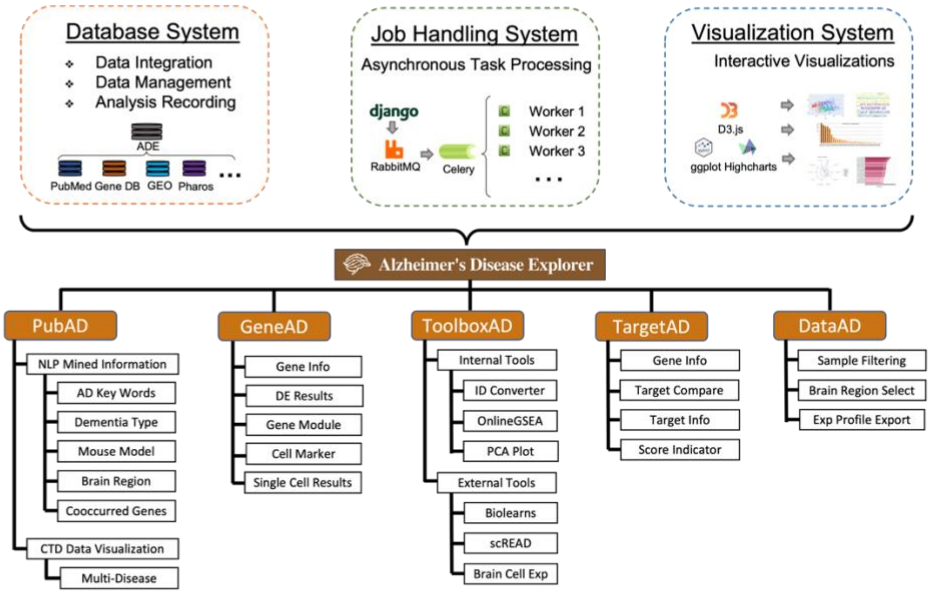


Figure 20. Systematic function and architecture design of ADE.

5.2.3 Database design

We carefully designed a modularized database structure to support the special needs of different functional modules of ADE (Figure 21).

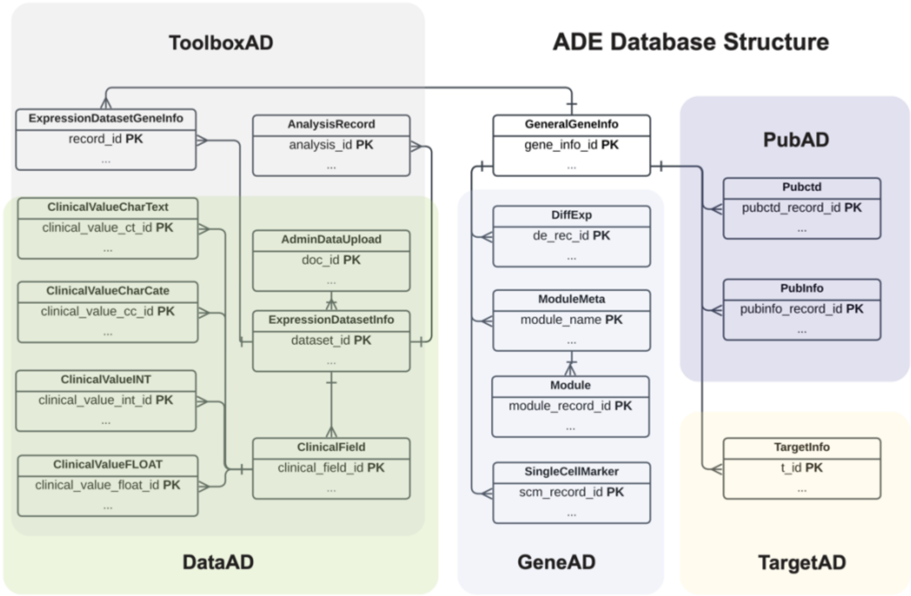


Figure 21. ADE database structure.

The database system is designed around the queried gene information which is shared in all function modules of ADE. The gene information includes gene identifiers, gene description, gene aliases, etc. To manage the data used in PubAD, we utilized the json field in the database to handle the processed data from our in-house NLP pipeline (ref). Due to the highly diversified data sources and types in GeneAD, all datasets in GeneAD are categorized into three categories: the DE results (both bulk RNASeq and scRNASeq), the gene modules and cell markers. The TargetAD has a similar database structure as GeneAD but includes different information according to the source of the data.

Other than the ability to integrate diverse data collected from various resources, the database structure of ADE also enables the automatic importing of transcriptomic datasets, specifically, meta data of mRNA gene expression profiles together with the accompanied clinical information. Transcriptomic datasets captured by the database can be directly utilized by different modules in ADE such as ToolboxAD and DataAD.

Two features are implemented by the database system while handling AD/ND related transcriptomic data. One is the inclusion of brain region information which can often be found in the clinical information, inclusion of multiple brain regions for a single sample will result in the difference between number of columns in gene expression profile and the number of samples in clinical information data. This has been handled by recording the mapping between brain region and gene expression file column separately from clinical information. Another feature is that the system can accommodate a variety length of the clinical attributes that usually differs among. To implement this feature, we stratified the sample attributes according to their data types and manage them in different database tables according to the data type, four main data types are used for stratification: float, integer, categorical and text (Figure 21). While loading new dataset into the database, clinical attributes in the meta data file are assigned to different tables according to the data type of the attribute. Some other important information regarding the transcriptomic dataset is also recorded in the database, including dataset description, link to data source, uploaded date, sequencing platform, genes in detected in the datasets, etc.

5.2.4 Data source

ADE integrated many different data resources in biomedical research area to provided bioinformatics support for AD/ND researchers, these data resources covered a wide range of data types, including literature data, multi-omics data such as bulk RNASeq data and single cell RNA-Seq data, clinical data and target druggability data.

In the PubAD module, we designed and implemented an NLP pipeline to process gene specific and AD/ND related publications in PubMed database, significant amount of useful information are extracted using the NLP pipeline and extracted data are visualized with intuitive demonstrations, the detailed data processing are described in a separated publication[89]. In addition to the information extracted from our inhouse NLP pipeline, we also collected AD/ND related information from the Comparative Toxicogenomics Database (CTD)[90] to provide literature-based and gene-specific information on environmental chemicals exposure's impact on AD/ND.

In the GeneAD module, we collected and processed gene related information from multiple resources. The detailed gene information, including aliases, gene location and description etc., are obtained from NCBI Gene database[91]. The differentially expressed (DE) genes results are processed by our inhouse pipeline, the source data are mainly collected from published studies and were directly downloaded from GEO database. The GeneAD also included gene co-expression network modules and clinical traits-correlated functional gene modules collected from several recent publications[92-95]. Due to the diverse of cell types in brain, we collected information about genes that can be potential cell markers for various cell types in brain. For the potential cell markers inferred from mouse models, we matched the corresponding gene to human genome and annotated the origin species in our curated GeneAD database. The DE results generated from scRNA sequencing datasets are also included in GeneAD to provide users with cell type specific DE information related to AD/ND diseases[95].

5.2.5 System implementation

The ADE is implemented using Django which is a python-based web application development framework. Django follows the Model-View-Controller (MVC) schema[96]; therefore, the database system as mentioned the database design of ADE is the main component of Model in Django framework, visualization system serves as the View and job handling system becomes part of Controller in Django (Figure 22).

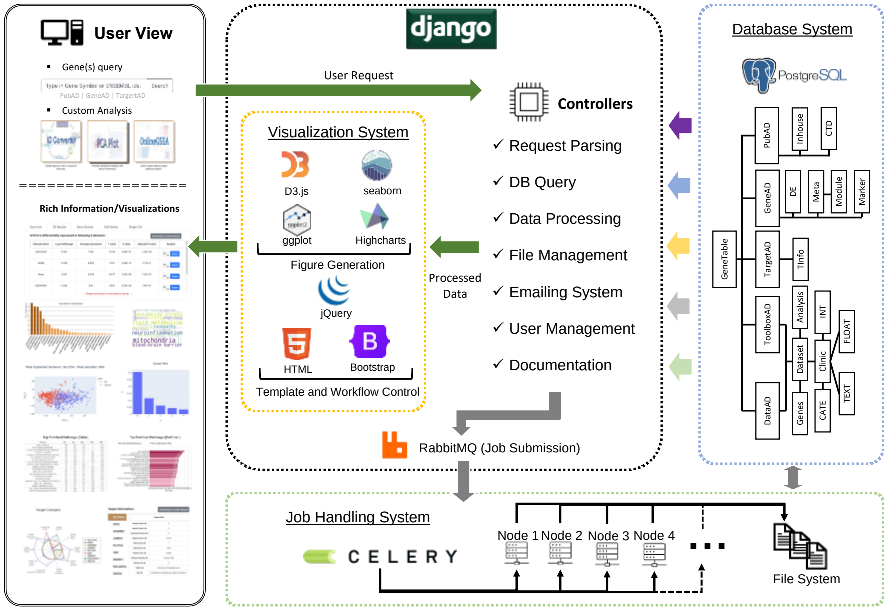


Figure 22. Details of ADE system architecture.

PostgreSQL [97] is used for the production version of ADE because of several key advantages of PostgreSQL system, including the support of complex data type storage, high performance of both read and write and its open-source nature. According to the database design, the data structure is implemented using APIs provided in Django.

The job handling system is implemented using RabbitMQ and Celery[98]. RabbitMQ serves as the message broker between the Django project and Celery workers, once a job is created and submitted in the Django app, the corresponding command and parameters for the job will be passed to Celery scheduler using RabbitMQ’s messaging service, then Celery will assign the job to a certain Celery worker according to the availability of Celery workers. Number of Celery workers is determined automatically according to number of available computational nodes of the server system. Once the job is finished, Celery worker will save the outputs in the designated location on the server and communicate with the database system to update the job status.

The visualization system of ADE is implemented with various technologies including Bootstrap v5, D3.js, ggpot, Highcharts.js, etc[52, 99]. The web page layout is controlled by Bootstrap v5 to utilize its well implemented griding system and other interactive features. Jquery is used in different user scenario to enhance the user experience. For example, when user tries to set up sample groups using gene’s expression level in OnlineGSEA, a few steps are provided for users to follow, various functions in these steps are implemented with different APIs of Jquery, such as the autocomplete in gene search box, the display of detailed gene information and selection button of brain regions of a certain dataset.

ADE system was developed and tested on local server first, then deployed it to Jetstream[100] which is a cloud environment built for computation intense research projects. Nginx and uwsgi are used in the Jetstream instance to provide all the server-side support for ADE system.

5.3 Results

With the architecture design, function design and database design, we implemented five functional modules in ADE: PubAD, GeneAD, ToolboxAD, TargetAD and DataAD. An overview of each module in ADE is summarized in Table 2. These five modules provide a complete tool set for AD/ND researchers to obtain bioinformatics support while conducting AD/ND research activities such as target identification, hypothesis generation, hypothesis testing. Main functions of each module are described below:

5.3.1 PubAD

Based on the extracted information from inhouse NLP pipeline which incorporated five common ND disease (Alzheimer's, and Huntington's disease, Parkinson Disease, Lewy Body Dementia and Frontotemporal dementia), various visualizations are created. Users can use the search box to query the gene they are interested in. The top of result page shows the basic gene information and a yearly count of publications related to the query gene in AD/ND research field. Then five tabs show different categories of information including keywords, dementia types, brain regions, mouse models and co-occurred genes. Users can check all recorded publications on PubMed website by click the bars of bar plots in each tab.

The gene-disease inference data was downloaded from CTD database, the dataset is filtered by MeSH IDs of five common ND diseases as mentioned above. The gene-chemical pairs and corresponding PMIDs are extracted from the filtered data and converted to tabular format. On the lower part of the PubAD result page, bar plots are provided for NDs with available information from CTD, the plots give intuitive illustration of manually curated literature information on environmental chemicals effects on AD/ND. The inference score indicates the degree of similarity between CTD chemical–gene–disease networks and a similar scale-free random network. The higher the score, the more likely the inference network has atypical connectivity. Users can use this data to investigate the relationship among the AD/ND, a certain gene and environmental chemicals inferred from literature data.

Table 2 Overview of function modules in ADE.

ADE module	**Short Description**	**Functions**	**Main Data Sources**
PubAD	PubAD provides neurodegenerative disease-related publication information sorted by various criteria.	Single gene query	PubMed, CTD
GeneAD	GeneAD offers single/multiple gene query and provides AD related transcriptomics analysis results.	Single or multiple genes query	GEO, NCBI Gene Database, publications
ToolboxAD	ToolboxAD provides commonly used bioinformatic tools with integrated datasets.	Customizable analysis: ID Convert, PCA, GSEA	GEO, Pharos, ChEMBL
TargetAD	TargetAD offers single/multiple gene query to provide AD target assessment information.	Single or multiple genes query	ChEMBL. Pharos, OpenTarget
DataAD	DataAD generates data subsets based on user-defined clinical traits for further analysis on public AD datasets.	mRNA expression profile filtering and exporting	GEO

5.3.2 GeneAD

GeneAD provides single gene or gene set query for users to investigate various gene related genomic and transcriptomic information, all available information is categorized by information type and arranged in different panels of the result page. These categories include basic gene information, differentially expressed (DE) gene results, functional network gene module information, cell type marker information and single cell data DE results. With the appearance of future publications related to such information, we will regularly update the database to incorporate the most up-to-date information.

Basic gene information panel shows the basic gene information, providing similar gene information as in PubAD. The DE Results panel shows the differentially expressed gene results collected from published studies[94]. The result table lists all datasets within which the query gene is differentially expressed between AD vs normal samples. The P value cutoff we used for defining significant DE genes is 0.05. The last column of the result table provides a button which allows users to generate two box plots in real time, using the query gene’s expression level comparing AD vs normal samples in one dataset.

Gene module panel shows a variety of published network module study results in the AD field, currently it includes: the frequent gene co-expression network modules generated from AD bulk tissue transcriptomic data[92-94] and clinical traits-correlated functional gene modules[95]. In the near future, metabolic network module will also be included[101]. Data sources of gene modules are provided in the result and the button in the last column can be used to check all genes in a specific module.

Cell Marker panel provides information on whether the query gene is a potential marker for a certain cell type in brain, we collected the information from several recently published studies[102-105] and the data source is provided in the first column of the result table. Since some cell markers are inferred from mouse models, the last column of the table indicates the origin species together with important notes related to the experiment setup.

Single Cell panel provides the DE results generated from AD/ND single cell datasets. The result table lists all significant DE results across all datasets integrated in ADE, the DE results are arranged in different panels according to the cell types. The last column of the result table indicates the DE experiment condition.

Other than single gene query, GeneAD also provides the capability of querying a list of genes by selecting the Multiple Gene option on top of the search box. The query results are similar as the single gene query, but with extra information. In the Gene Module tab, GeneAD implemented a similar method as DAVID enrichment analysis[106] to perform fisher exact test using the query gene list against all available gene modules in the database, only significantly overlapped gene modules are shown in the result table.Venn diagrams are provided to visualize the overlap between the query gene list and gene modules in the system. A Download button is implemented in each panel for users to download the query results into csv files for future reference. Please also note that the query results of the same gene are cross-linked on GeneAD and PubAD, to allow users quick check PubAD query results from GeneAD and vice versa.

5.3.3 ToolboxAD

ToolboxAD provides two set of bioinformatic tools for performing customizable bioinformatics analysis, the first set of tools are developed in the ADE system by utilizing all functions and sources in ADE, the other set of tools are well-implemented external bioinformatic tools that can provide AD/ND bioinformatics analysis. During the development of ToolboxAD, many reusable modules and APIs are implemented to support various designed functions of our inhouse tools, these modules and APIs eased the future development effort of new tools within the ADE system. Currently, there are three inhouse tools available in the ToolboxAD, they are ID Converter, Principal Component Analysis (PCA)[107] Plot and OnlineGSEA. Five external tools are list in ToolboxAD, they are Agora, Biolearns App[108], scREAD[82], RNA Expression In Cell Types and scFLUX[101]. The main functions of inhouse tool are:

The ID Converter is a gene ID converting engine with a user-friendly user interface. With well-designed layout of the page, users can easily convert one gene ID to other gene IDs by a few clicks. If the gene in the user input cannot be identified, ID Converter can recommend similar genes for users. ID Converter incorporated most of commonly used gene IDs, including HUGO gene symbol, CHEMBL, ENSEMBL, UniProt, RefSeq, ENTREZ. The successfully converted gene IDs can be downloaded as csv file.

The PCA Plot is an exploratory tool for investigating the variance of AD/ND related transcriptomic datasets with available clinical information. By using the integrated transcriptomic datasets in the ADE system, users can easily generate 2D and 3D PCA plots with customizable parameters, such as color-coding schema and number of principle components. All resulting plots are interactive and can be downloaded into various static image files.

The OnlineGSEA provides the capability to run GSEA[8] with integrated data in ADE system, for investigating the pathway alterations under different conditions. The OnlineGSEA in the ToolboxAD offers two methods for designing the GSEA experiment, one is using clinical attributes, another one is using a gene’s expression level to divide sample into different groups. If clinical attributes are used, users need to add filters of clinical features to a dataset to assemble two groups of samples they want to compare.

Some restrictions are automatically applied while adding filters to each sample group, such as the restriction of non-overlapping samples between two groups. If gene's expression values are used, users can choose to use median, quartile, or a custom range cutoff to group samples into different cohorts. Once sample groups are defined using either method, the GSEA settings page allows users to adjust GSEA parameters such as the pathway database to be used for the analysis. Once the experiment design and GSEA parameters are confirmed, the analysis will be submitted to ADE's job handling system, users will be provided with a unique analysis ID for retrieving analysis results. If users optioned for reveiving emails, they would receive a notification email with the unique analysis ID, experiment design and experiment parameters. The homepage of ToolboxAD provides the View Result function which can be used to view analysis results, the analysis results can also be download into a zip file.

5.3.4 TargetAD

TargetAD integrated several data sources that can be used for evaluating a gene or gene set's drug target feasibility and other characteristics. It provides intuitive visualizations for novel drug target profiling with multi-dimensional information. The search box on the TargetAD's homepage can be used for single gene or gene list query. The first section of the query result shows gene information. The last column of the gene information table provides the direct link to PubAD and GeneAD of the corresponding gene. The next section of the query result shows a spider plot displaying ten categories of information integrated in the TargetAD database, including Pharos Novelty score, Pharos Target Development Level (TDL), CHEMBL assay counts, etc. All categories of information of identifiable genes are shown on the spider plot for users to easily compare drug targets. A Target Data table is displayed on the right of the spider plot for users to check on the accurate information of each data category. The definition of each data category is provided in the table. A download button showing on top of the table allows users to download the query result into csv files. It's worth noting that some interactive features are provided, such as when the user's cursor is placed on the gene name on left of the Target Data table, the corresponding gene's information will be highlighted on the spider plot. The last section of the query result shows a Score Indicator Dashboard (SID), the SID generates boxplots of several types of scores in the TargetAD using all available data, the query gene's specific score is highlighted with red diamond on the boxplot. The SID provides users an interactive tool for visualizing the query gene's druggability scores with the background of all available drug targets, it enables the quick and precise drug target profiling.

5.3.5 DataAD

DataAD addresses the challenge of parsing transcriptomic data for users to perform downstream analysis. It allows users to build sample sub-cohorts by using user-selected clinical attributes and export the resulting gene expression profiles. In the DataAD module, users can browse all available datasets in the system using the left panel. Currently there are six datasets from publicly available large cohort studies on human brains. More datasets will be added in the future as we continuously processing datasets. The detailed description of the dataset will be shown below the dataset list when the cursor is placed on a dataset button. Upon the selection of a dataset, the right panel of the page will be enabled to apply filters to the selected dataset. All available clinical attributes of the dataset are displayed on top, users can choose one of them and use the pop-up window to input the desired range of clinical attribute's value or select available categories. Once filters are applied, the count of remaining samples will be displayed on the. A summary of all filters added to the current dataset is shown below the sample count. Brain region selection is also provided right before the data export. The Export Data button on the bottom allows users to download the filtered gene expression profile into csv file. Then they can be further analyzed outside the web portal for user's own purpose, such as hypothesis generation, testing and validation.

5.4 Discussion

In the ADE, the database system, the job handling system and the visualization system work together harmoniously to support all functional modules. ADE tries to address the full AD/ND research life cycle support by providing the most crucial bioinformatic information to AD/ND researchers. During the design and implementation process of ADE, we emphasized on three capabilities of the system, namely expandability, scalability, and flexibility, to make ADE a sustainable bioinformatic eco-system.

The expandability of the system is achieved by specially designed database system and modularized functions, so that the system can easily incorporate new data and functions. For example, the transcriptomic dataset handling data structure enables ADE to incorporate new transcriptomic datasets with minimal preprocessing, the added dataset can be used directly by tools in ToolboxAD and functions of DataAD; the function of sample cohort construction using clinical attributes is modularized and is used by several tools in ADE, it can be reused for other tools in the future development as needed.

The scalability of the ADE system comes from the modularized system structure and APIs. Depends on the demand of users, functions of ADE can be easily modified to fit users' needs. The system architecture of ADE is simple yet fully supportive of all designed functions. The system can be easily deployed to various environments like cloud-based systems and high-performance computers (HPCs). When deployed to HPCs, with minimal configuration of job handling system, ADE can utilize HPC's computation resources to run computation intense analysis workflows in ToolboxAD.

The Flexibility of the system describes the reusability of ADE system's function design and system architecture. During the development of ADE, we generalized a high-level framework for guiding the development of similar bioinformatic platforms. With this framework and further effort of collecting and processing domain specific knowledge and datasets, new bioinformatic systems can be developed to provide bioinformatics support in other research domains.

By focusing on these capabilities, the longevity of ADE system can be maximized and the effort of maintaining and further developing the system remains minimal. Moreover, developing ADE with these capabilities makes it possible to perform the fast development of similar bioinformatic platforms which can address challenges in different research areas than AD/ND.

Aside from all features provided in ADE, there are a few limitations in the current version of ADE. Currently, ADE can only incorporate processed transcriptomic datasets with clinical attributes, the capability of handling other types of omics data such as genomic data or proteomics data has not been implemented yet. The transcriptomic dataset import function is currently limited to administrative users of the system and not available to all users. The current ADE does not provide the mechanism to aggregate user interested information to generate a single gene-centric integrated evidence profile. However, with all available modules and APIs in ADE, these limitations can be easily cleared with future development effort.

5.5 Conclusion

In this study, we successfully designed and implemented ADE web server, a first-in-kind bioinformatics server and a one-stop web portal on AD/ND research, providing rich bioinformatics support from literature, omics and chemical data to greatly facilitate researchers in ND drug development field. In fact, it starts to be used by NIH funded AD Drug Discovery Center for potential drug target screening research. By providing users with an easy-to-use experience for conducting customizable bioinformatics analysis and drug target identification for AD/ND diseases, ADE is aimed to address the gene-centric informatic search needs from comprehensive AD/ND research and will accelerate the drug discovery process to finally stop or reverse the AD progression.

Chapter 6. Conclusions and Discussions

6.1 Conclusions

In my thesis, I firstly explored advanced methods for analyzing transcriptomic datasets and applied these methods to address challenges of colon cancer patient biomarker discovery using transcriptomic data, which provided insights on the crucial role of TFs in the cancer prognosis study. Then, I developed a patient-oriented colorectal cancer CDSS prototype by integrating advanced risk factor evaluation algorithms with user-friendly interface to help general population gain better understanding of colorectal cancer risks, the CDSS focused on lifestyle risk factors and provided actionable items for high CRC risk users. With the knowledge of advanced transcriptomic data analysis methods and skills of web-based application development, I developed CGPE which is an integrative bioinformatic platform that provides support for cancer related literature survey, gene set enrichment analysis using integrated transcriptomic datasets and cell line evaluation by gene. After the release of CGPE, thousands of users have used tools inside CGPE to help with their cancer research, outcomes generated from CGPE have been included in a few cancer studies which have been published in high impact biomedical joiurnals. Other than the functions provided by CGPE, a generalized framework for developing disease specific bioinformatic platforms was also generated. Finally, by using the generalized framework, I designed and implemented ADE which is a platform that provides bioinformatics support for AD/ND research and drug development. By exploring analytical bioinformatics algorithms and developing disease specific bioinformatic platforms, I have shown that integrated disease specific bioinformatic platforms can provide great value to the research community by allowing 1.) fast and accurate investigation of currently available literature, 2.) quick hypothesis generation and validation using transcriptomic datasets, 3.) multi-dimension drug target evaluation and 4) fast querying of published bioinformatic outcomes. These disease specific platforms significantly eased biomedical researchers' effort of collecting, processing, and analyzing critical bioinformatic datasets. The development of such platforms will dramatically accelerate the treatment development process of certain diseases.

6.2 Future Directions

With all functions in the current systems, a few improvements can be achieved by future development efforts. For example, in the current platforms, the literature data analysis pipeline only covers the title and abstract of related publications, further analysis of full text of publications may add more insights on certain research topics. With all the datasets and resources integrated in the disease specific platforms, users may find it difficult to gather all the information related to their research interest in the current platforms. Thus, a mechanism of managing available data in the system to generate an evidence base for a user interested research topic could further improve the usability of current platforms. While conducting customizable analysis, the current systems only provide transcriptomic datasets with corresponding algorithms that can be applied to them. In the future development, it will be useful to incorporate other type of omics data and corresponding algorithms to address a wider field of bioinformatics analytics.

Based on the current stage of the disease specific bioinformatics platforms, develop a user login system to allow users upload their own data to take advantage of all the functions in the current system would be an interesting next step. However, implementation of the user-controlled platform would add extensive complexity to the system and may bring new concerns to the user of the platform, such as whether the server is HIPPA compliance to fulfill data usage restrictions of certain datasets. With growing interest in the usage of disease specific platforms, I will explore the potential of incorporating user control system based on our currently available platforms.

Reference

1. Conesa A, Beck S. Making multi-omics data accessible to researchers. Scientific data. 2019;6(1):1-4.
2. Savage N. Collaboration is the key to cancer research. Nature. 2018;556(7700):S1-S.
3. Prins P, De Ligt J, Tarasov A, Jansen RC, Cuppen E, Bourne PE. Toward effective software solutions for big biology. Nature biotechnology. 2015;33(7):686-7.
4. Liu J, Dong C, Jiang G, Lu X, Liu Y, Wu H. Transcription factor expression as a predictor of colon cancer prognosis: a machine learning practice. BMC medical genomics. 2020;13(9):1-10.
5. Liu J, Li C, Xu J, Wu H. A patient-oriented clinical decision support system for CRC risk assessment and preventative care. BMC medical informatics and decision making. 2018;18(5):45-53.
6. Huang S, Yang J, Fong S, Zhao Q. Artificial intelligence in cancer diagnosis and prognosis: Opportunities and challenges. Cancer letters. 2020;471:61-71.
7. Baxevanis CN, Fortis SP, Perez SA, editors. The balance between breast cancer and the immune system: Challenges for prognosis and clinical benefit from immunotherapies. Seminars in cancer biology; 2021: Elsevier.
8. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proceedings of the National Academy of Sciences. 2005;102(43):15545-50.
9. Liu J, Dong C, Liu Y, Wu H. CGPE: an integrated online server for C ancer G ene and P athway E xploration. Bioinformatics. 2021;37(15):2201-2.
10. Plascencia-Villa G, Perry G. Status and future directions of clinical trials in Alzheimer's disease. International review of neurobiology. 2020;154:3-50.
11. Siegel RL, Miller KD, Jemal A. Cancer statistics, 2017. Ca A Cancer Journal for Clinicians. 2017;67(1):5.
12. Ponz dLM, Sassatelli R, Benatti P, Roncucci L. Identification of hereditary nonpolyposis colorectal cancer in the general population. The 6-year experience of a population-based registry. Cancer. 1993;71(11):3493-501.

13. Lee TI, Young RA. Transcription of eukaryotic protein-coding genes. Annual Review of Genetics. 2000;34(1):77-137.

14. Latchman DS. Transcription factors: an overview. International Journal of Experimental Pathology. 1997;74(5):1305-12.

15. Wang S, Liu Z, Wang L, Zhang X. NF-κB Signaling Pathway, Inflammation and Colorectal Cancer. Cellular and Molecular Immunology. 2009;6(5):327-34.

16. Lin L, Liu A, Peng Z, Lin HJ, Li PK, Li C, et al. STAT3 is necessary for proliferation and survival in colon cancer-initiating cells. Cancer Research. 2011;71(23):7226-37.

17. Wan LY, Deng J, Xiang XJ, Zhang L, Yu F, Chen J, et al. miR-320 enhances the sensitivity of human colon cancer cells to chemoradiotherapy in vitro by targeting FOXM1. Biochemical & Biophysical Research Communications. 2015;457(2):125-32.

18. Kourou K, Exarchos TP, Exarchos KP, Karamouzis MV, Fotiadis DIJC, journal sb. Machine learning applications in cancer prognosis and prediction. 2015;13:8-17.

19. Long NP, Park S, Anh NH, Nghi TD, Yoon SJ, Park JH, et al. High-Throughput Omics and Statistical Learning Integration for the Discovery and Validation of Novel Diagnostic Signatures in Colorectal Cancer. 2019;20(2):296.

20. Vafaee F, Diakos C, Kirschner MB, Reid G, Michael MZ, Horvath LG, et al. A data-driven, knowledge-based approach to biomarker discovery: application to circulating microRNA markers of colorectal cancer prognosis. 2018;4(1):20.

21. Xu J, Zhao J, Zhang RJSr. Four microRNAs signature for survival prognosis in colon cancer using TCGA data. 2016;6:38306.

22. Xu G, Zhang M, Zhu H, Xu JJG. A 15-gene signature for prediction of colon cancer recurrence and prognosis based on SVM. 2017;604:33-40.

23. Barrett T, Wilhite SE, Ledoux P, Evangelista C, Kim IF, Tomashevsky M, et al. NCBI GEO: archive for functional genomics data sets—update. 2012;41(D1):D991-D5.

24. Goldman M, Craft B, Swatloski T, Cline M, Morozova O, Diekhans M, et al. The UCSC Cancer Genomics Browser: update 2015. Nucleic Acids Research. 2015;43(Database issue):D812-D7.

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

26. Liaw A, Wiener MJRn. Classification and regression by randomForest. 2002;2(3):18-22.

27. Liu G, Dong C, Wang X, Hou G, Zheng Y, Xu H, et al. Regulatory activity based risk model identifies survival of stage II and III colorectal carcinoma. 2017;8(58):98360.

28. Ishwaran H, Kogalur UB. randomForestSRC: Random Forests for Survival, Regression and Classification (RF-SRC). 2016.

29. Ishwaran H, Kogalur UB, Gorodeski EZ, Minn AJ, Lauer MS. High-Dimensional Variable Selection for Survival Data. Publications of the American Statistical Association. 2010;105(489):205-17.

30. Kaplan EL, Meier PJJotAsa. Nonparametric estimation from incomplete observations. 1958;53(282):457-81.

31. Peto R, Peto JJJotRSSSA. Asymptotically efficient rank invariant test procedures. 1972;135(2):185-98.

32. Vaquerizas JM, Kummerfeld SK, Teichmann SA, Luscombe NMJNRG. A census of human transcription factors: function, expression and evolution. 2009;10(4):252.

33. Kanehisa M, Goto SJNar. KEGG: kyoto encyclopedia of genes and genomes. 2000;28(1):27-30.

34. Miller GJ, Miller HL, van Bokhoven A, Lambert JR, Werahera PN, Schirripa O, et al. Aberrant HOXC expression accompanies the malignant phenotype in human prostate. Cancer Res. 2003;63(18):5879-88.

35. Ji M, Feng Q, He G, Yang L, Tang W, Lao X, et al. Silencing homeobox C6 inhibits colorectal cancer cell proliferation. Oncotarget. 2016;7(20):29216-27.

36. Kuo KK, Jian SF, Li YJ, Wan SW, Weng CC, Fang K, et al. Epigenetic inactivation of transforming growth factor-β1 target gene HEYL, a novel tumor suppressor, is involved in the P53-induced apoptotic pathway in hepatocellular carcinoma. 2015;45(7):782-93.

37. Stevens SJ, van Essen AJ, van Ravenswaaij CM, Elias AF, Haven JA, Lelieveld SH, et al. Truncating de novo mutations in the Krüppel-type zinc-finger gene ZNF148 in patients with corpus callosum defects, developmental delay, short stature, and dysmorphisms. 2016;8(1):131.

38. Ching T, Zhu X, Garmire LXJPcb. Cox-nnet: An artificial neural network method for prognosis prediction of high-throughput omics data. 2018;14(4):e1006076.

39. Siegel RL, Miller KD, Fedewa SA, Ahnen DJ, Meester RG, Barzi A, et al. Colorectal cancer statistics, 2017. CA: a cancer journal for clinicians. 2017;67(3):177-93.

40. Edwards BK, Noone AM, Mariotto AB, Simard EP, Boscoe FP, Henley SJ, et al. Annual Report to the Nation on the status of cancer, 1975-2010, featuring prevalence of comorbidity and impact on survival among persons with lung, colorectal, breast, or prostate cancer. Cancer. 2014;120(9):1290-314.

41. Force* UPST. Screening for colorectal cancer: recommendation and rationale. Annals of internal medicine. 2002;137(2):129-31.

42. Whitlock EP, Lin JS, Liles E, Beil TL, Fu R. Screening for colorectal cancer: a targeted, updated systematic review for the US Preventive Services Task Force. Annals of internal medicine. 2008;149(9):638-58.

43. Giovannucci E. Modifiable risk factors for colon cancer. Gastroenterology Clinics. 2002;31(4):925-43.

44. Haggar FA, Boushey RP. Colorectal cancer epidemiology: incidence, mortality, survival, and risk factors. Clinics in colon and rectal surgery. 2009;22(04):191-7.

45. Selvachandran S, Hodder R, Ballal M, Jones P, Cade D. Prediction of colorectal cancer by a patient consultation questionnaire and scoring system: a prospective study. The Lancet. 2002;360(9329):278-83.

46. Freedman AN, Slattery ML, Ballard-Barbash R, Willis G, Cann BJ, Pee D, et al. Colorectal cancer risk prediction tool for white men and women without known susceptibility. Journal of clinical oncology. 2009;27(5):686.

47. Romano MJ, Stafford RS. Electronic health records and clinical decision support systems: impact on national ambulatory care quality. Archives of internal medicine. 2011;171(10):897-903.

48. Garg AX, Adhikari NK, McDonald H, Rosas-Arellano MP, Devereaux PJ, Beyene J, et al. Effects of computerized clinical decision support systems on practitioner performance and patient outcomes: a systematic review. Jama. 2005;293(10):1223-38.

49. Kawamoto K, Houlihan CA, Balas EA, Lobach DF. Improving clinical practice using clinical decision support systems: a systematic review of trials to identify features critical to success. Bmj. 2005;330(7494):765.

50. Quinlan JR. Induction of decision trees. Machine learning. 1986;1(1):81-106.

51. Krasner GE, Pope ST. A description of the model-view-controller user interface paradigm in the smalltalk-80 system. Journal of object oriented programming. 1988;1(3):26-49.

52. Bostock M, Ogievetsky V, Heer J. D^3 data-driven documents. IEEE transactions on visualization and computer graphics. 2011;17(12):2301-9.

53. MySQL A. MySQL reference manual. 2001.

54. Wells BJ, Kattan MW, Cooper GS, Jackson L, Koroukian S. Colorectal cancer predicted risk online (CRC-PRO) calculator using data from the multi-ethnic cohort study. The Journal of the American Board of Family Medicine. 2014;27(1):42-55.

55. Wright A, Sittig DF, Ash JS, Sharma S, Pang JE, Middleton B. Clinical decision support capabilities of commercially-available clinical information systems. Journal of the American Medical Informatics Association. 2009;16(5):637-44.

56. Velmourougan S, Dhavachelvan P, Baskaran R, Ravikumar B, editors. Software development Life cycle model to build software applications with usability. 2014 International Conference on Advances in Computing, Communications and Informatics (ICACCI); 2014: IEEE.

57. Chen H, Sun X, Ge W, Qian Y, Bai R, Zheng S. A seven-gene signature predicts overall survival of patients with colorectal cancer. Oncotarget. 2017;8(56):95054.

58. Heroku [Available from: https://www.heroku.com/about

59. Cheang MC, Voduc D, Bajdik C, Leung S, McKinney S, Chia SK, et al. Basal-like breast cancer defined by five biomarkers has superior prognostic value than triple-negative phenotype. Clinical cancer research. 2008;14(5):1368-76.

60. Ciriello G, Gatza ML, Beck AH, Wilkerson MD, Rhie SK, Pastore A, et al. Comprehensive molecular portraits of invasive lobular breast cancer. Cell. 2015;163(2):506-19.

61. Liu C, Srihari S, Lal S, Gautier B, Simpson PT, Khanna KK, et al. Personalised pathway analysis reveals association between DNA repair pathway dysregulation and chromosomal instability in sporadic breast cancer. Molecular oncology. 2016;10(1):179-93.

62. Pollack JR, Sørlie T, Perou CM, Rees CA, Jeffrey SS, Lonning PE, et al. Microarray analysis reveals a major direct role of DNA copy number alteration in the transcriptional program of human breast tumors. Proceedings of the National Academy of Sciences. 2002;99(20):12963-8.

63. Micheel CM, Sweeney SM, LeNoue-Newton ML, André F, Bedard PL, Guinney J, et al. American Association for Cancer Research Project Genomics Evidence Neoplasia Information Exchange: from inception to first data release and beyond—lessons learned and member institutions' perspectives. JCO clinical cancer informatics. 2018;2:1-14.

64. Jensen MA, Ferretti V, Grossman RL, Staudt LM. The NCI Genomic Data Commons as an engine for precision medicine. Blood, The Journal of the American Society of Hematology. 2017;130(4):453-9.

65. Barretina J, Caponigro G, Stransky N, Venkatesan K, Margolin AA, Kim S, et al. The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. Nature. 2012;483(7391):603-7.

66. Marx V. The big challenges of big data. Nature. 2013;498(7453):255-60.

67. Gao J, Aksoy BA, Dogrusoz U, Dresdner G, Gross B, Sumer SO, et al. Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal. Science signaling. 2013;6(269):pl1-pl.

68. Goldman M, Craft B, Swatloski T, Cline M, Morozova O, Diekhans M, et al. The UCSC cancer genomics browser: update 2015. Nucleic acids research. 2015;43(D1):D812-D7.

69. Tang Z, Li C, Kang B, Gao G, Li C, Zhang Z. GEPIA: a web server for cancer and normal gene expression profiling and interactive analyses. Nucleic acids research. 2017;45(W1):W98-W102.

70. Mullard A. Can you trust your cancer cell lines? Nature Reviews Drug Discovery. 2018;17(9):613-4.

71. Edgar R, Domrachev M, Lash AE. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. Nucleic acids research. 2002;30(1):207-10.

72. Lin DY, Wei L-J. The robust inference for the Cox proportional hazards model. Journal of the American statistical Association. 1989;84(408):1074-8.

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

74. Croft D, O'kelly G, Wu G, Haw R, Gillespie M, Matthews L, et al. Reactome: a database of reactions, pathways and biological processes. Nucleic acids research. 2010;39(suppl_1):D691-D7.

75. Wei C-H, Kao H-Y, Lu Z. PubTator: a web-based text mining tool for assisting biocuration. Nucleic acids research. 2013;41(W1):W518-W22.

76. Li S, Wu L, Zhang Z. Constructing biological networks through combined literature mining and microarray analysis: a LMMA approach. Bioinformatics. 2006;22(17):2143-50.

77. Taylor JP, Hardy J, Fischbeck KH. Toxic proteins in neurodegenerative disease. science. 2002;296(5575):1991-5.

78. Gitler AD, Dhillon P, Shorter J. Neurodegenerative disease: models, mechanisms, and a new hope. The Company of Biologists Ltd; 2017. p. 499-502.

79. Cooper-Knock J, Kirby J, Ferraiuolo L, Heath PR, Rattray M, Shaw PJ. Gene expression profiling in human neurodegenerative disease. Nature Reviews Neurology. 2012;8(9):518-30.

80. Tsuji S. Genetics of neurodegenerative diseases: insights from high-throughput resequencing. Human molecular genetics. 2010;19(R1):R65-R70.

81. Ahmadi A, Gispert JD, Navarro A, Vilor-Tejedor N, Sadeghi I. Single-cell Transcriptional Changes in Neurodegenerative Diseases. Neuroscience. 2021;479:192-205.

82. Jiang J, Wang C, Qi R, Fu H, Ma Q. scREAD: a single-cell RNA-seq database for Alzheimer's disease. Iscience. 2020;23(11):101769.

83. Wang Z, Feng X, Li SC. SCDevDB: a database for insights into single-cell gene expression profiles during human developmental processes. Frontiers in Genetics. 2019:903.

84. Courtney E, Kornfeld S, Janitz K, Janitz M. Transcriptome profiling in neurodegenerative disease. Journal of neuroscience methods. 2010;193(2):189-202.

85. Dugger BN, Dickson DW. Pathology of neurodegenerative diseases. Cold Spring Harbor perspectives in biology. 2017;9(7):a028035.

86. Greenwood AK, Gockley J, Daily K, Aluthgamage D, Leanza Z, Sieberts SK, et al. Agora: An open platform for exploration of Alzheimer’s disease evidence: Genetics/omics and systems biology. Alzheimer's & Dementia. 2020;16:e046129.

87. Lin C-X, Li H-D, Deng C, Erhardt S, Wang J, Peng X, et al. AlzCode: a Platform for Multiview Analysis of Genes Related to Alzheimer’s Disease. Bioinformatics. 2022.

88. Wörheide MA, Krumsiek J, Nataf S, Nho K, Greenwood AK, Wu T, et al. An Integrated Molecular Atlas of Alzheimer's Disease. medRxiv. 2021.

89. Liu J, Wu H, Robertson DH, Zhang J. Text mining and portal development for gene-specific publications on Alzheimer's disease and other neurodegenerative diseases. bioRxiv. 2022.

90. Davis AP, Grondin CJ, Johnson RJ, Sciaky D, McMorran R, Wiegers J, et al. The comparative toxicogenomics database: update 2019. Nucleic acids research. 2019;47(D1):D948-D54.

91. Brown GR, Hem V, Katz KS, Ovetsky M, Wallin C, Ermolaeva O, et al. Gene: a gene-centered information resource at NCBI. Nucleic acids research. 2015;43(D1):D36-D42.

92. Milind N, Preuss C, Haber A, Ananda G, Mukherjee S, John C, et al. Transcriptomic stratification of late-onset Alzheimer's cases reveals novel genetic modifiers of disease pathology. PLoS Genetics. 2020;16(6):e1008775.

93. Wan Y-W, Al-Ouran R, Mangleburg CG, Perumal TM, Lee TV, Allison K, et al. Meta-analysis of the Alzheimer’s disease human brain transcriptome and functional dissection in mouse models. Cell reports. 2020;32(2):107908.

94. Johnson TS, Xiang S, Dong T, Huang Z, Cheng M, Wang T, et al. Combinatorial analyses reveal cellular composition changes have different impacts on transcriptomic changes of cell type specific genes in Alzheimer’s Disease. Scientific Reports. 2021;11(1):1-19.

95. Mathys H, Davila-Velderrain J, Peng Z, Gao F, Mohammadi S, Young JZ, et al. Single-cell transcriptomic analysis of Alzheimer’s disease. Nature. 2019;570(7761):332-7.

96. Deacon J. Model-view-controller (mvc) architecture. Online][Citado em: 10 de março de 2006] http://www jdl co uk/briefings/MVC pdf. 2009.

97. PostgreSQL [Available from: https://www.postgresql.org.

98. Solem A. Celery: Distributed task queue. URL http://docs celeryproject org/en/latest/index html. 2013.

99. Wickham H, Wickham MH. The ggplot package. URL: https://cran r-project org/web/packages/ggplot2/index html. 2007.

100. Hancock DY, Fischer J, Lowe JM, Snapp-Childs W, Pierce M, Marru S, et al. Jetstream2: Accelerating cloud computing via Jetstream. Practice and Experience in Advanced Research Computing2021. p. 1-8.

101. Alghamdi N, Chang W, Dang P, Lu X, Wan C, Gampala S, et al. A graph neural network model to estimate cell-wise metabolic flux using single-cell RNA-seq data. Genome research. 2021;31(10):1867-84.

102. Keren-Shaul H, Spinrad A, Weiner A, Matcovitch-Natan O, Dvir-Szternfeld R, Ulland TK, et al. A unique microglia type associated with restricting development of Alzheimer’s disease. Cell. 2017;169(7):1276-90. e17.

103. Habib N, McCabe C, Medina S, Varshavsky M, Kitsberg D, Dvir-Szternfeld R, et al. Disease-associated astrocytes in Alzheimer’s disease and aging. Nature neuroscience. 2020;23(6):701-6.

104. Liddelow SA, Guttenplan KA, Clarke LE, Bennett FC, Bohlen CJ, Schirmer L, et al. Neurotoxic reactive astrocytes are induced by activated microglia. Nature. 2017;541(7638):481-7.

105. Zamanian JL, Xu L, Foo LC, Nouri N, Zhou L, Giffard RG, et al. Genomic analysis of reactive astrogliosis. Journal of neuroscience. 2012;32(18):6391-410.

106. Huang DW, Sherman BT, Tan Q, Kir J, Liu D, Bryant D, et al. DAVID Bioinformatics Resources: expanded annotation database and novel algorithms to better extract biology from large gene lists. Nucleic acids research. 2007;35(suppl_2):W169-W75.

107. Abdi H, Williams LJ. Principal component analysis. Wiley interdisciplinary reviews: computational statistics. 2010;2(4):433-59.

108. Huang Z, Han Z, Shao W, Xiang S, Salama P, Rizkalla M, et al. TSUNAMI: translational bioinformatics tool suite for network analysis and mining. Genomics, proteomics & bioinformatics. 2021.

Curriculum Vitae

Jiannan Liu

Education

- Indiana University (IUPUI campus)

  Ph.D. in Informatics with a specialization in Bioinformatics

- University of Florida

  M.S. in Industrial Engineering

- Nanjing Agricultural University

  B.S. in Industrial Engineering

Teaching Experience

- Co-Instructor of INFO-B573 Programming for Science Informatics

  2019/2020/2021

- Teach Assistant of INFO-B556 Biological Database Management

  2018/2019 spring

- Instructor of HIM-M200 Database Design for Health Information Management

  2018 fall

- Teach Assistant of HIM-M110 Computer Concepts for Health Information

  2018 summer

Work Experience

- Bioinformatics Co-op – Merck Research Laboratories

  2021 Feb. – Aug.

- Knowledge Engineering Internship – LabCorp

  2020 Summer

Publications

1. **Liu, Jiannan**, Chuanpeng Dong, Yunlong Liu, and Huanmei Wu. "CGPE: an integrated online server for C ancer G ene and P athway E xploration." *Bioinformatics* 37, no. 15 (2021): 2201-2202.
2. Rob Quick, Suresh Marru, Eroma Abeysinghe, Marlon Pierce, Yi Zhao, Sha Cao, Xiaowen Liu, Huanmei Wu, **Jiannan Liu**, Li Chen and Chi Zhang. "Alzheimer's Disease Drug Discovery Center Data Sharing Platform". *Gateways 2020*.
3. **Liu, Jiannan**, Chuanpeng Dong, Guanglong Jiang, Xiaoyu Lu, Yunlong Liu, and Huanmei Wu. "Transcription factor expression as a predictor of colon cancer prognosis: A machine learning practice." *BMC Medical Genomics* 13,no.9(2020):1-10.
4. Chuanpeng, **Jiannan Liu**, Steven X. Chen, Tianhan Dong, Guanglong Jiang, Yue Wang, Huanmei Wu, Jill L. Reiter, and Yunlong Liu. "Highly robust model of transcription regulator activity predicts breast cancer overall survival." *BMC medical genomics* 13 (2020): 1-10.
5. Wu, Huanmei, Parth Kothiya, and **Jiannan Liu**. "Family-HealthVault: A Group Caring and PHI Sharing Framework among Family Members." 2019 IEEE International Conference on Healthcare Informatics (ICHI). IEEE, 2019.
6. **Liu, Jiannan**, Chenyang Li, Jing Xu, and Huanmei Wu. "A patient-oriented clinical decision support system for CRC risk assessment and preventative care." *BMC medical informatics and decision making* 18, no. 5 (2018): 118.