UNIVERSITY OF HAWAII AT MANOA

PATHWAY-BASED MULTI-OMICS DATA INTEGRATION FOR BREAST CANCER DIAGNOSIS AND PROGNOSIS

BY

SIJIA HUANG

A THESIS SUBMITTED IN PARTIAL FULFILLMENT FOR THE

DEGREE OF DOCTOR OF PHILOSOPHY

COMMITTEE MEMBERS:

LANA X. GARMIRE

DANIEL JENKINS

QING LI

STEVEN SEIFRIED

HERBERT YU

MOLECULAR BIOSCIENCES AND BIOENGINEERING

NOVEMBER 2017

Acknowledgements

Completion of this research would not have been possible without my advisor, Dr. Lana Garmire and my committee members, Dr. Daniel Jenkins, Dr. Qing Li, Dr. Steven Seifried and Dr. Herbert Yu.

In addition, I would like to thank all of my fellow lab members, including graduate students: Travers Ching, Xun Zhu, Liangqun Lu and Runmin Wei, the postdocs: Olivier Poirion, Kumardeep Chaudhary, Fadhl Alakwaa, Paula Benny, Thomas Wolfgruber and Michael Ortega, and undergraduate alumni: Cameron Yee and Nicole Chong.

Special thanks to my dear family, my mom and my dad, for their love support during these years. Also to my best friends and my brothers and my sisters in Hawaii: Jie Bai, Yunjie Rao, Xuxiao Li, Yiqin Liu, Wei Yu, Junyao Heng, Wei Zhang, Chenchen Zhao, Han Lee, Biyu Wu, Rui Mao, Vicky Zhang, Hongying Zhong, Xinyi Chen, Baiyi Wang, Baien Wang, Meiyun Luo, Xia He, Ran Chen, Irene Wu, Min Zhu, Ning Li, Yaying Liu, Tingting Pan, Jane Zhang, Huiling L. Lee, Xiao Luo, Miaochan Li, Hong Wang, Xuxiao's Mom and Lihang's Mom; Kuangye Yu, Wang Liang, Lining Han, Lihang Chen, Haitong Chen, Huashan Lin, John Hu, William Pan, Hui Zhang, Zheng Lan, Lihang's Dad, Yongcheng Lee, for their generous help and encouragement, especially during my hard times.

Abstract

With the increasing awareness of heterogeneity in breast cancers, better predictions of breast cancer diagnosis and prognosis are important components of precision medicine. High-throughput profiles have been explored extensively in the last decades for diagnostic and prognostic biomarkers in breast cancer. However, different omics-based studies show little overlap results. With the abundance of multi-omics measurements for cancer patients, there is pressing need for integrative methods that can take advantage of biological information at different biological layers and extract the concerted mechanism in breast cancer.

Towards this goal we propose a new class of pathway-based diagnosis and prognosis prediction models, which emphasize individualized pathway-based risk measurement using the pathway dysregulation scores. We hypothesize that higher-level pathway-based models will consistently perform better than gene- or metabolites- based models. Towards this we have obtained some promising preliminary results, using pathway-based features from transcriptomics data to predict breast cancer prognosis, as well as from metabolomics data to predict breast cancer diagnosis. Next we applied this methodology together with deep learning approach to integrate multi-omics data (gene expression, methylation and copy number variation) for breast cancer patients from public resources such as TCGA and METABRIC, for the purposes of identifying breast cancer subpopulations with prognosis differences. Our results showed that not only our pathway-based prediction consistently performs better than raw data based prediction, but also our deep-learning based integration method gives a better characterization of different cancer subgroups compared to current state-of-art method.

In this thesis the significance of pathway-based biomarkers in breast cancer was characterized, from genomics, metabolomics to multi-omics level. In chapter 1, I further explain the breast cancer diagnosis and prognosis background relevant to the projects contained in this dissertation. Chapter 2 is a research paper published in Genome Medicine, using pathway-based approach on metabolomics data to discover biomarkers for breast cancer diagnosis. In Chapter 3, we applied our pathway-based pipeline on transcriptomics data, to predict for breast cancer prognosis; this work is published in PLOS Computation Biology. Chapter 4 is a trial of integrating clinical traits with biomarkers to evaluate the risk of bladder cancer diagnosis, published in Cancer Epidemiology, Biomarkers and Prevention. This work brings the promising value of integrating more than one levels of information to predict the cancer outcome. Chapter 5 is a review paper published in Frontiers in Genetics, focusing on the current work of multi-omics data integration, summarizing the diverse computational tools developed over the years, their advantages and limitations. In Chapter 6, I extend the pathway-based pipeline to multi-omics data based a deep-learning model, in order to predict patient survival, and to elucidate the biological pathways

relevant to each patient. Finally, in Chapter 7, I discuss what these research projects have accomplished in the grand scheme of the breast cancer research field, and explain what further work needs to be accomplished to follow up. In the future, we plan to validate the significant pathway biomarkers and discover the relationship of medicines with pathways to predict for better and personalized therapeutics treatment in breast cancer.

Specific Aims:

The goal of this work is to discover consistent pathway-based biomarkers for breast cancer diagnosis and prognosis. With an observation that there is little overlap across different studies for biomarker discovery in breast cancer, the hypothesis of this work is that higher-level representation of biomarkers in pathways are more robust and consistent. Through applying different machine learning methods on single-omics or multiple omics data of breast cancer. I will apply different classification and regression algorithms to elucidate novel biological insights in breast cancer.

To demonstrate and verify the hypothesis of this work,
1. Pathway-based metabolomics diagnosis model for breast cancer
2. Pathway-based transcriptomics (mRNA) and clinic-based prognosis model for breast cancer
3. Pathway-based multi-omics prognosis model for breast cancer

Manuscript included:

1. **Huang S**, Yee C, Ching T, Yu H, Garmire LX: Combining clinic and pathway-based features to predict prognosis of breast cancer, PLOS Computational Biology. Sep 18;10(9):e1003851.
2. **Huang S**, Chong N, Lewis NE, Jia W, Xie G, Garmire LX. Pathway-based metabolomics classification models reveal key metabolic pathways for breast cancer diagnosis and progression, Genome Medicine 2016 8:34
3. **Huang S***, Kou L*, Furuya H, Yu CH, Kattan M, Goodison S, Garmire LX, Rosser CJ, A nomogram derived by combination of demographic and biomarker data improves the non-invasive evaluation of patients at risk for bladder cancer, Cancer Epidemiology, Biomarkers and Prevention, 2016 Jul 6.
4. **Huang S**, Chaudhary K, Garmire LX . More is better: Recent progress in multi-omics data integration methods (Front Genet. 2017 Jun 16;8:84.)
5. **Huang S**, Poirion O, Garmire LX . Integration of multi-omics data with deep learning method to predict the prognosis of breast cancer (in preparation)

Table of Contents

LIST OF TABLES

CHAPTER 1

CHAPTER 2

CHAPTER 3

CHAPTER 4

CHAPTER 5

CHAPTER 6

LIST OF FIGURES

CHAPTER 2

CHAPTER 3

CHAPTER 4

CHAPTER 5

CHAPTER 6

Chapter 1. Background and Introduction

Breast Cancer Statistics

Breast cancer is the most frequently diagnosed cancer in women, and ranks second (after lung cancer) in the deaths of women. In 2016, a total number of 249,260 cases of breast cancer (2,600 male and 246,660 female) are diagnosed, with 40890 (440 males and 40450 females) deaths from breast cancer (M. Garcia et al., 2016).

From the data provided by American Cancer Society, the number of deaths from breast cancer has been around 40,000 since 2003 (American Cancer, 2003; A Jemal, Siegel, & Ward, 2009; Society, 2008). However, the number of diagnosed cancer increases from 212,600 cases to 249,260, together with the breast cancer mortality rate decreases slightly over the past decade. This is partially due to the progress of more accurate diagnostic screening technology.

Breast Cancer Screening and Diagnosis

Breast screening is considered when women are without any symptoms of breast cancer to ensure early diagnosis (Bevers et al., 2009). However, this is impacted by a range of factors including family history and risk assessment, physical examination and patient's familiarity with breasts. Breast awareness as noticing breast changes is recommended for women of all ages. These changes include: lumps inside breast; swelling, warmth, redness or darkening of the breast; changes in shape or size of the breast; itchy or rash in the nipple; pulling in of the nipple or other parts of breast; sustaining pain in the breast.

Typically, for women with normal risk, breast cancer checks are determined by age ranges. For women with ages between 20 and 40, a clinical breast exam is suggested every 1 to 3 years. For women after 40, an annual check of clinical breast exam together with annual mammography screening is considered.

Compared to normal risk group, increased risk group of women includes women with one of the following conditions: thoracic irradiation history, a family history of breast cancer, lobular carcinoma in situ and prior history of breast cancer. Upon these conditions, a combination with practical breast awareness by self, annual clinical breast exam, annual mammogram, annual MRI and risk reduction strategies should be applied.

Diagnostic breast evaluation is different from normal breast screening/checks. A current uniform breast cancer risk assessment model is developed to identify those at increased risk with ages 35 and older (http://www.cancer.gov/bcrisktool/Default.aspx). This model is also called Gail model, which is based on several basic risk factors for breast cancer. Those factors includes age, mutation in either BRCA1 or BRCA2 gene, age of the patient's first menstrual period, age of given birth to the first child and family history of breast cancer etc. This model gives an approximate 5-year probability of developing invasive breast cancer, and the risk result will suggest the frequency of breast screening for a women.

Basically, mammograms and MRI are the most common ways to detect breast cancer in different age groups. However, mammograms suffer from low sensitivity around 60% and MRI suffers from low specificity less than 40%. Low sensitivity is associated with more undiagnosed cases, leading to patients diagnosed at a later stage with worse survival. Low specificity leads to a higher rate of false positive findings, meaning more normal patients will be diagnosed with high-risk of breast cancer, provoking unnecessary treatments, financial burden and psychological stresses for the misdiagnosed group.

When a woman is suspected with a high risk of breast malignancy from screening, breast biopsy is recommended as a further check. Three types of biopsy can be offered: Fine needle aspiration biopsy, core needle biopsy and excisional biopsy. Fine needle biopsy is less invasive with minimum cost, but it requires pathologist's expertise to interpret the results and follow-up tissue

biopsy. Core needle biopsy is more accurate, but the procedure is very invasive with multiple needle insertions on breast. Excisional biopsy is removing the entire suspicious breast mass by a surgeon and is the most invasive biopsy approach.

In summary, current technology on breast cancer diagnosis including screening, biopsy and risk assessment has defects in accuracy and invasiveness. An accurate, robust and non-invasive diagnostic approach is in pressing need.

Breast cancer molecular subtypes and classification

Breast cancer is a heterogeneous disease. There is increasing evidence showing there is diversity between tumors, within tumors and among different individuals (Polyak, 2011). This diversity/heterogeneity in breast cancer contributes to different risks of tumor progression and leads to different treatment responses (Blows et al., 2010). Thus, there is urgent calling to accurately classify breast cancer patients into clinically relevant subtypes (Dai et al., 2015). Sorlie et al. reported, by using immunohistochemistry (IHC) markers including ER, PR, HER2 and KI67, breast cancers can be categorized into five molecular subtypes: Luminal A, Luminal B, HER2, Basal and Normal-like (Sørlie et al., 2001). The basic IHC expression patterns and prognosis prediction of these subtypes are summarized in Table I.

Table I. Summary of breast cancer molecular subtypes

Molecular Subtypes	IHC marker	Prognosis	Prevalence
Luminal A	[ER+|PR+]HER2-KI67-	Good	23.7%
Luminal B	[ER+|PR+]HER2-KI67+ [ER+|PR+]HER2+KI67+	Intermediate Poor	38.8% 14%
Basal	[ER-PR-]HER2-, basal+	Poor	11.2%
Her2	[ER-PR-]HER2-	Poor	12.3%
Normal-like	[ER+|PR+]HER2-KI67-	Intermediate	7.8%

Luminal type breast cancers consist the most prevalent tumor types of breast cancer. Luminal A and Luminal B subtypes both express hormonal receptors patterns. While Luminal A subtypes have higher expression of ER-related genes, Luminal B subtypes have higher expression in proliferative genes (Sørlie et al., 2003). The prognosis for Luminal B type of patients is worse than those with Luminal A subtype. The hormonal receptor patterns of luminal subtypes determine the benefit of hormonal treatment for this group of patients, and people found that chemotherapy works poor in treating luminal breast cancers, compared with hormonal therapy.

HER2 over-expression tumors are over-expressing genes in HER2 amplicon or over-expressing HER2 protein as receptors on breast cells (Perou et al., 2000). TP53 mutation is associated with 40%~80% of HER2 tumors. HER2 subtype breast cancer has a poorer prognosis compared to Luminal A and Luminal B subtypes, with most of HER2 tumors of grade 3. Chemotherapy works significantly better in HER2 subtype compared to Luminal breast cancers. Molecular level HER2 protein targeted treatments, like Trastuzumab, have also been developed for HER2 breast cancers.

Basal tumors have been known as the worst prognosis breast cancer subtypes. Low expression levels of hormonal receptors and HER2, together with a high expression on basal markers (EGFR etc.) and proliferative genes, characterizes basal subtype (Perou et al., 2000). Basal tumors have been reported to account for 60%~90% triple negative breast cancer, which is very aggressive and lacking systematic targeted clinical therapy (Cheng Fan et al., 2006). The triple negative status of immunohistochemistry markers such as ER, PR and Her2 determine that targeted treatments on these markers are not applicable and chemotherapy is left as the only treatment option.

Other following studies tried to identify these molecular subtypes through gene expression profiling and found a diversity of gene-based signatures. PAM (Prediction Analysis of

Microarray) 50 is a well-known collection of 50 genes which are mostly related to hormonal receptor and proliferation (Parker et al., 2009). PAM 50 has been widely studied and proved to be clinically valuable and applicable in classifying patients for prognosis (Ades et al., 2014; Dowsett et al., 2013).

In summary, these subtypes are found to be representative of the molecular-level differences among breast cancers, and are proved to be effective in differentiating clinical outcomes. However, more accurate and personalized risk prediction and management strategy for breast cancer prognosis are needed and current effort for personalized prognosis prediction is discussed in the following section.

Breast cancer prognosis and prediction

Commercial genomic prognostic assays for breast cancer, including Mammaprint and Oncotype DX etc., are summarized in Table II (Weigel & Dowsett, 2010).

Table II. Commercially available prognostic tools for breast cancer.

Tool Name	Mammaprint	MapQuant Dx	MapQuant Dx simplified	Oncotype DX	Theros	Veridex
Platform	DNA microarray	DNA microarray	qRT-PCR	qRT-PCR	qRT-PCR	DNA microarray
Assay	70-gene signature	97-gene signature	8-gene signature	21-gene signature	2-gene ratio	76-gene signature
Availability	Europe, US	Europe	Europe	Europe, US	US	
FDA approval	Yes	No	No	No	No	No
Tissue	Fresh, Frozen	Fresh, Frozen	FFPE	FFPE	FFPE	Fresh, Frozen

Discovery Set	78 ER+/-, N0, < 5 cm diameter cancers; age<55 years	64 ER+ cancers	64 ER+ cancers	447 ER+ cancers	60 ER+, tamoxifen only treated cancers	115 ER+/-, N0 cancers
Predicted Outcome	Distant metastasis at 5 years	Good (GGI I) or poor (GGI III) prognosis	Good (GGI I) or poor (GGI III) prognosis	Disease free relapse at 10 years	Relapse-free and overall survival	Distant metastasis at 5 years
Results representation	Dichotomous; good or poor	Dichotomous; good or poor	Dichotomous; good or poor	Continuous recurrence score	Continuous recurrence score	Dichotomous; good or poor
Predictive value	Chemotherapy response (poor prognosis group)	Chemotherapy response (poor prognosis group)	Chemotherapy response (poor prognosis group)	Chemotherapy response (High recurrence score)	Chemotherapy response (High recurrence score)	Chemotherapy response (poor prognosis group)
Citation	(Van't Veer et al., 2002)	(Sotiriou et al., 2006)	(Toussaint et al., 2009)	(Paik et al., 2004)	(Ma et al., 2004)	(Yixin Wang et al., 2005)

FDA: US Food and Drug Administration; qRT-PCR: quantitative real-time reverse transcription polymerase chain reaction; FFPE: formalin-fixed paraffin-embedded; GGI: Genomic grade index

Mammaprint is the first genomic breast cancer prognostic assay which has been fully developed and approved by the US Food and Drug Administration (FDA) (Van't Veer et al., 2002). It is used for breast cancer prognostic prediction for patient with early stage (stage I, II) negative node and small tumors (sizes < 5 cm).

MapQuant is another microarray-based biomarker assay focusing on classification of ER+ grade II tumors into grade I-like and grade III-like (Sotiriou et al., 2006). In contrast, the simplified version of MapQuant, MapQuant Dx simplified uses the technique of qRT-PCR and consist of only eight genes with the accuracy performance comparable to its pioneer MapQuant (Toussaint et al., 2009).

Oncotype calculated recurrence score based on the expression of 21 genes. The recurrence score is latter used to classify distant relapse of ER+, lymph node negative tumor in 10 years (Paik et al., 2004). This test has been included by the National Comprehensive Cancer Network for recurrence prediction and therefore guiding the therapeutic decisions for early ER+ and lymph node negative breast cancer patients (Reis-Filho & Pusztai, 2011).

The following array called Veridex is developed focusing on prognostic markers independently discovered in ER+ and ER- groups (Yixin Wang et al., 2005). For ER+ breast cancer group, 60-gene array is found to be predictive for distant metastasis. In contrast, 16-genes array is discovered to be predictive for distant metastasis in ER- breast cancers.

Theros, a unique two-gene ratio (HOXB13 to IL17R) predictor of relapse-free survival and overall survival for ER+ patients (Ma et al., 2004). The response to endocrine treatment is also predicted with higher two-gene ratio suggesting a higher risk of recurrence.

However, among these existing predicting assays, only very few genes are found to be overlapped, due to a large number of highly correlated genes (Sotiriou & Pusztai, 2009). For example, only one gene (SCUBE2) in common between Mammaprint and Oncotype Dx. Furthermore, it has been shown that different predictive signatures have comparable and concordant risk assignments, in spite of the few shared genes (Cheng Fan et al., 2006). This leads to the pathway-based approach which will be discussed in the later chapters.

Chapter 2. A novel model to combine clinical and pathway-based transcriptomic information for the prognosis prediction of breast cancer

Abstract

Background

More accurate diagnostic methods are pressingly needed to diagnose breast cancer, the most common malignant cancer in women worldwide. Blood-based metabolomics is a promising diagnostic method for breast cancer. However, many metabolic biomarkers are difficult to replicate among studies.

Methods

We propose that higher-order functional representation of metabolomics data, such as pathway-based metabolomic features, can be used as robust biomarkers for breast cancer. Towards this, we have developed a new computational method that uses personalized pathway dysregulation scores for disease diagnosis. We applied this method to predict breast cancer occurrence, in combination with correlation feature selection (CFS) and classification methods.

Results

The resulting all-stage and early-stage diagnosis models are highly accurate in two sets of testing blood samples, with average AUCs of 0.968 and 0.934, sensitivities of 0.946 and 0.954, and specificities of 0.934 and 0.918. These two metabolomics-based pathway models are further validated by RNA-Seq based TCGA breast cancer data, with AUCs of 0.995 and 0.993. Moreover, important metabolic pathways such as taurine and hypotaurine metabolism and alanine, aspartate and glutamate pathway are revealed as critical biological pathways for early diagnosis of breast cancer.

Conclusions

We have successfully developed a new type of pathway-based model to study metabolomics data for disease diagnosis. Applying this method to blood-based breast cancer metabolomics data, we have discovered crucial metabolic pathway signatures

for breast cancer diagnosis, especially early diagnosis. Further, this modeling approach may be generalized to other omics data types for disease diagnosis.

Introduction

Breast cancer is the most frequently diagnosed cancer in women worldwide excluding skin cancer, and it is ranked second for the deaths among cancer patients (Society, 2015). Early diagnosis of breast cancer is crucial for patients' prognosis. However, current clinically diagnosed breast tumors have a median size of 2 to 2.5 cm (Singletary et al., 2002), which are likely to be later stage (stage III) breast tumors already metastasized to axillary lymph nodes. A highly accurate diagnostic test for breast cancer is currently lacking. The standard mammography test has sensitivities of merely 54% to 77% (Guth et al., 2008). Other diagnostic tools such as ultrasound, computed tomography (CT) and magnetic resonance imaging (MRI) are slightly more sensitive, however, they are costly. There is pressing need for more accurate, cost efficient and non-invasive alternative methods for breast cancer diagnosis.

Meeting the criteria above, metabolomics has quickly risen as a new method in the cancer biomarker field. As the final products of various biological processes, metabolites hold the promise as accurate biomarkers that reflect upstream biological events such as genetic mutations and environmental changes (Fiehn, 2002). Discoveries of altered metabolites and pathways will help to gain better understanding of dysregulated metabolism in tumor initiation and progression. Previous metabolomics studies have shown that certain metabolites can successfully differentiate patients from normal controls, or even classify sub-populations of certain diseases including breast cancer (Blasco et al., 2014; Budczies et al., 2014; Cai et al., 2010; Y. Fan et al., 2011;

E. Garcia et al., 2011; Pasikanti et al., 2010; Qiu et al., 2010; Tenori et al., 2015; J. Wei et al., 2011). For example, glutamate was found enriched in breast cancer patients and the glutamate-to-glutamine ratio was significantly correlated with ER status (Budczies et al., 2014). Serum profiles of breast cancer patients showed that histidine, glucose and lipids were strongly correlated with breast cancer relapse with a predictive accuracy of 75% (Tenori et al., 2015). However, similar to other types of biomarkers, metabolomics biomarker results are difficult to duplicate among different studies, due to a combination of reasons, such as the heterogeneity of the populations and study sizes, variability of the experimental protocols, noise in the metabolomics data, as well as the biological variations in the turnover rates of metabolites.

Given the observation that metabolites and enzymes involved in the same biological processes are often dysregulated together in cancer (F. Zhang & Du, 2012), we hypothesize that higher-order quantitative representations of metabolomics features, such as pathway-based metabolomics features, are coherent surrogates of metabolomics biomarkers and with more information of biological functions. To our knowledge, this idea had not been implemented in the context of metabolomics data, although proposed before in other types of omics data analysis, such as transcriptomics and genetics (GWAS and Exome-Sequencing) data. Towards this, we have developed a completely personalized, novel computational method for pathway-based metabolomics data analysis, using the non-parametric principle curve approach (Hastie & Stuetzle, 1989). We integrate metabolite features as pathway features, and subject them to feature selection and machine-learning classifications. This methodology is applied to identify breast cancer diagnosis biomarkers, especially for early pathological stages. The resulting classification models are highly accurate for breast cancer all-

stage diagnosis (AUC=0.986) and early-stage diagnosis (AUC=0.995) in the plasma training set. Moreover, these models predict equally impressively in plasma testing and serum validation samples, with AUCs of 0.923 and 0.995 for the all-stage diagnosis, and AUCs of 0.905 and 0.902 for early-stage diagnosis. We have discovered several critical pathways for breast cancer early diagnosis, including taurine and hypotaurine metabolism and alanine, aspartate and glutamate metabolism.

Methods

Study population. Three data sets are used in this study: two metabolomics data from our own group, and one RNA-Seq data set from TCGA breast cancers. The first metabolomics cohort is composed of 132 breast cancer and 76 control plasma samples and the second independent set has 103 breast cancer and 31 control serum samples. All samples were obtained from City of Hope Hospital. This study was approved by the institutional review boards (IRB) of City of Hope National Medical Center. All participants signed an informed consent before they participated in the study. Additionally, we downloaded TCGA breast cancer RNA-Seq data from 1082 tumor and 98 tumour adjacent normal controls (Cancer Genome Atlas, 2012), from the TCGA data portal https://tcga-data.nci.nih.gov/tcga/. Patient characteristics, staging of disease and other parameters are shown in Table I.

Data set configurations for diagnostic model training, validation, and testing. For the all-stage diagnosis model, we used 80% of the plasma (106) and 80% of the control (61) samples as the training data. We employed three testing data sets including: (1) the remaining 20% of the plasma (26) and 20% of the control (15) samples as the first hold-out testing data; (2) the entire 103 breast cancer and 31 control serum samples; (3) a cohort of 98 pairs of age-matched breast cancer TCGA RNA-Seq data. There is no sample overlap between the training and testing set. To train the early stage diagnosis model, we used the

subset of Stage I (15 samples) and II (37 samples) of the training data in the all-stage diagnosis model described above, in combination with the 61 healthy control samples.

Collection and storage of blood serum and plasma. Fasting serum and plasma specimens were collected in the morning before breakfast from all the participants. The samples from controls were obtained from healthy volunteers. The breast cancer patients were newly diagnosed and were not recurrent or on any medication prior to sample collection. All samples were placed into clean tubes and immediately stored within two hours of collection at -80 °C until analysis.

Metabolic profiling. LC-TOFMS and GC-TOFMS were used for the metabolomics profiling of all blood samples in the study. The profiling procedure (sample preparation, metabolite separation and detection, metabolomics data pre-processing, metabolite annotation, and, finally, statistical analysis for biomarker identification) was performed. To eliminate batch effect, all of the plasma samples were processed in one batch, so were all of the serum samples. Experimental details are provided in the appendix. All annotated metabolites from GC-TOFMS and LC-TOFMS datasets were combined and exported to SIMCA-P+ 12.0 software (Umetrics, Umeå, Sweden) for multivariate statistical analysis.

Pathway mapping of metabolites. The names of metabolites are standardized by linking them to Human Metabolome Database (HMDB) IDs, with consideration of synonyms. A comprehensive master file was created, which contains the mapping information between 310 human metabolic pathways and affiliated metabolites. Pathway and metabolite information is extracted from HMDB (Wishart et al., 2013), Small Molecule Pathway Database (SMPDB) (Jewison et al., 2014), Kyoto Encyclopedia of Genes and Genomes (KEGG) (M. Kanehisa & S. Goto, 2000), Recon 2 (Thiele et al., 2013), IPA (QIAGEN's Ingenuity® Pathway Analysis, IPA®, QIAGEN Redwood City, www.qiagen.com/ingenuity), FLink (FLink: Frequency weighted links. Available from: http://ncbi.nlm.nih.gov/Structure/flink/flink.cgi) and PubChem (Bolton, Wang, Thiessen,

& Bryant, 2008). Most of the metabolites could be mapped to pathways by the master file. The remaining unmapped metabolites were manually searched in literatures.

Pathifier algorithm. We used the R package *pathifier* (Y. Drier, M. Sheffer, & E. Domany, 2013) to perform pathway-based metabolite sets analysis. The details about *pathifier* were described elsewhere (Y. Drier et al., 2013; S. Huang, C. Yee, T. Ching, H. Yu, & L. X. Garmire, 2014). Briefly, this algorithm transfers the information from metabolites-level to pathway-level by inferring the pathway deregulation score (PDS) for each sample in each pathway. This PDS score is an individualized pathway-level measurement of abnormality. The normal condition samples are utilized to construct a principal curve, which is then smoothed. Every sample will be projected to the smoothed principal curve and the PDS score is the normalized projection distance for each pathway of each sample. If the sample differs from others more in a particular pathway, then the projection distance to the curve is larger and leads to a higher PDS score for this pathway.

Feature selection and evaluation of classification models. For feature selection from the training data, we used the correlation feature selection (CFS) method implemented in Weka (Hornik, Buchta, & Zeileis, 2009) with 10-fold cross validation. CFS is a machine learning method that selects features with the highest correlation to responses and lowest correlation with other selected features (Hall, 1999). In the 10-fold cross validation step, training data is split into 10 parts and 9 of them were used as the actual training set while the rest 1 part as validation set, such that a set of features were selected by CFS. We repeated this process 10 times among different parts, and kept the features that were selected 10 out of 10 times (100%). To select the best suited classifier, we evaluated the performance of three classification methods: logistic regression, SVM and random forest on training dataset on the same set of CFS-selected features. We used a comprehensive list of metrics that include AUC, sensitivity, specificity, Matthew's correlation coefficient (MCC) and F-statistic.

TCGA RNA-Seq analysis. Breast cancer TCGA RNA-Seq data were downloaded from the data portal https://tcga-data.nci.nih.gov/tcga/ on 10/23/2015 (Cancer Genome Atlas, 2012). We included 1082 breast cancer samples with 98 control samples. For pathway level analysis, we implemented *pathifier* algorithm on the RNA-Seq data and applied limma's differential t-test to compare the pathway level results with our study. For metabolite level analysis, the enzyme (gene) information for featured metabolites were extracted from KEGG and SMPDB. Limma's differential t-tests were used for calculation of the p-values for each enzyme (gene). Barplots were used for comparison between metabolites and the related enzyme (gene) in breast cancer and normal samples.

Metabolite based model comparison. We built the metabolites-based model on the same plasma training data set. We conducted feature selection and classification the same way as pathway-based models, so that they are comparable. Specifically, we used the correlation feature selection (CFS) method implemented in Weka with a 10-fold cross-validation for feature selection. We implemented logistic regression models for all-stage and early-stage classification to compare with pathway-based model.

Power analysis of diagnosis model. To ensure the adequacy of our pathway-based metabolomics model, we calculated the sample size and statistical power using the module implemented in MetaboAnalyst (Xia, Sinelnikov, Han, & Wishart, 2015), where the implementation was described by van Iterson et al (van Iterson, van de Wiel, Boer, & de Menezes, 2013).

Data availability. All the input metabolomics data used for this study are deposited in Metabolomics Workbench: http://metabolomicsworkbench.org/ (Project ID PR000284). Additionally, the metabolites mapped to pathways are included in Supplementary File mapped_metabolites_names.csv. The R scripts for pathway mapping, PDS matrix generation and logistic regression are available at

http: //www2.hawaii.edu/~lgarmire/MetaboloPathwayModel.htm.

Results

Data sets and the analysis workflow

Three cohorts are used in this study: two of them are our own metabolomics profiling data sets from independent plasma and serum samples, and the third cohort is the TCGA breast cancer RNA-Seq data (to test the generalization of the pathway-based model across data types). The metabolomics data include newly diagnosed pre-treatment samples from (1) 132 breast cancer and 76 control plasmas, and (2) 103 breast cancer and 31 control serums. For the two cohorts of plasma and serum samples, we conducted metabolomics experiments by both liquid chromatography time-of-flight mass spectrometry (LC-TOFMS) and gas chromatography time-of-flight mass spectrometry (GC-TOFMS). According to the power analysis tool in MetaboAnalyst (Xia et al., 2015), the study achieves a power of 0.84 (supplementary figure 1), supporting the adequacy of the metabolomics data. The physiological and clinical information, such as age, ethnicity and tumor stage for the plasma, serum data and TCGA sets are summarized in Table I.

To analyse the metabolomics data, we have developed a novel computational pipeline that identifies pathway-based biomarkers for blood-based breast cancer diagnosis (Figure 1). The essence of the approach is to transform metabolite-level information to completely personalized pathway-level information. The overall workflow of the pathway-based model and the analysis process is as follows:

First, metabolites are mapped to their standardized Human Metabolome Database (HMDB) IDs, and the pathway-metabolite relationships are summarized in a master file from multiple resources, including Human Metabolome Database (HMDB), Kyoto Encyclopedia of Genes and Genomes (KEGG), Small Molecule Pathway Database (SMPDB), IPA, FLink, Recon 2 and PubChem. Next, we used the *pathifier* algorithm to convert the raw metabolite-based data matrix to the pathway-based matrix that contains pathway dysregulation scores (PDS). Pathifier

is a non-parametric method for dimension reduction, where a one-dimensional Principle Curve is derived from a cloud of data points in the high-dimensional space; The PDS is a metric for the degree of pathway abnormality per patient, and it is the distance on the Principle Curve from the starting point to the point projected by a particular and individualized pathway (Y. Drier et al., 2013; Hastie & Stuetzle, 1989). A PDS ranges from 0 to 1, where a score closer to 1 indicates a more aberrant pathway. Then, we used the PDS matrix from 80% of qualified plasma set to train classification models. We selected plasma set to train the classification models, as it has a larger sample size and more complete information of tumor stages. The details of feature selection and classification to train the models, and model testing with three different data sets are described in the following.

Metabolic-pathway based all-stage diagnostic model for breast cancer

We first investigated the metabolomics-based pathways as biomarkers to predict breast cancers composed of all stages of tumors (Figure 2). To select the best set of features that are maximally relevant and minimally redundant, we used CFS with 10-fold cross-validation on the plasma training dataset, which is composed of 80% of breast cancer and 80% healthy controls. With these selected features (Figure 2C), we evaluated three widely used classification methods: logistic regression, SVM and random forest on the plasma training data set. The resulting performance metric AUC (0.986) show that logistic regression performs the best among the three methods (Supplementary Table I). We thus used logistic model as the model of choice to evaluate three other testing datasets: the 20% hold-out plasma testing samples, the entire serum sample set, and a cohort of 98 pairs of age-matched breast cancer RNA-Seq data from TCGA. Note for TCGA data, we generated the pathway dysregulation scores (PDS) and extracted the values for the same features as the training data set. Although these three data sets are generated from different populations and technology platforms, our hypothesis is that pathway-based features should represent true biology and therefore the model based on metabolomics data should generally predictive.

The resulting metabolic pathway-based diagnostic model performs very well in all three testing data sets, with AUCs of 0.923, 0.995 and 0.9946 in the hold-out plasma testing samples, serum samples and TCGA RNA-Seq set, respectively (Figure 2A). Moreover, other statistical metrics such as the sensitivity, specificity, MCC and F-statistic are also outstanding, confirming the robustness and generality of the pathway-based model (Figure 2B). The even superior performances of the model on serum metabolomics data and TCGA RNA-Seq data are surprising. This may be due to the more complete lists of metabolites in serum and genes in RNA-Seq data, compared to the plasma samples. The good AUC obtained from the age-matched TCGA RNA-Seq data suggest that age is unlikely a driving factor leading to accuracy of the classification from the metabolomics based pathway-model. Nevertheless, we further examined if age is a dominant confounding factor in the metabolomics training data. For this, we divided the plasma data into subset 1 with 35 pairs of age-comparable samples and the other subset 2 with 97 breast cancer and 41 age-incomparable controls. If diagnosis signals were driven by age, then a model trained on age-incomparable subset 2 would have very poor prediction on subset 1 where the ages among these samples are comparable. However, a new model on age-incomparable subset 2 still achieves a very high AUC of 0.913 on age-comparable subset 1. Thus the pathway features (Figure 2C) in the earlier model are predictive of breast cancer diagnosis.

The relevance of these eight pathway features to diagnosis, as measured by Mutual Information (MI), is listed in the following descending order: taurine and hypotaurine metabolism, glutathione metabolism, methionine metabolism, glycine serine and threonine metabolism, phospholipid biosynthesis, propanoate metabolism, cAMP signaling pathway and mitochondrial beta-oxidation of medium chain saturated fatty acids. Interestingly, none of the pathways has an MI more than 0.5, indicating the complexity of the disease and the significance of pathways collectively. Among them, taurine and hypotaurine metabolism stands out as the most important pathway (MI=0.386). Hypotaurine is a product of enzyme cysteamine

dioxygenase in this pathway, involved in protecting against oxidative stress and cancer-induced membrane damage (Brand, Leibfritz, Hamprecht, & Dringen, 1998; Gossai & Lau-Cam, 2009). Taurine and hypotaurine metabolic pathway has been shown to be relevant to multiple types of cancers, such as ovarian, lung, colon and renal cancers (Fong, McDunn, & Kakar, 2011; Pradhan, Desai, & Palakal, 2013; Roy et al., 2014; Tiruppathi, Brandsch, Miyamoto, Ganapathy, & Leibach, 1992). Here for the first time, we have discovered that taurine and hypotaurine metabolism is also dysregulated in the blood samples of breast cancer. In order to confirm the significance of each pathway at the transcriptome level, we crosschecked pathway-level expression results using TCGA RNA-Seq data. The pathway level results of two data types are consistent overall as expected (Supplementary Table II). For example, taurine and hypotaurine metabolism pathway has a significant p-value of 1.01E-25 for the differential test in the metabolomics data, and it is also a top-ranked pathway with a p-value of 7.40E-9 in the RNA-Seq data.

Next, we identified the measurable metabolites in these selected pathways from both plasma and serum samples and presented their average log fold changes in tumor versus control samples (Figure 2D and Supplementary Table III (A)). Hypotaurine is the primary metabolite in the leading significant taurine and hypotaurine pathway, and it has 2.41-fold (0.0086 vs. 0.0025) amount in the tumor as in normal plasma samples. Pyruvate, the most central metabolite in the cell and a common component of glycine, serine and threonine metabolism and taurine and hypotaurine metabolism pathway, is consistently higher in breast cancer blood samples (Figure 2D and Supplementary Table III (A)). From control to cancer conditions, it has 1.82-fold increase in the plasma, and 2.89-fold in the serum samples (Figure 2D and Supplementary Table III (A)). Interestingly, several amino acids are lower in cancer samples compared to controls, including succinate (1.69-fold decrease in plasma, 4.58-fold decrease in serum), choline (1.23-fold decrease in plasma, 4.58-fold decrease in serum), serine (2.72-fold decrease in plasma, 1.13-fold decrease in serum), glycine (1.25-fold decrease in plasma, and

1.83-fold decrease in serum) and alanine (1.11-fold decrease in plasma, and 1.62-fold in serum (Supplementary Table III (A)). Decrease of glycine and alanine levels in plasma and serum of breast cancer have been reported before (Miyagi et al., 2011; J. Shen, Yan, Liu, Ambrosone, & Zhao, 2013). Choline, serine and glycine are the major component of glycine, serine and threonine metabolism, glutathione metabolism and methionine metabolism, whereas succinate is the major component of propanoate metabolism and cAMP signalling pathway. Similarly, glycerol-3-phosphate in phospholipid biosynthesis is significantly lower in the cancer samples, with a 6-fold decrease in plasma. The comparisons between some key metabolites in our metabolomics study and the corresponding enzymes from TCGA RNA-Seq data are shown in Supplementary Figure 2. Overall, the directions of changes in metabolites are consistent with those of corresponding enzymes.

Metabolic pathway based early-stage diagnostic model for breast cancer

Early detection of breast cancer is critical to improve patients' survival. Due to the small sample size (n=16) in Stage I, we combined the samples in stage I and II as early-stage cancers, and constructed a sub-model to diagnose early-stage breast cancer, similar to the previous all-stage diagnosis model. As expected, the pathway-based early-stage diagnostic model performs very well on the training data set, with AUCs of 0.995. Moreover, it also predict very well on the three testing data sets, with AUCs of 0.905, 0.902 and 0.999 in the 20% hold-out plasma testing, serum, and TCGA breast cancer samples (Figure 3A). Other model performance metrics also yield satisfactory results in both data sets, supporting the excellence of the early diagnostic model (Figure 3B).

Eight key pathways are identified as diagnostic features for early stage breast cancer detection (Figure 3A), namely taurine and hypotaurine metabolism, alanine aspartate and glutamate metabolism, protein digestion and absorption, purine metabolism, malate-aspartate shuttle, cAMP signalling pathway, propanoate metabolism and biosynthesis of unsaturated fatty acids,

in the descending order of significance. Similar to the all-stage diagnosis model, taurine and hypotaurine metabolism is again the top-ranked pathway (MI=0.414, Figure 3C), indicating its significance as a new signature for early stage breast cancer detection. Alanine, aspartate and glutamate metabolism is a new pathway feature selected by the early stage diagnosis model, largely due to the increase of aspartate from 0.063 to 0.182 and decrease of asparagine from 0.091 to 0.038 in the cancer and control plasma samples, respectively. This implies a transformation relationship from aspartate to asparagine from normal to cancer. The cAMP signalling pathway has been intrinsically linked to a variety of pathways such as PI3K pathway, and antibodies directed against the soluble adenylyl cyclase that catalyses cAMP have been shown as highly specific markers for melanoma (Desman, Waintraub, & Zippin, 2014; Rodriguez & Setaluri, 2014). To further confirm the significance of our finding, we calculated the differences of the above eight feature pathways between tumor vs. control sample, using the metabolomics data and TCGA RNA-Seq data. The pathway level results from the two data types are both significant (Supplementary Table II).

At the metabolite level, some key metabolites are preserved in the early-stage diagnosis sub-model (Figure 3D), compared to the all-stage model (Figure 2D). They include cysteine, glutamine and asparagine, which have higher concentrations in early-stage tumor samples; as well as alanine and aspartate, which are decreased during early tumorigenesis. The finding that aspartate, the precursor of beta-alanine (Marshall, 1965), is significantly and robustly lower even in early stage breast cancers is a very interesting finding, and this further confirms that dysregulations of amino acid metabolism and metabolites are early events associated with breast cancer tumorigenesis (Miyagi et al., 2011). We summarized the averaged expression of the key metabolites and the differential test p-values in Supplementary Table III (B). We also compared the relationship of the key metabolites from our study and the enzymes transforming those metabolites from TCGA RNA-Seq data in Supplementary Figure 3. Both sets of results show consistent trends in general.

Integrative analysis of key pathways and metabolites

Metabolic regulation is elaborately related with cancer initiation and progression, as proliferating cells demand nutrients for energy production as well as synthesis of genetic materials, proteins and lipids (Fiehn, 2002; F. Zhang & Du, 2012). Although the feature pathways identified by diagnostic and early diagnostic models are different, they are nevertheless interconnected in the cellular context (Figure 4). Alanine, glutamine and aspartate metabolisms are interconnected, and we observe consistent decreasing trends of alanine, glutamine and aspartate in cancer vs. normal samples. Moreover, the amino acid, glucose and phospholipid metabolisms can be inter-connected through glutaminolysis, a process that supplies carbon and nitrogen resources to the growing and proliferating cancer cells (Dang, 2010).We also summarize the overlap of metabolites from pathways featured in the all-stage diagnosis and early-stage diagnosis. Common metabolites important to the two models are beta-alanine, glycine, serine, lactate, succinate, oxoglutarate, alanine, 3-hydroxybutyrate, methionine, valine, cadaverine and pyruvate, all functionally linked to glutaminolysis (Supplementary Figure 4).

Comparison of pathway-based and metabolite-based metabolomics models

To evaluate the pathway-based metabolomics diagnosis modeling approach with the commonly used metabolite-based approach, we constructed a "baseline" metabolite-based model, using exactly the same CFS feature selection and logistic regression steps as done in our pathway-based method. Since the AUC values indicate that the early-stage model is less likely to have over-fitting, we use the early-stage breast cancer data to compare the pathway-based and metabolite-based diagnosis models. In the training data set, the pathway-based approach performs slighter better with an AUC of 0.995, compared 0.988 in the metabolite-based approach (Figure 5). Similar trend also exists in the testing data set, where the pathway-based model yields an AUC of 0.905, whereas the metabolite-based model has an AUC of 0.888 (Figure 5).

The FDA approval of biomarkers requires the demonstration of the biomarker candidate functions (Katz, 2004), we thus built single-variate logistic models to show the diagnostic potential of individual pathway or metabolite features selected by the models. Comparatively, the top pathway features show better disease association than the top metabolite features (Supplementary Table IV). In the pathway-based model, taurine and hypotaurine metabolism yields the most statistical significance ($p<2E-16$, t-test) followed by protein digestion and absorption pathway ($p= 3.5E-10$, t-test). On the other hand, in the metabolite-based model, the most significant metabolite cysteine (HMDB00192) has significant p-value of 2.22E-9. These results indicate that the top individual pathway feature may have better diagnostic performance than metabolites.

To investigate the effect of the number of pathways on the performance of pathway-based model, we conducted sensitivity analysis exemplified by early-stage diagnosis model. We randomly selected ½ (51) of the initial 101 pathways within exactly the same training sample sets, and applied the same CFS feature selection criteria with 10-fold cross validation. CFS selects 6 pathways for early-stage model (Supplementary Table V). We imposed logistic regressions on these selected features and compared the changes in AUCs due to changes in pathways. Reducing the initial number of pathways decreases the performance of the models, as expected. In the training data, the half-size pathway-based early stage diagnosis model has a slight decrease of AUC from 0.995 to 0.948. Such decrease is more pronounced in the serum testing data from 0.903 to 0.753. Similar trends are observed in the all-stage diagnosis model.

Discussion

Summary of discoveries

Metabolomics provides the most direct measurement of phenotypic changes, since it reflects the final molecular result of the combination of all upstream genetic, transcriptomic and proteomic changes (Denkert et al., 2012). The relative incomplete coverage of metabolomic measurements has been a challenge for their use for diagnostic classifiers. In this study, we

address this challenge by using a new metric of personalized pathway dysregulation score. This score can interpret the metabolomics data in the context of the metabolic pathways on individual patient level, thereby enables us to discriminate the differences in specific pathways between cancer and normal samples. This approach accurately predicted all-stage breast cancer patients from normal controls (AUC=0.968). It even detected early-stage (stage I and II) breast cancers with excellent accuracy (averaged AUC=0.904 in two testing sets). In addition to the increased power achieved by integrating concerted metabolic changes as described in this paper, our pathway-based classifiers can potentially offer deeper biological insights as to which cellular processes are dysregulated in breast cancer. We have discovered novel critical pathways, such as taurine and hypotaurine metabolism and alanine, aspartate and glutamate metabolism related to glutaminolysis, for the early diagnosis of breast cancer.

A new paradigm to use pathways as features of biomarker classification models

Conventionally, almost all metabolomics studies aim to identify metabolites as biomarkers. Even among the few studies that involve the systematic pathway approach (Borgan et al., 2010; Krumsiek, Suhre, Illig, Adamski, & Theis, 2011, 2012; Nam, Chung, Kim, Lee, & Lee, 2009)), none of them has developed a computational methodology to employ pathways as input features for the downstream statistical or machine learning modeling of biomarkers diagnosis or prognosis

The disadvantages of using metabolites as predictors of biomarker diagnosis or prognosis models are obvious: low reproducibility. This could be due to various reasons, such as the heterogeneity of the populations and small study sizes, variability of the experimental protocols and technical noise in the metabolomics data. In fact, we compared the multiple studies that had attempted to identify metabolites in blood as biomarkers for breast cancer previously (Jobard et al., 2014; Oakman et al., 2011; Poschke, Mao, Kiessling, & de Boniface, 2013; J. Shen et al., 2013), and found little overlap or even controversies among the studies (Supplementary Figure 5) (Asiago et al., 2010; de Leoz et al., 2011; Miller et al., 2015; Miyagi et al., 2011; Oakman et al., 2011; Poschke et al., 2013; J. Shen et al., 2013;

Tenori et al., 2015; C. Yang, Richardson, Smith, & Osterman, 2007). On the contrary, many metabolites in the featured pathways that we have found with our method coincide with previous reports, such as increases of alanine, pyruvate and lactate, as well as decrease of choline in cancer samples. Thus the pathway-based method is more tolerable to heterogeneity of the population, compare to the metabolite based biomarker approach. Furthermore, the tolerability of pathway-based method to population heterogeneity is also manifested through embracing age differences by the pathway features. The models predict fairly well on 3 different sets of testing data, even when the ages are matched. Moreover, biologically motivated feature selection approach offers systems level and biological level insights, which the metabolite-based models lack. Such system level knowledge is very critical as we move forward towards developing intervention strategies for cancer prevention or therapeutics strategies for cancer treatment. Biological system is highly robust with redundant components, and attacking the higher-level structures such as pathways offers a better strategy than changing the expression of lower-level components such as genes or metabolites.

The workflow that we propose here is a fully personalized pathway-based diagnostic modeling framework for metabolomics data. Moreover, it is compatible with conventional metabolite-based predictive modeling approach after the step of input matrix transformation. This methodology represents generalization of the pathway-based predictive modeling philosophy, which we had exemplified earlier using the transcriptomics and clinical data to predict breast cancer prognosis (S. Huang et al., 2014). The most distinguished characteristic of our method, is that it summarizes the contribution of potentially correlated metabolites in the same pathway into a single metric of PDS score, on a patient by patient basis. It not only preserves the individual patient information before classification, but also gives direct numerical value (rather than the rank) per pathway per patient. Doing so provides bountiful flexibilities to use pathways as features for various downstream analysis, exemplified here as diagnosis biomarker

modeling. The applications are far beyond disease diagnosis though. For example, one could also use the new data matrix of PDS scores to perform clustering or survival analysis. On the other hand, other bioinformatics tools for metabolomics analysis, such as MetaboAnalyst (Xia et al., 2015) and Metabolite Set Enrichment Analysis (Xia & Wishart, 2010), either use pathway enrichment *post hoc* or lose individual patient's value during the set enrichment analysis.

Perhaps the most powerful utility of this modeling approach, is that the pathway features may be generalized to other omics platforms, despite the differences in experimental protocols, masses that are measured (metabolites, mRNAs, proteins etc.) and their units. Here we have demonstrated that the pathway features obtained from metabolomics data have excellent predictive performance in TCGA breast cancer RNA-Seq data, where both the sample sources and technical platform are different from the metabolomics datasets. Moreover, by projecting metabolites profiles to pathway profiles, metabolomics data can be integrated with other types of omics data such as RNA-Seq gene expression, DNA methylation and copy number variation data.

Important discoveries of altered pathways during carcinogenesis

Our results demonstrate that taurine and hypotaurine metabolism is the most indicative pathway for breast cancer diagnosis. Taurine, converted from hypotaurine by hypotaurine dehydrogenase, is intricately linked with alanine and glutamate metabolism (Figure 4). Although it is the first time for us to report this significant pathway in breast cancer early diagnosis, many lines of evidence suggest this is a critical pathway in tumor development. Hypotaurine is known to modify the indices of oxidative stress and membrane damage, both of which are associated with cancers (Bucak et al., 2009; Gossai & Lau-Cam, 2009). Additionally, others have linked this pathway to worse prognosis in ovarian, kidney, colon and lung adenocarcinoma (Pradhan et al., 2013; Roy et al., 2014; Tiruppathi et al., 1992; W. Yang et al., 2014). Moreover, glutamate decarboxylase 1, a key enzyme in taurine and hypotaurine

metabolism, has been identified as a tissue biomarker for benign and malignant prostate cancer (Jaraj et al., 2011).

We also found alanine, aspartate and glutamate metabolism together with malate-aspartate shuttle to be significant pathways in early stage diagnosis model. Aspartate is the key metabolite that shows significant lowered level in breast cancer blood samples (Supplementary Table III (B). Aspartate is produced from oxaloacetate by a transamination process. It participates in urea cycle to facilitate the removal of ammonia and it also acts in the biosynthesis of pyrimidine for translocating NADH into mitochondria. Interestingly, the lower level of aspartate in the blood is reversely associated with increased aspartate in the breast cancer tissue and cell lines (Xie et al., 2015), suggesting that the aspartate pool in the blood is utilized to supply more aspartate in breast cancer cells. Consistent with this hypothesis, asparagine synthetase, the enzyme that generate asparagine from aspartate, was overexpressed under glucose deprivation in pancreatic cancer cells to protect against apoptosis(H. Cui et al., 2007).

Perspectives and future work

In this study, we have proposed a new and personalized pathway-based approach to integrate metabolites-level metabolomics data, in the application of breast cancer diagnosis. The success of this type of pathway models first relies on data obtained through a profiling (rather than targeted) approach where as many metabolites/genes as possible are recorded. Compared to other omics data types, metabolomics data are much less standardized across different studies, and data repositories are lacking (Berg et al., 2013; Johnson & Lange, 2015). A community effort needs to be devoted to improve data sharing, in order to accumulate statistically well-powered data sets to predict disease diagnosis and prognosis. To drive our modeling approach towards clinical diagnosis, we are planning to build a large database to store the metabolomics profiles as references. In the model construction step, samples will be labeled as cancer/normal classes are used, and their individual pathway scores (normalized scores between 0 and 1) will be calculated as inputs subject to feature selection and classification step. When a new sample arrives, the metabolite profile will be normalized relative to the database, a new vector of PDS

scores will be calculated after the same metabolite-to-pathway transformation. The classification model can then call for the probability for this new sample as being normal or cancerous. Depending on the accuracy of prediction in the new sample, we can elect to incorporate it into the training data set and re-train the model, thus improving the predictive power of the model over time. Moreover, from the new patient's PDS profile we can also infer the aberrant pathways, and identify problematic metabolites (and associated enzymes) for this specific patient. Therefore, the discoveries could be used for not only diagnosis prediction but also precision medicine.

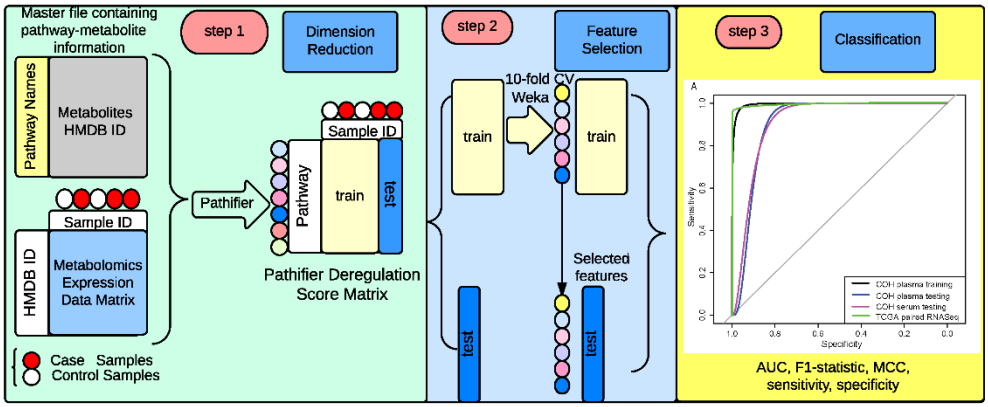
Conclusions

We have successfully developed a new type of pathway-based model that uses metabolomics data for disease diagnosis. Applying this method to blood-based breast cancer metabolomics data, we were able to discover crucial metabolic pathway signatures for breast cancer diagnosis, which may be valuable for diagnostic tests and therapeutic interventions (Yizhak et al., 2014). Further, this modeling approach can be broadly applicable to other omics data types for disease diagnosis.

Appendix A: Chapter 2 Figures

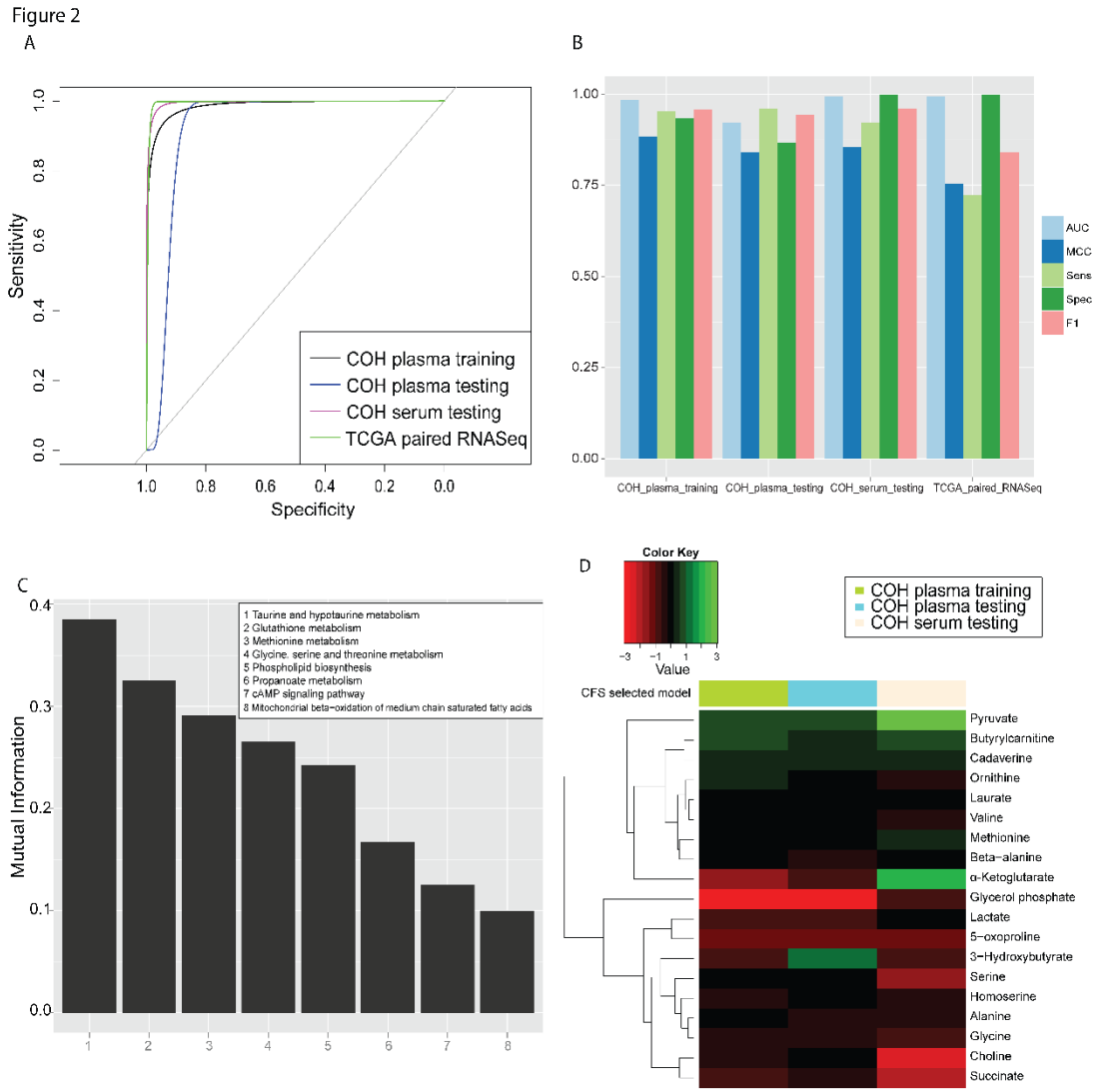
Figure 1. The workflow of pathway-based metabolomics data analysis.

Figure 1



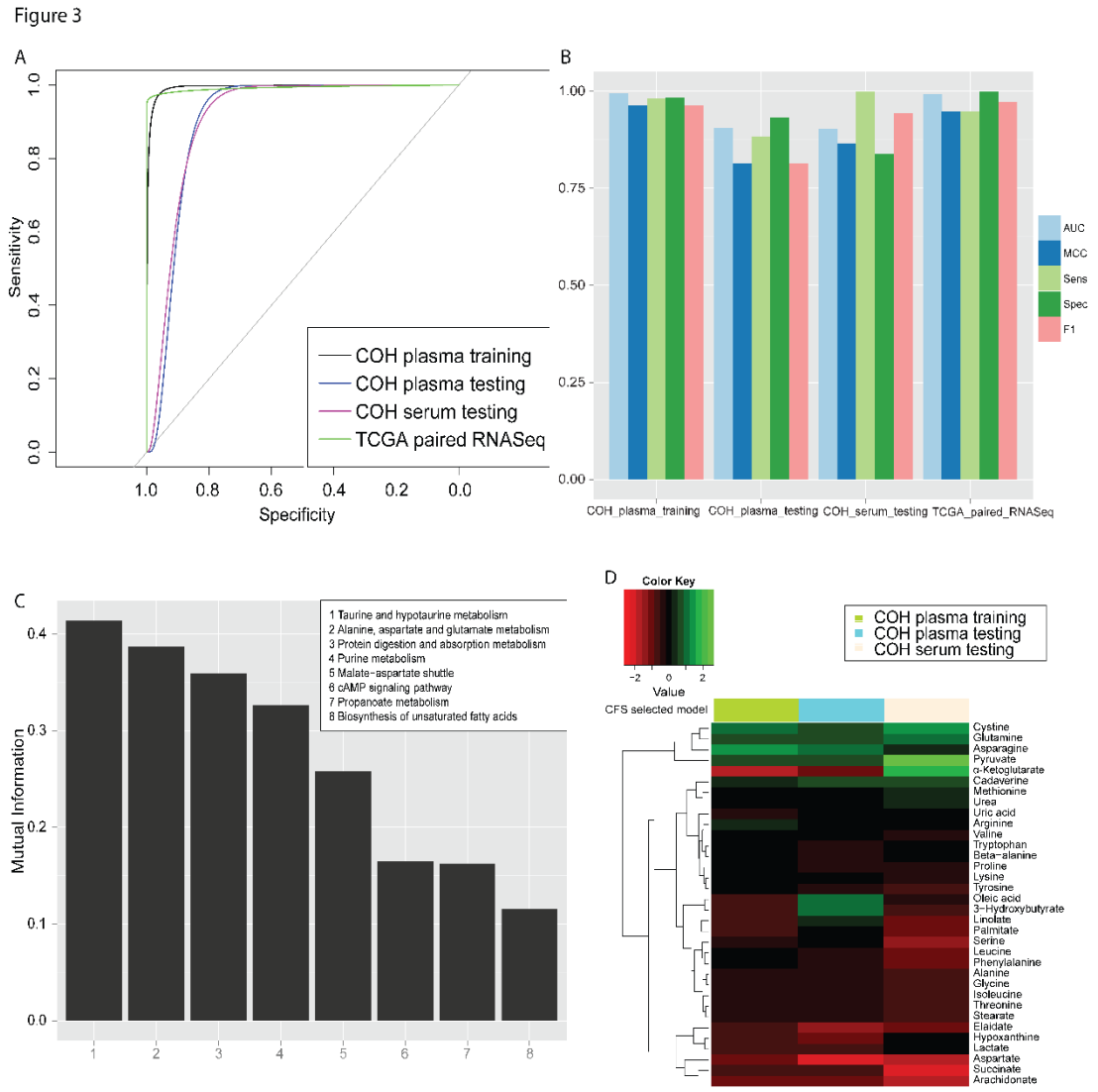
Step 1. Conversion from metabolite- to pathway-based metabolomics data. The input data include the master file containing pathway-metabolites mapping information, the metabolomics profiling data and the normal/tumor classification vector. The metabolomics-level data are transformed to pathway-level data by the *pathifier* algorithm. The output file of *pathifier* is the Pathway Deregulation Score matrix within which each score measures the deregulation of a specific pathway for a specific sample. **Step 2. Model construction**. Qualified COH plasma samples are split by 80/20 for training and holdout testing data. Correlation feature selection (CFS) is used for feature selection and the logistic regression model is used for classification. 10-fold cross-validation is applied with CFS feature selection in the plasma training dataset. Two models are constructed: all-stage diagnostic model and early-stage diagnostic model. **Step 3. Model evaluation.** The model performance is assessed using ROC curves and various metrics including AUC, MCC, Sensitivity, Specificity and F-statistic.

Figure 2. Analysis of the performance of the all-stage diagnosis model for breast cancer.



80% of controls and cases in COH plasma data set are used to train the model. The remaining COH plasma data (20%) and COH serum data set are used as the testing set and validation set. A. ROC curves for the all-stage breast cancer diagnosis from different data sets. B. AUC, MCC, Sensitivity, Specificity and F1-statistic to measure the performance of the all-stage diagnosis model. C. Mutual information for pathway features selected by the all-stage diagnosis model. D. Log fold change of metabolites associated with the selected pathway features, by comparing cases to the controls across different data sets.

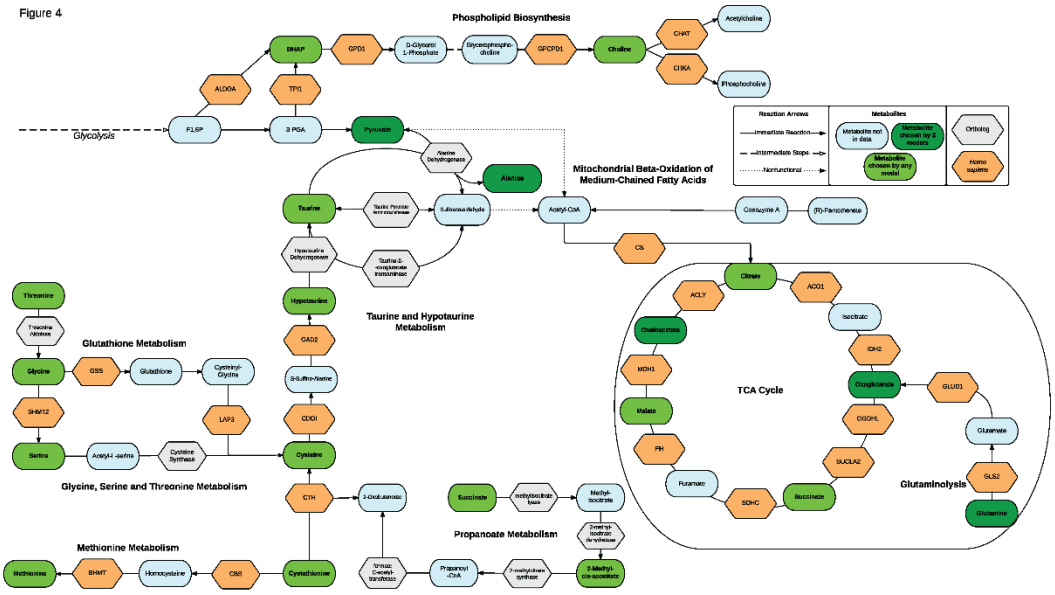
Figure 3. Analysis of the performance of the early-stage diagnosis model for breast cancer.



80% of controls and early-stage (stage I and II) cases in COH plasma are used to train the model. The remaining controls and early stage cases in COH plasma samples, as well as controls and early stage cases in COH serum data are used as the testing and validation set. A. ROC curves for the early-stage breast cancer diagnosis from different data sets. B. AUC, MCC, Sensitivity, Specificity and F1-statistic to measure the performance of the early-stage diagnosis model. C. Mutual information for pathway features selected by the all-stage

diagnosis model. D. Log fold change of metabolites associated with the selected pathway features, by comparing cases to the controls across different data sets.

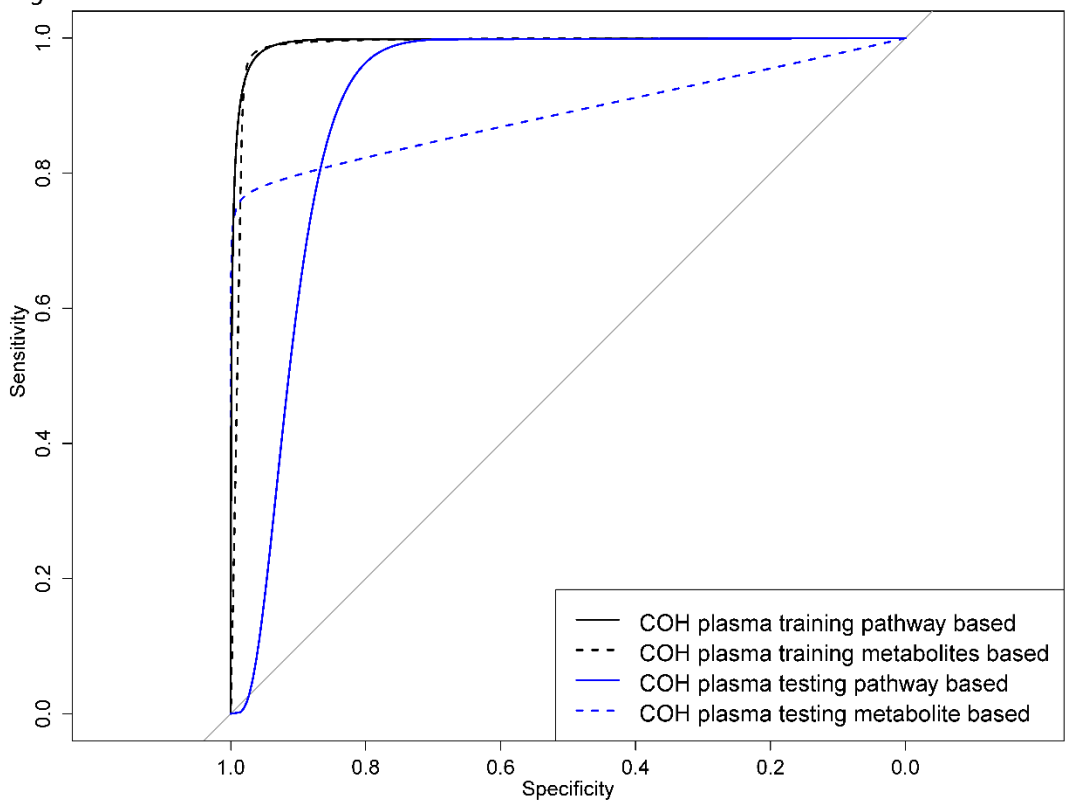
Figure 4. Integrative analysis of pathway features and the associated metabolites.



The key pathways and their intersections crucial for breast cancer diagnosis. Metabolites and enzymes are represented with nodes of different shapes and colors, and their relationships are represented by edges.

Figure 5. ROC curves comparison of pathway-based model and metabolites-based model among data sets

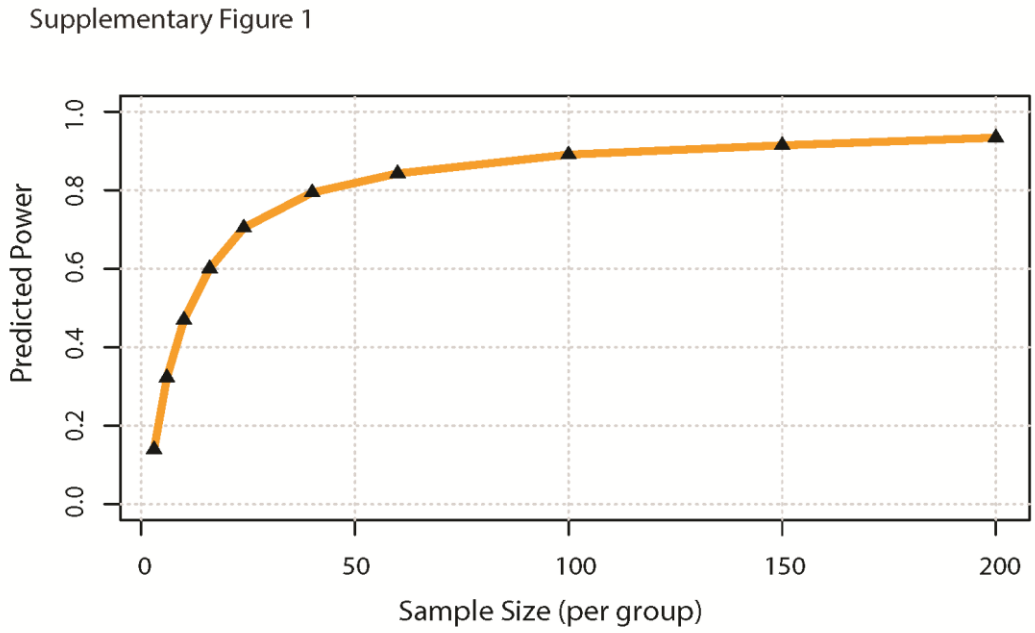
Figure 5



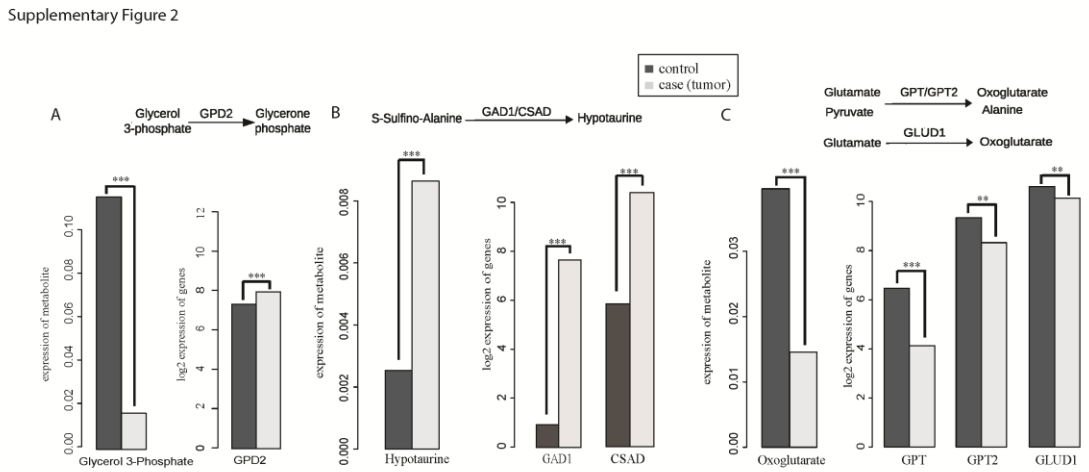
The same 80% of early stage (stage I and II) cases and controls from COH plasma from early stage diagnosis model stands for the plasma training set. The 20% of early stage (stage I and II) cases and controls testing represents the testing set. Metabolites based model is based on the same 10-fold cross-validation CFS selection on the plasma training set. ROC curves for training and testing sets are compared between plasma-based model and metabolites-based model among data sets.

Appendix B: Chapter 2 Supplementary Figures

Supplementary Figure 1. Power analysis and sample size estimation plot

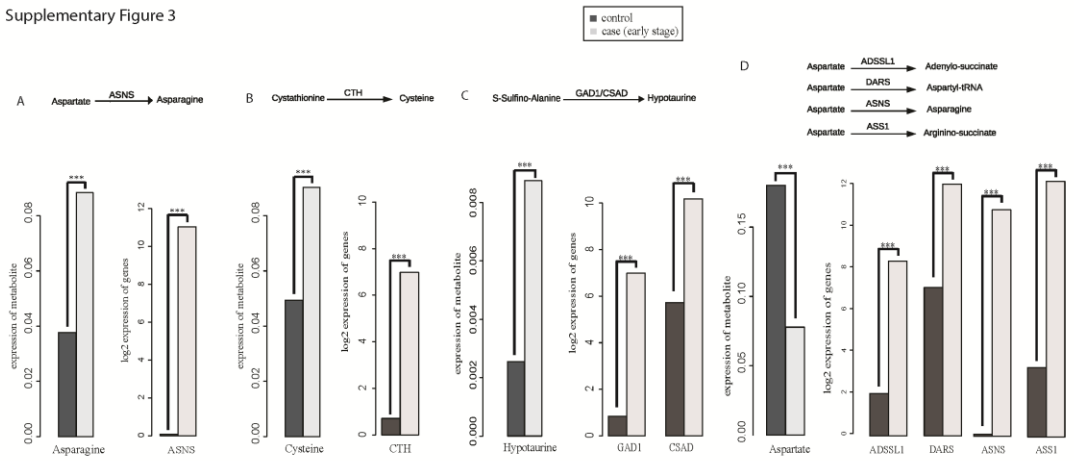


Supplementary Figure 2. Bar plot comparing the key metabolites in all-stage diagnosis model to the expressions of corresponding enzymes in TCGA breast cancer RNA-Seq data.



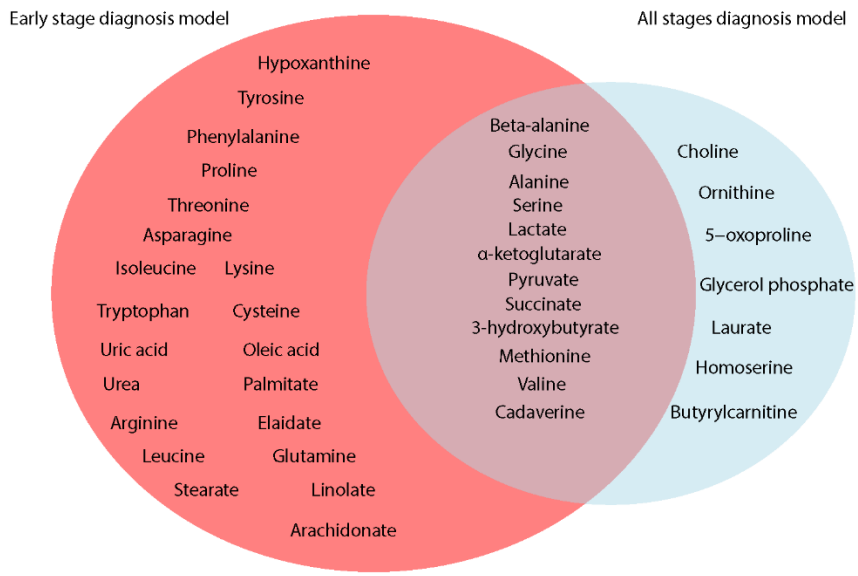
The enzymes (genes) for these metabolites were extracted from KEGG and SMPDB. P-values were calculated using differential tests in *Limma*. ***: P<0.001.

Supplementary Figure 3. Bar plot comparing the key metabolites in early-stage prediction model to the expressions of corresponding enzymes in TCGA breast cancer RNA-Seq data.

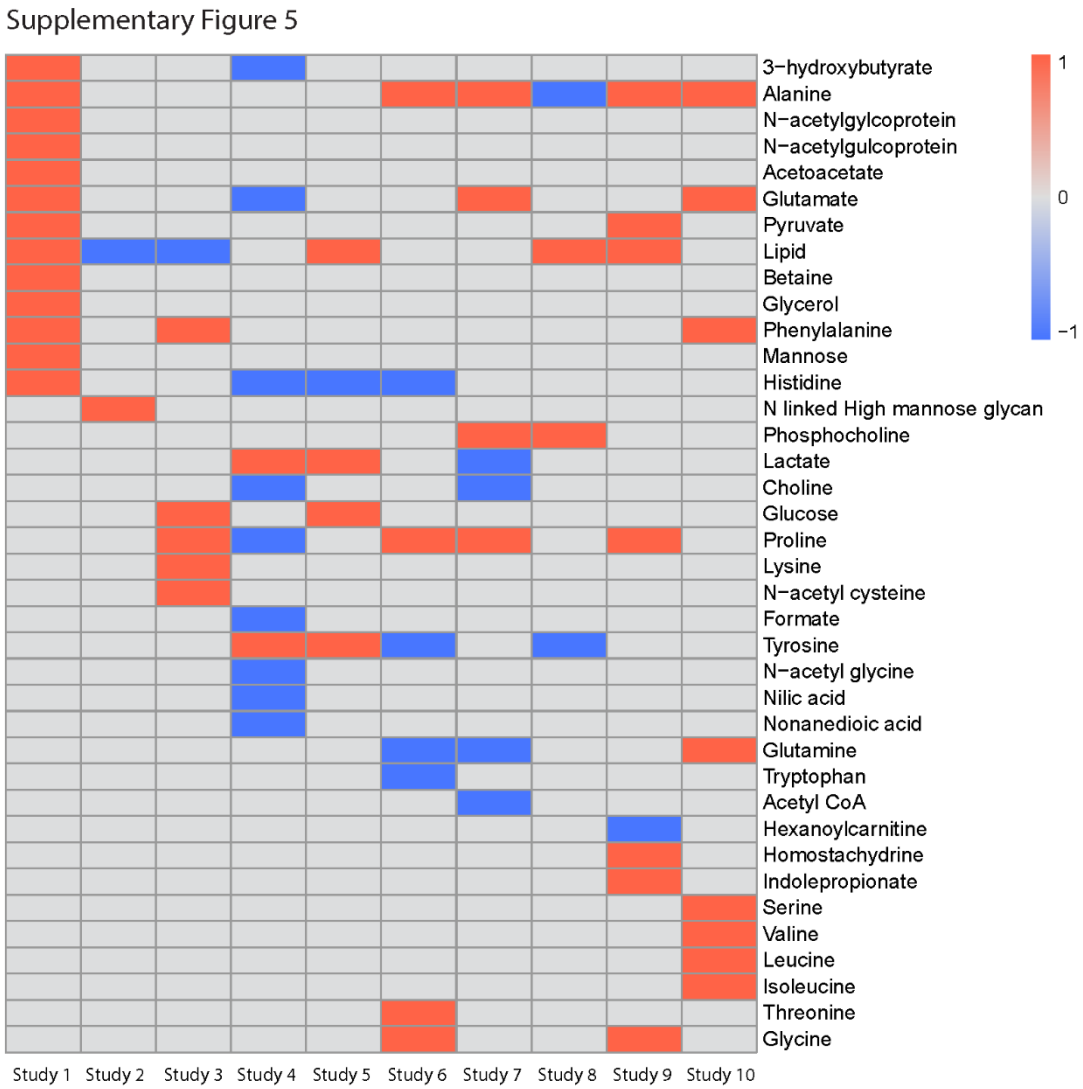


The enzymes (genes) for these metabolites were extracted from KEGG and SMPDB. P-values were calculated using differential tests in *Limma*. ***: P<0.001

Supplementary Figure 4. Venn diagram of the metabolites from the selected pathways in two models (all-stage diagnosis and early-stage diagnosis)



Supplementary Figure 5. Metabolites detected as biomarkers for breast cancers by different studies



Study1: (serum) E Jobard et.al (2014)
Study2: (serum) ML de Leoz et.al (2011)
Study3: (serum) C Oakman et.al (2011)
Study4: (serum) VM Asiago et.al (2010)
Study5: (serum) L Tenori et.al (2015)
Study6: (plasma) Y Miyagi et.al (2011)
Study7: (cell line) C Yang et.al (2007)
Study8: (plasma) J Shen et.al (2013)
Study9: (plasma) JA Miller et.al (2015)
Study10: (serum) I Poschke et.al (2013)

Appendix C: Chapter 2 Tables

Table I: Summarization of patient and clinic characteristics

Source		COH Plasma				COH serum		TCGA paired RNASeq	
Division		Training Set		Testing set		Testing set		Testing Set	
		Breast Cancer	Healthy Control	Breast Cancer	Healthy Control	Breast Cancer	Healthy Control	Breast Cancer	Healthy Control
Number of samples		106	61	26	15	103	31	98	98
Age (median, range)		53, 31-73	34, 21-40	54.5, 36-72	37, 21-40	52, 32-72	36, 18-49	56, 30-90	56, 30-90
Stage I		16		3		18		16	
Stage II		40		9		49		60	
Stage III		38		8		54		21	
Stage IV		11		6		19		1	
Race	Asian	14		3		14		1	1
	Black	5	12	4	1	6	5	6	6
	White	76	28	18	9	69	21	90	90
	Latino		21		5		5		
	Native					1			
	Others	10		1		13		1	1

Chapter 3. A novel model to combine clinical and pathway-based transcriptomic information for the prognosis prediction of breast cancer

Abstract

Breast cancer is the most common malignancy in women worldwide. With the increasing awareness of heterogeneity in breast cancers, better prediction of breast cancer prognosis is much needed for more personalized treatment and disease management. Towards this goal, we have developed a novel computational model for breast cancer prognosis by combining the Pathway Deregulation Score (PDS) based pathifier algorithm, Cox regression and L1-LASSO penalization method. We trained the model on a set of 236 patients with gene expression data and clinical information, and validated the performance on three diversified testing data sets of 606 patients. To evaluate the performance of the model, we conducted survival analysis of the dichotomized groups, and compared the areas under the curve based on the binary classification. The resulting prognosis genomic model is composed of fifteen pathways (e.g. P53 pathway) that had previously reported cancer relevance, and it successfully differentiated relapse in the training set (log rank p-value = 6.25e-12) and three testing data sets (log rank p-value < 0.0005). Moreover, the pathway-based genomic models consistently performed better than gene-based models on all four data sets. We also find strong evidence that combining genomic information with clinical information improved the p-values of prognosis prediction by at least three orders of magnitude in comparison to using either genomic or clinical information alone. In summary, we propose a novel prognosis model that harnesses the pathway-based dysregulation as well as valuable clinical information. The selected pathways in our prognosis model are promising targets for therapeutic intervention.

Introduction

Breast cancer is the second (after skin cancer) most frequently diagnosed cancer in women, and ranks second (after lung cancer) in the deaths of women in year 2013 (Society, 2013). Most clinical studies categorize breast cancer into four molecular subtypes: Luminal A, Luminal B, Triple Negative/ Basal like and Her2(Carey et al., 2006; O'Brien et al., 2010). The survival outcomes differ significantly among the clinical subtypes. Luminal A and B subtypes have a relatively good prognosis, whereas triple negative or basal like tumors, and Her2 tumors have very poor prognosis with much higher recurrence and metastasis rates(Carey et al., 2006; Haque et al., 2012; O'Brien et al., 2010). Furthermore, it is increasingly being realized that breast cancers are much more heterogeneous diseases than what is determined by the clinical subtypes, and that better prediction of prognosis is needed early on for more personalized treatment and management. Towards this goal, prognosis biomarkers of breast cancers have been investigated in many studies (Cancer Genome Atlas, 2012; van de Vijver et al., 2002; Y. Wang et al., 2005), based on signatures from high-throughput platforms such as gene expression profiles. Some signature panels such as the NKI 70 test are currently in commercial use with decent prediction of metastasis (van 't Veer et al., 2002).

However, transcriptomic data are usually poorly dimensioned with many more genes than the number of samples, thus methods that reduce the dimension by incorporating higher-order information of functional units, such as gene sets, pathways and network modules, have been recently explored (G. Abraham, A. Kowalczyk, S. Loi, I. Haviv, & J. Zobel, 2010; Efron & Tibshirani, 2007; E. Lee, H. Y. Chuang, J. W. Kim, T. Ideker, & D. Lee, 2008; S. Ma, M. R. Kosorok, J. Huang, & Y. Dai, 2011; F. Reyal et al., 2008; Subramanian et al., 2005; A. E. Teschendorff et al., 2010; van den Akker et al., 2013). This methodology is based on the observation that multiple genes involved in the same biological processes are often dysfunctional all together in cancers (Bild et al., 2006), therefore features selected from representative functional units are presumably more robust with better biological annotations(Bild et al., 2006; van den Akker et al., 2013). Currently, two main approaches to define functional units have been proposed. One

approach is to identify *de novo* functional units from the data. For example, van Vliet used an unsupervised module discovery method to identify gene modules, scored them and use them as features in a Bayes classifier (van Vliet, Horlings, van de Vijver, Reinders, & Wessels, 2012). Teschendorff et al. reported improved prognostic classification of breast cancers via a novel strategy to discover the activated pathways from the modules of "expression relevance network" (A. E. Teschendorff et al., 2010). Similarly, network analysis with combination of all the useful gene information has been developed and utilized to measure the coordination among the genes (S. Ma et al., 2011). The other main approach uses the existing pathway information to build functional units. For example, Lee et al used the MsigDB C2 gene sets to select feature sets using the t-test, and represented the pathway activity level by a subset of genes whose combined expression delivered optimal discriminative power for the disease phenotype (E. Lee et al., 2008). Abraham et. al used a set statistic that aggregated the expression levels of all genes in a set, and constructed prognostic gene sets that were as predictive as individual genes, yet more stable and interpretable within the biological context (G. Abraham et al., 2010).

However, most of these methods model the prognosis as binary outcomes, and *post hoc* analyze the performance of the methods using survival information; or individualized information of pathway deregulation is lost during information extraction before deriving statistical metrics. More importantly, the merits of combining clinical features and genomic features together have not been adequately addressed in most studies, where the models were only built upon the genomic information. In this study, we use a novel pathway-based deregulation scoring matrix to transform the gene-based genomic features in combination with the Cox regression and L1-LASSO regularization to model survivals. With this pathway deregulation score matrix as inputs, we constructed a pathway-based genomic model consisting of fifteen cancer relevant pathways that successfully predicted relapse difference (log rank p-value=6.25e-12, and AUC=0.80) and validated them on three breast cancer data sets with diversified clinical profiles (log rank p-value<0.0005, and average AUC=0.68). The pathway-based genomic models consistently performed better than gene-based models on all four data sets. Moreover, combining genomic level information with clinical

information improved prognosis prediction and classification by at least three orders of magnitudes of p-values, in comparison to either genomic or clinical information alone.

Materials and Methods

Study Population

We used four publicly available data sets of breast cancer samples from National Center for Biotechnology Information (NCBI) Gene Expression Omnibus (GEO) GSE4922 (A. V. Ivshina et al., 2006), GSE1456 (Y. Pawitan et al., 2005), GSE3494 (L. D. Miller et al., 2005) and GSE7390 (C. Desmedt et al., 2007). All four data sets are based on Affymetrix HG-U133A microarray platform, and have relapse-free survival information as well as some other clinical information, as shown in Table 1. For data set GSE7390 (C. Desmedt et al., 2007), all patients are lymph node negative. The GSE3494 data set was used as the training set as it has more clinical information, and all others were used as testing data sets.

Microarray Gene Expression Data Processing and Analysis

We mapped original probe IDs to Gene IDs using R package *biomaRt* (Kasprzyk, 2011). In order to relate the probe ID to the Gene ID, we downloaded the array annotation file and used the RefSeq IDs as the intermediates to map to the Gene ID. When a gene has multiple probes, we computed the geometric mean of log2 transformed probe intensities as the gene expression. All the data sets were normalized independently between array using *limma* package (Smyth, 2004). To minimize batch effects across different data sets, we used the CONOR package with the Bayesian method (Rudy & Valafar, 2011).

We generated the PAM50 heatmap of the gene expression data and the correlation heatmap with hierarchical clustering, where Euclidean distance measure was employed. For the clinical factors, we correlated their associations with the relapse in the training data set with both Chi-square test and Wilcoxon log-rank test for survival curves.

Prognostic Pathway-based Classifier Selection

The pathway information was obtained from the GSEA (http://www.broadinstitute.org/gsea/) curated gene sets that include a total of 403 pathways from Biocarta (http://www.biocarta.com) (Nishimura, 2001a)and

KEGG(M. Kanehisa & S. Goto, 2000). To perform gene sets analysis, we used R package *Pathifier* (Y. Drier et al., 2013), an algorithm that transforms the information from the gene level to pathway level and infers pathway deregulation scores for each pathway within each sample. The pathway deregulation score (PDS) in each sample is a measure of degrees of the deviation of a specific pathway from the "normal status" located on the principle curve. The concept of principle curve was proposed by Hastie and Stuetzle (Hastie & Stuetzle, 1989) as a nonparametric nonlinear extension of the PCA (Principle Component Analysis) in which the assumptions of dependence in the data are avoided. A principle curve is a one-dimensional curve that is derived from the local average of p-dimensional points and goes through the cluster of p-dimensional principle components. It sensibly captures the information of variation in all the samples. Specifically, the single parameter λ varies tracing the whole data along the curve (Hastie & Stuetzle, 1989). The curve $f(\lambda)$ is defined to be a principal curve if $E(\mathrm{X}|\lambda_f(\mathrm{X})=\lambda)=f(\lambda)$ for arbitrary λ. The principle curve is built through iterations of smoothed procedure in the local average of data points. If one sample differs from others in one specific pathway, the distance to the curve is further and it leads to a higher PDS score and vice versa.

In the model selection stage, we used Cox-Proportional Hazards (Cox-PH) model based on L1 – penalized (LASSO) estimation (J. J. Goeman, 2009; R. Tibshirani, 1997; Robert Tibshirani, 2011), with the R package *penalized* (J. J. Goeman, 2009). With the input of both PDS score containing the gene sets information and survival information of time and relapse, a tuning parameter lambda was used to restrict the number of parameters in the model. The optimal lambda was selected after running 250 simulations through likelihood cross-validation. A prognostic genomic model was thus generated with specific pathways and coefficients. We then computed a Prognosis Index (PI) score which is the logarithm of hazard ratio. We divided the samples into two groups of higher risk and lower risk with a 3 to 1 ratio, based on the 3rd quartile of PI. We used this cutoff to reflect the relapse/non-relapse ratio in the training data set.

We tested the above model in three other data sets. To do so we used the same PI cutoff above and separated samples into predicted high risk and low risk groups. We then used Kaplan-Meier curve together with

Wilcoxon log rank test to evaluate the performance of our model. To generate the receiver operating characteristic (ROC) curves, PIs are used as predicted values in comparison to the "truth" values of relapse/non-relapse information. The confusion matrix with sensitivity and 1- specificity is calculated for each division in ROC curves and the areas under the curve (AUC) is shown along with the ROC plot.

Combined Molecular and Clinical Model

To determine whether the clinical factors improve the prognosis of genomic pathway-based model, we re-normalize the clinical factors and molecular PDS independently to ensure that each factor has the standard normal distribution. We then combined the normalized clinical and molecular factors into the LASSO penalized step and built the combined model using the optimized lambda through 250 simulations, similar to the construction of the genomic model as described earlier. The model performance comparisons were also done similarly to those of the genomic model.

Survival Analysis

We used survival analysis to compare the relapse-free-survival results in the training and testing data sets. Patients without these events during the study were considered censored. We used the Cox-PH model to associate the risk of relapse to selected pathway features and clinical features by L1- LASSO. The Cox model is a semi-parametric model that is widely used to analyze the survival data. The non-parametric portion comes from the fact that no assumptions are made about the form of the baseline hazard. However, it has the assumption that the log hazard ratios are constant over the time for each feature. Assume that we obtained *p* features to be related with breast cancer relapse for each patient $X^J = (X_1^J, X_2^J, X_3^J, ..., X_p^J)'$, Cox-PH model represents the relationship between the risk of relapse and *X* features as:

$$h(t \mid \mathbf{X}) = \mathrm{h}_0(t) \exp(\boldsymbol{\beta}' \mathbf{X})$$

Here $h_0(t)$ is the baseline hazard (instantaneous risk) which only depends on time. The ratio of hazard (HR) between two pathways or two clinical features $\boldsymbol{X_m}$ and $\boldsymbol{X_n}$ is:

$$\frac{h(t \mid \boldsymbol{X}_m)}{h(t \mid \boldsymbol{X}_n)} = \exp(\boldsymbol{\beta}'(\mathbf{X_m} - \mathbf{X_n}))$$

The relative hazard between any two features is constant over time and only depends on the differences of the values in features. The PI for each patient J's features is calculated as

$$\boldsymbol{PI}^J = \hat{\boldsymbol{\beta}}' \boldsymbol{X}^J$$

This risk factor can be easily transformed to hazard ratio for different features, assuming that we have a baseline feature. The weights $\hat{\boldsymbol{\beta}}'$ for different features were calculated from the training data set using the Cox-LASSO model.

For the genomic, clinical and combined models, we used Kaplan Meier curves to present the prognosis performance in classified high risk and low risk groups. The data set was dichotomized into two groups, and the higher risk group is assumed to have higher hazard of relapse compared to the lower risk group. We used the Wilcoxon log-rank test to check the survival difference between these two groups. To find the significance of an individual factor's impact on relapse, we fit individual predictor with a univariate Cox-PH model. We then calculated the hazard ratio by computing the exponential of the coefficients in the Cox-PH model. All survival analysis was conducted using the R package *Survival* (Li, 2003).

Sensitivity Analysis of Pathway-based Models

To examine the effect of input pathways on model performance, We randomly select 1/2, 1/4, 1/8 and 1/16 of all input KEGG and BioCarta pathways, then generated the PDS Matrices for 18 times under each case. For each simulation, we built the model with the workflow in Figure 2 and computed the Wilcoxon log-rank test p-value between the survival curves of two risk groups, as well as the AUC of the classification results. We then used boxplots to demonstrate the differences of –log10 (p-values) and AUCs due to different total pathway counts.

To estimate the statistical confidence of comparisons of each model, we used leave one out cross validation (LOOCV) to compute p-values and AUCs across all simulations. In the *i*th simulation (*i*=1,…,total sample size of the data set), we deleted the *i*th patient sample, modified the PI threshold by the remaining sample ratio of recurrence to non-recurrence and finally calculated the Wilcoxon log-rank test p-value as well as the AUC of the classification results. We then used boxplots to demonstrate the comparisons between the pathway-based and the gene-based models, and among the genomic, clinical, and the combined models.

Comparison to the NKI70 model

We tested the NKI70 method to our training data set (Miller data). We mapped the NKI70 gene signatures from to the genes in the U133A array. We correlated the gene-expression profile with the good-prognosis/poor prognosis data from the NKI study (van 't Veer et al., 2002), and then classified the samples into good and poor clusters as done by others (van de Vijver et al., 2002). For consistency, we used the Wilcoxon log-rank test p-value from survival analysis and the AUC of the ROC classification to assess the results.

Results

Data Summary

We used four individual gene expression microarray data sets for the testing and validation of the pathway-based prognosis model (Table 1), all of which were measured by Affymetrix HG-U133A array and had relapse and survival information. We used the data set of 236 patients in Miller et. al.(L. D. Miller et al., 2005) as the training data mainly because this data set contains the most abundant clinical information, including ER status, PG status, tumor size, grade, lymph node status and P53 mutation.

PAM50 is a list of 50 genes initially proposed to successfully differentiate the breast cancer subtypes and it was later found that PAM50 also harbors good prognosis information on breast cancer (Chia et al., 2012). Therefore, we first present the testing data summary results and correlate relapse with PAM50 and other

clinical factors (Figure 1). Although tumor molecular subtypes are unknown due to the missing Her2 marker information, we nevertheless observed a good correlation between PAM50 matrix and relapse. Based on the hierarchical clustering results of PAM50 heatmap, we dichotomized the samples into high and low risk groups, This grouping approach, without any supervised learning, results in a fairly good association to relapse status (Chi-square test p=7.46e-5). Additionally, grade and lymph node have significant associations to relapse, with Chi-square test p-values of 0.018 and 9.146e-6 respectively. Single clinical factor based survival analysis also confirms such significant relevance to relapse: p-values of Wilcoxon log rank tests for the p53, grade, tumor size and lymph node status based survival differences are 0.0152, 0.00181, 1.92e-7 and 4.93e-8, respectively. Similar to previous observations (C. Fan et al., 2011), ER and PG status are not good prognosis indicators, with the log rank test p-values of 0.819 and 0.227, respectively. There are a total of around 600 samples in the three testing data sets, 2.5 times the size of samples in the training set. Testing set 1 (Ivshina data) (A. V. Ivshina et al., 2006) and testing set 2 (Pawitan data) (Y. Pawitan et al., 2005) have very similar distribution pattern to the training data (Miller data) (L. D. Miller et al., 2005). However testing set 3 (Desmedt data) (C. Desmedt et al., 2007) has very different distribution compared to other three data sets, as the samples were all lymph node negative tumors. We include set 3 as an extension to the other two testing data sets to exam the performance of the pathway-based genomic model for prognosis.

Building the pathway-based genomic model

We have developed a novel pathway-based prognosis prediction model, unlike most other models that are gene-based (Figure 2). We transformed a conventional gene-based matrix into a new pathway-based matrix of reduced numbers of rows, where each row represents a KEGG or BIOCARTA pathway-based scores over all samples (columns). Instead of using log2 transformed intensities as elements of the matrix, we used Pathway Dysregulation Scores (PDS) (Y. Drier et al., 2013) that measure the distance of a particular pathway to the "normal condition" curve in a hyperspace. PDS ranges from 0 to 1, and the higher PDS score signifies more "abnormity". This pathway-based PDS matrix was used as the initial input to select

featuring pathways that are predictive of survival, based on the multi-variate Cox-PH model (Gill, 1992). We used L1-LASSO penalization method (J. J. Goeman, 2009; R. Tibshirani, 1997; Robert Tibshirani, 2011) to constrain the featuring pathways to be selected. To be consistent, we conducted 250 simulations to select the best set of pathways.

We first evaluated the featuring pathways selected by the model, in relation to other clinical factors and relapse status in the training data set (Figure 3). Comparing the heatmap of selected featuring pathways to that of the PAM 50 genes (Figure 3A), the selected pathways are more prognostic for relapse. This is supported by two observations: (1) Dichotomized samples of high risk and low risk groups through hierarchical clustering of PDS scores have a higher correlation to relapse status (Chi-square test p=1.99e-6), compared to those of PAM50 gene matrix (Chi-square test p=7.46e-5) and (2) The median PDS scores over fifteen selected pathways have a correlation coefficient of 0.17 to relapse, in comparison to 0.08 for the median expression intensities over PAM50 genes. Thus the selected pathways by our model are better prognostic features than PAM50 genes, in terms of the correlation to disease relapse.

To investigate the performance of the model, we used the PI value which is the logarithm of hazard ratio from the fitted Cox-PH model to dichotomize the samples, similar to others (C. Fan et al., 2011) (Sveen et al., 2012). We divided the samples into higher and lower risk groups with a 3 to 1 ratio (3^{rd} quartile in PI), in order to match the relapse versus non-relapse sample ratio in the training data. Samples with larger PDS scores are expected to have higher PI scores, and are more likely to have relapsed diseases. The same PI threshold was applied to dichotomize the training data set as well as multiple independent testing data sets. The performance of the genomic model was then evaluated by two approaches: (1) the Wilcoxon log rank test p-values of the Kaplan-Meier survival curves from the two risk groups in each data set, and (2) the AUCs of ROC curve based on binary classification.

The pathway-based genomic model is predictive on multiple testing data sets

Instead of combining all four data sets for meta-analysis, we kept them as individual data sets to validate the robustness of our model. As expected, the pathway-based genomic model is highly accurate at

differentiating the risks of breast cancer relapse within the training data, with a Wilcoxon log rank p-value of 6.25e-12 (Figure 4A). The model yields very decent predictive results with the p-value of 1.52e-4 in testing set 1 and 3.91e-5 in testing set 2 (Figure 4B and 4C). The predictive performances are expected to drop in the testing data sets, since they have different patient populations and clinical characteristics from the training set (Table 1). Impressively, the model gives a very significant p-value of 3.73e-4 for testing data set 3 (Figure 4D), which are all early stage lymph node negative tumors whose prognosis is very difficult to predict. Additionally, we evaluated the performance of models using binary classification. We used the relapse/non-relapse information in the data sets as truth measures, and the model's high vs. low risk classification as predictions. As shown in Figure 4E, the ROC curve in the training set gives an AUC value of 0.80, and AUCs of 0.73 (testing set 1, Pawitan data), 0.67 (testing set 2, Ivshina data), 0.65 (testing set 3, Desmedt data), consistent with the results in Kaplan-Meier curves (Figure 4A-D).

To examine the effect of total number of input pathways on model performance, we randomly kept 1/2, 1/4, 1/8 and 1/16 of all input KEGG and BioCarta pathways in the training dataset, and then generated the PDS Matrices for 18 simulations under each scenario. For each simulation, we built the model with the same workflow as in Figure 2 and computed the Wilcoxon log-rank test p-value between the survival curves of the two risk groups, as well as the AUCs of the classification results. The boxplot in Figure S1 shows a gradual decrease of AUCs due to the input pathways, in the order of 1/2 > 1/4 > 1/8 > 1/16 pathway-based models. The difference between 1/2 and 1/4 pathways is significant (p-value<0.05). All AUCs, however, are in the range between 0.69 and 0.81.

The pathway-based genomic model is superior to the gene-based genomic model

Our earlier results of selected pathway features vs. PAM 50 genes suggested that pathway-based features may be better than gene-based features. To validate this, we trained the four data sets individually and compared within the same data set the performance of pathway-based models and gene-based genomic models which do not have the PDS matrix generation step (Figure 2). In order to test the risk differentiation power of the model, the cutoff PI value in each data set was set to match the ratio of relapse vs. non-relapse

patients in that particular set. The results of Kaplan-Meier survival curves and ROC plots based on classification all consistently show that pathway-based genomic models are superior to the gene-based models (Figure 5A-H). For example, in Miller data set the log-rank p-value is 6.25e-12 for the pathway-based model (Figure 5B), compared to that of 1.75e-9 for the gene-based model (Figure 5A). In the Desmedt data set, the p-value of the pathway-based model is even more significant than that of gene-based model (5.12e-36 vs. 8.84e-12, Figure 5H and 5G). Similarly, pathway-based genomic models have better ROC curves than gene-based genomic models (Figure 5I), with AUCs of 0.80 vs. 0.78 in Miller data, 0.85 vs. 0.77 in Pawitan data, 0.74 vs. 0.70 in Ivshina data, and 0.92 vs. 0.76 in Desmedt data. To estimate the statistical significance of comparisons among the pathway-based and gene-based models, we performed leave-one-out cross validation (LOOCV) simulations to compute the Wilcoxon log-rank test p-values and AUCs of ROC classification curves. The cross validation results show that statistically the pathway-based models perform better than the gene-based models (Figure S2, all t-test p-values <0.001). These results are consistent with the observations from previous studies (E. Lee et al., 2008; A. E. Teschendorff et al., 2010), and support the hypothesis that including higher-order secondary information yields better prognostic values.

NKI70 (Mammaprint) is one of the most commonly used model for breast cancer prognosis prediction, and it has been approved by FDA for commercially use in clinics. To demonstrate the potential clinical utilities of our model, we compared the NKI70 method with ours, and applied the NKI70 method to our training data set (Miller data). We first mapped the NKI70 gene signatures (van 't Veer et al., 2002) to the genes in the U133A array, then correlated the gene-expression profile with the good-prognosis/poor prognosis data from the NKI study and classified the samples into good and poor clusters as done previously(van de Vijver et al., 2002). The NKI70 test gives a Wilcoxon log-rank test p-value of 2.58e-3 for the survival analysis, in contrast to the p-value of 6.25e-12 obtained by our pathway-based model; it only yields an AUC of 0.62 for classification, in contrast to 0.80 from our model (Figure S3).

The combined model with pathway-based genomic and clinical features is superior to the genomic or clinical model alone

Previous studies suggested that clinical information of breast cancers provides additional values to a genomic model that was built on lists of genes (C. Fan et al., 2011). To test if such merit of clinical information also applies to our genomic model of fifteen pathway features, we investigated the performances of the genomic, clinical and genomic-clinical combined models.

Since the scales of PDS and clinical features vary significantly, we re- normalized PDS and clinical features independently to have the standard normal distribution, so that they are subject to the same selection criteria. The resulting clinical model is composed of four selected features: grade, tumor size, p53 and lymph node. This is not surprising, as they are also significant factors in the univariate Cox-PH models (Table 2 and Figure 1B-E). The combined model keeps ten of the fifteen pathways (Table 2) and about 60% of genes that were selected by the genomic model. It also selects tumor size and lymph node status as additional features (Table 2). This is expected given their highly significant p-values (1.92e-7 and 4.93e-8, respectively) in the univariate Cox-PH models (Figure 1B and 1E), as well as relatively large coefficients in the clinical model (0.27 and 0.36, respectively). Since only testing data set 2 has both tumor size and lymph node information, we used this data set and the testing data set to demonstrate the performances of genomic, clinical, and combined models.

The comparisons present the compelling advantage of combining clinical and genomic information in a model (Figure 6). As shown in the training data, selected clinical features are undoubtedly important: the Wilcoxon log rank test p-value of the clinical model is 2.21e-10 (Figure 6E), slightly less significant than the pathway-based genomic features by two orders of magnitude. Most importantly, the combined model is much better than either genomic model (p-value=6.25e-12) or clinical model alone, with a p-value of 1.88e-24 (Figure 6C). This trend of significances is consistent in the testing set 2, with the p-values of 1.12e-7 in the combined model (Figure 6D), 1.52e-4 in the genomic model (Figure 6B), and 2.7e-3 in the clinical model (Figure 6F). Moreover, the ROC curve comparisons of these three models also show the

same order of performances: combined model > genomic model > clinical model , with AUCs of 0.83, 0.80, and 0.74 in the training set, and 0.71, 0.68 and 0.65 in the testing set 2 (Figure 6G).

To demonstrate the statistical significance of comparisons among the pathway-based, clinical and combined model in the training set and the testing set 2, we performed leave-one-out cross validation (LOOCV) simulations to compute the Wilcoxon log-rank test p-values and AUCs of ROC classification curves. The cross validation results show that statistically the combined model performs better than the pathway-based model, and the pathway-based model performs better than the clinic model (Figure S4, all p-values <0.001 between pathway-base/clinical models and combined models).

Biological relevance of featured pathways and genes

We expect that the consensus pathways selected both in our genomic model and combined model convey important cancer-related functions. To test this we examed the annotations of this subset of ten pathways (Table 2). Interestingly, KEGG_MELANOGENESIS is selected as a feature, probably due to inclusion of many cancer relevant genes in this pathway: such as protein kinase genes PRKACB, PRKACG, PRKCB, PRKCA; phosphorylase kinase genes CALM1, CALM2, CALM3; G-protein related gene GNAQ, HRAS; mitogen-activated protein kinases MAPK1, MAPK3, MAP2K1; and other oncogenes like RAS (Tian et al., 2013; Yong et al., 2011). Many of these genes have been shown to function in breast cancer progression (Yong et al., 2011). Impressively, multiple signaling pathways are selected, including BIOCARTA_P53_PATHWAY, BIOCARTA_SRCRPTP_PATHWAY, BIOCARTA_PYK2_PATHWAY, BIOCARTA_VIP_PATHWAY, BIOCARTA_RARRXR_PATHWAY, and BIOCARTA_AKAP13_PATHWAY. They are well-known to be associated with breast cancers prognosis (Driggers, Segars, & Rubino, 2001; Fu et al., 2014; K. H. Lee et al., 2014; Pham, Angus, & Johnson, 2013; Rubino et al., 1998; Tao et al., 2011; Valdehita et al., 2010). The best example is BIOCARTA_P53_PATHWAY, the dysregulation of p53 Signaling Pathway is well-documented, and the tumor-suppressor gene p53 has one of the highest mutation rates in breast cancer (Cancer Genome Atlas, 2012; L. D. Miller et al., 2005).

In addition, some pathways related to basic cell functions are selected as prognostic features. For example, G1_PATHWAY is selected, and the G1/S cell cycle checkpoint controls are well known to be dysfunctional in many cancers including breast cancer (Guille, Chaffanet, & Birnbaum, 2013). FATTY_ACID_METABOLISM is also selected by the model, and many studies have showed that fatty acid metabolism is involved in breast cancer (Puig et al., 2008). In particular, Fatty acid synthase (FASN) is highly expressed in breast cancer with a poor prognosis compared to others (Puig et al., 2008). Interestingly, BIOCARTA_RNA_PATHWAY is also selected, largely due to its members TP53 and MAP3K14 that are closely related to breast cancer.

A total of 265 genes are overlapped between the selected pathways of the genomic model and the combined model. Table 3 summarizes the top 30 genes that are involved in the selected pathways. They are ranked by weighted sum of both occurrences in selected pathways (counts) and weights measured by the hazard ratio of each pathway. Among them, many genes encode protein kinases that are well-known to be involved in breast cancers, such as PRKACB, PRKACG, MAPK1 and CALM1. Some other genes encode transcription factors that are well-known for their close relationship to cancer, such TP53, RB1, HRAS, RAF1, GRB2, E2F1, and SRC (Engelmann & Pützer, 2012; Fan et al., 2014; Hagan et al., 2005; Tian et al., 2013). We therefore conclude that the selected pathways are prognostic features of significant cancer relevance.

Discussion

The heterogeneity of cancers is being increasingly recognized, suggesting more personalized care decisions with treatment for individual patients are needed. As a result, prognosis prediction of breast cancers with high-throughput data has been a growing topic in recent years. Many statistical and machine learning methods have been developed to analyze various types of high-throughput cancer genomics data, by taking advantage of higher-order relationships among genes. The hypothesis is that the highly correlated gene-based markers often represent identical biological processes; therefore by including higher-order

representative features, such as Gene Ontology sets, pathways and network modules, the prediction will be more stable (G. Abraham et al., 2010; J. J. Goeman & Buhlmann, 2007; E. Lee et al., 2008; S. Ma et al., 2011; F. Reyal et al., 2008; A. E. Teschendorff et al., 2010; van den Akker et al., 2013). Our novel method of prognosis prediction presented in this study belongs to this class of methods. However, unlike some other methods where individual pathway information is lost due to summarization or transformation, the pathway features proposed in this study explicitly measure the degrees of pathway dysregulation for cancer recurrence. Comparing selected pathways and the PAM50 genes which were demonstrated to be prognostic (Chia et al., 2012), the PDS-based pathway approach has better correlation to breast cancer relapses. Moreover, when comparing gene-based with the pathway-based genomic models, where the only difference between them was the input matrix, pathway-based models uniformly performed better than gene-based models in all the data sets we tested. Our results are consistent with several other gene-set/pathway-based models (G. Abraham et al., 2010; E. Lee et al., 2008), where different summarization metrics were used. It will be very interesting to compare the prediction results based on these different metrics in a follow-up study.

To demonstrate the robustness in predicting differential risks of relapse from the pathway-based genomic model, we chose to train and test on independent study samples, rather than combining them together as a large data set (C. Fan et al., 2006; C. Fan et al., 2011), which would diminish the effect of population heterogeneity. Despite population difference and much bigger testing data size relative to the training data size, the method still achieved good performance on all three testing data sets, including a data set of all early stage lymph node negative tumors where prognosis is particularly difficult to predict. Another merit of our method is that it enables combining the important clinical information with the pathway-based genomic information. Even though the clinical model by itself is the least predictive, compared to the genomic model and the combined model, it is nevertheless significant and informative, as shown by tumor size and lymph node status. The genomic model is better than clinical model alone. However, the combined model of clinical and genomic features performs the best. Our conclusions agree and extend the earlier

work from Fan et al. (C. Fan et al., 2011) who focused on prognosis prediction of all node-negative and systemically untreated breast cancer patients, since we include both node-negative and node-positive samples. The results of the genomic model (AUC=0.80 and p-value=6.25e-12 in training data, and AUC=0.68 and p-value=1.52e-4 in test data 2) and the combined model (AUC= 0.83 and p-value=1.88e-24 in the training set, and AUC=0.79 and p-value=1.12e-7 in test data set 2) are better than what was recently reported by Vilinia S et al (Volinia & Croce, 2013). They obtained an AUC =0.74 for the training set and 0.65 for the testing set, in a model that combined signatures of mRNA and microRNAs deriving from the TCGA IDC cohort sequencing data. This suggests the advantages of combining PDS based pathway score inputs with a Cox-PH model and LASSO penalization approach: even though the genomic data in our study are based on microarrays that have more noise and smaller sample sizes, they still yield better predictive results in comparison to the combined mRNA and microRNA sequencing signatures obtained from a larger sample size. It will be of great interest to apply our models to the TCGA breast cancer mRNA and microRNA sequencing data in the future.

The pathways selected by the model show biological relevance to breast cancer prognosis. The fatty acid metabolism pathway is found to be crucial to maintain the cancer cell malignant phenotype, and higher expression of fatty acid synthase has been discovered as a common phenotype in breast cancer with a poorer prognosis (Puig et al., 2008); As another example, Src kinase activation by protein tyrosine phosphatase alpha (SRCRPTP_PATHWAY), has been discovered in invasive breast cancer with compelling evidences. Src inhibitors are being considered as potential therapy to treat invasive breast cancers, as inhibition of c-src was recently found to be involved in E2-induced stress which would finally result in apoptosis in breast cancer cells (Fu et al., 2014). Increasing evidence shows that vasoactive intestinal peptide (VIP) in BIOCARTA_VIP_PATHWAY is highly expressed in breast cancer cells along with its receptor (Fu et al., 2014), and VIP-targeted nanomedicine is under study as therapy for breast cancer (Valdehita et al., 2010). Pyk2 in BIOCARTA_PYK2_PATHWAY is linked to map kinases MAPK, which has wealthy records in breast cancer studies (K. H. Lee et al., 2014). RARRXR_PATHWAY is the RAR/RAR nuclear receptor

complex that is co-activators to facilitate initiation of transcription in carcinoma cells (Tao et al., 2011). And BRX, the truncated form of Rho-Selective Guanine Exchange Factor AKAP13 in the BIOCARTA_AKAP13_PATHWAY, has been identified to function as an ER cofactor (Driggers et al., 2001).

Although the workflow proposed in this study is generic and the pathway features are clearly significant, we should point out a few potential limitations of the model. First of all, the pathway-based model is trained and tested on gene expression data from the U133A platform. We suspect that direct application of the model to other platforms, such as RNA-Seq, is not desirable, and some additional re-processing work has to be done additionally. The reason is that data distributions maybe very different between various platforms. One notorious example is that biomarkers identified by high-throughput microarray platform often had poor correlations in qPCR platform. Thus we recommend that when researchers use the workflow in Figure 2 on different data types, they may increase the predictive power by retraining the model with their own data. Another limit of our approach is that we only used the information from genes that compose the 403 pathways that we considered, thus some gene-level information is unavoidably lost. In our case, over 4500 genes were enlisted in the pathways, and among them over 3200 genes are probably expressed (averaged log 2 expression intensities > 7). On the other hand, the raw U133A array has results of over 14,000 genes within which over 10,000 genes are probably expressed. Therefore our model captures about 1/3 of the gene-level information overall. One can certainly use other curated gene sets, such as the MsigDB C2 gene sets, to increase the coverage of the genes by the pathways. However, from the sensitivity analysis that we have performed (Figure S1), we only observed a slight decrease of model performance based on AUCs, which are in the range of 0.69 and 0.81.

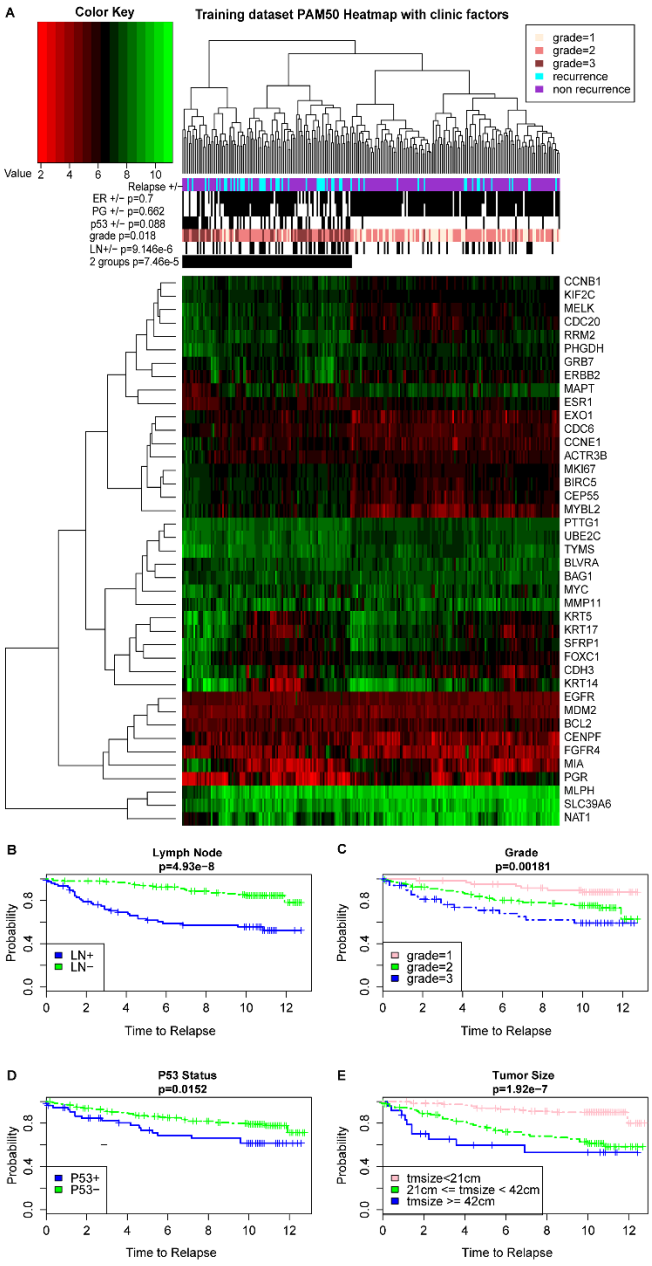
In conclusion, we propose a novel pathway-based genomic model that measures the pathway-based deregulation score and shows significant prognosis values. This pathway-based genomic model performs better than the gene-based genomic model. Additionally, we found that combining the clinical information of lymph node status and tumor size improves the performance of the prognosis model. Many selected

pathways in our study present values for breast cancer prognosis prediction, and they are also promising therapeutic targets for future investigations.

Appendix D: Chapter 3 Figures

Legends

Figure 1. The PAM50 gene signatures and their association with clinical information in the training data set.



A, The heatmap of the log2 transformed gene expression for PAM50 signatures. Green and red colors represent higher and lower expression levels, respectively. The samples are further categorized into two major groups based on the hierarchical clustering. The p-values of the clinical features such as ER, PG, P53, Grade, lymph node (LN) and dichotomized groups with relation to relapse status are calculated using Chi-square tests. B-E, Kaplan Meier survival estimates of relapse free survivals according to major clinical

features: (B) Lymph node status, (C) Grade, (D) P53 mutation status and (E) Tumor size. P-values are calculated using Wilcoxon log-rank tests and (+) denotes the censored observations in the study.

Figure 2. The workflow of the pathway-based genomic model.

Step 1. Transform the input data in the training set: the gene-based expression data are transformed into the pathway-based data input through the *pathifier* algorithm, using the pathway information from KEGG and BIOCARTA. The new input matrix is represented by Pathway Deregulation Scores (PDS). Step 2. Build the prognosis prediction model. The PDS matrix is integrated with the survival information via a Cox-PH model under penalized feature selection using the L1- LASSO method. Featuring pathways are selected and the coefficients (or weights) of these pathways are estimated using log likelihood cross validation. Step 3. Set the relapse risk threshold from the model. The prognostic index (PI) cutoff value is determined from the model to match the ratio of relapse/non-relapse in the training set. This PI is used as the relapse risk threshold on all the testing sets where the sum of weighted PDS is calculated on the pathways selected in Step 2. The input PDS matrices of testing data sets are computed the same as in Step 1. Step 4. Evaluate the performance of the prognostic model. The performance is evaluated through Kaplan-Meier curves of the dichotomized risk groups by PI scores, as well as the ROC curves and AUC values.

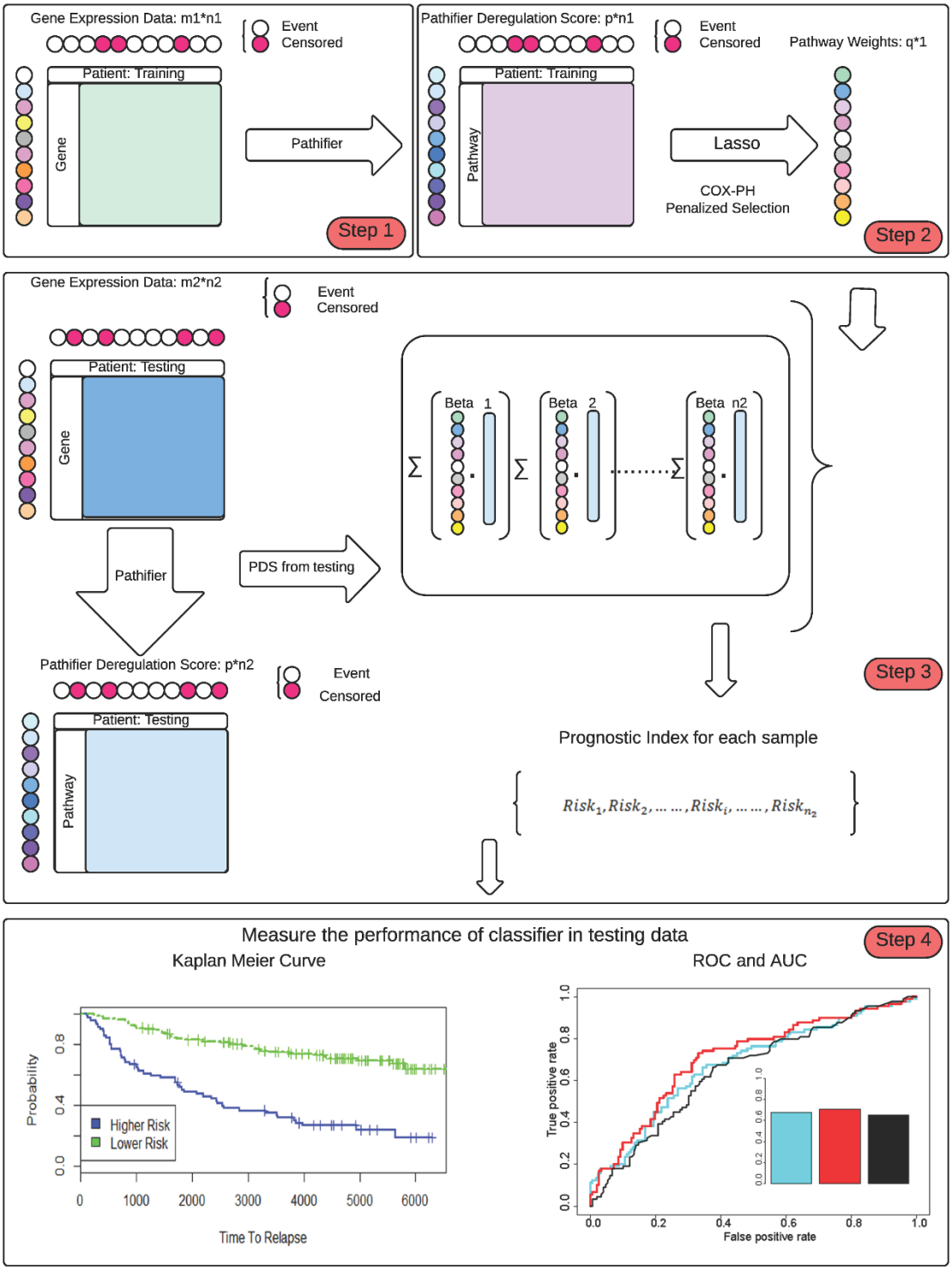

Gene Expression Data: m1*n1
Event
Censored
Patient: Training
Gene
Pathifier
Step 1
Pathifier Deregulation Score: p*n1
Event
Censored
Pathway Weights: q*1
Patient: Training
Pathway
Lasso
COX-PH
Penalized Selection
Step 2
Gene Expression Data: m2*n2
Event
Censored
Patient: Testing
Gene
Beta 1
Beta 2
Beta n2
Pathifier
PDS from testing
Pathifier Deregulation Score: p*n2
Event
Censored
Patient: Testing
Pathway
Step 3
Prognostic Index for each sample
$Risk_1, Risk_2, \ldots\ldots, Risk_i, \ldots\ldots, Risk_{n_2}$
Measure the performance of classifier in testing data
Step 4
Kaplan Meier Curve
Probability
Higher Risk
Lower Risk
Time To Relapse
ROC and AUC
True positive rate
False positive rate

Figure 3. The selected pathway signatures and their association with clinical information in the training data set.

The heatmap shows the patterns of Pathway Deregulation Score (PDS) of selected pathways in the genomic model. Green and red colors represent higher and lower PDS scores, respectively. The samples are further categorized into two major groups from hierarchical clustering, as in Figure 1. The p-values of the clinical features such as ER, PG, P53, Grade, lymph node (LN) and dichotomized groups with relation to relapse status are calculated using Chi-square tests.

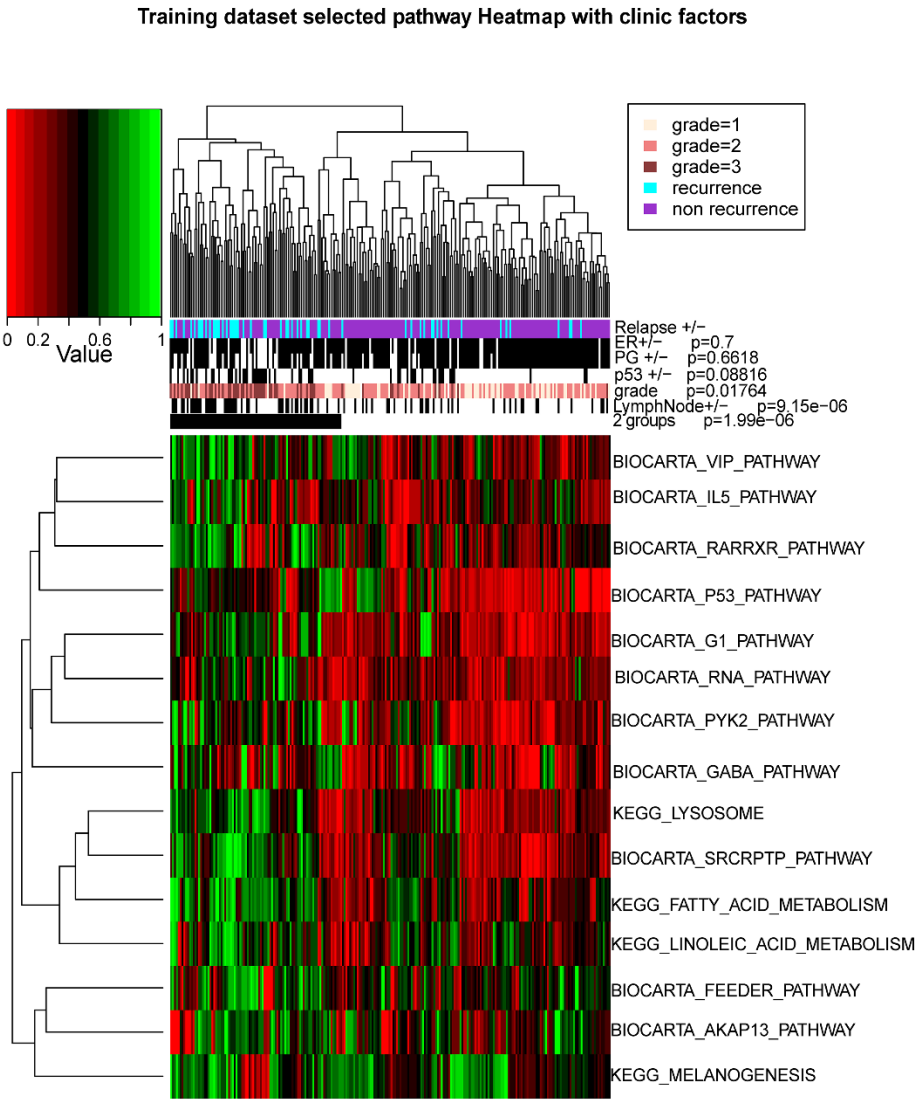


Figure 4. Prognosis performance of the pathway-based genomic model.

A-D. A prognosis index (PI) is calculated from the training data set and applied to dichotomize samples in training (A) and testing data sets (B-D). Higher risk and lower risk groups determined by the PI cutoff are compared by Kaplan-Meier curves. P-values of the survival difference between the two groups are calculated using Wilcoxon log-rank tests and (+) denotes the censored observations in the study. E. ROC curves are generated using PI values as predictions in comparison to the relapse/non-relapse information. AUCs are listed as the insert.

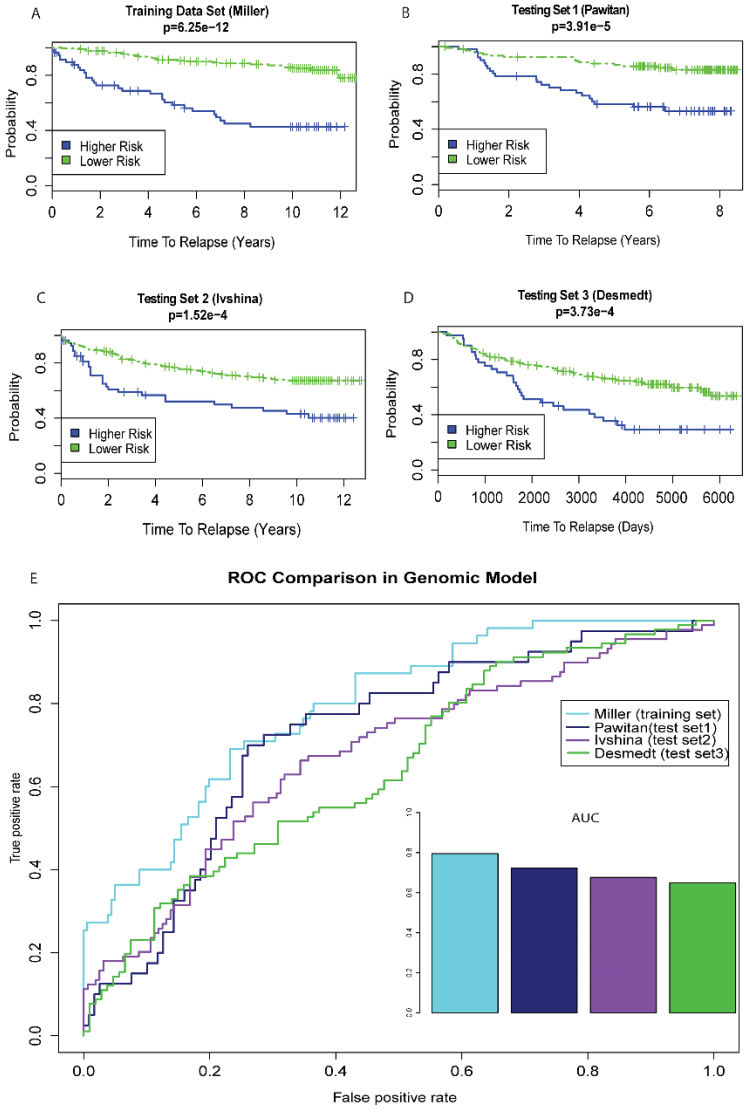
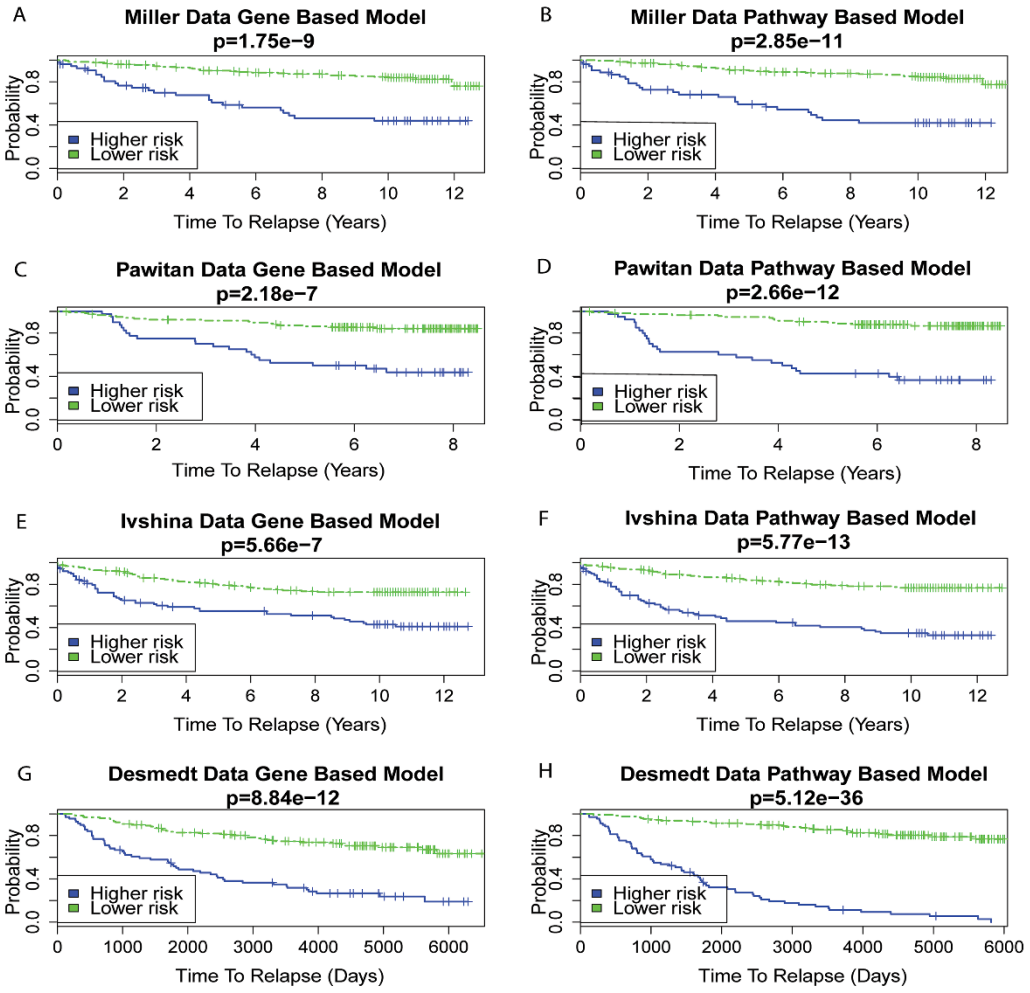


Figure 5. Comparing the prognosis performance between the gene-based and the pathway-based genomic models.

A-H. Gene-based and pathway-based genomic models are trained individually on the data sets. The PI is calculated to match the ratio of relapse to non-relapse on each data set and used to dichotomize the samples into higher risk and lower risk groups, similar to Figure 4. The associated p-values in Kaplan Meier curves are calculated using the Wilcoxon log-rank tests, as in Figure 4. Pathway-based genomic models consistently outperform alternative gene-based genomic models in all data sets. I. ROC curves are generated from PI based classification predictions in comparison to reported relapse information, similar to Figure 4. AUCs are listed as the insert. The ROC curves and AUC results also show that pathway-based models are better than gene-based models.



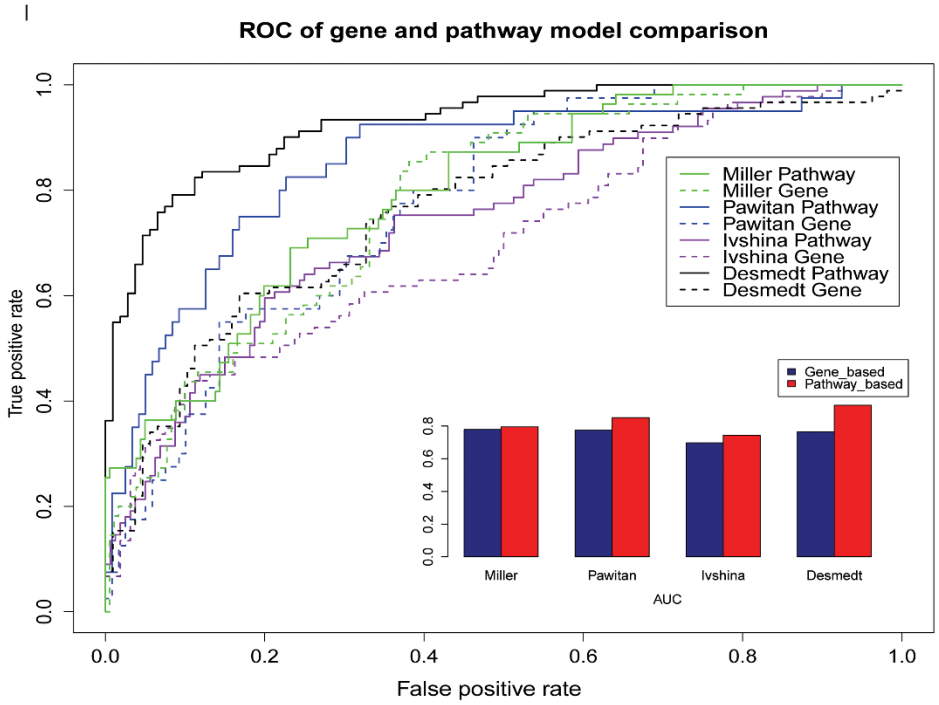
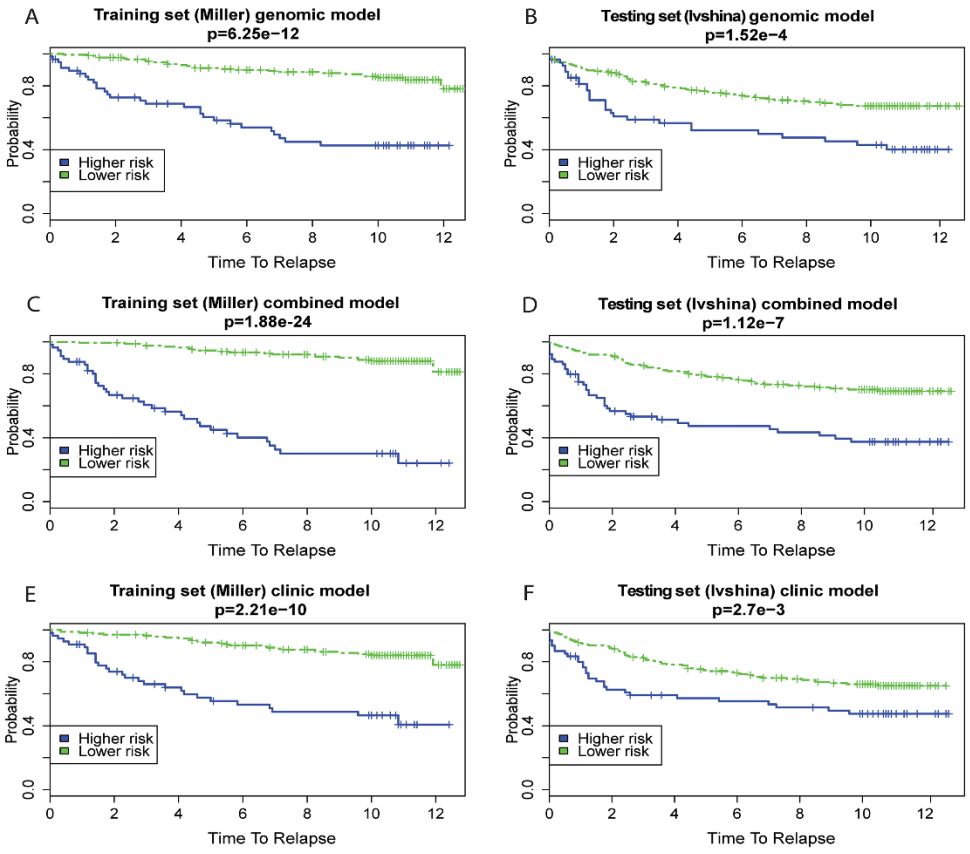
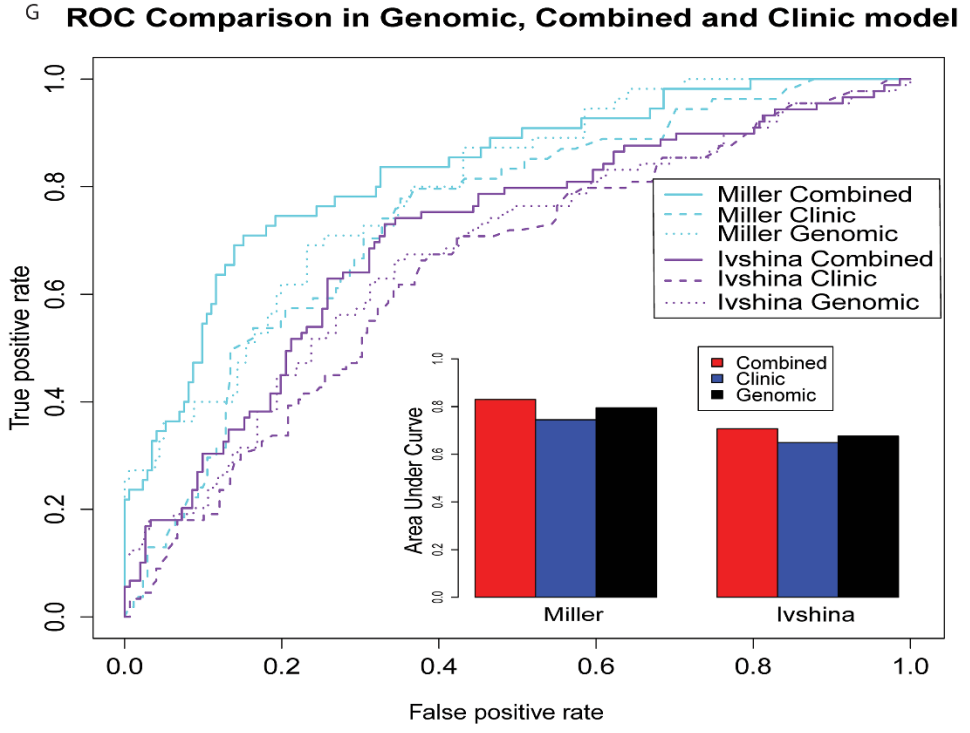


Figure 6. Comparing the prognosis performance from the pathway-based genomic model, the clinical model, and the combined model.

Higher risk and lower risk group are determined by the same PI cutoff as in Figure 4. The p-values in Kaplan-Meier curves are calculated using the Wilcoxon log-rank tests. In both the training data set and testing data set 2 (Ivshina data) that have full clinical information, the combined models outperform the pathway-based genomic model, and the pathway-based genomic model outperform the clinical model.


A
Training set (Miller) genomic model
p=6.25e−12
B
Testing set (Ivshina) genomic model
p=1.52e−4
C
Training set (Miller) combined model
p=1.88e-24
D
Testing set (Ivshina) combined model
p=1.12e−7
E
Training set (Miller) clinic model
p=2.21e−10
F
Testing set (Ivshina) clinic model
p=2.7e−3
Probability
Time To Relapse
Higher risk
Lower risk


G
ROC Comparison in Genomic, Combined and Clinic model
True positive rate
False positive rate
Miller Combined
Miller Clinic
Miller Genomic
Ivshina Combined
Ivshina Clinic
Ivshina Genomic
Area Under Curve
Combined
Clinic
Genomic
Miller
Ivshina

Appendix E: Chapter 3 Supplementary Figures

Figure S1. The effect of removing pathways on model performance (both P-values and AUCs).

A fraction (1/2, 1/4, 1/8 and 1/16) of the initial 403 pathways are randomly selected to generate PDS matrices over 18 simulations, followed by the flowchart in Figure 2. Boxplots of AUCs from ROC curves are shown.

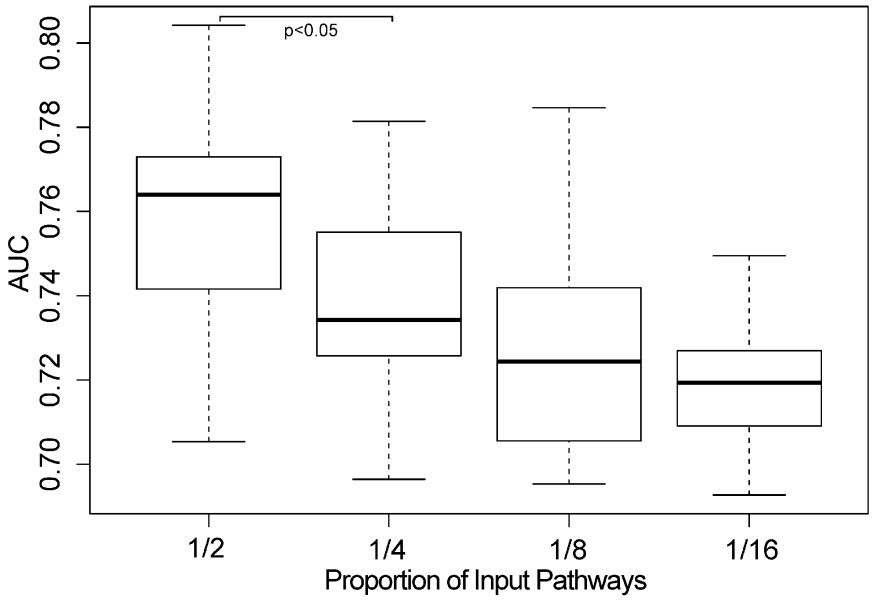


Figure S2. Cross validation results to compare the pathway-based and gene-based models on the 4 data sets in Figure 5.

Leave-one-out cross validation (LOOCV) was performed to compute the Wilcoxon log-rank test p-values (A) and AUCs (B) across all simulations. All pairs have t-test p-values < 0.001.

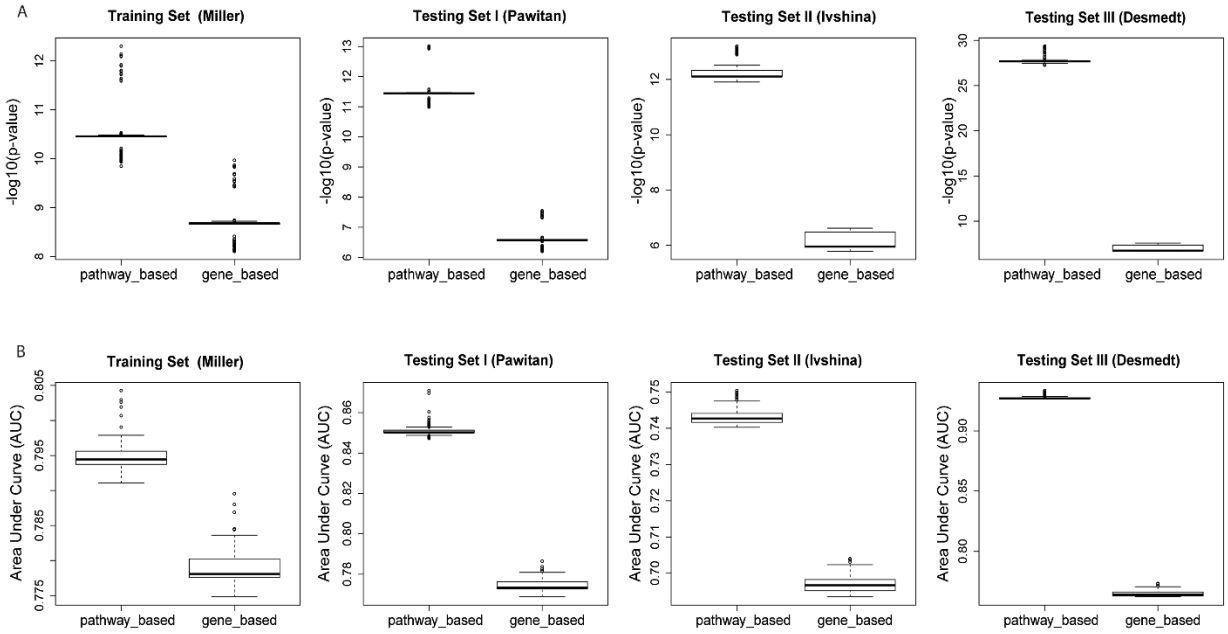


Figure S3. Comparison of ROC performance between the NKI70 method and our method on Miller dataset.

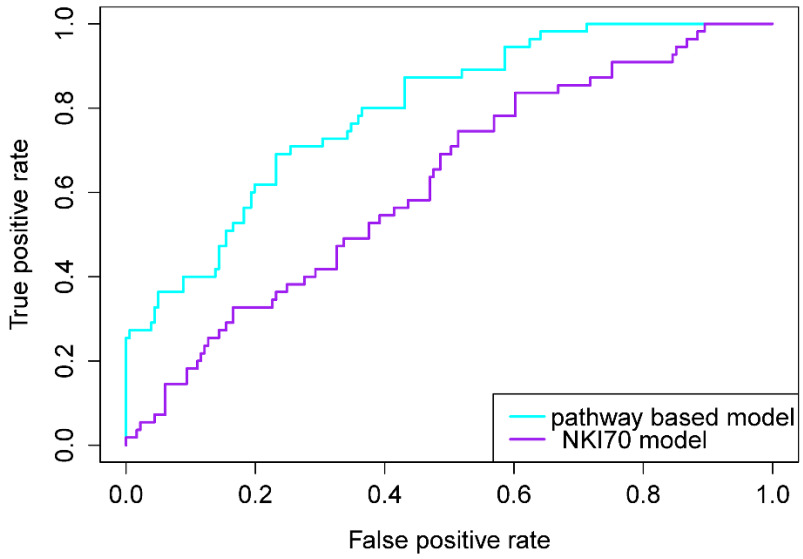
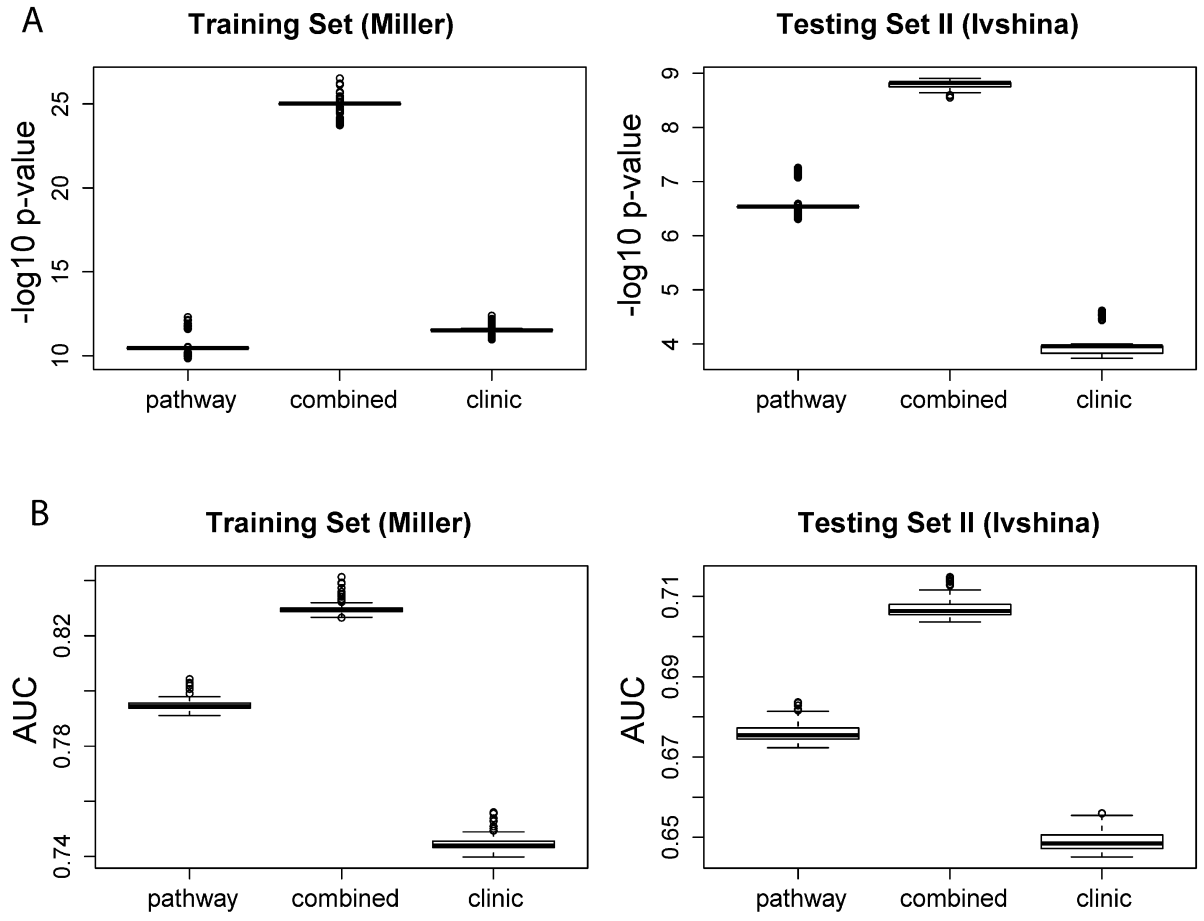


Figure S4. Cross validation results to compare the genomic, clinical, and combined models on the 2 data sets in Figure 6.

Leave-one-out cross validation (LOOCV) was performed to compute the Wilcoxon log-rank test p-values (A) and AUCs (B) across all simulations. All pairs have t-test p-values < 0.001.



Appendix F: Chapter 3 Tables

Table Legends

Table 1. Summary of patient and tumor characteristics of training and validation data sets in this study

Table 1. Summary of patient and tumor characteristics

Characteristics	Training Miller LD	Testing Set1 Pawitan Y	Testing Set2 Ivshina AV	Testing Set3 Desmedt D
No. of patients				
	236	159	249	198
Relapse, No. (%)				
Relapse	55 (23%)	40 (25%)	89 (35%)	91 (46%)
Non-relapse	181 (77%)	119 (75%)	160 (64%)	107 (54%)
Mean Relapse Free Survival (y)				
	8.167	5.959	7.142	9.312
Mean Age (year)				
	62.51		62.12	46.39
ER status, No. (%)				
Positive	201 (85%)		211 (85%)	134 (67%)
Negative	31 (13%)		34 (13%)	64 (33%)
NA	4 (2%)		4 (2%)	0
PG status, No. (%)				
Positive	57 (24%)			
Negative	179 (76%)			
NA	0			
Tumor Size(mm)				
<10 (T_{1a}, T_{1b})	13 (6%)		14 (6%)	9 (4%)
10-20 (T_{1c})	92 (40%)		95 (38%)	59 (30%)
20-50 (T_2)	123 (52%)		129 (52%)	129 (65%)
>50 (T_3)	5 (2%)		10 (4%)	1 (1%)
Grade, No. (%)				
1	62 (26%)	28 (18%)	68 (27%)	30 (15%)
2	121 (51%)	58 (36%)	126 (51%)	83 (42%)
3	51 (22%)	61 (38%)	55 (22%)	83 (42%)
NA	2 (1%)	12 (8%)	0	2 (1%)
Lymph Node Status, No. (%)				
Positive	78 (33%)		81 (32%)	
Negative	149 (63%)		159 (64%)	198 (100%)
NA	9 (4%)		9 (4%)	
P53 Mutation Status, No. (%)				
Mutated	55 (23%)		58 (23%)	
Wild Type	181 (77%)		189 (76%)	
NA	0		2 (1%)	

Table 2. Selected features in the genomic, clinical and combined models.

Table 2. Selected features in the models

Features	Coefficients	Hazard Ratio	p-values in univariate COX-PH model
Pathway-based genomic model			
KEGG_MELANOGENESIS*	1.075908	2.93266	0.00188
BIOCARTA_SRCRPTP_PATHWAY*	0.914698	2.49602	1.01e-7
BIOCARTA_AKAP13_PATHWAY*	0.828364	2.28957	0.00351
BIOCARTA_RARRXR_PATHWAY*	0.670795	1.95579	9.58e-6
BIOCARTA_VIP_PATHWAY*	0.635108	1.88723	2.15e-5
KEGG_FATTY_ACID_METABOLISM *	0.520653	1.68313	2.53e-6
BIOCARTA_G1_PATHWAY*	0.520446	1.68278	2.66e-6
KEGG_LINOLEIC_ACID_METABOLISM	0.368615	1.44573	3.55e-4
KEGG_LYSOSOME	0.300587	1.35065	2.2e-6
BIOCARTA_P53_PATHWAY*	0.239062	1.27006	8.74e-4
BIOCARTA_PYK2_PATHWAY*	0.158405	1.17164	1.29e-4
BIOCARTA_GABA_PATHWAY	0.139229	1.14939	0.0162
BIOCARTA_FEEDER_PATHWAY	0.110334	1.11665	0.0218
BIOCARTA_RNA_PATHWAY*	0.037978	1.03871	7.67e-5
BIOCARTA_IL5_PATHWAY	0.012039	1.01211	0.00895
Clinical model			
Lymph Node Status*	0.375874	1.456264	4.46e-7
Tumor Size*	0.270893	1.311135	6.03e-7
Grade	0.126814	1.135206	4.92e-4
P53	0.043517	1.044478	0.0171

*: these pathways and clinical parameters are also selected by the combined model

Table 3. Top 30 most frequent genes in the pathways of the genomic model and the combined model.

Gene ID	Genomic Model Counts	Combined Model Counts	Weighted Genomic Model Counts *	Weighted Combined Model Counts *
PRKACB	3	4	7.109452	4.46549683
PRKACG	3	4	7.109452	4.46549683
PRKCB	3	6	6.600318	6.54794229
PRKCA	3	6	6.600318	6.54794229
CALM1	3	6	5.991523	6.32738762
CALM2	3	6	5.991523	6.32738762
CALM3	3	6	5.991523	6.32738762
GNAQ	3	4	5.991523	4.26250968
SRC	3	5	4.817049	5.3291015
GSK3B	2	3	4.615435	3.23989145
CDC25A	2	2	4.178799	2.25477095
CDK1	2	2	4.178799	2.25477095
PRKAR2A	2	3	4.176796	3.26764027
PRKAR2B	2	3	4.176796	3.26764027
HRAS	2	4	4.104297	4.27393317
MAP2K1	2	4	4.104297	4.27393317
MAPK1	2	4	4.104297	4.27393317
MAPK3	2	4	4.104297	4.27393317
RAF1	2	4	4.104297	4.27393317
MAP2K2	2	4	4.104297	4.27393317
TP53	3	3	3.991545	3.03875447
GRB2	2	4	3.667662	4.32604023
CYCSP35	2	5	3.058867	5.12953106
PLCG1	2	4	3.058867	4.13411496
MAP3K1	2	3	3.058867	3.06465312
E2F1	2	2	2.952836	2.0232666
RB1	2	2	2.952836	2.0232666
CCND1	2	2	2.952836	2.0232666
CDK4	2	2	2.952836	2.0232666
CCNE1	2	2	2.952836	2.0232666

Chapter 4. A nomogram derived by combination of demographic and biomarker data improves the non-invasive evaluation of patients at risk for bladder cancer

ABSTRACT

Purpose: Improvements in the non-invasive clinical evaluation of patients at risk for bladder cancer (BCa) would be of benefit both to individuals and to healthcare systems. We investigated the potential utility of a hybrid nomogram that combined key demographic features with the results of a multiplex urinary biomarker assay in hopes of identifying patients at risk of harboring BCa. If proven accurate and reliable, the application of such a nomogram may better inform the decision to perform invasive diagnostic procedures.

Patients and Methods: Logistic regression analysis was used to model the probability of BCa burden in a cohort of 686 subjects (394 with BCa) using key demographic features alone, biomarker data alone and the combination of demographic features and key biomarker data. Demographic data included age, race, and tobacco history, and biomarker data included the urinary levels of 10 BCa-associated diagnostic proteins that we have previously described. We examined discrimination, calibration and decision curve analysis techniques to evaluate prediction model performance.

Results: Area under the receiver operating characteristic curve (AUROC) analyses revealed that demographic features alone predicted tumor burden with an accuracy of 0.806 [95% CI: 0.76-0.85], while biomarker data had an accuracy of 0.835 [95% CI: 0.80-0.87]. The addition of molecular data into the nomogram improved the predictive performance to 0.891 [95% CI: 0.86-0.92]. Decision curve analyses showed that the hybrid nomogram performed better than demographic or biomarker data alone.

Conclusion: A nomogram construction strategy that combines key demographic features with biomarker data may facilitate the accurate, non-invasive evaluation of patients at risk of harboring BCa. Further research is needed to evaluate the BCa risk nomogram for potential clinical utility.

INTRODUCTION

With an estimated 70,980 newly diagnosed cases of bladder cancer (BCa) and 14,330 deaths from BCa in 2015, cancer of the urinary bladder is the second most common genitourinary malignancy in the US and among the five most common malignancies worldwide (Madeb & Messing, 2008; Messing, 2007). When detected early (*i.e.,* non-muscle invasive), the 5-year survival rate of BCa is > 90%, however at later stages (*i.e.,* muscle invasive and beyond) the 5-year survival rate is < 50%. Thus, early BCa identification, both at the initial diagnosis and at recurrence can dramatically affect outcomes(Khadra, Pickard, Charlton, Powell, & Neal, 2000). Urine based assays that can noninvasively detect BCa have the potential to improve the rapid diagnosis of BCa. As such, several urine-based commercial molecular tests have been FDA-approved for BCa detection and surveillance. These tests include the measurement of soluble proteins such as bladder tumor antigen (BTA), and nuclear matrix protein 22 (NMP22), or proteins detected on fixed urothelial cells (ImmunoCyt), and chromosomal aberrations detected by fluorescent *in situ* hybridization (Urovysion) (Edwards, Dickinson, Natale, Gosling, & Mcgrath, 2006). Because of their marginal detection performance, these urine-based assays have a limited role in the management of patients at risk for, or with BCa, and thus the search for non-invasive urine-based tests with clinical utility for BCa continues.

We and others, have described the diagnostic capabilities of urine-based molecular signatures to non-invasively detect BCa (Bundix & Wauters, 1997; Elias, Svatek, Gupta, Ho, & Lotan, 2010; Lokeshwar et al., 2005; Nakamura et al., 2009; Têtu, 2009; Trivedi & Messing, 2009; Van Rhijn, Van der Poel, & van Der Kwast, 2005). We have refined and validated a multiplex protein biomarker panel (MMP9, MMP10, IL8, VEGFA, SERPINE1, SERPINA1, CA9, APOE, ANG and SCD1) in a series of independent cohorts (Aaboe et al., 2005; Hanke, Kausch, Dahmen, Jocham, & Warnecke, 2007; Holyoake et al., 2008). Given the utility of key demographic features (*e.g.,* age, race, sex, tobacco history) in stratifying patients, in this study we investigated the potential utility of a hybrid nomogram that incorporates key demographic features with the results of the BCa-associated diagnostic signature in hopes of improving the evaluation of risk for

harboring BCa. If proven accurate and reliable, the application of such a nomogram may guide the decision to perform invasive diagnostic procedures.

MATERIALS AND METHODS

Study Subjects

Demographic, clinical and biomarker data from 686 subjects (394 BCa subjects and 292 subjects with benign urologic conditions) were extracted from our series of independent cohorts previously published, **Table 1** (Aaboe et al., 2005; Hanke et al., 2007; Holyoake et al., 2008). All molecular data were normalized to creatinine. Based on the total distribution of each biomarker's concentration, cut-points were identified deriving low/high expression status.

Primary End Point and Baseline Information

The primary end point of the study was to predict the histologic presence of transitional cell carcinoma of the bladder, which was confirmed by biopsy. Tumor grade (2002 WHO classification) (Mengual et al., 2010) and tumor stage (2002 TNM classification)(Bartoletti et al., 2006) were noted for each case. No central pathology review was obtained.

Statistical Analysis

The distributions of the key demographic data as well as molecular data were examined. Multivariate logistic regression analysis was used to examine the association between these predictor variables and detection of BCa. All decisions with respect to the coding of the nomogram variables were made prior to modeling, as making these decisions afterwards can have deleterious effects on the predictive ability of the model (N. Yang et al., 2011). A logistic regression model based on disease status was the basis for our nomograms, which included only key demographic data, only key biomarker data and the combination of key demographic data and biomarker data.

Nomogram validation contained two components. The nomogram was subjected to bootstrapping as a means of calculating a relatively unbiased measure of its ability to discriminate among subjects. Briefly,

we compared the predicted probability of diagnosis *vs.* actual diagnosis (*i.e.*, nomogram calibration) on the 686 subjects, using 200 bootstraps to reduce overfit bias, which would otherwise overstate the accuracy of the nomogram. We quantified the discrimination ability of the risk calculator by calculating the concordance index (C-index), which is a surrogate of the nonparametric area under the receiver operating characteristic curve (AUC)(Urquidi, Goodison, Cai, Sun, & Rosser, 2012). C-index gives the probability that, in a randomly selected pair of subjects in which one has BCa and the other does not, the subject with the BCa will be assigned the worse predicted risk (Goodison, Chang, Dai, Urquidi, & Rosser, 2012). C-index ranges from 0.5 (no discrimination) to 1.0 (perfect discrimination). To test the significance between the AUCs of the three nomograms (demographic data only, biomarker data only and combination of key demographic data and biomarker data), we created 2,000 C-indices for each model by using bootstrapping analysis and then calculated the differences between the paired C-indices. Lastly, nonparametric bootstrap test (Rosser et al., 2013) was used to calculate the p-value for each pair of the nomograms.

The calibration of the three nomograms was compared by plotting the prediction on the X-axis and the observed outcomes on the Y-axis in the same plot (L.-M. Chen et al., 2014). In the calibration plot, the 45-degree line represents the perfect predictions. Due to binary outcomes, a smoothing technique was used to generate the observed probabilities of BCa on the X-axis. We also applied the decision curve analysis (Rosser et al., 2014) on our proposed nomograms and compared the net-benefits of different examine actions. All statistical analyses were performed using S-Plus software (PC Version 3.3, Redmond, WA) and R software version 3.2.3 with additional functions. All *p* values were calculated by two-sided statistical tests, unless notified otherwise.

RESULTS

Of the 686 subjects available for analysis, 394 had BCa while 292 were healthy volunteers/benign controls. Over 84% of the BCa subjects was >55 years (60% of controls), 92% of the BCa subjects were Caucasian (66% of controls), 83% of the BCa subjects were male (79% of controls). Nineteen

percent of BCa subjects denied tobacco history while 37% of controls denied tobacco use (**Table 2**). Of the subjects with BCa, 240 of the tumors were non-invasive (Ta, Tis, T1) and 147 were muscle invasive and 7 did not have a stage reported. In addition, 134 were low-grade, 251 were high-grade and 9 did not have a grade reported.

Logistic regression analysis identified key demographic risk factors (*e.g.*, age, race, sex and tobacco use) and molecular biomarkers (MMP9, MMP10, IL8, VEGFA, SERPINE1, SERPINA1, CA9, APOE, ANG and SCD1) associated with BCa. The key demographic factors were used to generate a demographic only model with AUROC of 0.81 [95% CI: 0.76-0.85]. The key biomarker data were used to generate a biomarker only model with AUROC of 0.84 [95% CI: 0.80-0.87]. Under the likelihood ratio test, the biomarker model performed better than the demographic model (p = 6.745e-4). Subsequently, these two nomograms were combined to create a hybrid nomogram that incorporated key demographic and biomarker data (**Figure 1)**. The AUROC of the hybrid nomogram was 0.891 [95% CI: 0.86-0.92], which based on the nonparametric bootstrap test, was significantly improved from the demographic model (0.81 [95% CI: 0.76-0.85], $p < 0.0001$) and the biomarker model (0.84 [95% CI: 0.80-0.87], $p < 0.0001$) (**Figure 2**). Using the hybrid nomogram, we were able to calculate the sensitivity and specificity for a range of probability for BCa (**Table 3**).

Figure 3 illustrates how the predictions from the hybrid nomogram compare with actual outcomes for the 686 subjects. The X-axis is the prediction calculated with use of the hybrid nomogram and the Y-axis is the actual freedom from cancer for our subjects. The dashed line represents the performance of an ideal nomogram, in which predicted outcome perfectly corresponds with actual outcome. Our hybrid nomogram performance after adjusting the over-fitting bias with bootstrap is plotted as the solid line Note that, because the solid line is relatively close to the dashed reference line, the predictions calculated with the use of our hybrid nomogram approximate the actual outcomes. In general, the performance of the hybrid nomogram appears to be within 10% of actual outcome, and possibly slightly more accurate at very high

levels of predicted probability.

We also applied decision curve analysis to measure the performance of our hybrid nomogram for BCa (**Figure 4**). We tested the theoretical net benefits of all actions in a range of threshold probabilities for BCa. Basically, the net benefit is measuring how our action can affect the examined relative value of false positives and false negatives, which is, when our hybrid nomogram is compared to cystoscopy and biopsy. The decision curve analyses showed that the hybrid nomogram performed better than demographic data alone above the risk threshold of 6% as well as biomarker data alone above 24% to 88%.

DISCUSSION

Predictive and prognostic nomograms in bladder cancer have been published in both nonmuscle-invasive (Bossuyt et al., 2003) and muscle-invasive bladder cancer (Edge, 2010; Montironi & Lopez-Beltran, 2005). Specifically, non-muscle-invasive nomograms of precystoscopy urinary levels of NMP22 improved the ability of age, gender and VUC to predict tumor stage and grade as well as tumor recurrence (Bossuyt et al., 2003). While in muscle-invasive nomograms, precystectomy clinical and pathologic factors pT and pN stages at the time of cystectomy (Montironi & Lopez-Beltran, 2005) and to estimate the probabilities of recurrence and all-cause and bladder cancer-specific survival (Montironi & Lopez-Beltran, 2005) after cystectomy.

To the best of our knowledge, this is the first study to evaluate and internal validate a BCa diagnostic nomogram composed of pertinent demographic features and our BCa-associated diagnostic signature. Previously, we have reported and confirmed in voided urines our BCa-associated diagnostic signature comprised of 10 biomarkers in three separate studies(Aaboe et al., 2005; Hanke et al., 2007; Holyoake et al., 2008). The first study was a case-control study of 127 patients (64 tumor bearing subjects) in which we reported a sensitivity of sensitivity 92% and specificity 97%, significantly outperforming voided urinary cytology (Aaboe et al., 2005). Subsequently, in another case-control study, we tested the

BCa diagnostic signature in 308 patients (102 tumor bearing subjects and 206 subjects with varying urological disorders, *e.g.*, uroilthiasis, gross hematuria, urinary tract infection, moderate to severe voiding symptoms), recording a sensitivity of 74% and specificity of 90%, which outperformed voided urinary cytology and the UroVysion® cytogenetic test (Holyoake et al., 2008). Recently, we published a multicenter, international case control study of 320 patients (183 tumor bearing subjects) and demonstrated continued diagnostic performance with a sensitivity of 79% and a specificity of 79%(Hanke et al., 2007) .

The gold standard for initial clinical diagnosis and staging of BCa involves cystoscopic examination of the bladder together with cytologic examination for malignant cells in the urine. Cystoscopy is an unpleasant invasive procedure, which may involve anesthetizing the patient and resection of biopsies for histopathological diagnosis and staging. Cystoscopy may also have certain side effects such as urinary tract infection, voiding symptoms and stenosis of the urethra. Voided urine cytology (VUC) remains the method of choice for the noninvasive detection of bladder cancer, with its main use being to recognize the presence of recurrence and early progression in stage and grade. VUC can be used to diagnose new malignancy, yet while it has a specificity of >93%, its sensitivity is only 25-40%, especially for low-grade and low-stage tumors (Edwards et al., 2006; Hanley & McNeil, 1982; Harrell, Lee, & Mark, 1996). Thus, current methods to non-invasively detect BCa leave much to be desired. The inadequate power of these single markers must partly explain this. The concept that the presence or absence of one molecular marker will aid diagnostic or prognostic evaluation has not proved to be the case. A number of molecular signatures have been derived and are being made commercially available as clinical assays, especially in the breast cancer field (Kattan, Eastham, Stapleton, Wheeler, & Scardino, 1998; Vickers & Elkin, 2006). We have employed a range of genomic (C.-L. Chen et al., 2013; Seigel, Naishadham, & Jemal, 2012) and proteomic (Ahmedin Jemal et al., 2011; Lokeshwar et al., 2005) profiling approaches to study voided urine samples in hopes of identifying a unique, yet accurate, molecular signature associated with BCa.

In biomarker research, though a variable maybe statistically significant in a multivariate model, it does not necessarily equate to the biomarker improving the model's predictive accuracy. For example, a

biomarker with an odds ratio of 3 may be a poor classifier and thus an odds ratio of 10 or more may be required. In addition, a single measure of association such as an odds ratio may not meaningfully describe a biomarker's ability to risk classify patients (Silverman, Hartge, Morrison, & Devesa, 1992). Thus, it is critical to determine if the addition of biomarker(s) to an existing clinical and pathologic model possesses the ability to improve the predictive accuracy of this model. The accuracy of the hybrid nomogram improved to 0.89 [95% CI: 0.86-0.92] compared to key demographic model (0.81 [95% CI: 0.76-0.85], p = 5.886e-8) and biomarker model (0.84 [95% CI: 0.80-0.87], p < 7.707e-5) (**Fig. 2**). In general, the performance of the hybrid nomogram appears to be within 10% of actual outcome, and possibly slightly more accurate at very high levels of predicted probability. We also applied decision curve analysis to measure the performance of our hybrid nomogram for BCa. The decision curve analyses showed that the hybrid nomogram performed better than demographic data alone above the risk threshold of 6% as well as biomarker data alone above 24% to 88%.

The main clinical utility of a hybrid nomogram in the described setting is to facilitate the decision on whether a patient requires cystoscopy with subsequent bladder biopsy. The hybrid nomogram would provide a probability of harboring BCa. For example if the probability of harboring BCa is < 10% perhaps the patient and physician would forego an invasive procedure. However if the risk was substantial (*i.e.,* > 70%) then mostly likely the patient would be compelled to undergo confirmative diagnostic procedure. In the absence of definitive risk thresholds, it would be important to provide a range of threshold probabilities (**Table 2**).

We acknowledge that this study is limited due to its retrospective design, to the analysis of banked urine samples collected from high volume centers and so may not be representative of the general population at risk for BCa and to limitations of available data (*e.g.,* tobacco history). Nevertheless, this cohort reflects a contemporary cohort of BCa patients, which enabled the derivation of a hybrid nomogram for testing in larger, more diverse prospective studies.

In this study, we developed a hybrid nomogram that facilitates the accurate prediction of the

probability of a patient harboring BCa. The hybrid nomogram has been constructed by combining readily available key demographic factors with key biomarker data. If such a nomogram is proven to be reliable, adoption may assist the physician and patient in deciding whether or not further evaluation is needed.

Appendix G: Chapter 4 Figures

FIGURE LEGENDS

Figure 1 Diagnostic nomogram for predicting bladder cancer.

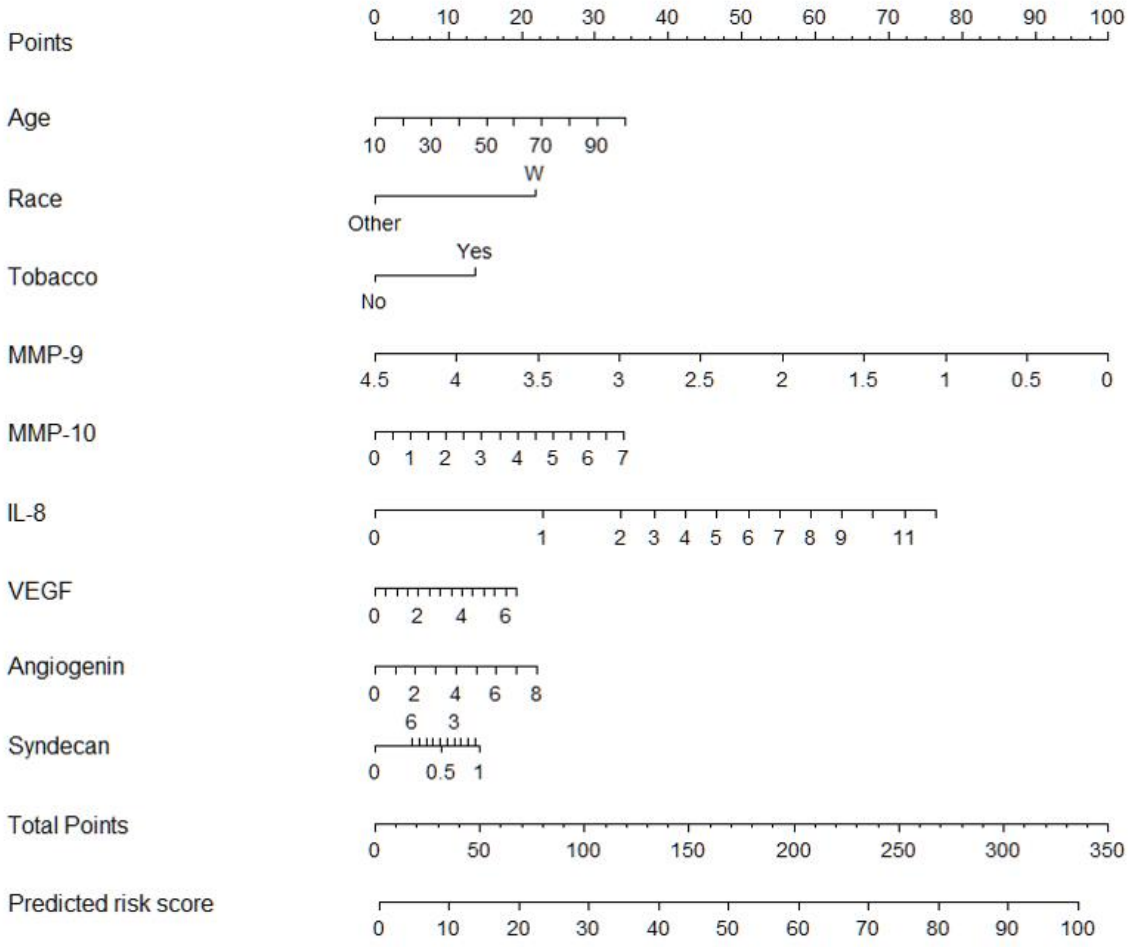


Figure 2 Receiver operating characteristic (ROC) curves for key demographic data, key biomarker data, and the combination of both for predicting the presence of bladder cancer.

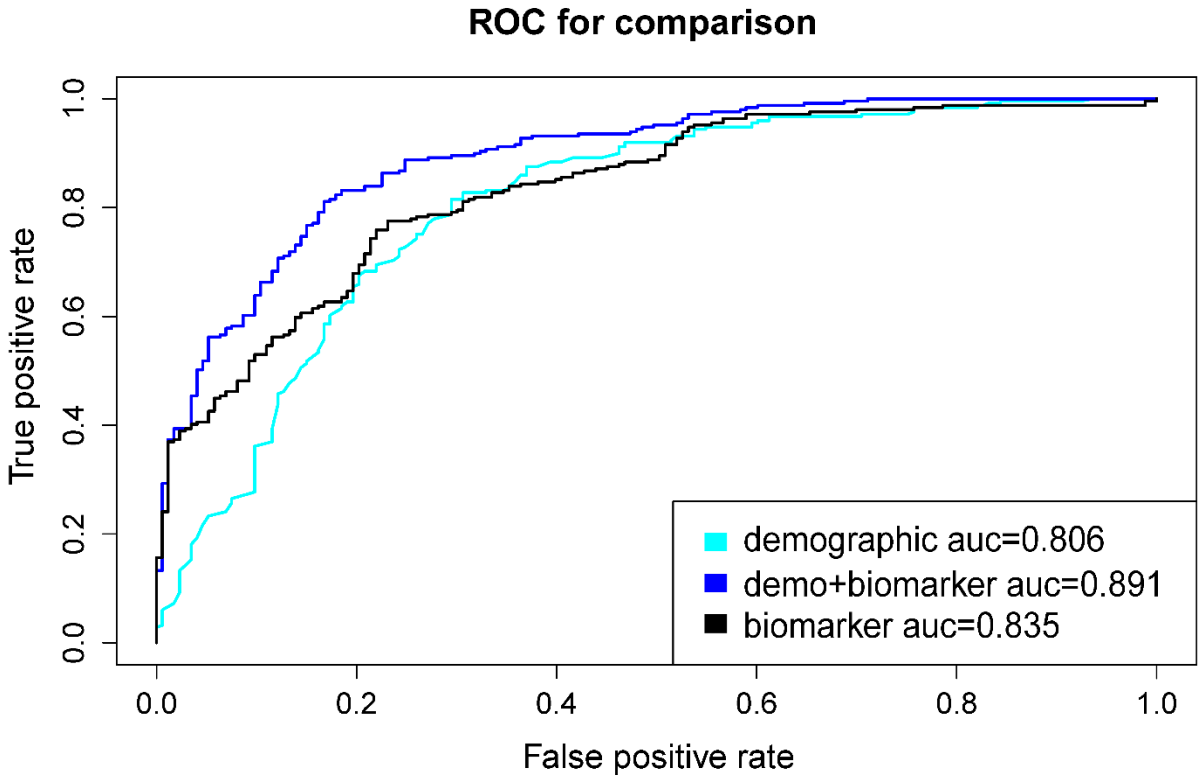


Figure 3 Calibration of the hybrid nomogram for bladder cancer. Dashed line is reference line where an ideal nomogram would lie. Dotted line is the performance of hybrid nomogram, while the solid line corrects for any bias in hybrid nomogram.

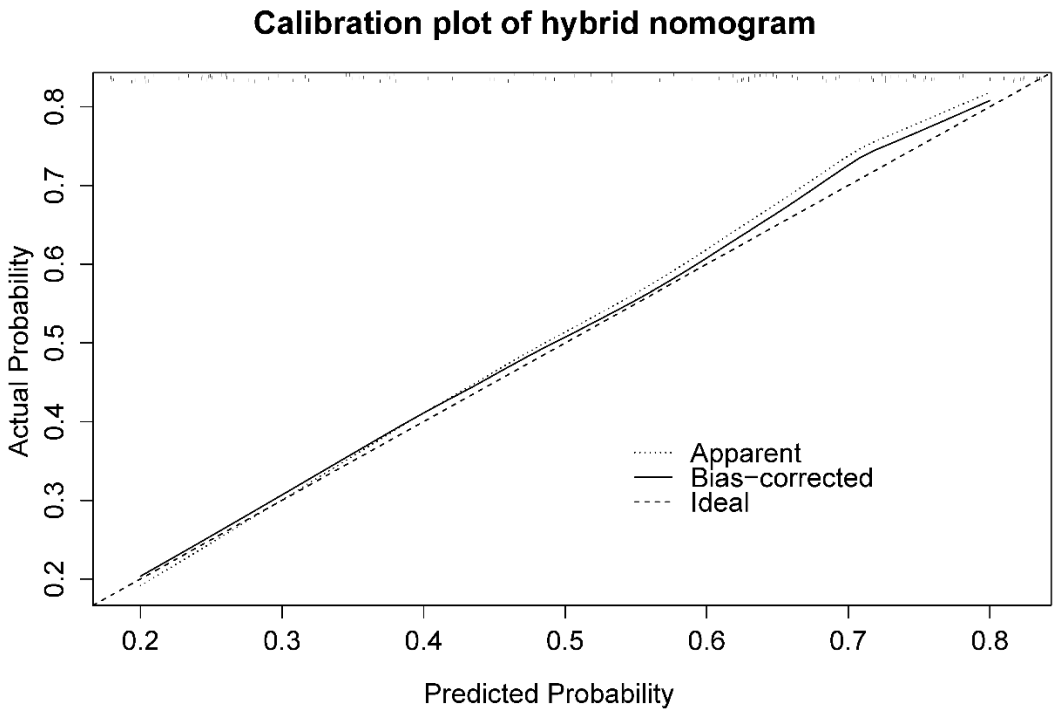
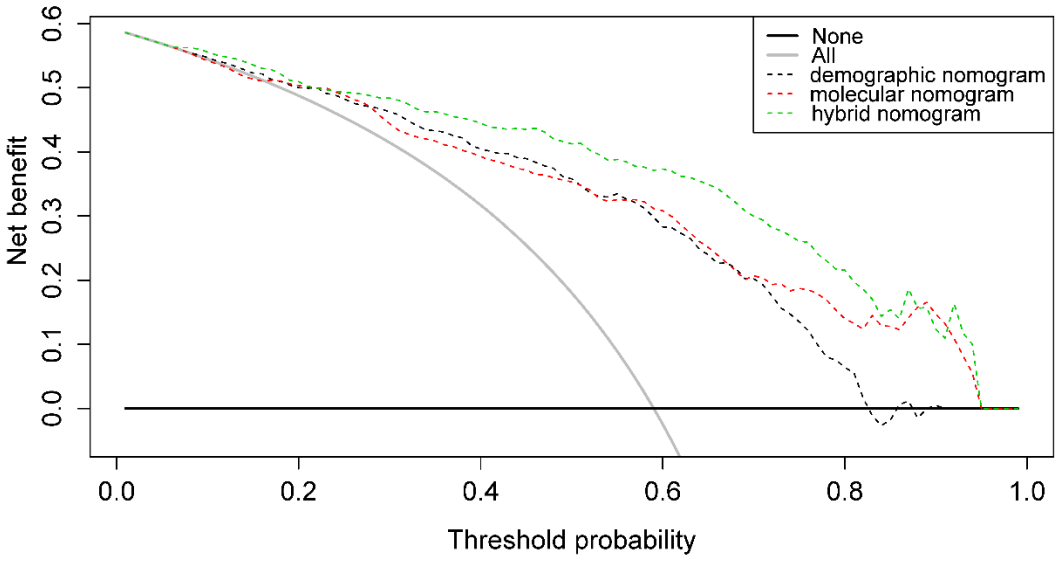


Figure 4 Decision curve analysis of hybrid nomogram. The Y-axis represents the net benefit, which is calculated by summing the benefits (gaining true positives) and subtracting weighted harms (deleting false positives). A model is of clinical value if it has the highest net benefit.



Appendix G: Chapter 4 Tables

Table 1 Multivariate Logistic Regression Analysis of Factors Associated with Bladder Cancer

	Frequency Distribution						
	Case		Controls				
Factor	**No.**	**%**	**No.**	**%**	**OR**	**95% CI**	***p***
Age, years							
≤ 55	40	16%	79	46%	0.23	(0.15,0.36)	3.17e-11
56-74	128	51%	75	43%	1.38	(0.94,2.04)	0.104
≥ 75	81	33%	19	11%	3.91	(2.27,6.74)	3.16e-07
Ethnicity							
White	221	89%	95	55%	6.48	(3.95,10.62)	3.33e-15
Other	28	11%	78	45%	0.15	(0.09, 0.25)	
Sex							
Male	204	81.9%	152	87.9%	0.63	(0.36, 1.09)	0.099
Female	45	18.1%	21	12.1%	1.60	(0.91, 2.79)	
Tobacco history							
Absent	75	30.1%	108	62.4%	0.26	(0.17, 0.39)	4.75e-11
Present	174	69.9%	65	37.6%	3.85	(2.56, 5.81)	
Biomarkers							
IL-8							
Low	75	30.1%	136	71.9%	0.12	(0.07, 0.18)	< 2.20e-16
High	174	69.9%	37	28.1%	8.53	(5.42, 13.42)	
MMP9							
Low	97	39.0%	114	65.9%	0.33	(0.22, 0.50)	5.41e-08
High	152	61.0%	59	34.1%	3.03	(2.02, 4.54)	
MMP10							
Low	118	47.4%	95	54.9%	0.74	(0.50, 1.09)	0.129
High	131	52.6%	78	45.1%	1.35	(0.92, 1.99)	
VEGF							
Low	101	40.6%	124	71.7%	0.27	(0.18, 0.41)	3.10e-10
High	148	59.4%	49	28.3%	3.71	(2.45, 5.62)	
CA9							
Low	118	47.4%	93	53.8%	0.78	(0.53, 1.14)	0.198
High	131	52.6%	80	46.2%	1.29	(0.88, 1.90)	
APOE							
Low	105	42.2%	106	61.3%	0.46	(0.31, 0.68)	1.16e-04
High	144	57.8%	67	38.7%	2.17	(1.46, 3.22)	
A1AT							
Low	80	32.1%	131	75.7%	0.15	(0.09, 0.24)	< 2.20e-16
High	169	67.9%	42	24.3%	6.59	(4.25, 10.21)	
ANG							
Low	103	41.4%	108	62.4%	0.42	(0.29, 0.63)	2.13e-05

High	146	58.6%	65	37.6%	2.36	(1.58, 3.51)	
Syndecan							
Low	113	45.4%	98	56.6%	0.64	(0.43, 0.94)	2.30e-02
High	136	54.6%	75	43.4%	1.57	(1.06, 2.32)	
PAI1							
Low	111	44.6%	100	57.8%	0.59	(0.40, 0. 87)	7.60e-03
High	138	55.4%	73	42.2%	1.70	(1.15, 2.52)	

Table 2 Sensitivity, Specificity, PPV, NPV for a Range of Probability for Bladder Cancer

Nomogram Probability (%)	Sensitivity (%)	Specificity	PPV (%)	NPV (%)
Test Characteristics for Predicting Any Cancer				
10	0.283	1.000	1.000	0.668
15	0.376	0.988	0.956	0.695
25	0.503	0.952	0.879	0.734
40	0.659	0.912	0.838	0.794
50	0.775	0.863	0.798	0.846
75	0.902	0.639	0.634	0.903
Test Characteristics for Predicting High-grade or High Stage Cancer				
10	0.008	1.000	1.000	0.399
15	0.057	1.000	1.000	0.411
25	0.455	0.840	0.812	0.504
40	0.764	0.667	0.777	0.651
50	0.862	0.494	0.721	0.702
75	0.959	0.160	0.634	0.722

Chapter 5. More is better:

Recent progress in multi-omics data integration methods

Abstract

Multi-omics data integration is one of the major challenges in the era of precision medicine. Considerable work has been done with the advent of high-throughput studies, which have enabled the data access for downstream analyses. To improve the clinical outcome prediction, a gamut of software tools has been developed. This review outlines the progress done in the field of multi-omics integration and comprehensive tools developed so far in this field. Further, we discuss the integration methods to predict patient survival at the end of the review.

Introduction

A new era of personalized medicine has arrived, which proposes an individualized health care model with tailored medical target treatment and management for each patient (Chin, Andersen, & Futreal, 2011). Under this regime, not only clinical profiles of patients but also their molecular profiles are personally managed to drive for advanced treatment. Cancer studies that are focused on one-dimensional omics data have only provided limited information regarding the etiology of oncogenesis and tumor progression. To overcome this, tremendous efforts have been made to obtain multi-platform based genomic data from biospecimen.

The Cancer Genome Atlas (TCGA) is by far the largest endeavor in the USA to collect and analyze the tumor specimens from over 10,000 cancer patients (Weinstein et al., 2013). Measurements of these specimens include tissue exome sequencing, copy number variation (CNV), DNA methylation, gene expression and microRNA (miRNA) expression, as well as some physiological and clinical data such as race, tumor stage, relapse, and treatment response. However, relative to the genomic data of different levels that are available to the public, the clinical information is more limited. A scale-up of TCGA is the International Cancer Genome Consortium (ICGC), which provides the information of genomic, transcriptomic and epigenomic abnormalities and somatic mutations over 50 different cancer types (Hudson

et al., 2010). These consortia have created unprecedented opportunities to reveal underlying oncogenic molecular signatures beneath phenotypes.

However, human genomes are complex and regulated at multiple levels, which can be manifested by various genomic assays mentioned above. While each of these assays offers a peek of the complex system, these events are rather interdependent (or interactive). Thus, when combining several different omics data to discover the coherent biological signatures, it is challenging to incorporate different biological layers of information to predict phenotypic outcomes (tumor/normal, early/late stage, survival, etc.). It is herein our goal to address the pressing and challenging issues for developing novel algorithms and theoretical methods for multi-omics data integration, in the hope to extract biologically meaningful information of clinical relevance.

The outline of this review is as follows. First, we will discuss the unsupervised data integration algorithms. Among them, we will highlight matrix factorization methods, Bayesian methods, and network-based methods. Next, we will review in-depth the supervised data integration methods, including network-based models, multiple kernel learning methods, and multi-step analysis based models. Subsequently, we will elaborate semi-supervised data integration methods. Finally, we will discuss the advancement of data integration methods for the aim of prognosis prediction and the biological insights underneath the data integration methods.

Unsupervised data integration

Unsupervised data integration refers to the cluster of methods that draw an inference from input datasets without labeled response variables. The different approaches under the umbrella of unsupervised data integration are presented in Figure 1 and Table 1. We have categorized them below into five areas: matrix factorization methods, Bayesian methods, network-based methods and multiple kernel learning and multi-step analysis.

Matrix factorization methods

NMF (Joint non-negative matrix factorization): The most straightforward method for unsupervised data

integration falls into the matrix factorization category, which focuses on the projection of variations among data sets onto dimension-reduced space (D. D. Lee & Seung, 2001) . Zhang et al. proposed NMF framework for multi-omics data integration(S. Zhang, Li, Liu, & Zhou, 2011; S. Zhang et al., 2012). This method is based on decomposing a non-negative matrix into non-negative loadings and non-negative factors:

$$min\|X - WH\|^2, W \geq 0, H \geq 0 \qquad (1)$$

where X is the matrix of mRNA transcriptome, methylome or other omics data that has M × N dimensions, W is the common factor for M × K dimension matrix and H is the K × N dimension coefficient matrix. Rather than simple correlation, the rationale is to project data onto common basis space, so that one can detect the coherent patterns among data, by examining the elements having significant z-scores. However, NMF is quite time-consuming and requires bulk memory space. For NMF, it is worth noting that not only it requires non-negative input matrices, but also proper normalization step for these input data sets as they have quite different distributions and variability.

iCluster: Like NMF, iCluster (R. Shen et al., 2012; R. Shen, Olshen, & Ladanyi, 2009) assumes a regularized joint latent variable, which is similar to W in NMF but without non-negative constraints. H is the loading factor (coefficient), the imposed sparsity with different types of penalty functions for various data types. iCluster uses E to represent the error/noise term, and the underlying decomposition equation is:

$$X = WH + E \qquad (2)$$

iCluster+: The upgraded iCluster+ expands iCluster by making the assumption of different modeling approaches for the relationships of X and W within different data platforms. It allows for diverse data types including binary, continuous, categorical and sequential data with different modeling assumptions including logistic, normal linear, multilogit and Poisson distributions (Mo et al., 2013). The common latent variable vector W represents the underlying driving factors that can be used for disease subtype assignment. Least absolute shrinkage and selection operator (LASSO) penalty is introduced to address the sparsity issue in H (Robert Tibshirani, 1996) . Since this approach requires high computational complexity, it is necessary to preselect the features critical for clustering results (Speicher & Pfeifer, 2015; Wang et al., 2014). Both

iCluster and iCluster+ do not require non-negative input data, unlike NMF.

JIVE (Joint and Individual Variation Explained): Another variation of NMF category is Joint and Individual Variation Explained (JIVE) method. JIVE decomposes the original data of each layer into three parts, including an approximation of joint variation across data types, approximation of specific structured variation for each data type, and residual noise. In other words, JIVE factors the original data input matrix (gene expression etc.) into two lower ranked representative portions W^c (shared factor) and W^s (data-specific factor), dependent on H^c and H^s (Lock, Hoadley, Marron, & Nobel, 2013). H matrix is contributed from one sub-matrix H^c common for all data types, and the other sub-matrix H^s specific to each data type.

$$X = W^c H^c + W^s H^s + E \qquad (3)$$

It should be noted that there can be separate loading factors (H^c and H^s) for the shared factor and data-specific factor (W^c and W^s). The ranks of the two loading factors can be different. An application of JIVE on gene expression data and microRNA data on Glioblastoma (GBM) samples provided information to better characterize samples into different subtypes and strong clues for associations between each input layer (gene expression and microRNA). Based on PCA for factorization, JIVE suffers from outliers, thus the robustness of JIVE is a major concern. L1 penalties are also placed to reduce the dimensions in JIVE, giving non-zero loadings representing larger and significant contributions to the variation of data.

Joint Bayes Factor: On the other hand, an alternate called Joint Bayes Factor, assumes a common factor loadings H for both shared and data-specific factor W^c and W^s (Ray, Zheng, Lucas, & Carin, 2014). Like JIVE, the original data input (e.g. gene expression data matrix) is decomposed into shared common factors across data types, data-type specific factors, and residual noise. However, unlike JIVE, which introduces sparsity using L1 penalties, the Joint Bayes Factor model assumes a beta-Bernoulli process for both the common factors and data specific factors (W^c and W^s) (Griffiths & Ghahramani, 2005; Thibaux & Jordan, 2007). For factor loadings (H), the model uses the student-t sparseness-promoting prior, to impose sparsity (Tipping, 2001). As a result, both shared features from each data type and unique features for individual

layers can be identified for further analysis. One limitation of Joint Bayes Factor lies within the linear relationship between the latent space and the observational space, and it also assumes very close relationship for different levels of data. Joint analysis of gene expression data with copy number variation data through this approach identified experimentally validated key drivers, as well as important candidates for further validation for ovarian cancer.

$$X = (W^c + W^s)H + E \qquad (4)$$

Correlation-based analysis

CCA (Canonical correlation analysis), a traditional method to investigate the relationship between two sets of variables, has been modified and applied to the data integration field. In CCA, two datasets can be decomposed as:

$$X = W_x H_x + E \qquad (5)$$

$$Y = W_y H_y + E \qquad (6)$$

H_x and H_y stand for loading factors for each data set. CCA aims to find the loading factors ($h_x^i and\ h_y^i$ representing the i^{th} column for loading factors) which maximize the correlation:

$$argmax_{H_x,H_y}\, corr(Xh_x^i,\ Yh_y^i) \qquad (7)$$

Traditional CCA doesn't account for dimension reduction techniques to compute the inverse of a covariance matrix. For the integration purpose, penalization and regularization terms are added cooperatively to create more stable and sparse solutions of loading factors. L1-penalized sCCA (sparse CCA) together with elastic net CCA were proposed to filter the number of variables to make the results more biologically interpretable (Parkhomenko, Tritchler, & Beyene, 2009; Witten & Tibshirani, 2009). Recent research on CCA includes consideration of grouped effects of features as structures embedded within the data sets, such as ssCCA (structure-constrained CCA) and CCA-sparse group (Jun Chen, Bushman, Lewis, Wu, & Li, 2013; Lin et al., 2013).

PLS (partial least squares) is focused on maximizing covariance and can potentially avoid the issue of

sensitivity to outliers. It projects variables onto a new hyperplane while maximizing the variance to find the fundamental relationship between the two sets of data.

$$X = W_x H_x + E \qquad (8)$$

$$Y = W_y H_y + E \qquad (9)$$

H_x and H_y stand for loading factors for each data set. The aim of PLS is to find the loading factors which maximize the covariance between W_x and W_y:

$$argmax_{H_x,H_y}\ cov(W_x\ ,\ W_y) \qquad (10)$$

However, in some cases such as in high dimensional biological omics data, it is desired to obtain sparse solutions for better interpretations of the result. More recently, sparse solutions of PLS such as sPLS has been shown to perform equivalently with that of the CCA-elastic net (Lê Cao, Martin, Robert-Granié, & Besse, 2009). Other implementations of PLS with different objective functions and various constraints were also reported. For example, sMBPLS (sparse Multi-Block Partial Least Squares) overcomes the limit of two data block computation through redefining the objective function as a weighted sum of latent variables in different layers (n>=2)(Ramskold et al., 2012). And SNPLS (Sparse Network regularized Partial Least Square) is specialized in identification of gene expression and drug-response relationship co-modules through incorporating gene interaction network structures (Jinyu Chen & Zhang, 2016). It showed significantly better performance in accuracy compared to sPLS in simulated data.

Bayesian methods

Bayesian methods have been applied to data integration for over a decade (Imoto et al., 2004; Zhao, Rubinstein, Gemmell, & Han, 2012). The main advantage of Bayesian methods in data integration is that they can make assumptions not only on different types of data sets with various distributions but also on the correlations among data sets. We briefly overview these methods below:

MDI (Multiple Dataset Integration): It offers to model each data set using the Dirichlet-Multinomial

Allocation (DMA) mixture model, thus can explore the shared information through deriving statistical dependencies (Kirk, Griffin, Savage, Ghahramani, & Wild, 2012). In this approach, the allocation of genes from one data set has an influence on those in another set. Apart from bi-clustering (clustering two dimensions from the same data set simultaneously), MDI can cluster a single dimension (e.g. genes) across multiple data sets, under the assumption that these genes are measured in all different levels. It can be extended flexibly by allowing variable associations from different groups of genes across data types. This method excels in identifying genes having their protein products in the same complex, apart from the co-regulated genes. Finally, after learning the similarity of clusters in different data sets, MDI obtains a single-dimension cluster among all the input data sets.

Prob_GBM is another probabilistic framework to construct patient similarity network, where patients are represented by nodes and phenotypic similarities among the patients are edges (Cho & Przytycka, 2013). This method uses the genetic phenotype, which is the gene expression data of each patient, to assign corresponding disease subtype. Explanatory features (e.g. CNVs, mutations and miRNA expression) are used to explain phenotypic similarities constructed from gene expression data, among patients. Thus, each disease subtype is modeled by a distribution of these features, and each patient is characterized as the mixture of the genetic characteristic of each subtype. Finally, patients are labeled by the most likely subtype assignment. This method considers the biological relationships among several genomic layers including mutation, CNVs, and miRNA expression data, but it is limited in terms of the types of input data.

PSDF (Patient-Specific Data Fusion): It is based on a two-level hierarchy of Dirichlet Process model, a widely used Bayesian nonparametric model for clustering (Yinyin Yuan, Savage, & Markowetz, 2011). It checks the concordance between expression and the copy number variation for each patient. Moreover, it also selects informative features and estimates the number of disease subtypes from the given data. However, this method limits the input for only two types of data (gene expression and CNV), thus reduces its flexibility within multi-platform analysis.

BCC (Bayesian Consensus Clustering): This method is a flexible clustering approach capable of

simultaneously modeling the dependence and the heterogeneity of various data sources (Lock & Dunson, 2013). It allows for separate clustering of the objects from each data source and performs post hoc integration of separated clusters. Consensus clustering is applied to model the source-specific structures as well as to determine the overall clustering.

CONEXIC (COpy Number and EXpression In Cancer): It is a Bayesian network-based method to integrate copy number variation and gene expression data (Akavia et al., 2010). A score-guided search is applied to identify the combination of modulators (genes). A ranked list of high-scoring modulators (candidate driver genes) is produced, representing genes that are both correlated with differential gene expression modules across tumor samples and are present in significantly amplified/deleted regions. The key feature of the CONEXIC goes beyond identifying mutation drivers, as it provides the insights into the roles of drivers and associated genes.

Network-based methods

Network-based approaches can identify modules, symbolic representations of the disease-associated mechanisms. In this regime, nodes represent genes and edges are links between two genes if there exists interaction between them. Under the unsupervised category, network-based methods are mostly applied for detecting significant genes within pathways, discovering sub-clusters or finding co-expression network modules (Bonnet, Calzone, & Michoel, 2015; Vaske et al., 2010; Wang et al., 2014).

PARADIGM (PAthway Representation and Analysis by DIrect reference on Graphical Models): It is a probabilistic graphical model framework to infer patient-specific genetic variations, with the incorporation of curated pathway interactions among genes (Vaske et al., 2010). PARADIGM converts each pathway in National Cancer Institute (NCI) Pathway Interaction Database (PID) into a distinct probabilistic model, represented as a factor graph with both hidden and observed states. Variables in the graph are used to describe molecules, protein-coding genes and complexes (all three assigned as physical entities) apart from gene families and abstract processes. A pathway is modeled as a directed acyclic graph where edges are defined as either positive or negative influence on the downstream nodes, and the nodes are determined by

combining all input signals. The output of PARADIGM includes the integrated pathway activity (IPA) score, representing a patient specific measure for the degree of alteration for a specific pathway, through summarizing information from input data sets such as gene expression and copy number variations. PARADIGM claims to provide more robust and consistent signatures for subgrouping patients through demonstration in breast cancer and glioblastoma samples. However, in PARADIGM pathways are measured independently, and interactions among pathways are not considered.

SNF (Similarity Network Fusion): This approach aims at discovering the patient subgroup clusters. SNF integrates different data types by constructing a network of samples (rather than genomic features) for each data type, and then fusing these networks into one comprehensive network(Wang et al., 2014). It has two main steps for data integration: First, it constructs a sample-by-sample similarity matrix for each data type, acting as an individual network. Similarity matrices help to identify universal clusters and networks. It also detects different types of data that give support to each connection in the network. Then, by using the nonlinear method of message passing theory (KNN and graph diffusion), SNF fuses different similarity matrices and networks, making the combined networks more coherent during each iteration. As a result, weak similarities (e.g. noises) are removed, and strong similarities are added. SNF is relative flexible without constraints for input data format and but only matched samples across different omics layers. By outputting combined similarities among patients across various layers, SNF offers deeper insight into the comprehensive biological relationship, beyond the scope of basic classification and subtyping methods.

Lemon-Tree: It is another unsupervised method focused on reconstructing module networks (Bonnet et al., 2015). After finding co-expressed clusters from the expression data matrix, Lemon-Tree helps to identify consensus modules and upstream regulatory programs through ensemble methods. First, a gene expression matrix is employed to infer co-expressed gene clusters through a model-based Gibbs sampler. Consensus modules of co-expressed genes are merged through spectral edge clustering algorithm with an ensemble of the gene cluster results. On the other side, additional candidate regulator types of data such as miRNA

expression, CNV and methylation data are combined with the consensus module to infer a regulatory score calculated by a decision tree structure. The above separation of module learning and regulator assignment steps provides much more flexibility allowing combination with the other methods. According to the authors, Lemon-Tree has the advantage of inferring more closely related short-path networks with more significant gene ontology-related categories, in comparison to CONEXIC. However, it limits the input data types to be only gene expression and additional one data type, as it is focused on finding co-expressed clusters.

Multiple kernel learning and Multi-step analysis

Multi-step (or multi-stage) methods are commonly used to find relationships between the different data types first, and then between the data types and the trait or phenotypes (Ritchie, Holzinger, Li, Pendergrass, & Kim, 2015). Kernel methods are defined by the use of kernel functions, which enables to operate in a high-dimensional feature space by simply computing the inner products among the images of all pairs of data in the feature space (Hofmann, Schölkopf, & Smola, 2008). Kernel-based data integration methods are usually multi-steps, thus we exemplify multi-kernel and multi-step methods together.

rMKL-LPP (Regularized Multiple Kernel Learning Locality Preserving Projections): This approach can deal with multiple omics data integration such as gene expression, DNA methylation, and microRNA expression profiles (Speicher & Pfeifer, 2015). It is an extension of the current MKL-DR (multiple kernel learning with dimensional reduction) method, where the data are projected into a lower dimensional and integrative subspace. A regularization term is added to avoid overfitting during the optimization procedure, and it allows using several different kernel types. The Locality Preserving Projections (LPP) is applied to conserve the sum of distances for each sample's k-Nearest Neighbors. The finalized clustering is done through applying k-means on the distance summation. Compared to SNF, rMKL-LPP claims to offer comparable results with much more flexibility, as it provides different choices of dimension reduction methods and a variety of kernels per data type.

CNAmet: It is a state-of-the-art multi-step integration tool for CNV, DNA methylation, and gene expression

data (Louhimo & Hautaniemi, 2011). The major goal of CNAmet is to identify genes that are both amplified and upregulated or both deleted and downregulated. This tool integrates CNV and DNA methylation data through their functions on gene regulation. The underlying hypothesis is that the gene upregulation is due to both amplified copy number and hypomethylation, whereas gene downregulation is the result of deleted copy number and hypermethylation. It uses three steps to detect the significant genes: weight calculation, score calculation, and significance evaluation. During the first weight calculation step, the signal-to-noise statistics is calculated to measure the copy number and methylation aberrations relative to the expression values. In the second score calculation step, the weight values are combined to infer a deterministic score, which informs the causes of the alterations in the gene expression. Finally, the permutation test is performed on the combined scores and the P-values are corrected. Identification of the genes which are synergistically regulated by methylation and copy number variation data leads to better characterization of these genes and better understanding of biological process underlying cancer progression.

iPAC (in-*trans* process associated and *cis*-correlated): It is a multi-step method to identify genes that are in-*cis* correlated through integrating gene expression and CNV data, as well as genes that are in-*trans* associated to the biological processes (Aure et al., 2013). The novelty of this method is the capability to adjust for confounding effects of co-occurring copy number aberrations. This analysis module combines correlation analysis, regression, gene set enrichment, and adjustment for co-occurring copy number aberrations with avoidance of confounding effects. In the in-*cis* correlation, it proposes a linear model where log gene expression is a linear function of log copy number and noise. In the in-*trans* association, it imposes a direct integration through a statistical enrichment step to get the confidence level of in-*trans* associations between the genes and biological processes.

Supervised data integration

Contrary to the unsupervised data integration methods, the supervised methods consider the phenotype labels of samples (disease or normal), and invoke machine training approaches to evaluate the models. Supervised data integration methods are built via information of available known labels from the training

omics data. In the following section, we enlist representative network-based, multi-kernel and multi-step based methods (Figure 2 and Table 1).

Network-based methods

ATHENA (Analysis Tool for Heritable and Environmental Network Associations) is a neural network approach to integrate different omics data with a supervised model which can further be extended to do prognosis analysis (Kim, Li, Dudek, & Ritchie, 2013). In ATHENA, grammatical evolution neural networks (GENN) algorithm is utilized to train individual models from different data platforms. Based on neural networks, grammatical evolution algorithm is utilized to train the model with selected features that are less noisy and significantly associated with clinical outcomes. After selecting the features, individual models are summed up to a final integrative model, which can be utilized for multiple purposes including diagnosis and prognosis. Overall, ATHENA provides a comprehensive way of visualizing genomics data's correlation with clinical features such as survival outcomes, making it stand out compared to other network-based integration methods. One limitation of ATHENA lies in lacking interaction terms among different layers, as the features are selected from individual data type first and then combined into an integrated model.

jActiveModules: It is another network-based Cytoscape plug-in which seeks underlying network hotspots through the integration of gene expression, protein-protein interaction and protein-DNA interaction data (Ideker, Ozier, Schwikowski, & Siegel, 2002). This method is based on the hypothesis that molecular interactions linking the genes are more likely to correlate expression profiles than randomly chosen genes in the network. This method requires an external input of significance measurements over genes for significance calculation of sub-networks. The external filtering step is a supervised feature selection for genes based on the P-values in the differential expression tests, while the integration method itself doesn't require additional outcomes as inputs. Through random sampling approach and iterative calculations, jActiveModules determines the highest-scoring sub-network circuits in a full network of molecular interactions, leading to further biologically interesting discoveries (Cline et al., 2007). Compared to other

clustering methods, jActiveModules is subject to the molecular interaction network and can include genes without dramatically expression fold changes.

Another network-based integration method (Ruffalo, Koyutürk, & Sharan, 2015) claims to identify key proteins at sample-level using propagated protein networks, based on integrated mutation and differential gene expression (DGE) data sets. Propagated mutation and DGE profiles for each gene are generated with the help of prior knowledge in PPI framework (Schaefer et al., 2012). Feature selection is then done on these propagated profiles in a supervised fashion, with top features being most relevant to outcomes, and a final set of proteins is selected based on the network proximity across the samples. The final step involves logistic regression using the selected genes. This method is useful to find the hidden repertoire of genes/proteins at pathway level with impact on tumor progression/clinical outcome, which might be overlooked by individual mutational or differential expression analysis.

Multiple kernel learning

SDP/SVM (Semidefinite Programming/Support Vector Machine): It offers a pioneering kernel-based framework for data integration (Lanckriet, De Bie, Cristianini, Jordan, & Noble, 2004). Each data set is represented by a specific kernel function that defines similarity between pairs of entities. Then the kernel functions, derived from different omics data, are combined directly using the SDP (Semidefinite Programming) techniques to reduce the integration problem to a convex optimization problem. The SDP method outperforms the classifier trained with a naïve and unweighted combination of kernels. Different kernels correspond to different transformation of the data, with an extraction of a specific type of information from each data set. The FFT (Fast Fourier Transform) kernel is specific for the membrane protein recognition, by directly incorporating information of hydrophobicity patterns. Higher-order polynomials such as radial basis kernels can be used to capture higher-order non-linear associations of a trait with genotypes. Diffusion kernels are applied to exploit unlabeled data. SDP/SVM is a prototype work for kernel-based data integration methods (published in 2004) and doesn't include a programming package. FSMKL (Feature Selection Multiple Kernel Learning) is another method implementing the multiple kernel

learning-based supervised learning (Seoane, Day, Gaunt, & Campbell, 2014). This new scheme uses the statistical score for feature selection per data type per pathway. By employing additional kernels based on clinical covariates, it improves the prediction accuracy for cancer detection. Multiple kernel learning constructs classifiers with a decision function dependent on a variety of different types of input data (gene expression & CNV) using pathway-based kernels. Each type of data (omics) is encapsulated into an object called base kernel; a composite kernel is built as a linear combination of these base kernels. To further incorporate biological information into the algorithm, not only individual feature (such as genes) are independently used to construct kernels, but also specific groups of genes, which are known to have membership from a KEGG pathway, are combined to derive other base kernels. The most appropriate decision function over kernels is finalized after feature selection steps, contributing to an integrative function over base kernels. This method stands out among other kernel-based methods with the inclusion of pathway-based information to build kernels, as prior knowledge. Pathway membership is a central criterion for FSMKL to group samples into different clusters, bringing more biological knowledge compared to basic statistical priors from other methods. Combining clinical factors along with high-throughput profiles into the classifier also brings power for prediction accuracy. FSMKL claimed that this method competes with the winner methods from the DREAM challenge for breast cancer prognosis.

Multi-step analysis

iBAG (integrative Bayesian analysis of genomics data) is a flexible tool to integrate data from an arbitrary number of platforms (Jennings, Morris, Carroll, Manyam, & Baladandayuthapani, 2013). A hierarchical model is built to incorporate the information from different genomic layers with biological sense. Basically, this multi-step analysis consists of two-stage models. The first-stage mechanistic model is a regression model which is constructed to partition gene expression data into small segments including methylation principal component, copy number variation principal component and unknown components other than the previous two. In the second stage of developing clinical a model, clinical data such as binary outcome and survival information is modeled as the response of joint regression from those factors in the previous

regression. Normal-Gamma (NG) prior is applied to improve the effect size estimation and address sparsity. This study considers gene expression, methylation and CNV data, in specific, to identify genes having a significant impact on patient survival. The hypothesis of this research lies in the linear relationship between methylation data and copy number variation, together with the effect of gene expression on survival outcome. These relationships may not reflect the actual biological process underneath, thus the output prognostic genes may not be considered as causal factors. Independent functional experiments and other datasets are needed to validate the results.

MCD (Multiple Concerted Disruption): This method allows to integrate copy number variation, DNA methylation, and allelic (loss of heterozygosity) status to find genes representing key nodes in the pathways as well as genes which exhibit prognostic significance (Chari, Coe, Vucic, Lockwood, & Lam, 2010). For each differentially expressed gene, the copy number variation, methylation and allelic statuses are examined for whether the observed expression change would match the expected change in the DNA level. This multi-step tool can be broken down into several sequential steps: First, a set of most frequent differentially expressed genes is identified for each sample with a pre-defined frequency threshold. Next, this subset of genes is further checked according to the concerted pattern of the expression change and also in at least another DNA dimension (CNV, methylation or loss of heterozygosity). Finally, genes are selected which have a role in multiple disruption mechanisms and changes in expression. As a pioneering work in data integration field, MCD offers a biologically sensible way to select genes step wisely by incorporating parallel analysis in genomic and epigenomic layers. However, it is more like a filtering step to finalize a group of genes rather than a systematic way to integrate information embedded from multiple layers.

Anduril: It is a bioinformatics workflow proposed to generate integrative results from multiple platforms into a report for biologists for better comprehension (Ovaska et al., 2010). It is a flexible and intuitive analysis tool, which facilitates the integration of various data formats, bio-databases and analysis techniques to identify the genes and loci with high impact on survival. It supports data input including gene expression, miRNA expression, methylation, CNV, exome sequencing and array CGH data. The workflow maneuvers

to manage and automate the sequence of multi-platform analyses from importing the raw data to reporting and visualizing the results. The generated comprehensive website collects all the analyses results and thus facilitates the interpretation of the data. However, this framework is more of a platform to collect and process multiple types of data, rather than a package that performs data integration with sophisticated statistical or machine learning methods.

Semi-supervised data integration

Semi-supervised integration methods, lies between supervised and unsupervised methods, takes both labeled and unlabeled samples to develop learning algorithm. Most of the semi-supervised data integration methods are graph-based, as illustrated with a few examples below (Table 1).

GeneticInterPred: It is a tool to predict the genetic interactions through combining the protein-protein interaction, protein complex and gene expression data (You, Yin, Han, Huang, & Zhou, 2010). This method starts with building a high-coverage, high-precision weighted functional gene network by integrating gene expression, protein complex, and protein-protein interaction data. The topological properties of the protein pairs and gene expression in the function gene network are used as input for the subsequent classification step. A weight matrix is built summarizing the information among the edges in the graph, which is made symmetric. A similarity matrix is inferred from the weight matrix iteratively, using local connectivity in the gene network until convergence. Using connected weighted graph, the graph-based semi-supervised learning (SSL) method can infer the information of the unlabeled interactions in the graph. The final product is a classification matrix where all the unlabeled interactions are assigned. This method is specifically designed for prediction of genetic interaction from integrated functional gene networks. Moreover, the semi-supervised idea of inferring unlabeled data from labeled data in the connected graph of similarity matrix can be applied to clinical predictions like cancer diagnosis and prognosis.

Another pilot framework employing graph-based semi-supervised learning uses the multi-level genomic data sources (including CNV, gene expression, methylation and miRNA expression) for molecular classification of clinical outcomes (Kim, Shin, Song, & Kim, 2012). This method uses the genomic

relationship to define the edges (relationship) between the nodes (samples), and the unlabeled samples are influenced by the propagation of their annotated neighbors. In the end, diverse graphs from different layers are combined by the linear combination of coefficients for the individual graphs. It allows the flexibility to extend to integrate multiple levels of genomic data (n>3), while preserving the level-specific properties from the different and heterogeneous layers. In summary, this work pioneered in combining genomics, epigenomics and transcriptomics data to predict for cancer phenotypes. However, the interaction relationships among different layers were not considered, such as the regulatory role of methylation or microRNA on gene expression.

Biological insights from data integration methods

By now we discussed a variety of integration methods in three categories: unsupervised, supervised and semi-supervised. Unsupervised methods recruit different approaches (factorization, Bayesian, network etc.) to explore their biological profiles to assign objects into different subgroups (clusters). Supervised methods employ the biological information of labelled objects to derive patterns for different phenotypes and assign labels to unlabeled data by comparing the patterns. Semi-supervised methods are mostly building object-wise similarity networks through compiling omics data and assign labels to unknown objects through their relationship to labelled objects.

Interactions among different layers are major concerns for data integration strategies. The corresponding mapping relationship among different layers such as methylation to gene expression, microRNA to gene expression etc. should not only be considered independently but also together during the integrative process. At the initiating stage of data integration, many integrative methods are independently working on different layers (such as multi-step analysis) and then find the common subset of biological identities (e.g. genes)which are significantly differentially expressed in each layer. The more recent emerging state-of-the-art integrative tools are considering interactions while integrating different layers. SNF, for example, tries to integrate patient-wise similarities as a combined network, which both strengthens the coherent relationships from each network and reduces the noise of weak signals from the individual network.

iCluster+, on the other hand, aims to discover the common latent variable (structure) from all different omics layers with different modeling assumptions. Thus, the internal relationship of different layers is considered as the driving factor that acts in a concerted manner from each omics data.

Data integration for survival prediction

Nowadays cancer prognosis prediction is a keen point of interest for physicians, cancer patients, and healthcare-providers. Information about cancer prognosis helps all kinds of decisions regarding the patient management and therapeutic treatments etc. (Hagerty, Butow, Ellis, Dimitry, & Tattersall, 2005; Rabin et al., 2013). Prognostic biomarkers have been used for more effective selection of patient subgroups with different therapeutic strategies (Huang et al., 2016; S. Huang et al., 2014). Therefore, molecular data with increasing power to detect personalized molecular characteristics has been studied widely in the past decade (Kim & Ritchie, 2014; Van't Veer et al., 2002). However, methods to integrate multi-omics data optimized for prognosis prediction (rather than being post hoc evaluation) are far fewer (Table 1). We enlist some representative methods below:

CoxPath: It is a vector space integration methodology that can handle CNV, gene expression, DNA methylation and miRNA expression data (Mankoo, Shen, Schultz, Levine, & Sander, 2011). First, the Spearman rank correlations among different data types are computed, and separate cut-offs are used to filter the correlated data pairs. After the filtering, L1-penalty is combined with Cox proportional hazards model for feature selection and model shrinkage simultaneously. This metric is a typical multi-step analysis method to predict survival.

MKGIs (Metadimensional Knowledge-driven Genomic Interactions): This framework performs knowledge-based integration of multi-omics genomics data at pathway level (Kim et al., 2016; Kim & Ritchie, 2014), to predict the clinical outcome of patients. The strength of the framework lies in capturing genomic interactions by integrating pathways with the metadimensional models to achieve improved prognosis and diagnosis. In transformation phase, each genomic layer is converted to pathway-based knowledge-driven matrix. In modeling phase, an evolutionary algorithm-based method called grammatical

evolution neural networks (GENN) is used to develop knowledge-driven models for predicting clinical outcome. GENN is essentially an artificial neural network (ANN) based on grammar rules, which optimizes the high-dimensional input features, network structure, and weights. Further, different genomic interaction models are integrated to develop MKGI models to predict survival, stage and grade. This method concludes that knowledge-driven (pathway-based) genomic models overall perform better than single genomic-based models where gene expression is most contributing at the pathway level.

Conclusion

A plethora of data is accruing with the high-end experimental set-ups in the field of pathology. Advanced technologies are coupled with the computational challenges to deliver the most relevant biological interpretation of data. In this direction, a considerable number of tools have been developed to make the most out of the multi-tier data sets. This review summarizes the diverse computational tools developed over the years, their advantages and limitations. As this field flourishes, comparisons among different methods will be critical, to aid decision-making by investigators with big data needs. Despite these accomplishments, there needs to be more accurate and efficient tools, especially when clinical outcome (e.g. survival) is to be modeled. Biological knowledge guided integrative methods will continue to be desirable, with consideration of the interactive relationship among different omics layers. Moreover, given that most studies have only a single or a few omics layers, integrating heterogeneous data from multiple cohorts, rather than coupled samples, will need rigorous investigation (R. Wei et al., 2016). For the purpose of precision medicine, additional benefits may be obtained by integrating omics data with other data types, such as imaging and electronic health record (EHR) data.

Appendix I. Chapter 5 Figures

Figure 1: Unsupervised data integration methodology.

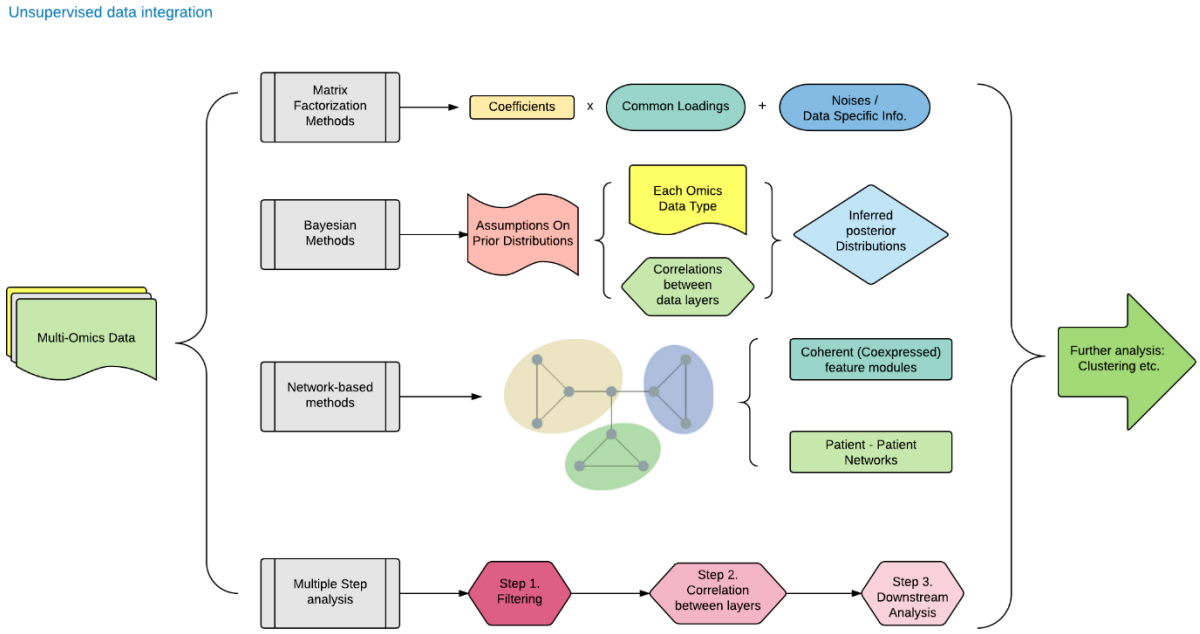
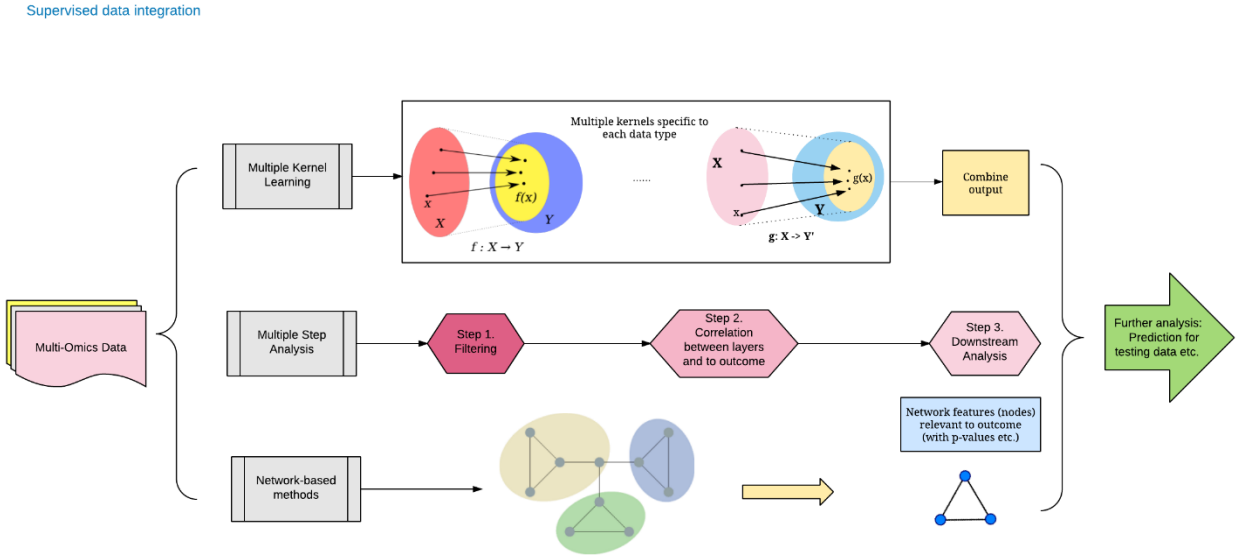


Figure 2: Supervised data integration methodology.



Appendix J. Chapter 5 Tables

Table 1: Summary of data integration tools.

Name	Category	Data Type	Output	Stats Method	FFS Method	Reference
Joint NMF	Unsupervised	Multi-data	subset of genes (modules)	Matrix factorization	NA	(Zhang et al., 2011, 2012)
iCluster	Unsupervised	EXP, CNV	cluster	matrix factorization	L1 penalty	(Shen et al., 2012)
iCluster+	Unsupervised	Multi-data	cluster	matrix factorization	L1 penalty	(Mo et al., 2013)
JIVE	Unsupervised	Multi-data	shared factors and unique factors	Matrix factorization	L1 penalty	(Lock et al., 2013)
Joint Bayes Factor	Unsupervised	EXP, MET, CNV	shared factors and unique factors	Matrix factorization	Student-t sparseness promoting prior	(Ray et al., 2014)
ssCCA	Unsupervised	Sequence data	Operational taxonomic unit and cluster	Canonical Correlation Analysis	L1 penalty	(Chen et al., 2013)

CCA sparse group	Unsupervised	Two types of data	Group of features with weights	Canonical Correlation Analysis	L1 penalty	(Lin et al., 2013)
sMBPLS	Unsupervised	Multi-data	Group of features as modules	Partial Least Squares	L1 penalty	(Li et al., 2012)
SNPLS	Unsupervised	EXP, drug response, gene network info.	Gene-drug co-module	Partial Least Squares	Network-based penalty	(Chen and Zhang, 2016)
MDI	Unsupervised	Multi-data	Cluster	Bayesian	NA	(Kirk et al., 2012)
Prob_GBM	Unsupervised	EXP, CNV, miRNA, SNP	Cluster	Bayesian	NA	(Cho and Przytycka, 2013)
PSDF	Unsupervised	EXP, CNV	Cluster	Bayesian	Binary indicator->likelihood of feature	(Yuan et al., 2011)
BCC	Unsupervised	EXP, MET, miRNA, proteomics	Cluster	Bayesian	NA	(Lock and Dunson, 2013)
CONEXIC	Unsupervised	EXP, CNV	Groups of genes	Bayesian	NA	(Akavia et al.,

			associated with modulators			2010)
PARADIGM	Unsupervised	Multi-data	gene score and significance in each pathway	pathway networks	NA	(Vaske et al., 2010)
SNF	Unsupervised	EXP, MET, miRNA	Cluster	similarity network fusion	NA	(Wang et al., 2014)
Lemon-Tree	Unsupervised	EXP, CNV/miRNA/methyl (only one type)	association network graphics	module network	NA	(Bonnet et al., 2015)
rMKL-LPP	Unsupervised	Multi-data	Cluster	Multiple kernel learning	Dimension reduction metric Locality Preserving Projections (LPP)	(Speicher and Pfeifer, 2015)
CNAmet	Unsupervised	EXP, MET, CNV	scores and p-values of genes	Multi-step analysis	NA	(Louhimo and Hautaniemi, 2011)
iPAC	Unsupervised	EXP, CNV	subset of genes	Multi-step analysis	Multiple filtering steps including common aberrant genes, in-cis correlation and in-trans functionality	(Aure et al., 2013)
ATHENA	Supervised	EXP, CNV, MET, miRNA	Final model with patient index	Grammatical Evolution Neural	Neural Networks	(Kim et al., 2013)

				Networks (GENN)		
jActiveModules	Supervised	EXP, PPI, protein-DNA interactions	Subnetwork (network hotspots)	Network simulated annealing	NA	(Ideker et al., 2002)
Network propagation	Supervised	Gene expression, mutation, PPI	Propagated network relative to differential expression of gene	Network	NA	(Ruffalo et al., 2015)
SDP/SVM	Supervised	EXP, protein sequence, protein interactions, hydropathy profile	Linear classifier based on combination of kernels	SDP/SVM	Recommends CCA (canonical correlation analysis)	(Lanckriet et al., 2004)
FSMKL	Supervised	EXP, CNV, Clinic feature (ER status)	Linear classifier based on combination kernel	Multiple kernel learning	SimpleMKL (gradient descent method)	(Seoane et al., 2014)
iBAG	Supervised	Multi-data	Subset of genes	Multi-step analysis	Bayesian	(Jennings et al., 2013)
MCD	Supervised	MET, CNV, LoH	Subset of genes	Multi-step analysis	NA	(Chari et al., 2010)
Anduril	Supervised	EXP, MET, miRNA, exon, aCGH, SNP	Comprehensive report	Multi-step analysis	NA	(Ovaska et al., 2010)

GeneticInterPred	Semi-supervised	EXP, PPI, protein complex data	Genetic interaction labels	Graph integration	NA	(You et al., 2010)
Graph-based learning	Semi-supervised	EXP, CNV, MET, miRNA	Patient scores for classification purpose	Graph integration	NA	(Kim et al., 2012)
CoxPath	Survival-driven	EXP, CNV, MET, miRNA	Prognosis index for each patient	Multi-step analysis	L1 penalty	(Mankoo et al., 2011)
MKGI	Survival-driven	EXP, CNV, MET, miRNA	Final model with patient index	Grammatical Evolution Neural Networks (GENN)	Neural Networks	(Kim et al., 2016)

#FS Method=Feature Selection Method, EXP= Expression, CNV= Copy Number Variation, MET=DNA Methylation, SNP= Single Nucleotide Polymorphism, aCGH= Array Comparative Genomic Hybridization, PPI= Protein-Protein Interaction, LoH=Loss of Heterozygosity

Chapter 6. Deep learning based pathway level multi-omics integration for breast cancer prognosis prediction

Abstract

Breast cancer is the most common malignancy in women worldwide. With the increasing awareness of heterogeneity in breast cancers, better prediction of breast cancer prognosis is much needed for more personalized treatment and disease management. In this chapter, we extend the work from previous chapters, and have developed a novel multi-omics computational model for breast cancer prognosis by combining the Pathway Deregulation Score (PDS) based pathifier algorithm with an improved deep version of learning integration framework DeepProg. We trained the model on METABRIC 2-omics set with gene expression and copy number variation (CNV) data, and validated the performance on four diversified independent testing data sets from GEO. To evaluate the performance of the model, we conducted survival analysis of the dichotomized groups, and compared the Kaplan-Meier curves and C-indexes of our prediction model. The resulting prognosis model successfully differentiated relapse in the training set (log rank p-value = 3.65e-20) and four testing datasets (log rank p-value < 0.05). Moreover, the pathway-based model consistently performed better than original data based models on all five data sets. Our deep-learning based multi-omics integration method outperforms the current state-of-art method SNF for patient survival prediction, on five benchmark datasets. In summary, we propose a novel prognosis model that harnesses the pathway-based dysregulation as well as deep-learning integration for breast cancer prognosis prediction. Our model is also flexible to predict future fewer-omics or individual omics level breast cancer patients' survival.

1 Introduction

Breast cancer (BRCA) is the most frequently diagnosed cancer in women in United States with 30% prevalence rate, and it is ranked second (14%) for the deaths among cancer patients in women in 2017(Society, 2017). It has been increasingly realized that breast cancer is a heterogeneous disease and

can’t be simply stratified by molecular subtypes only. More personalized identification and management tools are pressingly needed for breast cancer prognosis. Toward this goal, a variety of high-throughput based signatures have been explored for breast cancer prediction purposes, and some of the signature panels are currently in commercial use (Ma et al., 2004; Paik et al., 2004; Sotiriou et al., 2006; Van't Veer et al., 2002; Yixin Wang et al., 2005). However, there is consistency problem using the gene-based/ raw data-based platforms. Ein-Dor et.al showed that the NKI-70 signature is not unique and can be strongly influenced by the selection of training subset to derive candidate genes(Ein-Dor, Kela, Getz, Givol, & Domany, 2004). These fluctuations in signature genes are mostly due to the large pool of survival-correlated genes. The correlation difference among those genes are so small that randomly changing the training sets lead to different signature patterns with similar predict performances.

Moreover, the accumulation of multi-platform based genomic data offers measurements in genomic, transcriptomic and epigenomic levels on the same cancer patient. The Cancer Genome Atlas (TCGA) and the International Cancer Genome Consortium (ICGC) have created unprecedented opportunities to uncover the biological oncogenetic and progression processes underneath cancer phenotypes. The integration of multi-omics data has been applied to breast cancer to discover the mechanisms (Mosca et al., 2010). Mosca et. al presented a Genes-to-Systems Breast Cancer (G2SBC) Database which integrates all knowledge of protein-protein interactions (PPIs), protein structure, molecular pathways and systems modeling to study breast cancers(Mosca et al., 2010). However, this multi-level perspective only provides a summarized information rather than a quantified personalized measurement which will be more beneficial for personalized management. Similarity Network Fusion (SNF) analysis integrates different data platforms by constructing a network of samples (rather than genomic features) for each data type, and then fusing these networks into one comprehensive network(Wang et al., 2014). SNF helps to define 5 clusters for breast cancer with different molecular patterns in gene expression, methylation and miRNA profiles. However, it doesn’t allow for new data prediction based on the clustering results. More importantly, few integration methods have been proposed to link molecular features to predict survival phenotypes, the molecular

dysregulations at multiple levels haven't been combined systematically to identify the risky pathways in breast cancers(Huang, Chaudhary, & Garmire, 2017).

Given the observation that genes and other biological entities involved in the same biological processes are often dysregulated together in cancer, we hypothesize that higher-order quantitative representations of features, such as pathway-based features, are coherent surrogates of original data biomarkers and add more information of biological functions. Previously, we developed a personalized, novel computational pathway based model and have applied on transcriptomics and metabolomics data for breast cancer prognosis and diagnosis(Huang et al., 2016; Sijia Huang, Cameron Yee, Travers Ching, Herbert Yu, & Lana X Garmire, 2014). Recently, several studies used deep-learning approach to transform genomics data(Hongzhu Cui et al., 2017). In a previous study, we combined autoencoders and Cox-PH models to extract new features linked to survival and predict cancer subtypes for Hepatocellular carcinoma (HCC), using mRNA, microRNA and DNA methylation data (Chaudhary, Poirion, Lu, & Garmire, 2017). In this chapter, we modified this deep-learning based computational pipeline that takes multi-omics input features, and extended to the pathway-based level. This new workflow, named DeepProg, first creates an autoencoder for each omic to extract omic-specific survival features. Then, the model identifies the cancer subtypes using a clustering approach and further builds a machine-learning classification model to predict new samples' survival risks. Other characteristics of DeepProg include adopting a bagging approach that increases the robustness of the results and the prediction accuracy of the method. We built a model using 1981 breast cancer samples and predicted the survival subtypes for four external GEO datasets. Furthermore, we show that pathway features are superior to original features (genes, CNV) in predicting breast cancer prognosis. We also demonstrate the improved accuracy of survival-risk prediction of DeepProg tool with comparison to the current state-of-art SNF method.

2 Materials and Methods

2.1 Study Population

For breast cancer prognosis, we downloaded breast cancer (BRCA) samples from METABRIC data set as the training dataset to predict the breast cancer prognosis using our integration pipeline. We used the normalized data available from the Synapse repository: https://www.synapse.org/#!Synapse:syn1688369. METABRIC data set consist 1981 breast cancer samples. For each of these patients, matched DNA and RNA were extracted from each primary tumor specimen. Subjected copy-number and genotype analysis and transcriptional profiling were separately performed on the Affymetrix SNP 6.0 platform and the Illumina HT-12 v3 platform (Illumina-Human-WG-v3). High-quality follow-up clinical data including disease-free survival information are also available for the 1981 samples.

We also included four publicly available datasets of breast cancer samples from National Center for Biotechnology Information (NCBI) Gene Expression Omnibus (GEO) GSE4922(Anna V Ivshina et al., 2006), GSE1456(Yudi Pawitan et al., 2005), GSE3494(Lance D Miller et al., 2005) and GSE7390(Christine Desmedt et al., 2007). All four data sets are based on Affymetrix HG-U133A microarray platform, and have relapse-free survival information, as shown in Table 1.

2.2 Features transformation

For a given training dataset, we transformed the original omic features into either pathway expression, using the Pathifier algorithm, or using a normalization approach based on the Pearson correlation distance.

2.2.1 Features transformation using pathway dysregulation scores

To normalize the omic features of a training set, we retrieved the pathway information from broad institute GSEA (http://www.broadinstitute.org/gsea) curated gene sets which include a total of 403 pathways from KEGG (186 pathways) and Biocarta (217 pathways, http://www.biocarta.com)(Minoru Kanehisa & Susumu Goto, 2000; Nishimura, 2001b). We used R package Pathifier to perform pathway-based

transformation on multi-omics data(Yotam Drier, Michal Sheffer, & Eytan Domany, 2013). The details of the usage of Pathifier algorithm is described elsewhere, with applications on transcriptomics level and metabolomics level(Huang et al., 2016; Sijia Huang et al., 2014). This algorithm aims to summarize and transform information from gene level to pathway level, inferring individualized pathway deregulation scores (PDS) for each pathway. The PDS score basically measures the deviation from the normal status in each pathway. A principal curve is built summarizing the local average through a p-dimensional principal component dimension cluster of samples with smoothing procedures(Hastie & Stuetzle, 1989). Each sample point in the p-dimensional principal component dimension is projected onto the principal curve. The pathway deregulation score is calculated as the normalized distance of the sample's projected point to the normal centroid point on the curve. Basically, if one sample is more distant to the other samples in one specific pathway, the distance of the projected point to the normal centroid is greater and leads to a higher PDS score for this sample in this pathway. We transformed each test datasets independently but also fit the PDS model on a specific test set

2.2.2 Feature transformation using normalization

As an alternative approach to pathway expression, we normalized our features for a given training set and for each omic using the following procedure: For each sample, we inversely ranked the features according to their raw expression and used this rank as a measurement. We then normed these rank between 0 and 1 with division by the number of features. Thus the feature with the highest expression had the score 1, the second feature had the score 1 * (m-1) / m, with m the number of features, and so on. Then, in a second time we computed the per-sample Pearson correlation distance and obtained a squared matrix of size n, with n the number of training samples. Finally in a third and last time, we once again inversely ranked the n features of each sample, used these rank as new features, and normed these new features between 0 and 1. Thus, the final training matrix is a square matrix of size n. To normalize a new sample, with selected the common subset of features between the sample and the training set. We then applied the same procedure described for the training samples by first used the inverse rank normalization on both the new sample

features and for all the samples from the training set. We computed the Pearson distance between the new sample and all the training samples and used once again the inverse norm rank on the new sample. Thus, the new sample will be a vector of size n, with n the number of training samples.

2.3 Automatic inference of cancer subtypes and prediction for new sample

The DeepProg pipeline is a semi-supervised approach which consists in first inferring the cancer subtypes of a training dataset and in a second time building a supervised model using the labels inferred. Moreover, rather than constructing only one model, we used a boosting-like approach and built a model for different random splits of the training set. Finally, we inferred the final cancer subtypes by combining the results of all the instances.

2.3.1 Unsupervised identification of cancer subtypes

Given a multi-omics dataset, we used an autoencoder for each omic layer, to transform the original features (either pathway features or transformed omic features), to new abstract features. We then searched amongst these new features those significantly associated to survival. Finally, we stacked these survival features together and used them to perform a clustering analysis and identify the cancer subtypes.

2.3.1.1 Construction of the autoencoders

An autoencoder is a function f(v) = v' that reconstruct the original input vector v composed of mm features through multiple nonlinear transformations of its features, and such that size(v) = size(v') = m. For each omic, we created an autoencoder with one hidden layer of size h that corresponds to the following equation:

$$f(v) = \tanh(W'.s(W.v + b) + b')$$

W' and W are two weight matrices of size h x m and m x h, respectively, and b and b' are two bias vectors of size h and h'. Finally, tanh is a nonlinear, element-wise activation function defined as f(x) = (exp(x) - exp(-x)) / (exp(x) + exp(-x)). To train our autoencoders, we searched the optimal W*, W`*, b* and b'* that minimizes the logloss function.

We used the python (2.7) Keras package (1.2.2) with theano as tensor library, to build our autoencoders.

We constructed autoencoders with one hidden layer having h=100 and used the adam optimization algorithm to find W*, W`*, b* and b'*. We trained our autoencoder on 10 epochs and introduced 50% of dropout (i.e. 50% of the coefficients from W and W' will be randomly set to 0) at each training iteration.

2.3.1.2 Selection of new features linked to survival

For each omic, we searched amongst the 100 new features produced by the autoencoder, those significantly linked to the survival. For each new feature, we built an univariate Cox-PH model using the R package survival and identified features having a log-rank p-value (Wilcoxon test) < 0.01. Patients without relapse status events during the study time were considered censored. Moreover, we used Kaplan Meier curves to present the survival outcome of each classified/predicted group. We also implemented C-index to measure the performance of the prognosis prediction(Harrell et al., 1996). All survival analysis was conducted using the R package Survival(Therneau & Grambsch, 2000).

We finally extracted all the significant new features and stacked them together to form a new matrix Z of size n x hs. Here, n corresponds to the number of samples and hs to the total number of significant features in all the omics.

2.3.1.3 Cancer subtype detection

We then used a gaussian mixture to partition Z into clusters that we defined as the subtypes. We used the GaussianMixture function from the scikit-learn package with 1000 iterations, 100 initiations and a diagonal covariance matrix. Then, we sorted the clusters according to their median survival. Thus, the cluster labelled as "0" has the overall lowest survival while the last cluster "N" has the overall highest survival.

2.3.2 Supervised cancer subtype assignment

We used the cluster labels inferred by the gaussian mixture procedure to build several supervised models that can assign a label for any new sample having at least a subset of features in common with the features from the training set.

2.3.2.1 Supervised model construction

We computed a Kruskal-Wallis test for each omic and for each feature, in order to detect the most discriminative features with respect to the cluster labels. Then, we selected the 10 most discriminative features for each omic and stacked them together to form a new training matrix M. Finally, we used the Support Vector Machine (SVM) algorithm to construct a predictive model using M as input and the cluster labels as targets. To find the best hyper-parameters of the classifier, we performed a grid-search amongst a set of hyper-parameter values, using a 5-fold cross-validation on M, and searched to minimize the errors of the test folds. The different hyper-parameters tested are summarized in Suppl. Table XXX. We used the SVC module of the python scikit-learn library to build the SVM model. Finally, we used the predict_proba function from the SVC module to infer the label probability, which implements the Wu et. al. method (Wu, Lin, & Weng, 2004).

2.3.2.2 Label assignment for a new sample

To assign the label of new sample having only a subset of common features with the training set, we used the following procedure: We first transformed the features of the new sample into pathway features (see M&M 2.3.1), or into normalized feature (See M&M 2.3.2). In the case of different omics between the new sample and the training sample, we selected the omics having common features and identified the 10 most discriminative features using a Kruskal-Wallis test. Then, we used the procedure described above (M&M 2.4.2.1) to rebuild a classifier and predict the label.

2.3.3 Boosting procedure

Rather than building only one instance of the models described in M&M 2.4.1 and 2.4.2, we constructed several models using each time a random fraction of the total number of training samples. We constructed several instances of the clustering procedure (M&M 2.4.1) and the supervised model (M&M 2.4.2) selecting randomly, for each instance, 80% of the whole samples. We eliminated the instances for which we didn't obtain any new features linked to survival,or obtained only one cluster or having cluster labels not significantly linked with survival (log-rank p-value > 0.05). For a given sample, the probability to belong to a particular cancer subtype is the average of the probabilities given by all the instances. Using

this approach, we estimated the labels of all the samples from the training set or any new sample.

2.3.4 Number of clusters selection

To decide the number of clusters, we trained our training dataset with different number of clusters (c=2, 3, 4, 5) with the same seed and compared their performances on training data. We used different metrics including log-rank test p-values, C-indexes and Silhouette scores to evaluate and determine the best number of clusters.

2.4 Evaluation metrics

We used several metrics to evaluate the performances of the survival models.

2.4.1 Log-rank p-values of Cox-PH regression

For each experiment and for a given dataset, we used the labels inferred by DeepProg together with survival data to build a univariate Cox-PH model. We then computed the log-rank p-value of this model which tests the null hypothesis that all the coefficients of the Cox-PH model are zero. Moreover, instead of using the cancer subtype directly labels (discrete values), we considered the probability to belong to the cluster with the lowest survival median. We used this probability as univariate feature to construct a Cox-PH model and compute the log-rank p-value of the model.

2.4.2 C-indexes

We also computed C-indexes as a measurement of our model in training set and different validation sets(Triantaphyllou, 2000). C-index helps to measure a probability that given a pair of samples, the one with a higher predicted risk will experience the event earlier compared to the other low risk sample.

2.4.3 Clusters consistency

We also used adjusted rand index to measure the similarity of clustering during each boosting process(Rand, 1971). Rand index equals 0 means there is no consistency among clustering of different boosting iterations, whereas rand index equals 1 means the clustering results are exactly the same within different iterations. We used the adjusted rand indexes to correct for chance, and the adjusted rand index can be negative when

the observed consistency is lower than the expected consistency.

2.5 Survival analysis

We used survival analysis in both the feature selection step and the model evaluation step, with the comparison of relapse free survival in different groups. Patients without relapse status events during the study time were considered censored. We used the Cox-PH model to associate the risk of relapse with the transformed autoencoder features and predicted subgroups. Cox-PH (proportional hazards) model is a semi-parametric model. The nonparametric part comes from a time-dependent baseline hazard function where no parametric assumptions is made. For the parametric part, each feature (covariate) is multiplicatively related to the hazard. Assuming we selected p features from auto-encoders to be highly correlated with the risk of relapse in breast cancer. For sample J we have $X^J = (X_1^J, X_2^J, X_3^J, ..., X_p^J)'$. In the Cox-PH model, we estimate the relationship between the risk and features as:

$$h(t \mid \mathbf{X}) = \mathrm{h}_0(t) \exp(\boldsymbol{\beta}' \mathbf{X})$$

Here $h_0(t)$ is the baseline hazard only depending on time. Regarding the hazard ratio (HR) between two selected transformed pathway features $\boldsymbol{X_m}$ and $\boldsymbol{X_n}$, we have:

$$\frac{h(t \mid \boldsymbol{X_m})}{h(t \mid \boldsymbol{X_n})} = \exp(\boldsymbol{\beta}'(\mathbf{X_m} - \mathbf{X_n}))$$

The relative hazard between any two features is constant over time and depends only on the difference of the values of features.

3 Results

3.1 The workflow of DeepProg

After transforming the input features (eg. gene, CNV) into pathway features, we used our DeepProg pipeline (Figure 1) to infer the cancer subtypes based on the survival information. DeepProg is basically constituted in three steps: (1) Using a set of training samples, it first combines autoencoders with Cox-PH

models to transform the input features into new features and selects those linked to survival. Then, using these survival features, DeepProg infers the survival subtypes using a clustering procedure. (2) In a second step, DeepProg processes the labels inferred in step 1 and the input features to build a supervised model to assign a label and its probability to any new sample. If a new label doesn't share the same omics as the training samples or if it only have a subset of common features, then DeepProg rebuilds a supervised model and store it. (3) DeepProg adopts a bagging approach by constructing a series of sub-models from random subsets of the original training samples. The bagging procedure helps increase the overall robustness of the model by tackling the randomness during the construction of individual autoencoders or supervised models. Moreover, when increasing the number of sub-models, the bagging procedure helps to produce better results in term of the p-value and C-index.

3.1 Two survival subgroups are detected in METABRIC multi-omics breast cancer data

To determine the best number of clusters from the pathway based DeepProg model, we compared log-rank p-value, C-index and the rand index from the training set: METABRIC multi-omics breast cancer data set (Curtis et al., 2012). The results are shown in Table 2. It is very clear that when cluster is set to 2, the prediction model performs the best (log rank p-value = 3.65e-20, C-index = 0.710) in the training data set. When the number of clusters is set to be 3 or higher, the output p-values in separating the training data set is less significant and the consistency is lower. Based on these observations, we decided to use the 2-cluster prediction model to examine all the validation datasets.

3.3 The survival subgroups are consistently validated in 4 independent cohorts

We evaluated the performances of our training model on 4 transcriptomic based microarray validation datasets: GSE4922(Anna V Ivshina et al., 2006), GSE1456(Yudi Pawitan et al., 2005), GSE3494(Lance D Miller et al., 2005) and GSE7390(Christine Desmedt et al., 2007). According to Figure 2, METABRIC pathway based model predicts two classes in training dataset with a separation log-rank p-value of 3.65e-

20 and C-index of 0.710 (Figure 2B). The predictive performances are expected to drop in the testing data sets, since they have different patient populations and clinical characteristics from the training set (Table 1). Nevertheless, the model yields very decent predictive results with the p-value of 1.91e-5 in Anna dataset (Figure 3B), 1.22e-3 in Miller dataset (Figure 3D), 1.9e-6 in Pawitan dataset (Figure 3F) and 1.03e-2 in Desmedt dataset (Figure 3H). The C-index results in the validation sets gives C-index of 0.683 (Anna), 0.725 (Miller), 0.876 (Pawitan) and 0.570 (Desmedt), consistent with the results in Kaplan-Meier curves (Figure 3).

3.4 The pathway-based integration is better than original-data based integration

To evaluate the usefulness of pathway transformation normalization, we compared the results obtained with other types of data normalization on the original data. We evaluated 14 normalization procedures on the BRCA training set and reported the p-values of the inferred cluster labels and cluster probabilities. Overall, the rank-correlation-rank normalization (see M&M 2.3.2) presents the best performances. However, when applied on BRCA datasets, the pathway normalization proved to be much more efficient to highlight significant survival subtypes for the training dataset and all the validation datasets (Figure 2 and Figure 3). The results of Kaplan-Meier survival curves and C-indexes based on classification all consistently show that pathway-based genomic models are superior to the gene-based models (Figure 2 and Figure 3). For example, in the training set pathway based model predicts two clusters with log-rank p-value 3.65e-20 (Figure 2B), whereas the original data based normalization gives a log-rank p-value 3.33e-16 with a crossover in the Kaplan Meier curves (Figure 2A).The C-index for pathway based normalization is 0.710 compared to that of 0.674 in the original data based normalization approach.

Among the transcriptomics based testing sets, the pathway-based normalization is also validated to improve the prediction of separate clusters compared to the original level normalization. In Anna data set, the log-rank p-value is 1.91e-5 for the pathway-based model, compared to that of 9.72e-4 for the gene-based model (Figure 3A and Figure 3B). In the Miller data set, the p-value of the pathway-based model is also more significant than that of gene-based model (1.22e-3 vs.2.67e-2, Figure 3C and 3D). In the Pawitan data set,

again the prognosis prediction of the pathway-based model is more significant than that of gene-based model (1.9e-6 vs. 7.74e-6, Figure 3E and 3F). In the Desmedt data set, with the fact that all the samples are lymph node negative, which is very different from other datasets, the p-value of the pathway-based model is still significant (p=1.03e-2), compared to the non-significant result of the gene-based model (0.192, Figure 3G and 3H). The C-index comparisons between pathway vs. original-data based models on these data sets are: 0.683 vs. 0.632 in Anna data, 0.725 vs. 0.720 in Miller data, 0.876 vs. 0.725 in Pawitan data and 0.570 vs. 0.558 in Desmedt data.

3.5 The DeepProg methodology outperforms alternative data integration approaches

Similarity network fusion (SNF) is a state-of-art genomics data integration method, which constructed patient-patient similarity networks from each omics and then efficiently fuses the networks into one that represents the consensus underlying structure(Wang et al., 2014). This method has been proved to outperform a variety of integration methods including iCluster. We downloaded the 5 benchmark TCGA datasets from SNF paper including glioblastoma (211 samples), breast cancer (105 samples), colon cancer (92 samples), kidney cancer (122 samples) and lung cancer (106 samples). The details of the datasets are described in the SNF work.

We performed comparison of our *DeepProg* framework and SNF on the 5 benchmark datasets. We trained *DeepProg* in each dataset with 5-fold cross validation and decide the best number of clusters based on certain criteria including C-index, log-rank p-value and adjusted rand-index (Supplementary Figure 1-5, Supplementary Table 1-5). Then we compared the DeepProg predicted clusters to those from SNF in Figure 4 and Supplementary Figure 6. DeepProg achieved significance in survival analysis compared to SNF. In glioblastoma dataset, SNF predicts 3 clusters with p-value 3.87e-3. DeepProg predicts 2 clusters with p-value 1.3e-5, and it is clear from the Kaplan Meier curve that two cluster prediction is superior (Figure 4A and Figure 4B). In breast cancer dataset, DeepProg gives a prediction of 2 classes with p-value 1.37e-4, whereas SNF predicts 5 classes with p-value 1.35e-3, and their subclasses is too small with cross-overs in

Kaplan Meier curves (Figure 4C and Figure 4D). In colon cancer dataset, DeepProg and SNF both gives a prediction model of 3 clusters but DeepProg outperforms SNF with a more significant p-value 1.75e-4 compared to 3.6e-2 (Supplementary Figure 5A and 5B). In kidney cancer dataset, DeepProg outputs a 4 cluster model with p-value 1.86e-5, whereas SNF outputs a 3 cluster model with p-value 3.11e-3 (Supplementary Figure 5C and 5D). Finally, in lung cancer dataset, DeepProg predicts a 2 cluster model with p-value 2.09e-4. However, SNF predicts a 4 cluster model with p-value 1.78e-2 and the survival curves cross-over among 3 clusters at the beginning of the Kaplan-Meier curve, which indicates bad survival grouping (Supplementary Figure 5E and 5F).

4 Discussion

4.1 Pathway transformation is an efficient normalization to predict BRCA survival subtype

Our results show that pathway transformation infers BRCA survival subtypes better, compared to any other standard normalizations at the original data level (without pathway transformation). The results confirm the assumption that higher-order representative features, such as Gene Ontology sets, KEGG pathways and other network modules, allows better prediction of patient survival, compared to the original level information (Gad Abraham, Adam Kowalczyk, Sherene Loi, Izhak Haviv, & Justin Zobel, 2010; Akker et al., 2014; Jelle J Goeman & Bühlmann, 2007; E. Lee, H.-Y. Chuang, J.-W. Kim, T. Ideker, & D. Lee, 2008; Shuangge Ma, Michael R Kosorok, Jian Huang, & Ying Dai, 2011; Fabien Reyal et al., 2008; Andrew E Teschendorff et al., 2010). However, unlike some other methods in this category, where individual pathway information is lost due to summarization or transformation, the pathway features proposed in this study explicitly measure the degrees of pathway dysregulation for cancer recurrence. As demonstrated by our previous work in transcriptomics data, pathway-based feature transformation uniformly performs better than gene based models in breast cancer prognosis(Sijia Huang et al., 2014).

4.2 Multi-omics integration framework provides flexible prediction

To demonstrate the robustness in predicting differential risks of relapse from the pathway-based integration model, we chose to train and test on independent study samples representing population heterogeneity. Despite population difference and much diversified testing data platforms, the method still achieved good performance on all four data sets of microarray platform where prognosis is particularly difficult to predict from RNA-Seq training data.

Another merit of our method is that it enables to predict future single-omics samples. Our prognosis model provide independent submodels for each omics data, thus greatly generalize our model to predict future samples with fewer features compared to the multi-omics data. We include four validation sets from single-level transcriptomics data, and the prognosis prediction is very significant (Figure 3). Thus it greatly reduces the cost of multi-omics model measurement for clinical practice.

4.3 TCGA breast cancer dataset issue

At the initiate stage of this project, we considered using The Cancer Genome Atlas dataset as the training dataset as it includes around 1000 patients with more than two omics levels of data. We downloaded The Cancer Genome Atlas data with gene expression (UNC IlluminaHiSeq_RNASeqV2; Level 3), methylation and copy number variation data (SNP 6.0) using R package TCGA-assembler (v2.0.1) on 01/31/2017(Zhu, Qiu, & Ji, 2014). 793 samples from the BRCA dataset have survival information. We divided the 793 samples into training set with 680 samples and testing set with 213 samples. The training set (680 samples) methylation data is generated from Infinium HumanMethylation27 BeadChip Kit and the testing set (213 samples) is generated from Infinium HumanMethylation450 BeadChip Kit.

However after we trained the TCGA model on the 680 sample training set, the original level integration model was found not predictive of any other dataset. We did many parallel comparison of our DeepProg pipeline to other integration tools on original level data (5 benchmark datasets from TCGA), all of those experiments showed that DeepProg is a decent tool to predict prognosis at the original data level. Those experiments led us to suspect the quality of TCGA Breast Cancer samples may not be as good as some

other cohorts we used. Supporting this, other researchers also proposed that directly usage of TCGA molecular data is not very predictive of cancer prognosis (Yuan Yuan et al., 2014).

4.4 Limitations of our work and conclusion

A major limitation of our approach is that we only used the information from genes that compose the 403 pathways that we considered, thus some gene-level information is unavoidably lost. In our case, over 4500 genes were enlisted in the pathways. On the other hand, we only considered the methylation sites and CNV regions that are mappable to these 4500 genes, which also leads to a reduced level of features from other omics. Therefore our model captures about 1/3 of the gene-level information overall. One can certainly use other curated gene sets, such as the MsigDB C2 gene sets or self-defined sets, to increase the coverage of the genes by the pathways. Future work will be done in expanding the current defined pathway features to detection of consensus network modules from different omics levels.

Another limitation of our work is that our pathway based pipeline is currently limited to the mapping relationship among the molecular features. The features are firstly mapped to genes, and then genes are mapped to pathways. Currently microRNA features are not included in the pathway pipeline, as each microRNA feature has potentially multiple targeted genes and therefore it is not straightforward to determine the pathway assignment for the these features. We plan to integrate some target-prediction methods with our pathway based pipeline to evaluate the contribution of microRNA information.

In conclusion, we propose a novel pathway-based deep-learning integration multi-omics framework to predict breast cancer prognosis. This pathway-based genomic model performs better than the original level based model. Moreover, we found that our deep-learning based integration method outperforms the current state-of-art method SNF in five benchmark datasets from TCGA. This framework is very flexible and allows to predict individual-omics measurement of patients in the future.

Appendix K: Chapter 6 Figures

Legends

Figure 1. Workflow of pathway based DeepProg integration for breast cancer data

A) Using a subset of the training set, the input omic features are transformed using pathway and autoencoders. New features linked to survival are selected. Then a clustering procedure is conducted to identify the subtypes. B) For each omic, a supervised classification model is constructed on the top of the inferred label by A, which allows classifying any new sample having common features. C) The final model is a consensus of several models from steps A and B, and constructed each time using a different subset of the training sample.

Figure 1

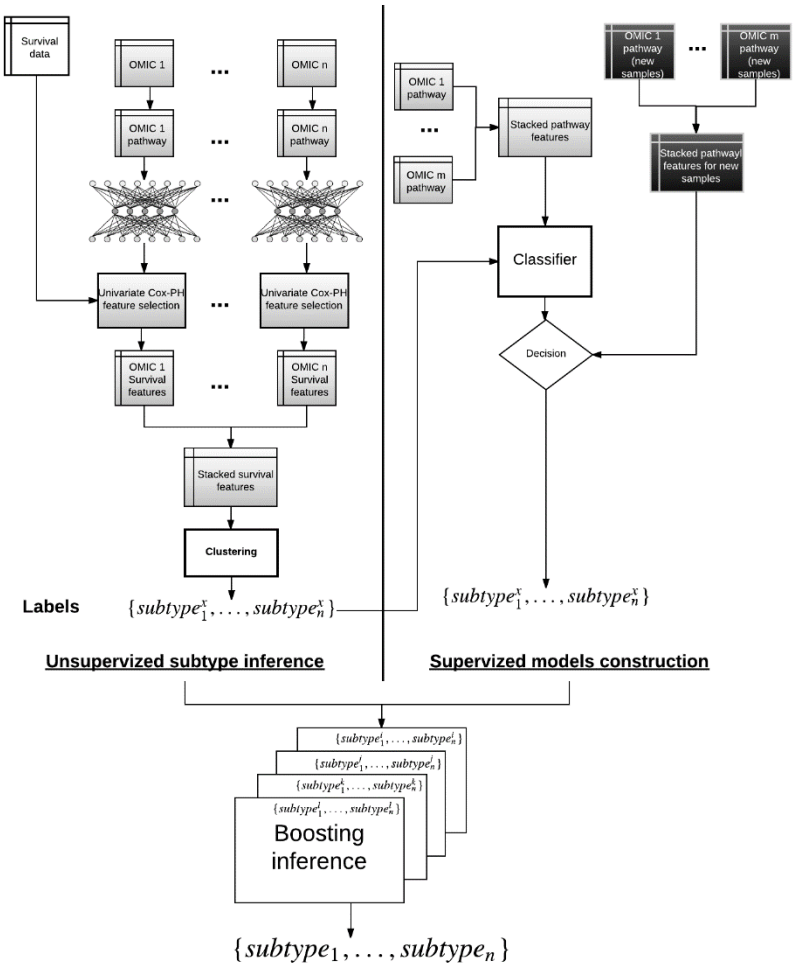


Figure 2. Comparing the prognosis performance between raw level and pathway based model in training dataset

The raw model and pathway model are independently built on METABRIC training dataset. The trained DeepProg model assigns training and validation multi-omics patients into higher risk and lower risk groups. The two groups are compared by Kaplan-Meier curves. P-values of the survival difference between the two groups are calculated using Wilcoxon log-rank tests and (+) denotes the censored observations in the study.

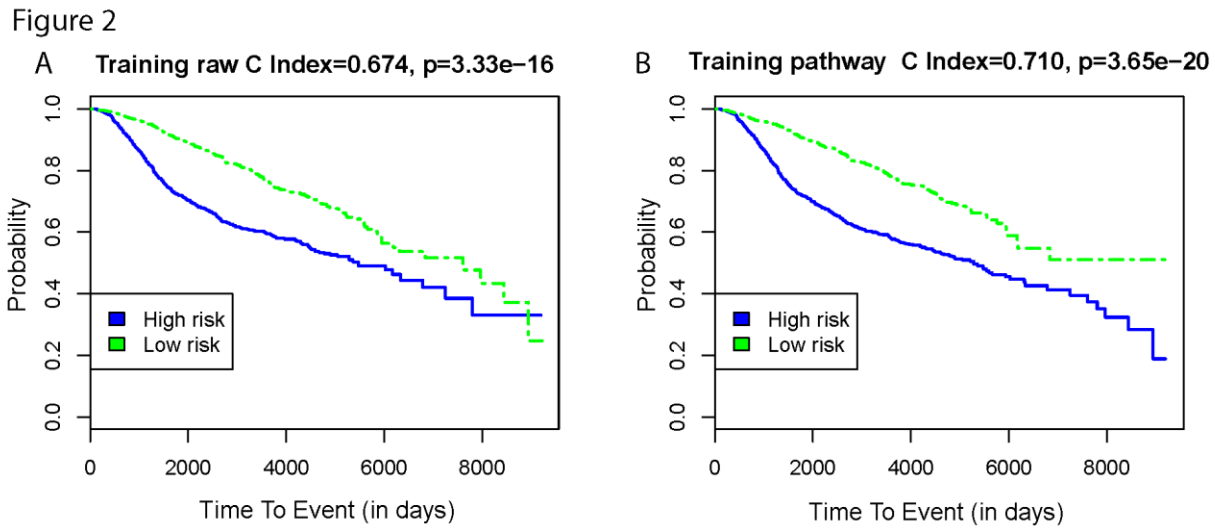


Figure 3. Comparing the prognosis performance between raw level and pathway based model in single-omics testing sets

The trained DeepProg model assigns the validation single-omics patients into higher risk and lower risk groups. The two groups are compared by Kaplan-Meier curves. P-values of the survival difference between the two groups are calculated using Wilcoxon log-rank tests and (+) denotes the censored observations in the study.

Figure 3

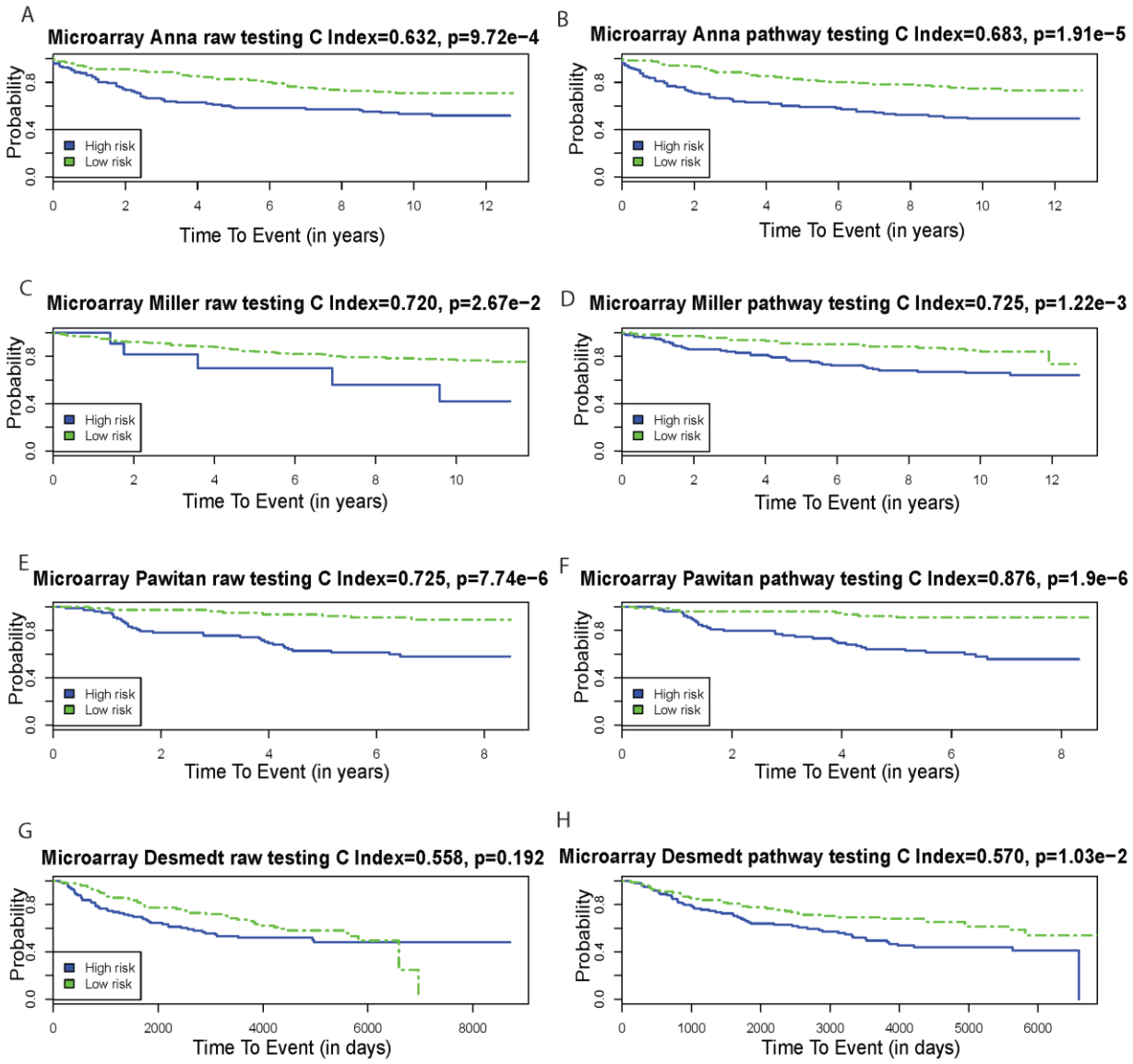
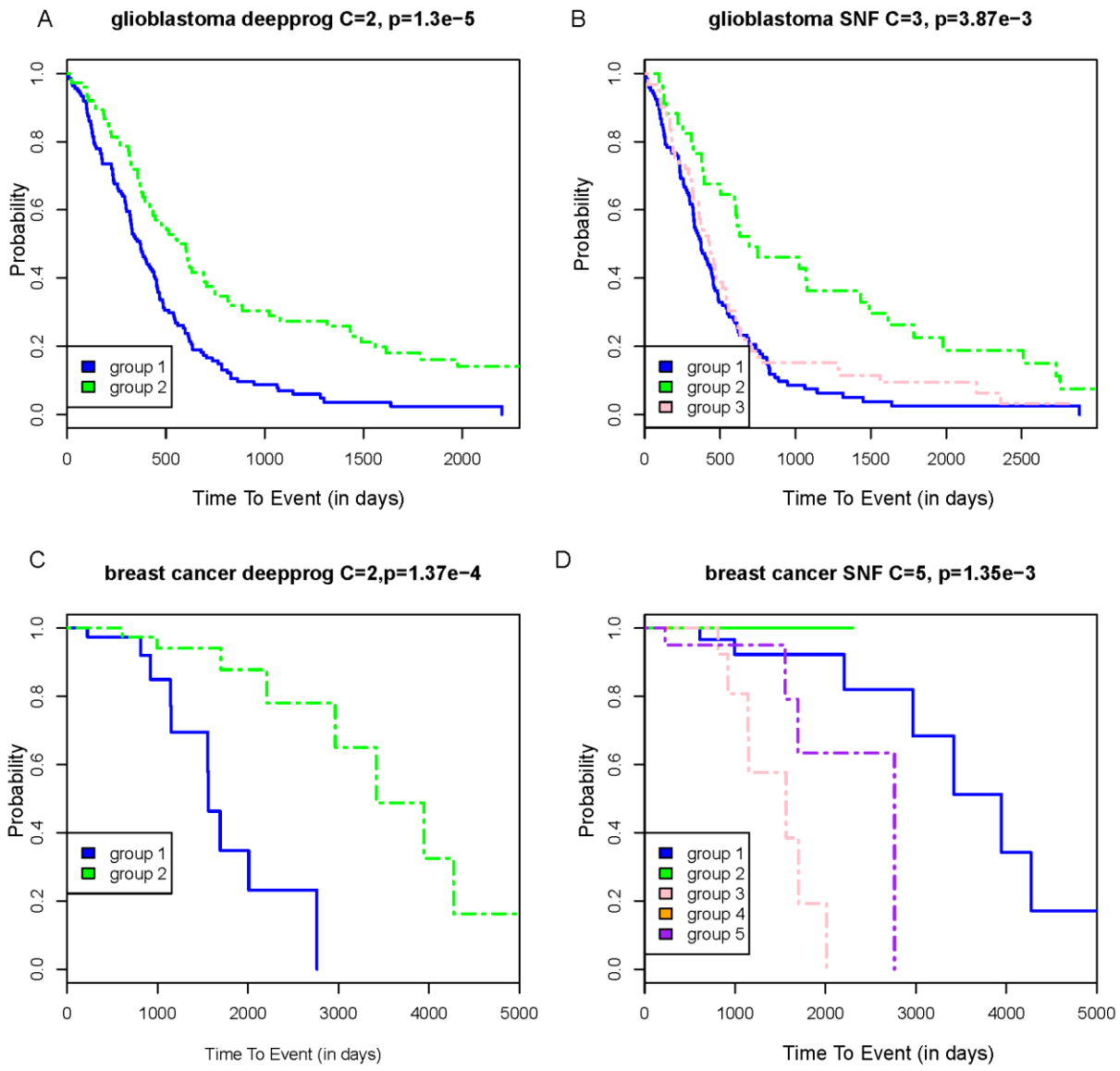


Figure 4. Comparing the prognosis performance between deepprog and SNF in two benchmark datasets from SNF: glioblastoma and breast cancer

For the TCGA glioblastoma dataset and breast cancer dataset, we built independent model by DeepProg and SNF to cluster the samples into two groups. The trained DeepProg/SNF model assigns the validation single-omics patients into higher risk and lower risk groups. The two groups are compared by Kaplan-Meier

curves. P-values of the survival difference between the two groups are calculated using Wilcoxon log-rank tests and (+) denotes the censored observations in the study.

Figure 4



Appendix L: Chapter 6 Tables

Table 1. Datasets summary: breast cancer

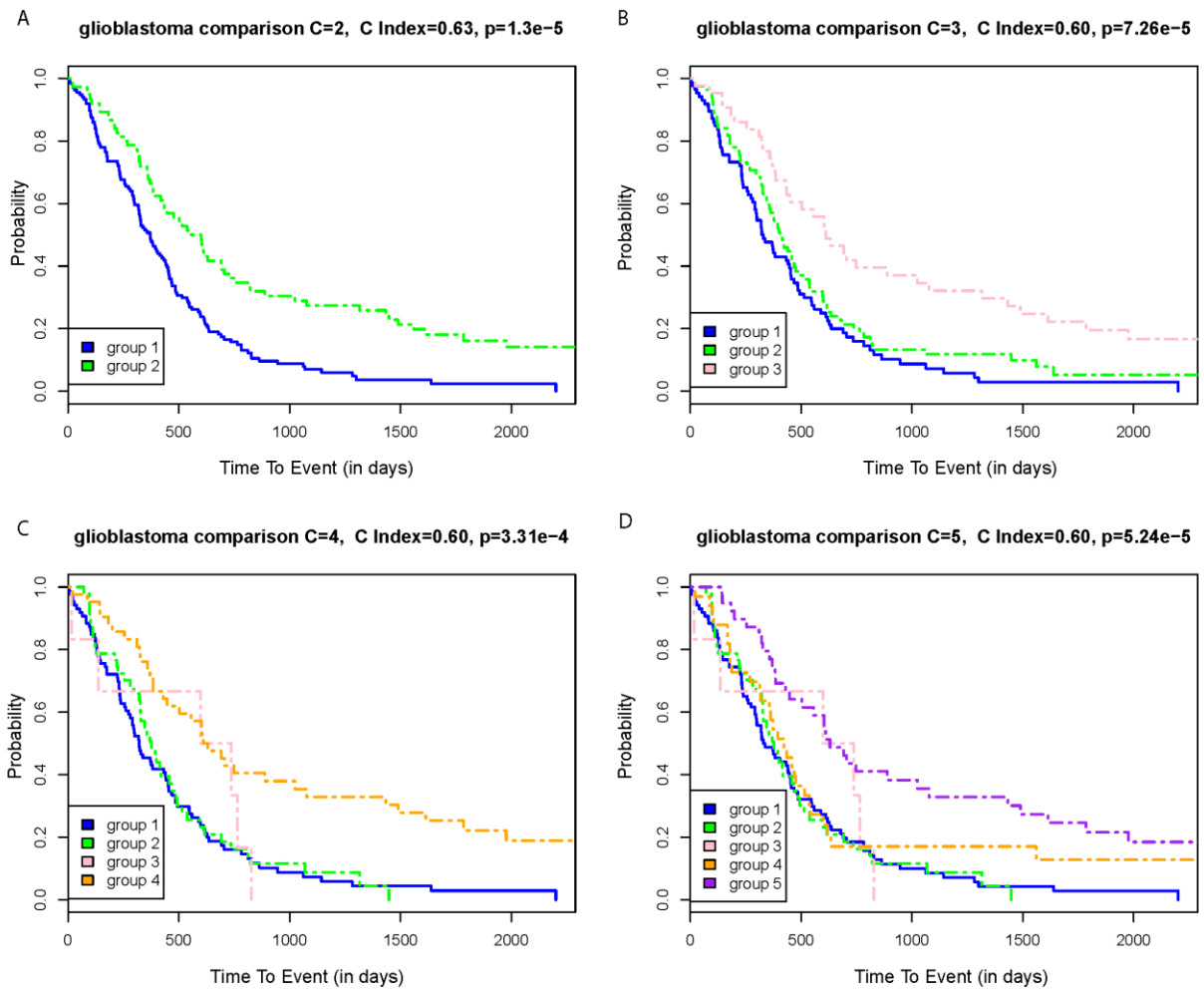
Dataset	METABRIC validation	Miller validation	Pawitan validation	Anna validation	Desmedt validation
Platforms					
	EXPR, CNV	EXPR (microarray)	EXPR (microarray)	EXPR (microarray)	EXPR (microarray)
# of Patients					
	1981	236	159	249	198
Relapse (%)					
Relapse	623 (31%)	55 (23%)	40 (25%)	89 (35%)	91 (46%)
Non-Relapse	1358 (69%)	181 (77%)	119 (75%)	160 (64%)	107 (54%)
Mean Relapse Free Survival (y)					
	8.085	8.167	5.959	7.142	9.312

Table 2. Selection of best cluster in METABRIC training dataset

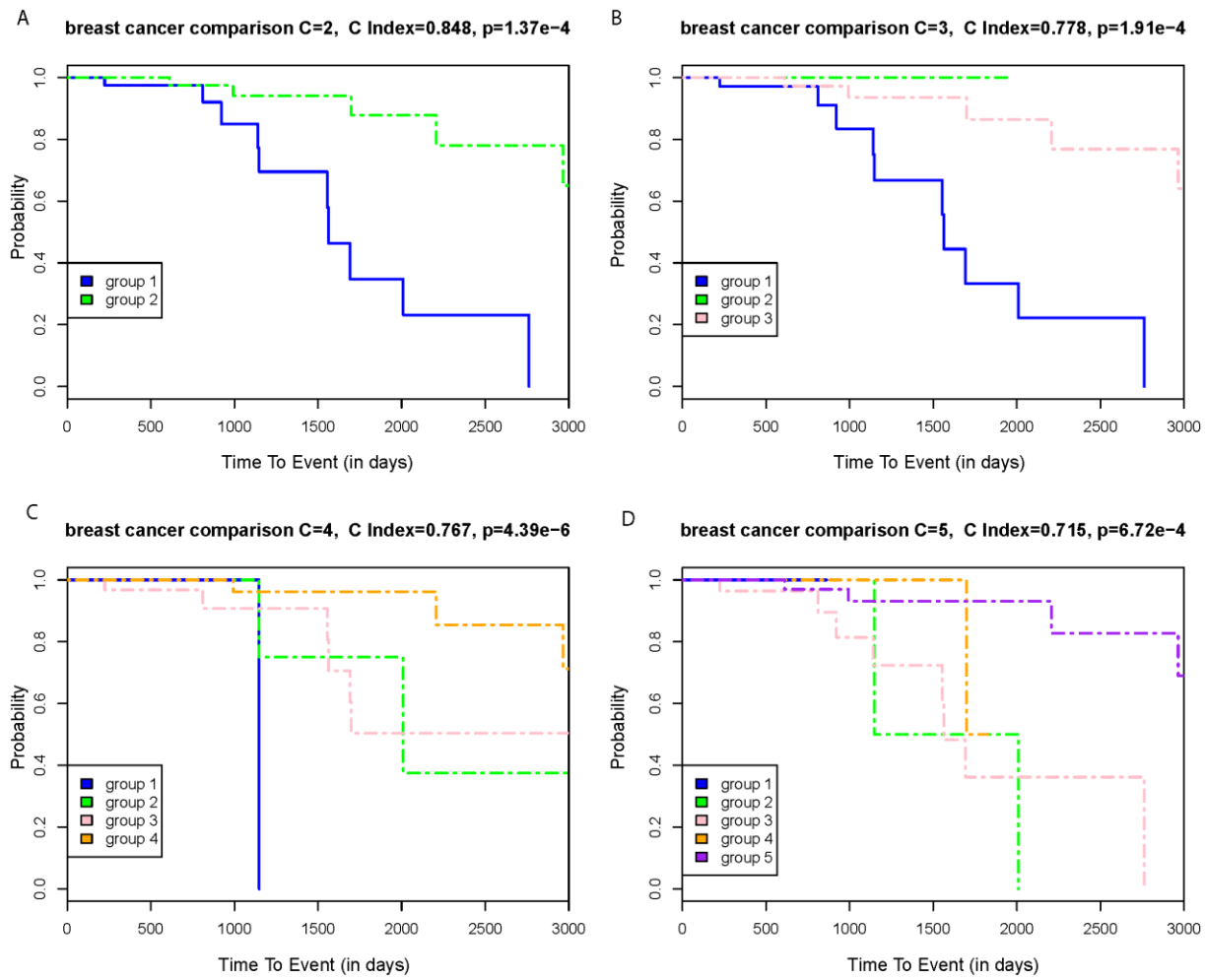
Metrics	C=2	C=3	C=4	C=5
Metabric training	3.65e-20	2.59e-14	2.46e-13	1.20e-9
C-index training	0.710	0.668	0.657	0.651
Adjusted Rand	0.031	0.021	0.021	0.097

Appendix M: Chapter 6 Supplementary Figures

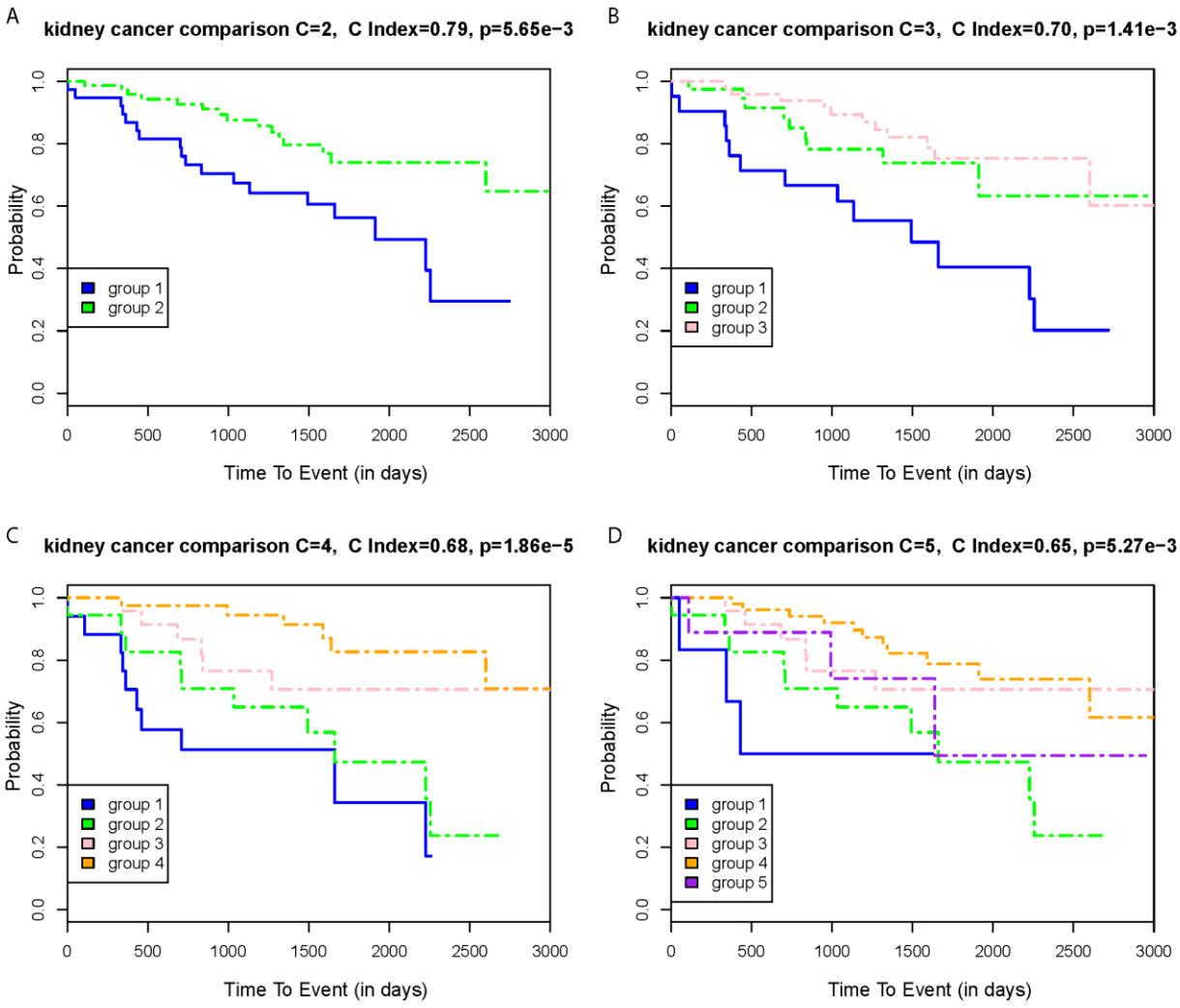
Supplementary Figure 1. Selection of clusters in glioblastoma dataset



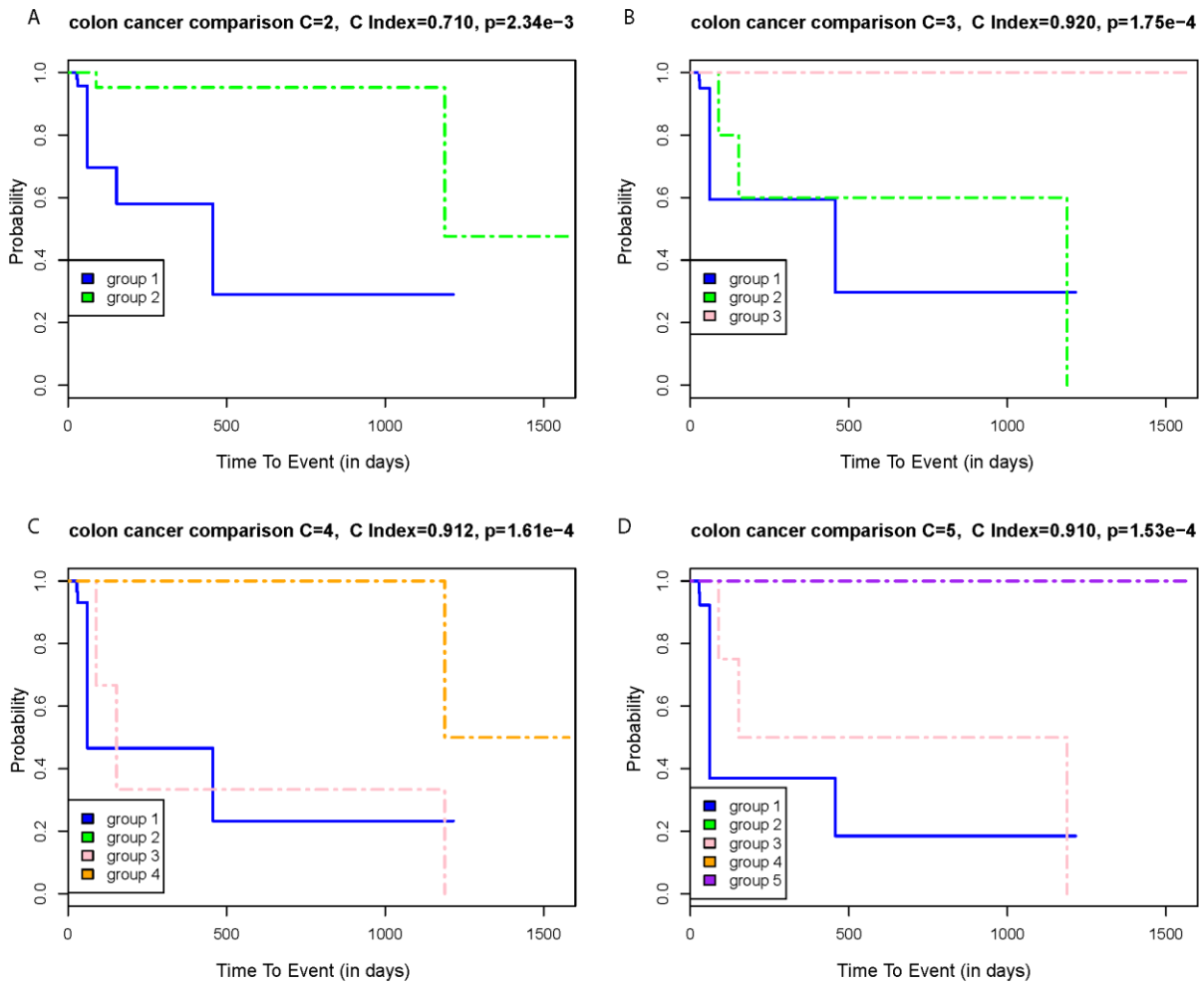
Supplementary Figure 2. Selection of clusters in breast cancer dataset



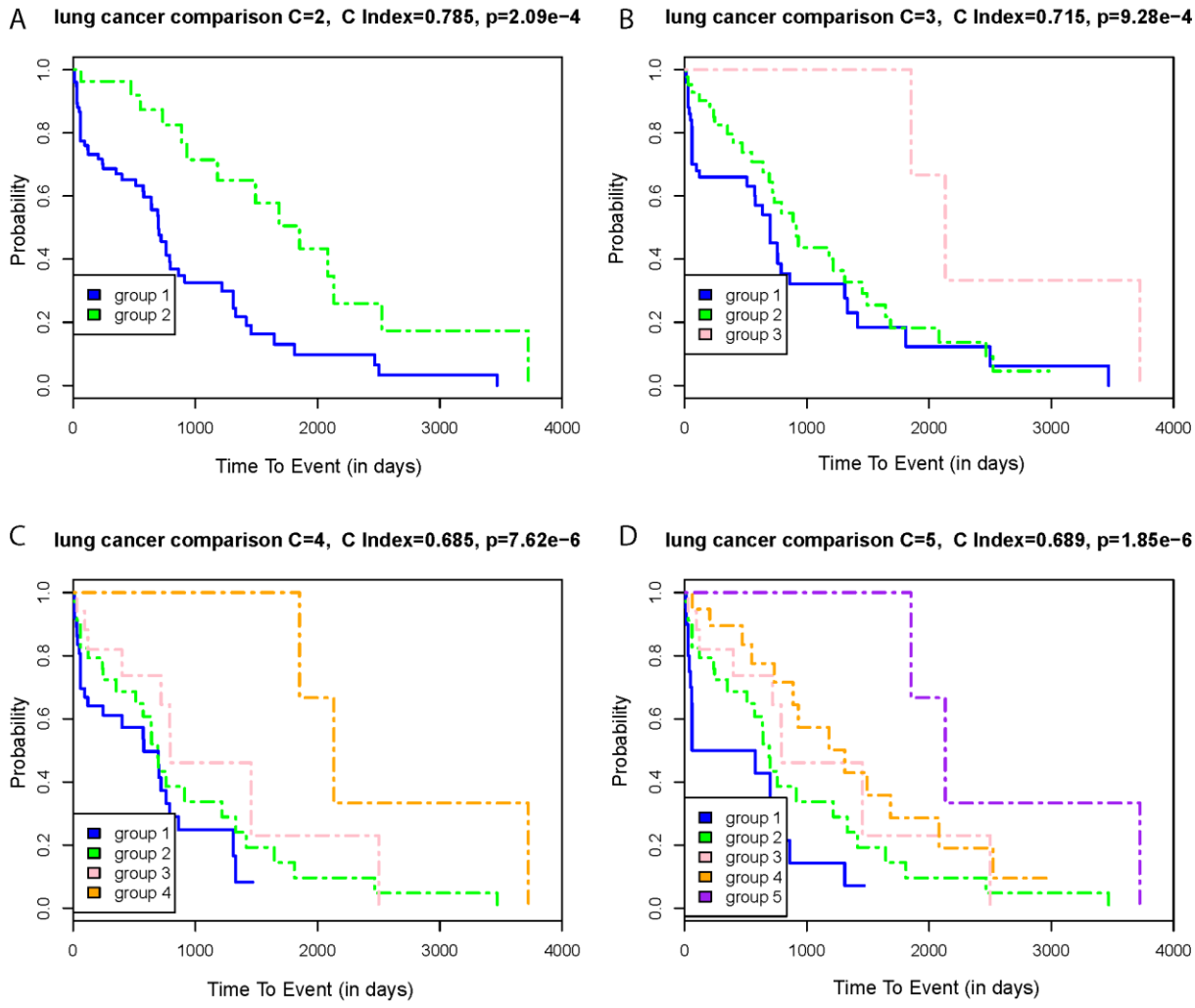
Supplementary Figure 3. Selection of clusters in kidney cancer dataset



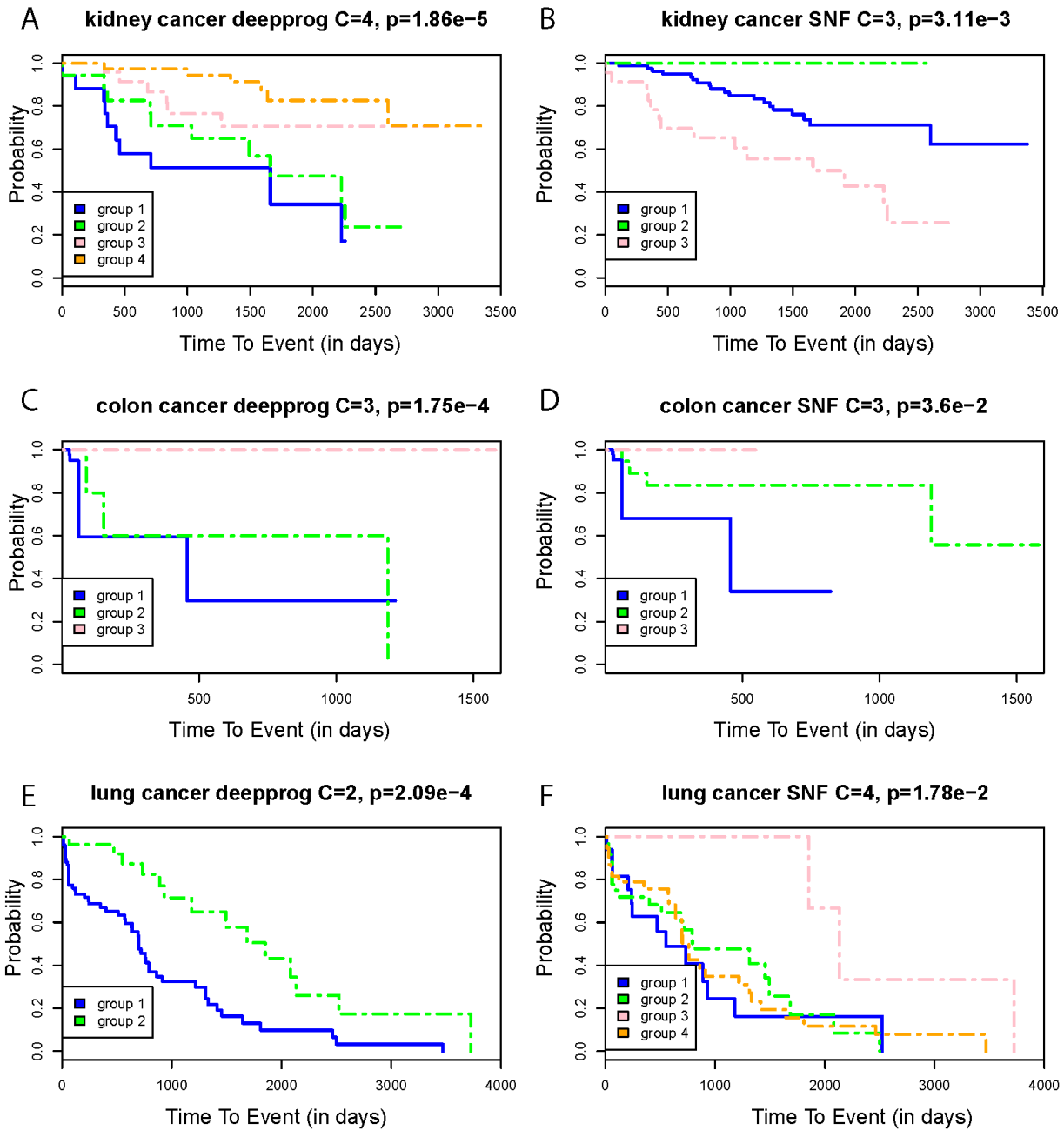
Supplementary Figure 4. Selection of clusters in colon cancer dataset



Supplementary Figure 5. Selection of clusters in lung cancer dataset


A
lung cancer comparison C=2, C Index=0.785, p=2.09e−4
Probability
Time To Event (in days)
group 1
group 2
B
lung cancer comparison C=3, C Index=0.715, p=9.28e−4
Probability
Time To Event (in days)
group 1
group 2
group 3
C
lung cancer comparison C=4, C Index=0.685, p=7.62e−6
Probability
Time To Event (in days)
group 1
group 2
group 3
group 4
D
lung cancer comparison C=5, C Index=0.689, p=1.85e−6
Probability
Time To Event (in days)
group 1
group 2
group 3
group 4
group 5

Supplementary Figure 6. Comparing the prognosis performance between deepprog and SNF in three benchmark datasets from SNF: kidney cancer, colon cancer and lung cancer



Appendix N: Chapter 6 Supplementary Tables

Supplementary Table 1. Glioblastoma cluster performance comparison

glioblastoma	C=2	C=3	C=4	C=5
p-value	1.30e-5	7.26e-5	3.31e-4	5.24e-5
C-index	0.625	0.603	0.597	0.604
Rand score	0.023	0.015	0.015	0.014

Supplementary Table 2. Breast cancer cluster performance comparison

Breast cancer	C=2	C=3	C=4	C=5
p-value	1.37e-4	1.91e-4	4.39e-6	6.72e-4
C-index	0.848	0.778	0.767	0.715
Rand score	0.180	0.122	0.182	0.098

Supplementary Table 3. Kidney cancer cluster performance comparison

Kidney cancer	C=2	C=3	C=4	C=5
p-value	5.65e-3	1.41e-3	1.86e-5	5.27e-3
C-index	0.738	0.699	0.674	0.655
Rand score	0.072	0.041	0.028	0.026

Supplementary Table 4. Colon cancer cluster performance comparison

Colon cancer	C=2	C=3	C=4	C=5
p-value	2.34e-3	1.75e-4	1.61e-4	1.53e-4
C-index	0.710	0.920	0.912	0.910
Rand score	0.425	0.437	0.337	0.311

Supplementary Table 5. Lung cancer cluster performance comparison

Lung cancer	C=2	C=3	C=4	C=5
p-value	2.09e-4	9.28e-4	7.62e-6	1.85e-6
C-index	0.785	0.715	0.685	0.689
Rand score	0.182	0.136	0.113	0.123

Chapter 7. Conclusions

Objectives

This project had three primary aims 1) To discover pathway-based metabolomic biomarkers for breast cancer diagnosis 2) To identify pathway-based transcriptomic biomarkers for breast cancer prognosis and 3) To extend to multi-omics data based on pathways and predict for breast cancer prognosis.

Completion of specific aims

I got interested in breast cancer research from a class project in epidemiology in University of Florida, which leads me to the current state. At the beginning on my doctoral project, I first researched the current state in breast cancer biomarkers discovery through next generation sequencing. For Aim I, I comprehensively compared the current identified metabolite biomarkers in breast cancer and found the result is so heterogeneous among different study groups. In the pathway-based metabolomics prediction paper (Chapter 2), I proposed a pathway-based diagnosis model which emphasizes on individualized pathway-based risk measurement using the pathway dysregulation score (PDS). I evaluated the performance of the pathway-based model compared to raw metabolite based model and found the pathway based model is more accurate and robust in breast cancer diagnosis prediction.

In Aim 2, to apply the pathway-based prediction approach on NGS data, we started to work on transcriptomics data to predict for the breast cancer prognosis (Chapter 3). Using several public datasets, I evaluated the performance of pathway-based prognosis model compared to gene-based model, and found that combining clinical information to the prediction model helps to differentiate the risk groups of breast cancer recurrence.

Following Aim 1 and Aim 2, as I found the integration of biomarkers and clinical traits helps to gain a better prediction, I did an investigation in bladder cancer looking into the combined benefit of gene biomarkers and demographic characteristics (Chapter 4). This experiment validates the value of looking into cancer subjects with multiple perspectives, which establish the direction of aim 3, multi-omics data.

For Aim 3 of this project, I firstly did a comprehensive collection of the current progress done in the field of multi-omics integration and computational tools developed so far (Chapter 5). I categorized the methods into unsupervised, supervised and semi-supervised subgroups. I compared the advantages and limitations for each method and found that there are few integration methods linking to survival analysis and making predictions of future dataset. The discovery leads me to the latest work for my project, based on multi-omics pathway-based data integration for breast cancer prognosis prediction (Chapter 6). I employ auto-encoders, a framework from deep-learning, to extract the information out from multi-layered pathway-based data from breast cancer and identify the survival-relevant features to classify the patients into subgroups. Those extracted and survival-relevant features will be used to predict for future samples.

Future work and directions

Overall, the proposed pathway-based approach is very promising to predict for breast cancer occurrence and recurrence both accurately and consistently. Regardless, more

Specifically, for Aim 1, we have received numerous feedbacks asking for a complete computational tool of our experiment. Further work will be building up a package performing the integrative analysis of metabolomics data, mapping the metabolites to pathways and predicting subgroups or survival.

To follow-up with Aim 2, the pathway-based model is trained and tested on gene expression data from the U133A platform. We suspect that direct application of the model to other platforms, such as RNA-Seq, is not desirable, and some additional re-processing work has to be done additionally. There is a need to compare the distribution of RNASeq data and microarray based data, and performing appropriate normalization methods to extract the critical information embedded within the data.

Aim 3 currently targets multi-omics on 3 levels: gene expression, methylation and copy number variation. We haven't looked into microRNAs because the mapping relationship between microRNAs and genes are very complex, leading to confusion of assigning microRNAs to pathways. Future work shall be using

model-based method to assign scores to the each microRNAs to different pathways and collectively compute the pathway dysregulation scores for miRNA-based data.

Another key limitation of our research is that the pathways in our research are already established and experimentally validated. Given the fact of heterogeneity of cancer development, there is a pressing need to identify novel modules or networks of biological entities (genes, metabolites etc.) for more personalized cancer occurrence and progression mechanism discovery. Through the incorporation of different biological entities dysregulated together, there is a new direction of subgrouping breast cancers and also for personalized therapeutics. Moreover, investigating the relationship of the effect of medications with pathway biomarkers in breast cancer will be another further step to go.

References

Aaboe, M., Marcussen, N., Jensen, K. M., Thykjaer, T., Dyrskjøt, L., & Ørntoft, T. (2005). Gene expression profiling of noninvasive primary urothelial tumours using microarrays. *British journal of cancer, 93*(10), 1182-1190.

Abraham, G., Kowalczyk, A., Loi, S., Haviv, I., & Zobel, J. (2010). Prediction of breast cancer prognosis using gene set statistics provides signature stability and biological context. *BMC bioinformatics, 11*(1), 277.

Abraham, G., Kowalczyk, A., Loi, S., Haviv, I., & Zobel, J. (2010). Prediction of breast cancer prognosis using gene set statistics provides signature stability and biological context. *BMC Bioinformatics, 11*, 277. doi:10.1186/1471-2105-11-277

Ades, F., Zardavas, D., Bozovic-Spasojevic, I., Pugliano, L., Fumagalli, D., De Azambuja, E., . . . Piccart, M. (2014). Luminal B breast cancer: molecular characterization, clinical management, and future perspectives. *Journal of clinical oncology, 32*(25), 2794-2803.

Akavia, U. D., Litvin, O., Kim, J., Sanchez-Garcia, F., Kotliar, D., Causton, H. C., . . . Pe'er, D. (2010). An integrated approach to uncover drivers of cancer. *Cell, 143*(6), 1005-1017.

Akker, E. B., Passtoors, W. M., Jansen, R., Zwet, E. W., Goeman, J. J., Hulsman, M., . . . Penninx, B. W. (2014). Meta‐analysis on blood transcriptomic studies identifies consistently coexpressed protein–protein interaction modules as robust markers of human aging. *Aging cell, 13*(2), 216-225.

American Cancer, S. (2003). Cancer facts and figures 2003. *Cancer facts and figures 2003*.

Asiago, V. M., Alvarado, L. Z., Shanaiah, N., Gowda, G. A., Owusu-Sarfo, K., Ballas, R. A., & Raftery, D. (2010). Early detection of recurrent breast cancer using metabolite profiling. *Cancer Res, 70*(21), 8309-8318. doi:10.1158/0008-5472.CAN-10-1319

Aure, M. R., Steinfeld, I., Baumbusch, L. O., Liestøl, K., Lipson, D., Nyberg, S., . . . Børresen-Dale, A.-L. (2013). Identifying in-trans process associated genes in breast cancer by integrated analysis of copy number and expression data. *PloS one, 8*(1), e53014.

Bartoletti, R., Cai, T., Dal Canto, M., Boddi, V., Nesi, G., & Piazzini, M. (2006). Multiplex polymerase chain reaction for microsatellite analysis of urine sediment cells: a rapid and inexpensive method for diagnosing and monitoring superficial transitional bladder cell carcinoma. *The Journal of urology, 175*(6), 2032-2037.

Berg, M., Vanaerschot, M., Jankevics, A., Cuypers, B., Breitling, R., & Dujardin, J. C. (2013). LC-MS metabolomics from study design to data-analysis - using a versatile pathogen as a test case. *Comput Struct Biotechnol J, 4*, e201301002. doi:10.5936/csbj.201301002

Bevers, T. B., Anderson, B. O., Bonaccio, E., Buys, S., Daly, M. B., Dempsey, P. J., . . . Harris, R. E. (2009). Breast cancer screening and diagnosis. *Journal of the National Comprehensive Cancer Network, 7*(10), 1060-1096.

Bild, A. H., Yao, G., Chang, J. T., Wang, Q., Potti, A., Chasse, D., . . . Nevins, J. R. (2006). Oncogenic pathway signatures in human cancers as a guide to targeted therapies. *Nature, 439*(7074), 353-357. doi:10.1038/nature04296

Blasco, H., Nadal-Desbarats, L., Pradat, P. F., Gordon, P. H., Antar, C., Veyrat-Durebex, C., . . . Corcia, P. (2014). Untargeted 1H-NMR metabolomics in CSF: toward a diagnostic biomarker for motor neuron disease. *Neurology, 82*(13), 1167-1174. doi:10.1212/WNL.0000000000000274

Blows, F. M., Driver, K. E., Schmidt, M. K., Broeks, A., Van Leeuwen, F. E., Wesseling, J., . . . Blomqvist, C. (2010). Subtyping of breast cancer by immunohistochemistry to investigate a relationship between subtype and short and long term survival: a collaborative analysis of data for 10,159 cases from 12 studies. *PLoS Med, 7*(5), e1000279.

Bolton, E. E., Wang, Y., Thiessen, P. A., & Bryant, S. H. (2008). PubChem: integrated platform of small molecules and biological activities. *Annual reports in computational chemistry, 4*, 217-241.

Bonnet, E., Calzone, L., & Michoel, T. (2015). Integrative multi-omics module network inference with Lemon-Tree. *PLoS Comput Biol, 11*(2), e1003983.

Borgan, E., Sitter, B., Lingjaerde, O. C., Johnsen, H., Lundgren, S., Bathen, T. F., . . . Gribbestad, I. S. (2010). Merging transcriptomics and metabolomics--advances in breast cancer profiling. *BMC Cancer, 10*, 628. doi:10.1186/1471-2407-10-628

Bossuyt, P. M., Reitsma, J. B., E Bruns, D., Gatsonis, C. A., Glasziou, P. P., Irwig, L. M., . . . De Vet, H. C. (2003). Towards complete and accurate reporting of studies of diagnostic accuracy: the STARD initiative. *Clinical chemistry and laboratory medicine, 41*(1), 68-73.

Brand, A., Leibfritz, D., Hamprecht, B., & Dringen, R. (1998). Metabolism of cysteine in astroglial cells: synthesis of hypotaurine and taurine. *J Neurochem, 71*(2), 827-832.

Bucak, M. N., Tuncer, P. B., Sariozkan, S., Ulutas, P. A., Coyan, K., Baspinar, N., & Ozkalp, B. (2009). Effects of hypotaurine, cysteamine and aminoacids solution on post-thaw microscopic and oxidative stress parameters of Angora goat semen. *Res Vet Sci, 87*(3), 468-472. doi:10.1016/j.rvsc.2009.04.014

Budczies, J., Pfitzner, B. M., Gyorffy, B., Winzer, K. J., Radke, C., Dietel, M., . . . Denkert, C. (2014). Glutamate enrichment as new diagnostic opportunity in breast cancer. *Int J Cancer*. doi:10.1002/ijc.29152

Bundix, F., & Wauters, H. (1997). The diagnostic value of macroscopic hematuria in diagnosing urological cancer. *Fam. Pract*, 1463-1468.

Cai, Z., Zhao, J. S., Li, J. J., Peng, D. N., Wang, X. Y., Chen, T. L., . . . Xie, D. (2010). A combined proteomics and metabolomics profiling of gastric cardia cancer reveals characteristic dysregulations in glucose metabolism. *Mol Cell Proteomics, 9*(12), 2617-2628. doi:10.1074/mcp.M110.000661

Cancer Genome Atlas, N. (2012). Comprehensive molecular portraits of human breast tumours. *Nature, 490*(7418), 61-70. doi:10.1038/nature11412

Carey, L. A., Cheang, M. C. U., & Perou, C. M. (2014). Genomics, prognosis, and therapeutic interventions *Diseases of the Breast: Fifth Edition*: Wolters Kluwer Health Adis (ESP).

Carey, L. A., Perou, C. M., Livasy, C. A., Dressler, L. G., Cowan, D., Conway, K., . . . Edmiston, S. (2006). Race, breast cancer subtypes, and survival in the Carolina Breast Cancer Study. *Jama, 295*(21), 2492-2502.

Carey, L. A., Perou, C. M., Livasy, C. A., Dressler, L. G., Cowan, D., Conway, K., . . . Millikan, R. C. (2006). Race, breast cancer subtypes, and survival in the Carolina Breast Cancer Study. *JAMA, 295*(21), 2492-2502. doi:10.1001/jama.295.21.2492

Chari, R., Coe, B. P., Vucic, E. A., Lockwood, W. W., & Lam, W. L. (2010). An integrative multi-dimensional genetic and epigenetic strategy to identify aberrant genes and pathways in cancer. *BMC systems biology, 4*(1), 67.

Chen, C.-L., Lin, T.-S., Tsai, C.-H., Wu, C.-C., Chung, T., Chien, K.-Y., . . . Chen, Y.-T. (2013). Identification of potential bladder cancer markers in urine by abundant-protein depletion coupled with quantitative proteomics. *Journal of proteomics, 85*, 28-43.

Chen, J., Bushman, F. D., Lewis, J. D., Wu, G. D., & Li, H. (2013). Structure-constrained sparse canonical correlation analysis with an application to microbiome data analysis. *Biostatistics, 14*(2), 244-258.

Chen, J., & Zhang, S. (2016). Integrative analysis for identifying joint modular patterns of gene-expression and drug-response data. *Bioinformatics, 32*(11), 1724-1732.

Chen, L.-M., Chang, M., Dai, Y., Chai, K. X., Dyrskjøt, L., Sanchez-Cabayo, M., . . . Jeronimo, C. (2014). External validation of a multiplex urinary protein panel for the detection of bladder

cancer in a multicenter cohort. *Cancer Epidemiology and Prevention Biomarkers*, cebp. 0029.2014.

Chia, S. K., Bramwell, V. H., Tu, D., Shepherd, L. E., Jiang, S., Vickery, T., . . . Nielsen, T. O. (2012). A 50-gene intrinsic subtype classifier for prognosis and prediction of benefit from adjuvant tamoxifen. *Clin Cancer Res, 18*(16), 4465-4472. doi:10.1158/1078-0432.CCR-12-0286

Chin, L., Andersen, J. N., & Futreal, P. A. (2011). Cancer genomics: from discovery science to personalized medicine. *Nature medicine, 17*(3), 297-303.

Cho, D.-Y., & Przytycka, T. M. (2013). Dissecting cancer heterogeneity with a probabilistic genotype-phenotype model. *Nucleic acids research*, gkt577.

Cline, M. S., Smoot, M., Cerami, E., Kuchinsky, A., Landys, N., Workman, C., . . . Gross, B. (2007). Integration of biological networks and gene expression data using Cytoscape. *Nature protocols, 2*(10), 2366-2382.

Cui, H., Darmanin, S., Natsuisaka, M., Kondo, T., Asaka, M., Shindoh, M., . . . Kobayashi, M. (2007). Enhanced expression of asparagine synthetase under glucose-deprived conditions protects pancreatic cancer cells from apoptosis induced by glucose deprivation and cisplatin. *Cancer Res, 67*(7), 3345-3355. doi:10.1158/0008-5472.CAN-06-2519

Cui, H., Zhou, C., Dai, X., Liang, Y., Paffenroth, R., & Korkin, D. (2017). Boosting Gene Expression Clustering with System-Wide Biological Information: A Robust Autoencoder Approach. *bioRxiv*. doi:10.1101/214122

Dai, X., Li, T., Bai, Z., Yang, Y., Liu, X., Zhan, J., & Shi, B. (2015). Breast cancer intrinsic subtype classification, clinical use and future trends. *American journal of cancer research, 5*(10), 2929.

Dang, C. V. (2010). Glutaminolysis: supplying carbon or nitrogen or both for cancer cells? *Cell Cycle, 9*(19), 3884-3886.

de Leoz, M. L., Young, L. J., An, H. J., Kronewitter, S. R., Kim, J., Miyamoto, S., . . . Lebrilla, C. B. (2011). High-mannose glycans are elevated during breast cancer progression. *Mol Cell Proteomics, 10*(1), M110 002717. doi:10.1074/mcp.M110.002717

Denkert, C., Bucher, E., Hilvo, M., Salek, R., Oresic, M., Griffin, J., . . . Fiehn, O. (2012). Metabolomics of human breast cancer: new approaches for tumor typing and biomarker discovery. *Genome Med, 4*(4), 37. doi:10.1186/gm336

Desman, G., Waintraub, C., & Zippin, J. H. (2014). Investigation of cAMP microdomains as a path to novel cancer diagnostics. *Biochim Biophys Acta, 1842*(12PB), 2636-2645. doi:10.1016/j.bbadis.2014.08.016

Desmedt, C., Piette, F., Loi, S., Wang, Y., Lallemand, F., Haibe-Kains, B., . . . d'Assignies, M. S. (2007). Strong time dependence of the 76-gene prognostic signature for node-negative breast cancer patients in the TRANSBIG multicenter independent validation series. *Clinical Cancer Research, 13*(11), 3207-3214.

Desmedt, C., Piette, F., Loi, S., Wang, Y., Lallemand, F., Haibe-Kains, B., . . . Consortium, T. (2007). Strong time dependence of the 76-gene prognostic signature for node-negative breast cancer patients in the TRANSBIG multicenter independent validation series. *Clin Cancer Res, 13*, 3207-3214. doi:10.1158/1078-0432.CCR-06-2765

Dowsett, M., Sestak, I., Lopez-Knowles, E., Sidhu, K., Dunbier, A. K., Cowens, J. W., . . . Cuzick, J. (2013). Comparison of PAM50 Risk of Recurrence Score With Onco type DX and IHC4 for Predicting Risk of Distant Recurrence After Endocrine Therapy. *Journal of clinical oncology, 31*(22), 2783-2790.

Drier, Y., Sheffer, M., & Domany, E. (2013). Pathway-based personalized analysis of cancer. *Proc Natl Acad Sci U S A, 110*(16), 6388-6393. doi:10.1073/pnas.1219651110

Drier, Y., Sheffer, M., & Domany, E. (2013). Pathway-based personalized analysis of cancer. *Proceedings of the National Academy of Sciences, 110*(16), 6388-6393.

Driggers, P. H., Segars, J. H., & Rubino, D. M. (2001). The proto-oncoprotein Brx activates estrogen receptor beta by a p38 mitogen-activated protein kinase pathway. *J Biol Chem, 276*(50), 46792-46797. doi:10.1074/jbc.M106927200

Edge, S. (2010). American joint committee on cancer., american cancer society., Teton data systems (firm). AJCC cancer staging handbook from the AJCC cancer staging manual.

Edwards, T. J., Dickinson, A. J., Natale, S., Gosling, J., & Mcgrath, J. S. (2006). A prospective analysis of the diagnostic yield resulting from the attendance of 4020 patients at a protocol‐driven haematuria clinic. *BJU international, 97*(2), 301-305.

Efron, B., & Tibshirani, R. (2007). On Testing the Significance of Sets of Genes. *Annals of Applied Statistics, 1*(1), 107-129. doi:Doi 10.1214/07-Aoas101

Ein-Dor, L., Kela, I., Getz, G., Givol, D., & Domany, E. (2004). Outcome signature genes in breast cancer: is there a unique set? *Bioinformatics, 21*(2), 171-178.

Elias, K., Svatek, R. S., Gupta, S., Ho, R., & Lotan, Y. (2010). High‐risk patients with hematuria are not evaluated according to guideline recommendations. *Cancer, 116*(12), 2954-2959.

Engelmann, D., & Pützer, B. M. (2012). The Dark Side of E2F1: In Transit beyond Apoptosis. *Cancer Research, 72*(3), 571-575. doi:10.1158/0008-5472.CAN-11-2575

Fan, C., Oh, D. S., Wessels, L., Weigelt, B., Nuyten, D. S., Nobel, A. B., . . . Perou, C. M. (2006). Concordance among gene-expression-based predictors for breast cancer. *N Engl J Med, 355*, 560-569. doi:10.1056/NEJMoa052933

Fan, C., Oh, D. S., Wessels, L., Weigelt, B., Nuyten, D. S., Nobel, A. B., . . . Perou, C. M. (2006). Concordance among gene-expression–based predictors for breast cancer. *New England Journal of Medicine, 355*(6), 560-569.

Fan, C., Prat, A., Parker, J. S., Liu, Y., Carey, L. A., Troester, M. A., & Perou, C. M. (2011). Building prognostic models for breast cancer patients using clinical variables and hundreds of gene expression signatures. *BMC Med Genomics, 4*, 3. doi:10.1186/1755-8794-4-3

10.1186/1755-8794-4-3.

Fan, P., Agboke, F. A., McDaniel, R. E., Sweeney, E. E., Zou, X., Creswell, K., & Jordan, V. C. (2014). Inhibition of c-Src blocks oestrogen-induced apoptosis and restores oestrogen-stimulated growth in long-term oestrogen-deprived breast cancer cells. *Eur J Cancer, 50*(2), 457-468. doi:10.1016/j.ejca.2013.10.001

Fan, Y., Murphy, T. B., Byrne, J. C., Brennan, L., Fitzpatrick, J. M., & Watson, R. W. (2011). Applying random forests to identify biomarker panels in serum 2D-DIGE data for the detection and staging of prostate cancer. *J Proteome Res, 10*(3), 1361-1373. doi:10.1021/pr1011069

Fiehn, O. (2002). Metabolomics--the link between genotypes and phenotypes. *Plant Mol Biol, 48*(1-2), 155-171.

Fong, M. Y., McDunn, J., & Kakar, S. S. (2011). Identification of metabolites in the normal ovary and their transformation in primary and metastatic ovarian cancer. *PLoS One, 6*(5), e19963. doi:10.1371/journal.pone.0019963

Fu, M., Maresh, E. L., Helguera, G., Kiyohara, M., Qin, Y., Ashki, N., . . . Wadehra, M. (2014). Rationale and pre-clinical efficacy of a novel anti-EMP2 antibody for the treatment of invasive breast cancer. *Mol Cancer Ther*. doi:10.1158/1535-7163.MCT-13-0199

Garcia, E., Andrews, C., Hua, J., Kim, H. L., Sukumaran, D. K., Szyperski, T., & Odunsi, K. (2011). Diagnosis of early stage ovarian cancer by 1H NMR metabonomics of serum explored by use of a microflow NMR probe. *J Proteome Res, 10*(4), 1765-1771. doi:10.1021/pr101050d

Garcia, M., Jemal, A., Ward, E., Center, M., Hao, Y., Siegel, R., & Thun, M. (2016). Global cancer facts & figures 2016. *Atlanta, GA: American cancer society, 1*(3), 52.

Gill, R. D. (1992). Multistate life-tables and regression models. *Math Popul Stud, 3*(4), 259-276. doi:10.1080/08898489209525345

Goeman, J. J. (2009). L1 penalized estimation in the Cox proportional hazards model. *Biom J, 52*, 70-84. doi:10.1002/bimj.200900028

10.1002/bimj.200900028.

Goeman, J. J., & Buhlmann, P. (2007). Analyzing gene expression data in terms of gene sets: methodological issues. *Bioinformatics, 23*(8), 980-987. doi:10.1093/bioinformatics/btm051

Goeman, J. J., & Bühlmann, P. (2007). Analyzing gene expression data in terms of gene sets: methodological issues. *Bioinformatics, 23*(8), 980-987.

Goodison, S., Chang, M., Dai, Y., Urquidi, V., & Rosser, C. J. (2012). A multi-analyte assay for the non-invasive detection of bladder cancer. *PloS one, 7*(10), e47469.

Gossai, D., & Lau-Cam, C. A. (2009). The effects of taurine, taurine homologs and hypotaurine on cell and membrane antioxidative system alterations caused by type 2 diabetes in rat erythrocytes. *Adv Exp Med Biol, 643*, 359-368. doi:10.1007/978-0-387-75681-3_37

Griffiths, T. L., & Ghahramani, Z. (2005). *Infinite latent feature models and the Indian buffet process.* Paper presented at the NIPS.

Guille, A., Chaffanet, M., & Birnbaum, D. (2013). Signaling pathway switch in breast cancer. *Cancer cell international, 13*. doi:10.1186/1475-2867-13-66

Guth, U., Huang, D. J., Huber, M., Schotzau, A., Wruk, D., Holzgreve, W., . . . Zanetti-Dallenbach, R. (2008). Tumor size and detection in breast cancer: Self-examination and clinical breast examination are at their limit. *Cancer Detect Prev, 32*(3), 224-228. doi:10.1016/j.cdp.2008.04.002

Hagan, S., Al-Mulla, F., Mallon, E., Oien, K., Ferrier, R., Gusterson, B., . . . Kolch, W. (2005). Reduction of Raf-1 kinase inhibitor protein expression correlates with breast cancer metastasis. *Clinical cancer research: an official journal of the American Association for Cancer Research, 11*, 7392-7397. doi:10.1158/1078-0432.CCR-05-0283

Hagerty, R., Butow, P., Ellis, P., Dimitry, S., & Tattersall, M. (2005). Communicating prognosis in cancer care: a systematic review of the literature. *Annals of Oncology, 16*(7), 1005-1053.

Hall, M. A. (1999). *Correlation-based feature selection for machine learning.* The University of Waikato.

Hanke, M., Kausch, I., Dahmen, G., Jocham, D., & Warnecke, J. M. (2007). Detailed technical analysis of urine RNA-based tumor diagnostics reveals ETS2/urokinase plasminogen activator to be a novel marker for bladder cancer. *Clinical chemistry, 53*(12), 2070-2077.

Hanley, J. A., & McNeil, B. J. (1982). The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology, 143*(1), 29-36.

Haque, R., Ahmed, S. A., Inzhakova, G., Shi, J., Avila, C., Polikoff, J., . . . Press, M. F. (2012). Impact of breast cancer subtypes and treatment on survival: an analysis spanning two decades. *Cancer Epidemiol Biomarkers Prev, 21*(10), 1848-1855. doi:10.1158/1055-9965.EPI-12-0474

Harrell, F. E., Lee, K. L., & Mark, D. B. (1996). Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Statistics in medicine, 15*(4), 361-387.

Hastie, T., & Stuetzle, W. (1989). Principal Curves. *Journal of the American Statistical Association, 84*(406), 502-516. doi:Doi 10.2307/2289936

Hofmann, T., Schölkopf, B., & Smola, A. J. (2008). Kernel methods in machine learning. *The annals of statistics*, 1171-1220.

Holyoake, A., O'Sullivan, P., Pollock, R., Best, T., Watanabe, J., Kajita, Y., . . . Kerr, N. (2008). Development of a multiplex RNA urine test for the detection and stratification of transitional cell carcinoma of the bladder. *Clinical Cancer Research, 14*(3), 742-749.

Hornik, K., Buchta, C., & Zeileis, A. (2009). Open-source machine learning: R meets Weka. *Computational Statistics, 24*(2), 225-232. doi:DOI 10.1007/s00180-008-0119-7

Huang, S., Chaudhary, K., & Garmire, L. X. (2017). More Is Better: Recent Progress in Multi-Omics Data Integration Methods. *Front Genet, 8*, 84.

Huang, S., Chong, N., Lewis, N. E., Jia, W., Xie, G., & Garmire, L. X. (2016). Novel personalized pathway-based metabolomics models reveal key metabolic pathways for breast cancer diagnosis. *Genome medicine, 8*(1), 34.

Huang, S., Yee, C., Ching, T., Yu, H., & Garmire, L. X. (2014). A novel model to combine clinical and pathway-based transcriptomic information for the prognosis prediction of breast cancer. *PLoS Computational Biology, 10*(9), e1003851.

Huang, S., Yee, C., Ching, T., Yu, H., & Garmire, L. X. (2014). A novel model to combine clinical and pathway-based transcriptomic information for the prognosis prediction of breast cancer. *PLoS Comput Biol, 10*(9), e1003851. doi:10.1371/journal.pcbi.1003851

Hudson, T. J., Anderson, W., Aretz, A., Barker, A. D., Bell, C., Bernabé, R. R., . . . Gerhard, D. S. (2010). International network of cancer genome projects. *Nature, 464*(7291), 993-998.

Ideker, T., Ozier, O., Schwikowski, B., & Siegel, A. F. (2002). Discovering regulatory and signalling circuits in molecular interaction networks. *Bioinformatics, 18*(suppl 1), S233-S240.

Imoto, S., Higuchi, T., Goto, T., Tashiro, K., Kuhara, S., & Miyano, S. (2004). Combining microarrays and biological knowledge for estimating gene networks via Bayesian networks. *Journal of bioinformatics and computational biology, 2*(01), 77-98.

Ivshina, A. V., George, J., Senko, O., Mow, B., Putti, T. C., Smeds, J., . . . Nordgren, H. (2006). Genetic reclassification of histologic grade delineates new clinical subtypes of breast cancer. *Cancer Res, 66*(21), 10292-10301.

Ivshina, A. V., George, J., Senko, O., Mow, B., Putti, T. C., Smeds, J., . . . Miller, L. D. (2006). Genetic reclassification of histologic grade delineates new clinical subtypes of breast cancer. *Cancer Res, 66*, 10292-10301. doi:10.1158/0008-5472.CAN-05-4414

Jaraj, S. J., Augsten, M., Häggarth, L., Wester, K., Pontén, F., Östman, A., & Egevad, L. (2011). GAD1 is a biomarker for benign and malignant prostatic tissue. *Scandinavian journal of urology and nephrology, 45*(1), 39-45.

Jemal, A., Bray, F., Center, M. M., Ferlay, J., Ward, E., & Forman, D. (2011). Global cancer statistics. *CA: a cancer journal for clinicians, 61*(2), 69-90.

Jemal, A., Siegel, R., & Ward, E. (2009). Cancer facts and figures 2009. *American Cancer Society, Technical Report*.

Jennings, E. M., Morris, J. S., Carroll, R. J., Manyam, G. C., & Baladandayuthapani, V. (2013). Bayesian methods for expression-based integration of various types of genomics data. *EURASIP Journal on Bioinformatics and Systems Biology, 2013*(1), 13.

Jewison, T., Su, Y., Disfany, F. M., Liang, Y., Knox, C., Maciejewski, A., . . . Wishart, D. S. (2014). SMPDB 2.0: big improvements to the Small Molecule Pathway Database. *Nucleic Acids Res, 42*(Database issue), D478-484. doi:10.1093/nar/gkt1067

Jobard, E., Pontoizeau, C., Blaise, B. J., Bachelot, T., Elena-Herrmann, B., & Tredan, O. (2014). A serum nuclear magnetic resonance-based metabolomic signature of advanced metastatic human breast cancer. *Cancer Lett, 343*(1), 33-41. doi:10.1016/j.canlet.2013.09.011

Johnson, S. R., & Lange, B. M. (2015). Open-access metabolomics databases for natural product research: present capabilities and future potential. *Front Bioeng Biotechnol, 3*, 22. doi:10.3389/fbioe.2015.00022

Kanehisa, M., & Goto, S. (2000). KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res, 28*(1), 27-30.
Kanehisa, M., & Goto, S. (2000). KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Research, 28*(1), 27-30.
Kasprzyk, A. (2011). BioMart: driving a paradigm change in biological data management. *Database (Oxford), 2011*, bar049. doi:10.1093/database/bar049
Kattan, M. W., Eastham, J. A., Stapleton, A. M., Wheeler, T. M., & Scardino, P. T. (1998). A preoperative nomogram for disease recurrence following radical prostatectomy for prostate cancer. *JNCI: Journal of the National Cancer Institute, 90*(10), 766-771.
Katz, R. (2004). Biomarkers and surrogate markers: an FDA perspective. *NeuroRx, 1*(2), 189-195. doi:10.1602/neurorx.1.2.189
Khadra, M., Pickard, R., Charlton, M., Powell, P., & Neal, D. (2000). A prospective analysis of 1,930 patients with hematuria to evaluate current diagnostic practice. *The Journal of urology, 163*(2), 524-527.
Kim, D., Li, R., Dudek, S. M., & Ritchie, M. D. (2013). ATHENA: Identifying interactions between different levels of genomic data associated with cancer clinical outcomes using grammatical evolution neural network. *BioData mining, 6*(1), 23.
Kim, D., Li, R., Lucas, A., Verma, S. S., Dudek, S. M., & Ritchie, M. D. (2016). Using knowledge-driven genomic interactions for multi-omics data analysis: metadimensional models for predicting clinical outcomes in ovarian carcinoma. *Journal of the American Medical Informatics Association*, ocw165.
Kim, D., & Ritchie, M. D. (2014). Data Integration for Cancer Clinical Outcome Prediction. *J Health Med Informat, 5*, e122.
Kim, D., Shin, H., Song, Y. S., & Kim, J. H. (2012). Synergistic effect of different levels of genomic data for cancer clinical outcome prediction. *Journal of biomedical informatics, 45*(6), 1191-1198.
Kirk, P., Griffin, J. E., Savage, R. S., Ghahramani, Z., & Wild, D. L. (2012). Bayesian correlated clustering to integrate multiple datasets. *Bioinformatics, 28*(24), 3290-3297.
Krumsiek, J., Suhre, K., Illig, T., Adamski, J., & Theis, F. J. (2011). Gaussian graphical modeling reconstructs pathway reactions from high-throughput metabolomics data. *BMC Syst Biol, 5*, 21. doi:10.1186/1752-0509-5-21
Krumsiek, J., Suhre, K., Illig, T., Adamski, J., & Theis, F. J. (2012). Bayesian independent component analysis recovers pathway signatures from blood metabolomics data. *J Proteome Res, 11*(8), 4120-4131. doi:10.1021/pr300231n
Lanckriet, G. R., De Bie, T., Cristianini, N., Jordan, M. I., & Noble, W. S. (2004). A statistical framework for genomic data fusion. *Bioinformatics, 20*(16), 2626-2635.
Lê Cao, K.-A., Martin, P. G., Robert-Granié, C., & Besse, P. (2009). Sparse canonical methods for biological data integration: application to a cross-platform study. *BMC bioinformatics, 10*(1), 34.
Lee, D. D., & Seung, H. S. (2001). *Algorithms for non-negative matrix factorization.* Paper presented at the Advances in neural information processing systems.
Lee, E., Chuang, H.-Y., Kim, J.-W., Ideker, T., & Lee, D. (2008). Inferring pathway activity toward precise disease classification. *PLoS Computational Biology, 4*(11), e1000217.
Lee, E., Chuang, H. Y., Kim, J. W., Ideker, T., & Lee, D. (2008). Inferring pathway activity toward precise disease classification. *PLoS Comput Biol, 4*(11), e1000217. doi:10.1371/journal.pcbi.1000217
Lee, K. H., Ho, W. Y., Wu, S. J., Omar, H. A., Huang, P. J., Wang, C. C., & Hung, J. H. (2014). Modulation of Cyclins, p53 and Mitogen-Activated Protein Kinases Signaling in Breast

Cancer Cell Lines by 4-(3,4,5-Trimethoxyphenoxy)benzoic Acid. *Int J Mol Sci, 15*(1), 743-757. doi:10.3390/ijms15010743

Li, J. C. A. (2003). Modeling survival data: Extending the Cox model. *Sociological Methods & Research, 32*(1), 117-120. doi:Doi 10.1177/0049124103031004005

Li, Q., Seo, J.-H., Stranger, B., McKenna, A., Pe'er, I., LaFramboise, T., . . . Freedman, M. L. (2013). Integrative eQTL-based analyses reveal the biology of breast cancer risk loci. *Cell, 152*(3), 633-641.

Lin, D., Zhang, J., Li, J., Calhoun, V. D., Deng, H.-W., & Wang, Y.-P. (2013). Group sparse canonical correlation analysis for genomic data integration. *BMC bioinformatics, 14*(1), 245.

Lock, E. F., & Dunson, D. B. (2013). Bayesian consensus clustering. *Bioinformatics*, btt425.

Lock, E. F., Hoadley, K. A., Marron, J. S., & Nobel, A. B. (2013). Joint and individual variation explained (JIVE) for integrated analysis of multiple data types. *The annals of applied statistics, 7*(1), 523.

Lokeshwar, V. B., Habuchi, T., Grossman, H. B., Murphy, W. M., Hautmann, S. H., Hemstreet, G. P., . . . Schmitz-Dräger, B. J. (2005). Bladder tumor markers beyond cytology: International Consensus Panel on bladder tumor markers. *Urology, 66*(6), 35-63.

Louhimo, R., & Hautaniemi, S. (2011). CNAmet: an R package for integrating copy number, methylation and expression data. *Bioinformatics, 27*(6), 887-888.

Ma, S., Kosorok, M. R., Huang, J., & Dai, Y. (2011). Incorporating higher-order representative features improves prediction in network-based cancer prognosis analysis. *BMC medical genomics, 4*(1), 5.

Ma, S., Kosorok, M. R., Huang, J., & Dai, Y. (2011). Incorporating higher-order representative features improves prediction in network-based cancer prognosis analysis. *BMC Med Genomics, 4*, 5. doi:10.1186/1755-8794-4-5

10.1186/1755-8794-4-5.

Ma, X.-J., Wang, Z., Ryan, P. D., Isakoff, S. J., Barmettler, A., Fuller, A., . . . Tuggle, J. T. (2004). A two-gene expression ratio predicts clinical outcome in breast cancer patients treated with tamoxifen. *Cancer Cell, 5*(6), 607-616.

Madeb, R., & Messing, E. M. (2008). Long-term outcome of home dipstick testing for hematuria. *World journal of urology, 26*(1), 19-24.

Mankoo, P. K., Shen, R., Schultz, N., Levine, D. A., & Sander, C. (2011). Time to recurrence and survival in serous ovarian tumors predicted from integrated genomic profiles. *PloS one, 6*(11), e24709.

Marshall, K. C. (1965). The role of beta-alanine in the biosynthesis of nitrate by Aspergillus flavus. *Antonie Van Leeuwenhoek, 31*(4), 386-394.

Mengual, L., Burset, M., Ribal, M. J., Ars, E., Marín-Aguilera, M., Fernández, M., . . . Alcaraz, A. (2010). Gene expression signature in urine for diagnosing and assessing aggressiveness of bladder urothelial carcinoma. *Clinical Cancer Research, 16*(9), 2624-2633.

Messing, E. (2007). *Markers of detection.* Paper presented at the Urologic Oncology: Seminars and Original Investigations.

Miller, J. A., Pappan, K., Thompson, P. A., Want, E. J., Siskos, A. P., Keun, H. C., . . . Chow, H. H. (2015). Plasma metabolomic profiles of breast cancer patients after short-term limonene intervention. *Cancer Prev Res (Phila), 8*(1), 86-93. doi:10.1158/1940-6207.CAPR-14-0100

Miller, L. D., Smeds, J., George, J., Vega, V. B., Vergara, L., Ploner, A., . . . Liu, E. T. (2005). An expression signature for p53 status in human breast cancer predicts mutation status, transcriptional effects, and patient survival. *Proc Natl Acad Sci U S A, 102*(38), 13550-13555.

Miller, L. D., Smeds, J., George, J., Vega, V. B., Vergara, L., Ploner, A., . . . Bergh, J. (2005). An expression signature for p53 status in human breast cancer predicts mutation status, transcriptional effects, and patient survival. *Proc Natl Acad Sci U S A, 102*, 13550-13555. doi:10.1073/pnas.0506230102

Miyagi, Y., Higashiyama, M., Gochi, A., Akaike, M., Ishikawa, T., Miura, T., . . . Okamoto, N. (2011). Plasma free amino acid profiling of five types of cancer patients and its application for early detection. *PLoS One, 6*(9), e24143. doi:10.1371/journal.pone.0024143

Mo, Q., Wang, S., Seshan, V. E., Olshen, A. B., Schultz, N., Sander, C., . . . Shen, R. (2013). Pattern discovery and cancer gene identification in integrated cancer genomic data. *Proceedings of the National Academy of Sciences, 110*(11), 4245-4250.

Moffat, F. L. (2014). Clinical and pathologic prognostic and predictive factors *Diseases of the Breast: Fifth Edition*: Wolters Kluwer Health Adis (ESP).

Montironi, R., & Lopez-Beltran, A. (2005). The 2004 WHO classification of bladder tumors: a summary and commentary. *International journal of surgical pathology, 13*(2), 143-153.

Mosca, E., Alfieri, R., Merelli, I., Viti, F., Calabria, A., & Milanesi, L. (2010). A multilevel data integration resource for breast cancer study. *BMC systems biology, 4*(1), 76.

Nakamura, K., Kasraeian, A., Iczkowski, K. A., Chang, M., Pendleton, J., Anai, S., & Rosser, C. J. (2009). Utility of serial urinary cytology in the initial evaluation of the patient with microscopic hematuria. *BMC urology, 9*(1), 12.

Nam, H., Chung, B. C., Kim, Y., Lee, K., & Lee, D. (2009). Combining tissue transcriptomics and urine metabolomics for breast cancer biomarker identification. *Bioinformatics, 25*(23), 3151-3157. doi:10.1093/bioinformatics/btp558

Nishimura, D. (2001a). BioCarta. *Biotech Software & Internet Report: The Computer Software Journal for Scient, 2*(3), 117-120.

Nishimura, D. (2001b). BioCarta. *Biotech Software & Internet Report, 2*(3), 4.

O'Brien, K. M., Cole, S. R., Tse, C.-K., Perou, C. M., Carey, L. A., Foulkes, W. D., . . . Millikan, R. C. (2010). Intrinsic breast tumor subtypes, race, and long-term survival in the Carolina Breast Cancer Study. *Clinical Cancer Research, 16*(24), 6100-6110.

O'Brien, K. M., Cole, S. R., Tse, C. K., Perou, C. M., Carey, L. A., Foulkes, W. D., . . . Millikan, R. C. (2010). Intrinsic breast tumor subtypes, race, and long-term survival in the Carolina Breast Cancer Study. *Clin Cancer Res, 16*(24), 6100-6110. doi:10.1158/1078-0432.CCR-10-1533

Oakman, C., Tenori, L., Claudino, W. M., Cappadona, S., Nepi, S., Battaglia, A., . . . Di Leo, A. (2011). Identification of a serum-detectable metabolomic fingerprint potentially correlated with the presence of micrometastatic disease in early breast cancer patients at varying risks of disease relapse by traditional prognostic methods. *Ann Oncol, 22*(6), 1295-1301. doi:10.1093/annonc/mdq606

Ovaska, K., Laakso, M., Haapa-Paananen, S., Louhimo, R., Chen, P., Aittomäki, V., . . . Karinen, S. (2010). Large-scale data integration framework provides a comprehensive view on glioblastoma multiforme. *Genome medicine, 2*(9), 65.

Paik, S., Shak, S., Tang, G., Kim, C., Baker, J., Cronin, M., . . . Park, T. (2004). A multigene assay to predict recurrence of tamoxifen-treated, node-negative breast cancer. *New England Journal of Medicine, 351*(27), 2817-2826.

Parker, J. S., Mullins, M., Cheang, M. C., Leung, S., Voduc, D., Vickery, T., . . . Hu, Z. (2009). Supervised risk predictor of breast cancer based on intrinsic subtypes. *Journal of clinical oncology, 27*(8), 1160-1167.

Parkhomenko, E., Tritchler, D., & Beyene, J. (2009). Sparse canonical correlation analysis with application to genomic data integration. *Statistical Applications in Genetics and Molecular Biology, 8*(1), 1-34.

Pasikanti, K. K., Esuvaranathan, K., Ho, P. C., Mahendran, R., Kamaraj, R., Wu, Q. H., . . . Chan, E. C. (2010). Noninvasive urinary metabonomic diagnosis of human bladder cancer. *J Proteome Res, 9*(6), 2988-2995. doi:10.1021/pr901173v

Pawitan, Y., Bjöhle, J., Amler, L., Borg, A.-L., Egyhazi, S., Hall, P., . . . Klaar, S. (2005). Gene expression profiling spares early breast cancer patients from adjuvant therapy: derived and validated in two population-based cohorts. *Breast cancer research, 7*(6), R953.

Pawitan, Y., Bjohle, J., Amler, L., Borg, A. L., Egyhazi, S., Hall, P., . . . Bergh, J. (2005). Gene expression profiling spares early breast cancer patients from adjuvant therapy: derived and validated in two population-based cohorts. *Breast Cancer Res, 7*, R953-964. doi:10.1186/bcr1325

Perou, C. M., Sørlie, T., Eisen, M. B., van de Rijn, M., Jeffrey, S. S., Rees, C. A., . . . Akslen, L. A. (2000). Molecular portraits of human breast tumours. *Nature, 406*(6797), 747-752.

Pham, T. T., Angus, S. P., & Johnson, G. L. (2013). MAP3K1: Genomic Alterations in Cancer and Function in Promoting Cell Survival or Apoptosis. *Genes & cancer, 4*, 419-426. doi:10.1177/1947601913513950

Polyak, K. (2011). Heterogeneity in breast cancer. *The Journal of clinical investigation, 121*(10), 3786-3788.

Poschke, I., Mao, Y., Kiessling, R., & de Boniface, J. (2013). Tumor-dependent increase of serum amino acid levels in breast cancer patients has diagnostic potential and correlates with molecular tumor subtypes. *J Transl Med, 11*, 290. doi:10.1186/1479-5876-11-290

Pradhan, M. P., Desai, A., & Palakal, M. J. (2013). Systems biology approach to stage-wise characterization of epigenetic genes in lung adenocarcinoma. *BMC Syst Biol, 7*, 141. doi:10.1186/1752-0509-7-141

Puig, T., Vázquez-Martín, A., Relat, J., Pétriz, J., Menéndez, J. A., Porta, R., . . . Colomer, R. (2008). Fatty acid metabolism in breast cancer cells: differential inhibitory effects of epigallocatechin gallate (EGCG) and C75. *Breast cancer research and treatment, 109*, 471-479. doi:10.1007/s10549-007-9678-5

Qiu, Y., Cai, G., Su, M., Chen, T., Liu, Y., Xu, Y., . . . Jia, W. (2010). Urinary metabonomic study on colorectal cancer. *J Proteome Res, 9*(3), 1627-1634. doi:10.1021/pr901081y

Rabin, B. A., Gaglio, B., Sanders, T., Nekhlyudov, L., Dearing, J. W., Bull, S., . . . Marcus, A. (2013). Predicting cancer prognosis using interactive online tools: a systematic review and implications for cancer care providers. *Cancer Epidemiology and Prevention Biomarkers*, cebp. 0513.2013.

Ramskold, D., Luo, S., Wang, Y. C., Li, R., Deng, Q., & Faridani, O. R. (2012). Full-length mRNA-seq from single-cell levels of RNA and individual circulating tumor cells. *Nat Biotechnol, 30*. doi:10.1038/nbt.2282

Rand, W. M. (1971). Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical Association, 66*(336), 846-850.

Ray, P., Zheng, L., Lucas, J., & Carin, L. (2014). Bayesian joint analysis of heterogeneous genomics data. *Bioinformatics, 30*(10), 1370-1376.

Reis-Filho, J. S., & Pusztai, L. (2011). Gene expression profiling in breast cancer: classification, prognostication, and prediction. *The Lancet, 378*(9805), 1812-1823.

Reyal, F., Van Vliet, M. H., Armstrong, N. J., Horlings, H. M., de Visser, K. E., Kok, M., . . . Caldas, C. (2008). A comprehensive analysis of prognostic signatures reveals the high predictive capacity of the proliferation, immune response and RNA splicing modules in breast cancer. *Breast cancer research, 10*(6), R93.

Reyal, F., van Vliet, M. H., Armstrong, N. J., Horlings, H. M., de Visser, K. E., Kok, M., . . . Wessels, L. F. (2008). A comprehensive analysis of prognostic signatures reveals the high

predictive capacity of the proliferation, immune response and RNA splicing modules in breast cancer. *Breast Cancer Res, 10*(6), R93. doi:10.1186/bcr2192

Ritchie, M. D., Holzinger, E. R., Li, R., Pendergrass, S. A., & Kim, D. (2015). Methods of integrating data to uncover genotype-phenotype interactions. *Nature Reviews Genetics, 16*(2), 85-97.

Rodriguez, C. I., & Setaluri, V. (2014). Cyclic AMP (cAMP) signaling in melanocytes and melanoma. *Arch Biochem Biophys, 563*, 22-27. doi:10.1016/j.abb.2014.07.003

Rosser, C. J., Chang, M., Dai, Y., Ross, S., Mengual, L., Alcaraz, A., & Goodison, S. (2014). Urinary protein biomarker panel for the detection of recurrent bladder cancer. *Cancer Epidemiology and Prevention Biomarkers*.

Rosser, C. J., Ross, S., Chang, M., Dai, Y., Mengual, L., Zhang, G., . . . Goodison, S. (2013). Multiplex protein signature for the detection of bladder cancer in voided urine samples. *The Journal of urology, 190*(6), 2257-2262.

Roy, D., Mondal, S., Wang, C., He, X., Khurana, A., Giri, S., . . . Shridhar, V. (2014). Loss of HSulf-1 promotes altered lipid metabolism in ovarian cancer. *Cancer Metab, 2*, 13. doi:10.1186/2049-3002-2-13

Rubino, D., Driggers, P., Arbit, D., Kemp, L., Miller, B., Coso, O., . . . Segars, J. (1998). Characterization of Brx, a novel Dbl family member that modulates estrogen receptor action. *Oncogene, 16*(19), 2513-2526. doi:10.1038/sj.onc.1201783

Rudy, J., & Valafar, F. (2011). Empirical comparison of cross-platform normalization methods for gene expression data. *Bmc Bioinformatics, 12*. doi:Artn 467

Doi 10.1186/1471-2105-12-467

Ruffalo, M., Koyutürk, M., & Sharan, R. (2015). Network-based integration of disparate omic data to identify" silent players" in cancer. *PLoS Comput Biol, 11*(12), e1004595.

Schaefer, M. H., Fontaine, J.-F., Vinayagam, A., Porras, P., Wanker, E. E., & Andrade-Navarro, M. A. (2012). HIPPIE: Integrating protein interaction networks with experiment based quality scores. *PloS one, 7*(2), e31826.

Seigel, R., Naishadham, D., & Jemal, A. (2012). Cancer statistics 2012. *CA Cancer J Clin, 62*(1), 10-29.

Seoane, J. A., Day, I. N., Gaunt, T. R., & Campbell, C. (2014). A pathway-based data integration framework for prediction of disease progression. *Bioinformatics, 30*(6), 838-845.

Shen, J., Yan, L., Liu, S., Ambrosone, C. B., & Zhao, H. (2013). Plasma metabolomic profiles in breast cancer patients and healthy controls: by race and tumor receptor subtypes. *Transl Oncol, 6*(6), 757-765.

Shen, R., Mo, Q., Schultz, N., Seshan, V. E., Olshen, A. B., Huse, J., . . . Sander, C. (2012). Integrative subtype discovery in glioblastoma using iCluster. *PloS one, 7*(4), e35236.

Shen, R., Olshen, A. B., & Ladanyi, M. (2009). Integrative clustering of multiple genomic data types using a joint latent variable model with application to breast and lung cancer subtype analysis. *Bioinformatics, 25*(22), 2906-2912.

Silverman, D. T., Hartge, P., Morrison, A., & Devesa, S. (1992). Epidemiology of bladder cancer. *Hematology/oncology clinics of North America, 6*(1), 1-30.

Singletary, S. E., Allred, C., Ashley, P., Bassett, L. W., Berry, D., Bland, K. I., . . . Greene, F. L. (2002). Revision of the American Joint Committee on Cancer staging system for breast cancer. *J Clin Oncol, 20*(17), 3628-3636.

Smyth, G. K. (2004). Linear models and empirical bayes methods for assessing differential expression in microarray experiments. *Stat Appl Genet Mol Biol, 3*, Article3. doi:10.2202/1544-6115.1027

Society, A. C. (2008). Breast cancer facts & figures 2007-2008: American Cancer Society Atlanta.

Society, A. C. (2013). *Cancer Facts & Figures 2013*. Retrieved from Atlanta: http://www.cancer.org/acs/groups/content/@epidemiologysurveilance/documents/document/acspc-036845.pdf
Society, A. C. (2015). *Cancer facts & figures*: The Society.
Society, A. C. (2017). Cancer facts and figures 2017: American Cancer Society Atlanta.
Sørlie, T., Perou, C. M., Tibshirani, R., Aas, T., Geisler, S., Johnsen, H., . . . Jeffrey, S. S. (2001). Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. *Proceedings of the National Academy of Sciences, 98*(19), 10869-10874.
Sørlie, T., Tibshirani, R., Parker, J., Hastie, T., Marron, J., Nobel, A., . . . Geisler, S. (2003). Repeated observation of breast tumor subtypes in independent gene expression data sets. *Proceedings of the National Academy of Sciences, 100*(14), 8418-8423.
Sotiriou, C., & Pusztai, L. (2009). Gene-expression signatures in breast cancer. *New England Journal of Medicine, 360*(8), 790-800.
Sotiriou, C., Wirapati, P., Loi, S., Harris, A., Fox, S., Smeds, J., . . . Haibe-Kains, B. (2006). Gene expression profiling in breast cancer: understanding the molecular basis of histologic grade to improve prognosis. *Journal of the National Cancer Institute, 98*(4), 262-272.
Speicher, N. K., & Pfeifer, N. (2015). Integrating different data types by regularized unsupervised multiple kernel learning with application to cancer subtype discovery. *Bioinformatics, 31*(12), i268-i275.
Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S., Ebert, B. L., Gillette, M. A., . . . Mesirov, J. P. (2005). Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci U S A, 102*(43), 15545-15550. doi:10.1073/pnas.0506580102
Sveen, A., Agesen, T. H., Nesbakken, A., Meling, G. I., Rognum, T. O., Liestol, K., . . . Lothe, R. A. (2012). ColoGuidePro: a prognostic 7-gene expression signature for stage III colorectal cancer patients. *Clin Cancer Res, 18*, 6001-6010. doi:10.1158/1078-0432.CCR-11-3302
10.1158/1078-0432.CCR-11-3302. Epub 2012 Sep 18.
Tao, M. H., Mason, J. B., Marian, C., McCann, S. E., Platek, M. E., Millen, A., . . . Freudenheim, J. L. (2011). Promoter methylation of E-cadherin, p16, and RAR-beta(2) genes in breast tumors and dietary intake of nutrients important in one-carbon metabolism. *Nutr Cancer, 63*(7), 1143-1150. doi:10.1080/01635581.2011.605982
Tenori, L., Oakman, C., Morris, P. G., Gralka, E., Turner, N., Cappadona, S., . . . Di Leo, A. (2015). Serum metabolomic profiles evaluated after surgery may identify patients with oestrogen receptor negative early breast cancer at increased risk of disease recurrence. Results from a retrospective study. *Mol Oncol, 9*(1), 128-139. doi:10.1016/j.molonc.2014.07.012
Teschendorff, A. E., Gomez, S., Arenas, A., El-Ashry, D., Schmidt, M., Gehrmann, M., & Caldas, C. (2010). Improved prognostic classification of breast cancer defined by antagonistic activation patterns of immune response pathway modules. *BMC cancer, 10*(1), 604.
Teschendorff, A. E., Gomez, S., Arenas, A., El-Ashry, D., Schmidt, M., Gehrmann, M., & Caldas, C. (2010). Improved prognostic classification of breast cancer defined by antagonistic activation patterns of immune response pathway modules. *BMC Cancer, 10*, 604. doi:10.1186/1471-2407-10-604
Têtu, B. (2009). Diagnosis of urothelial carcinoma from urine. *Modern Pathology, 22*, S53-S59.
Therneau, T. M., & Grambsch, P. (2000). Extending the Cox model. *Edited by P. Bickel, P. Diggle, S. Fienberg, K. Krickeberg*, 51.
Thibaux, R., & Jordan, M. I. (2007). *Hierarchical Beta Processes and the Indian Buffet Process.* Paper presented at the AISTATS.

Thiele, I., Swainston, N., Fleming, R. M., Hoppe, A., Sahoo, S., Aurich, M. K., . . . Palsson, B. O. (2013). A community-driven global reconstruction of human metabolism. *Nat Biotechnol, 31*(5), 419-425. doi:10.1038/nbt.2488

Tian, S., Simon, I., Moreno, V., Roepman, P., Tabernero, J., Snel, M., . . . Capella, G. (2013). A combined oncogenic pathway signature of BRAF, KRAS and PI3KCA mutation improves colorectal cancer classification and cetuximab treatment prediction. *Gut, 62*(4), 540-549. doi:10.1136/gutjnl-2012-302423

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 267-288.

Tibshirani, R. (1997). The lasso method for variable selection in the Cox model. *Stat Med, 16*, 385-395.

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

Tipping, M. E. (2001). Sparse Bayesian learning and the relevance vector machine. *Journal of machine learning research, 1*(Jun), 211-244.

Tiruppathi, C., Brandsch, M., Miyamoto, Y., Ganapathy, V., & Leibach, F. H. (1992). Constitutive expression of the taurine transporter in a human colon carcinoma cell line. *Am J Physiol, 263*(5 Pt 1), G625-631.

Toussaint, J., Sieuwerts, A. M., Haibe-Kains, B., Desmedt, C., Rouas, G., Harris, A. L., . . . Durbecq, V. (2009). Improvement of the clinical applicability of the Genomic Grade Index through a qRT-PCR test performed on frozen and formalin-fixed paraffin-embedded tissues. *BMC genomics, 10*(1), 424.

Triantaphyllou, E. (2000). Multi-criteria decision making methods *Multi-criteria decision making methods: A comparative study* (pp. 5-21): Springer.

Trivedi, D., & Messing, E. M. (2009). Commentary: the role of cytologic analysis of voided urine in the work-up of asymptomatic microhematuria. *BMC urology, 9*(1), 13.

Urquidi, V., Goodison, S., Cai, Y., Sun, Y., & Rosser, C. J. (2012). A candidate molecular biomarker panel for the detection of bladder cancer. *Cancer Epidemiology and Prevention Biomarkers, 21*(12), 2149-2158.

Valdehita, A., Bajo, A. M., Fernández-Martínez, A. B., Arenas, M. I., Vacas, E., Valenzuela, P., . . . Carmena, M. J. (2010). Nuclear localization of vasoactive intestinal peptide (VIP) receptors in human breast cancer. *Peptides, 31*, 2035-2045. doi:10.1016/j.peptides.2010.07.024

Van't Veer, L. J., Dai, H., Van De Vijver, M. J., He, Y. D., Hart, A. A., Mao, M., . . . Witteveen, A. T. (2002). Gene expression profiling predicts clinical outcome of breast cancer. *Nature, 415*(6871), 530-536.

van 't Veer, L. J., Dai, H., van de Vijver, M. J., He, Y. D., Hart, A. A., Mao, M., . . . Friend, S. H. (2002). Gene expression profiling predicts clinical outcome of breast cancer. *Nature, 415*(6871), 530-536. doi:10.1038/415530a

van de Vijver, M. J., He, Y. D., van't Veer, L. J., Dai, H., Hart, A. A., Voskuil, D. W., . . . Bernards, R. (2002). A gene-expression signature as a predictor of survival in breast cancer. *N Engl J Med, 347*(25), 1999-2009. doi:10.1056/NEJMoa021967

van den Akker, E. B., Passtoors, W. M., Jansen, R., van Zwet, E. W., Goeman, J. J., Hulsman, M., . . . Beekman, M. (2013). Meta-analysis on blood transcriptomic studies identifies consistently coexpressed protein-protein interaction modules as robust markers of human aging. *Aging Cell*. doi:10.1111/acel.12160

van Iterson, M., van de Wiel, M. A., Boer, J. M., & de Menezes, R. X. (2013). General power and sample size calculations for high-dimensional genomic data. *Stat Appl Genet Mol Biol, 12*(4), 449-467. doi:10.1515/sagmb-2012-0046

Van Rhijn, B. W., Van der Poel, H. G., & van Der Kwast, T. H. (2005). Urine markers for bladder cancer surveillance: a systematic review. *European urology, 47*(6), 736-748.

van Vliet, M. H., Horlings, H. M., van de Vijver, M. J., Reinders, M. J., & Wessels, L. F. (2012). Integration of clinical and gene expression data has a synergetic effect on predicting breast cancer outcome. *PLoS One, 7*(7), e40358. doi:10.1371/journal.pone.0040358

Vaske, C. J., Benz, S. C., Sanborn, J. Z., Earl, D., Szeto, C., Zhu, J., . . . Stuart, J. M. (2010). Inference of patient-specific pathway activities from multi-dimensional cancer genomics data using PARADIGM. *Bioinformatics, 26*(12), i237-i245.

Vickers, A. J., & Elkin, E. B. (2006). Decision curve analysis: a novel method for evaluating prediction models. *Medical Decision Making, 26*(6), 565-574.

Voduc, K. D., Cheang, M. C., Tyldesley, S., Gelmon, K., Nielsen, T. O., & Kennecke, H. (2010). Breast cancer subtypes and the risk of local and regional relapse. *Journal of Clinical Oncology, 28*(10), 1684-1691.

Volinia, S., & Croce, C. M. (2013). Prognostic microRNA/mRNA signature from the integrated analysis of patients with invasive breast cancer. *Proc Natl Acad Sci U S A, 110*, 7413-7417. doi:10.1073/pnas.1304977110

10.1073/pnas.1304977110. Epub 2013 Apr 15.

Wang, B., Mezlini, A. M., Demir, F., Fiume, M., Tu, Z., Brudno, M., . . . Goldenberg, A. (2014). Similarity network fusion for aggregating data types on a genomic scale. *Nat Methods, 11*(3), 333-337.

Wang, Y., Klijn, J. G., Zhang, Y., Sieuwerts, A. M., Look, M. P., Yang, F., . . . Yu, J. (2005). Gene-expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer. *The Lancet, 365*(9460), 671-679.

Wang, Y., Klijn, J. G., Zhang, Y., Sieuwerts, A. M., Look, M. P., Yang, F., . . . Foekens, J. A. (2005). Gene-expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer. *Lancet, 365*(9460), 671-679. doi:10.1016/S0140-6736(05)17947-1

Wei, J., Xie, G., Zhou, Z., Shi, P., Qiu, Y., Zheng, X., . . . Jia, W. (2011). Salivary metabolite signatures of oral cancer and leukoplakia. *Int J Cancer, 129*(9), 2207-2217. doi:10.1002/ijc.25881

Wei, R., De Vivo, I., Huang, S., Zhu, X., Risch, H., Moore, J. H., . . . Garmire, L. X. (2016). Meta-dimensional data integration identifies critical pathways for susceptibility, tumorigenesis and progression of endometrial cancer. *Oncotarget, 7*(34), 55249.

Weigel, M. T., & Dowsett, M. (2010). Current and emerging biomarkers in breast cancer: prognosis and prediction. *Endocrine-related cancer, 17*(4), R245-R262.

Weinstein, J. N., Collisson, E. A., Mills, G. B., Shaw, K. R. M., Ozenberger, B. A., Ellrott, K., . . . Network, C. G. A. R. (2013). The cancer genome atlas pan-cancer analysis project. *Nature genetics, 45*(10), 1113-1120.

Wishart, D. S., Jewison, T., Guo, A. C., Wilson, M., Knox, C., Liu, Y., . . . Scalbert, A. (2013). HMDB 3.0--The Human Metabolome Database in 2013. *Nucleic Acids Res, 41*(Database issue), D801-807. doi:10.1093/nar/gks1065

Witten, D. M., & Tibshirani, R. J. (2009). Extensions of sparse canonical correlation analysis with applications to genomic data. *Statistical Applications in Genetics and Molecular Biology, 8*(1), 1-27.

Wu, T.-F., Lin, C.-J., & Weng, R. C. (2004). Probability estimates for multi-class classification by pairwise coupling. *Journal of Machine Learning Research, 5*(Aug), 975-1005.

Xia, J., Sinelnikov, I. V., Han, B., & Wishart, D. S. (2015). MetaboAnalyst 3.0--making metabolomics more meaningful. *Nucleic Acids Res, 43*(W1), W251-257. doi:10.1093/nar/gkv380

Xia, J., & Wishart, D. S. (2010). MSEA: a web-based tool to identify biologically meaningful patterns in quantitative metabolomic data. *Nucleic Acids Res, 38*(Web Server issue), W71-77. doi:10.1093/nar/gkq329

Xie, G., Zhou, B., Zhao, A., Qiu, Y., Zhao, X., Garmire, L., . . . Jia, W. (2015). Lowered circulating aspartate is a metabolic feature of human breast cancer. *Oncotarget, 6*(32), 33369-33381. doi:10.18632/oncotarget.5409

Yang, C., Richardson, A. D., Smith, J. W., & Osterman, A. (2007). Comparative metabolomics of breast cancer. *Pac Symp Biocomput*, 181-192.

Yang, N., Feng, S., Shedden, K., Xie, X., Liu, Y., Rosser, C. J., . . . Goodison, S. (2011). Urinary glycoprotein biomarker discovery for bladder cancer detection using LC/MS-MS and label-free quantification. *Clinical Cancer Research, 17*(10), 3349-3359.

Yang, W., Yoshigoe, K., Qin, X., Liu, J. S., Yang, J. Y., Niemierko, A., . . . Yang, M. (2014). Identification of genes and pathways involved in kidney renal clear cell carcinoma. *BMC Bioinformatics, 15 Suppl 17*, S2. doi:10.1186/1471-2105-15-S17-S2

Yizhak, K., Gaude, E., Le Devedec, S., Waldman, Y. Y., Stein, G. Y., van de Water, B., . . . Ruppin, E. (2014). Phenotype-based cell-specific metabolic modeling reveals metabolic liabilities of cancer. *Elife, 3*. doi:10.7554/eLife.03641

Yong, H.-Y., Hwang, J.-S., Son, H., Park, H.-I., Oh, E.-S., Kim, H.-H., . . . Moon, A. (2011). Identification of H-Ras-Specific Motif for the Activation of Invasive Signaling Program in Human Breast Epithelial Cells. *Neoplasia (New York, N.Y.), 13*(2), 98-107.

You, Z.-H., Yin, Z., Han, K., Huang, D.-S., & Zhou, X. (2010). A semi-supervised learning approach to predict synthetic genetic interactions by combining functional and topological properties of functional gene network. *BMC bioinformatics, 11*(1), 343.

Yuan, Y., Savage, R. S., & Markowetz, F. (2011). Patient-specific data fusion defines prognostic cancer subtypes. *PLoS Comput Biol, 7*(10), e1002227.

Zhang, F., & Du, G. (2012). Dysregulated lipid metabolism in cancer. *World J Biol Chem, 3*(8), 167-174. doi:10.4331/wjbc.v3.i8.167

Zhang, S., Li, Q., Liu, J., & Zhou, X. J. (2011). A novel computational framework for simultaneous integration of multiple types of genomic data to identify microRNA-gene regulatory modules. *Bioinformatics, 27*(13), i401-i409.

Zhang, S., Liu, C.-C., Li, W., Shen, H., Laird, P. W., & Zhou, X. J. (2012). Discovery of multi-dimensional modules by integrative analysis of cancer genomic data. *Nucleic acids research*, gks725.

Zhao, B., Rubinstein, B. I., Gemmell, J., & Han, J. (2012). A bayesian approach to discovering truth from conflicting sources for data integration. *Proceedings of the VLDB Endowment, 5*(6), 550-561.

Zhu, Y., Qiu, P., & Ji, Y. (2014). TCGA-assembler: open-source software for retrieving and processing TCGA data. *Nat Methods, 11*(6), 599-600.

Aaboe, M., Marcussen, N., Jensen, K. M., Thykjaer, T., Dyrskjøt, L., & Ørntoft, T. (2005). Gene expression profiling of noninvasive primary urothelial tumours using microarrays. *British journal of cancer, 93*(10), 1182-1190.

Abraham, G., Kowalczyk, A., Loi, S., Haviv, I., & Zobel, J. (2010). Prediction of breast cancer prognosis using gene set statistics provides signature stability and biological context. *BMC Bioinformatics, 11*, 277. doi:10.1186/1471-2105-11-277

Abraham, G., Kowalczyk, A., Loi, S., Haviv, I., & Zobel, J. (2010). Prediction of breast cancer prognosis using gene set statistics provides signature stability and biological context. *BMC bioinformatics, 11*(1), 277.

Ades, F., Zardavas, D., Bozovic-Spasojevic, I., Pugliano, L., Fumagalli, D., De Azambuja, E., . . . Piccart, M. (2014). Luminal B breast cancer: molecular characterization, clinical management, and future perspectives. *Journal of clinical oncology, 32*(25), 2794-2803.

Akavia, U. D., Litvin, O., Kim, J., Sanchez-Garcia, F., Kotliar, D., Causton, H. C., . . . Pe'er, D. (2010). An integrated approach to uncover drivers of cancer. *Cell, 143*(6), 1005-1017.

Akker, E. B., Passtoors, W. M., Jansen, R., Zwet, E. W., Goeman, J. J., Hulsman, M., . . . Penninx, B. W. (2014). Meta‐analysis on blood transcriptomic studies identifies consistently coexpressed protein–protein interaction modules as robust markers of human aging. *Aging cell, 13*(2), 216-225.

American Cancer, S. (2003). Cancer facts and figures 2003. *Cancer facts and figures 2003*.

Asiago, V. M., Alvarado, L. Z., Shanaiah, N., Gowda, G. A., Owusu-Sarfo, K., Ballas, R. A., & Raftery, D. (2010). Early detection of recurrent breast cancer using metabolite profiling. *Cancer Res, 70*(21), 8309-8318. doi:10.1158/0008-5472.CAN-10-1319

Aure, M. R., Steinfeld, I., Baumbusch, L. O., Liestøl, K., Lipson, D., Nyberg, S., . . . Børresen-Dale, A.-L. (2013). Identifying in-trans process associated genes in breast cancer by integrated analysis of copy number and expression data. *PloS one, 8*(1), e53014.

Bartoletti, R., Cai, T., Dal Canto, M., Boddi, V., Nesi, G., & Piazzini, M. (2006). Multiplex polymerase chain reaction for microsatellite analysis of urine sediment cells: a rapid and inexpensive method for diagnosing and monitoring superficial transitional bladder cell carcinoma. *The Journal of urology, 175*(6), 2032-2037.

Berg, M., Vanaerschot, M., Jankevics, A., Cuypers, B., Breitling, R., & Dujardin, J. C. (2013). LC-MS metabolomics from study design to data-analysis - using a versatile pathogen as a test case. *Comput Struct Biotechnol J, 4*, e201301002. doi:10.5936/csbj.201301002

Bevers, T. B., Anderson, B. O., Bonaccio, E., Buys, S., Daly, M. B., Dempsey, P. J., . . . Harris, R. E. (2009). Breast cancer screening and diagnosis. *Journal of the National Comprehensive Cancer Network, 7*(10), 1060-1096.

Bild, A. H., Yao, G., Chang, J. T., Wang, Q., Potti, A., Chasse, D., . . . Nevins, J. R. (2006). Oncogenic pathway signatures in human cancers as a guide to targeted therapies. *Nature, 439*(7074), 353-357. doi:10.1038/nature04296

Blasco, H., Nadal-Desbarats, L., Pradat, P. F., Gordon, P. H., Antar, C., Veyrat-Durebex, C., . . . Corcia, P. (2014). Untargeted 1H-NMR metabolomics in CSF: toward a diagnostic biomarker for motor neuron disease. *Neurology, 82*(13), 1167-1174. doi:10.1212/WNL.0000000000000274

Blows, F. M., Driver, K. E., Schmidt, M. K., Broeks, A., Van Leeuwen, F. E., Wesseling, J., . . . Blomqvist, C. (2010). Subtyping of breast cancer by immunohistochemistry to investigate a relationship between subtype and short and long term survival: a collaborative analysis of data for 10,159 cases from 12 studies. *PLoS Med, 7*(5), e1000279.

Bolton, E. E., Wang, Y., Thiessen, P. A., & Bryant, S. H. (2008). PubChem: integrated platform of small molecules and biological activities. *Annual reports in computational chemistry, 4*, 217-241.

Bonnet, E., Calzone, L., & Michoel, T. (2015). Integrative multi-omics module network inference with Lemon-Tree. *PLoS Comput Biol, 11*(2), e1003983.

Borgan, E., Sitter, B., Lingjaerde, O. C., Johnsen, H., Lundgren, S., Bathen, T. F., . . . Gribbestad, I. S. (2010). Merging transcriptomics and metabolomics--advances in breast cancer profiling. *BMC Cancer, 10*, 628. doi:10.1186/1471-2407-10-628

Bossuyt, P. M., Reitsma, J. B., E Bruns, D., Gatsonis, C. A., Glasziou, P. P., Irwig, L. M., . . . De Vet, H. C. (2003). Towards complete and accurate reporting of studies of diagnostic accuracy: the STARD initiative. *Clinical chemistry and laboratory medicine, 41*(1), 68-73.

Brand, A., Leibfritz, D., Hamprecht, B., & Dringen, R. (1998). Metabolism of cysteine in astroglial cells: synthesis of hypotaurine and taurine. *J Neurochem, 71*(2), 827-832.

Bucak, M. N., Tuncer, P. B., Sariozkan, S., Ulutas, P. A., Coyan, K., Baspinar, N., & Ozkalp, B. (2009). Effects of hypotaurine, cysteamine and aminoacids solution on post-thaw microscopic and oxidative stress parameters of Angora goat semen. *Res Vet Sci, 87*(3), 468-472. doi:10.1016/j.rvsc.2009.04.014

Budczies, J., Pfitzner, B. M., Gyorffy, B., Winzer, K. J., Radke, C., Dietel, M., . . . Denkert, C. (2014). Glutamate enrichment as new diagnostic opportunity in breast cancer. *Int J Cancer*. doi:10.1002/ijc.29152

Bundix, F., & Wauters, H. (1997). The diagnostic value of macroscopic hematuria in diagnosing urological cancer. *Fam. Pract*, 1463-1468.

Cai, Z., Zhao, J. S., Li, J. J., Peng, D. N., Wang, X. Y., Chen, T. L., . . . Xie, D. (2010). A combined proteomics and metabolomics profiling of gastric cardia cancer reveals characteristic dysregulations in glucose metabolism. *Mol Cell Proteomics, 9*(12), 2617-2628. doi:10.1074/mcp.M110.000661

Cancer Genome Atlas, N. (2012). Comprehensive molecular portraits of human breast tumours. *Nature, 490*(7418), 61-70. doi:10.1038/nature11412

Carey, L. A., Perou, C. M., Livasy, C. A., Dressler, L. G., Cowan, D., Conway, K., . . . Millikan, R. C. (2006). Race, breast cancer subtypes, and survival in the Carolina Breast Cancer Study. *JAMA, 295*(21), 2492-2502. doi:10.1001/jama.295.21.2492

Chari, R., Coe, B. P., Vucic, E. A., Lockwood, W. W., & Lam, W. L. (2010). An integrative multi-dimensional genetic and epigenetic strategy to identify aberrant genes and pathways in cancer. *BMC systems biology, 4*(1), 67.

Chaudhary, K., Poirion, O. B., Lu, L., & Garmire, L. (2017). Deep Learning based multi-omics integration robustly predicts survival in liver cancer. *bioRxiv*, 114892.

Chen, C.-L., Lin, T.-S., Tsai, C.-H., Wu, C.-C., Chung, T., Chien, K.-Y., . . . Chen, Y.-T. (2013). Identification of potential bladder cancer markers in urine by abundant-protein depletion coupled with quantitative proteomics. *Journal of proteomics, 85*, 28-43.

Chen, J., Bushman, F. D., Lewis, J. D., Wu, G. D., & Li, H. (2013). Structure-constrained sparse canonical correlation analysis with an application to microbiome data analysis. *Biostatistics, 14*(2), 244-258.

Chen, J., & Zhang, S. (2016). Integrative analysis for identifying joint modular patterns of gene-expression and drug-response data. *Bioinformatics, 32*(11), 1724-1732.

Chen, L.-M., Chang, M., Dai, Y., Chai, K. X., Dyrskjøt, L., Sanchez-Cabayo, M., . . . Jeronimo, C. (2014). External validation of a multiplex urinary protein panel for the detection of bladder cancer in a multicenter cohort. *Cancer Epidemiology and Prevention Biomarkers*, cebp. 0029.2014.

Chia, S. K., Bramwell, V. H., Tu, D., Shepherd, L. E., Jiang, S., Vickery, T., . . . Nielsen, T. O. (2012). A 50-gene intrinsic subtype classifier for prognosis and prediction of benefit from adjuvant tamoxifen. *Clin Cancer Res, 18*(16), 4465-4472. doi:10.1158/1078-0432.CCR-12-0286

Chin, L., Andersen, J. N., & Futreal, P. A. (2011). Cancer genomics: from discovery science to personalized medicine. *Nature medicine, 17*(3), 297-303.

Cho, D.-Y., & Przytycka, T. M. (2013). Dissecting cancer heterogeneity with a probabilistic genotype-phenotype model. *Nucleic acids research*, gkt577.

Cline, M. S., Smoot, M., Cerami, E., Kuchinsky, A., Landys, N., Workman, C., . . . Gross, B. (2007). Integration of biological networks and gene expression data using Cytoscape. *Nature protocols, 2*(10), 2366-2382.

Cui, H., Darmanin, S., Natsuisaka, M., Kondo, T., Asaka, M., Shindoh, M., . . . Kobayashi, M. (2007). Enhanced expression of asparagine synthetase under glucose-deprived conditions protects pancreatic cancer cells from apoptosis induced by glucose deprivation and cisplatin. *Cancer Res, 67*(7), 3345-3355. doi:10.1158/0008-5472.CAN-06-2519

Cui, H., Zhou, C., Dai, X., Liang, Y., Paffenroth, R., & Korkin, D. (2017). Boosting Gene Expression Clustering with System-Wide Biological Information: A Robust Autoencoder Approach. *bioRxiv*. doi:10.1101/214122

Curtis, C., Shah, S. P., Chin, S.-F., Turashvili, G., Rueda, O. M., Dunning, M. J., . . . Yuan, Y. (2012). The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. *Nature, 486*(7403), 346-352.

Dai, X., Li, T., Bai, Z., Yang, Y., Liu, X., Zhan, J., & Shi, B. (2015). Breast cancer intrinsic subtype classification, clinical use and future trends. *American journal of cancer research, 5*(10), 2929.

Dang, C. V. (2010). Glutaminolysis: supplying carbon or nitrogen or both for cancer cells? *Cell Cycle, 9*(19), 3884-3886.

de Leoz, M. L., Young, L. J., An, H. J., Kronewitter, S. R., Kim, J., Miyamoto, S., . . . Lebrilla, C. B. (2011). High-mannose glycans are elevated during breast cancer progression. *Mol Cell Proteomics, 10*(1), M110 002717. doi:10.1074/mcp.M110.002717

Denkert, C., Bucher, E., Hilvo, M., Salek, R., Oresic, M., Griffin, J., . . . Fiehn, O. (2012). Metabolomics of human breast cancer: new approaches for tumor typing and biomarker discovery. *Genome Med, 4*(4), 37. doi:10.1186/gm336

Desman, G., Waintraub, C., & Zippin, J. H. (2014). Investigation of cAMP microdomains as a path to novel cancer diagnostics. *Biochim Biophys Acta, 1842*(12PB), 2636-2645. doi:10.1016/j.bbadis.2014.08.016

Desmedt, C., Piette, F., Loi, S., Wang, Y., Lallemand, F., Haibe-Kains, B., . . . d'Assignies, M. S. (2007). Strong time dependence of the 76-gene prognostic signature for node-negative breast cancer patients in the TRANSBIG multicenter independent validation series. *Clinical Cancer Research, 13*(11), 3207-3214.

Desmedt, C., Piette, F., Loi, S., Wang, Y., Lallemand, F., Haibe-Kains, B., . . . Consortium, T. (2007). Strong time dependence of the 76-gene prognostic signature for node-negative breast cancer patients in the TRANSBIG multicenter independent validation series. *Clin Cancer Res, 13*, 3207-3214. doi:10.1158/1078-0432.CCR-06-2765

Dowsett, M., Sestak, I., Lopez-Knowles, E., Sidhu, K., Dunbier, A. K., Cowens, J. W., . . . Cuzick, J. (2013). Comparison of PAM50 Risk of Recurrence Score With Onco type DX and IHC4 for Predicting Risk of Distant Recurrence After Endocrine Therapy. *Journal of clinical oncology, 31*(22), 2783-2790.

Drier, Y., Sheffer, M., & Domany, E. (2013). Pathway-based personalized analysis of cancer. *Proc Natl Acad Sci U S A, 110*(16), 6388-6393. doi:10.1073/pnas.1219651110

Drier, Y., Sheffer, M., & Domany, E. (2013). Pathway-based personalized analysis of cancer. *Proceedings of the National Academy of Sciences, 110*(16), 6388-6393.

Driggers, P. H., Segars, J. H., & Rubino, D. M. (2001). The proto-oncoprotein Brx activates estrogen receptor beta by a p38 mitogen-activated protein kinase pathway. *J Biol Chem, 276*(50), 46792-46797. doi:10.1074/jbc.M106927200

Edge, S. (2010). American joint committee on cancer., american cancer society., Teton data systems (firm). AJCC cancer staging handbook from the AJCC cancer staging manual.

Edwards, T. J., Dickinson, A. J., Natale, S., Gosling, J., & Mcgrath, J. S. (2006). A prospective analysis of the diagnostic yield resulting from the attendance of 4020 patients at a protocol‐driven haematuria clinic. *BJU international, 97*(2), 301-305.

Efron, B., & Tibshirani, R. (2007). On Testing the Significance of Sets of Genes. *Annals of Applied Statistics, 1*(1), 107-129. doi:Doi 10.1214/07-Aoas101

Ein-Dor, L., Kela, I., Getz, G., Givol, D., & Domany, E. (2004). Outcome signature genes in breast cancer: is there a unique set? *Bioinformatics, 21*(2), 171-178.

Elias, K., Svatek, R. S., Gupta, S., Ho, R., & Lotan, Y. (2010). High‐risk patients with hematuria are not evaluated according to guideline recommendations. *Cancer, 116*(12), 2954-2959.

Engelmann, D., & Pützer, B. M. (2012). The Dark Side of E2F1: In Transit beyond Apoptosis. *Cancer Research, 72*(3), 571-575. doi:10.1158/0008-5472.CAN-11-2575

Fan, C., Oh, D. S., Wessels, L., Weigelt, B., Nuyten, D. S., Nobel, A. B., . . . Perou, C. M. (2006). Concordance among gene-expression-based predictors for breast cancer. *N Engl J Med, 355*, 560-569. doi:10.1056/NEJMoa052933

Fan, C., Oh, D. S., Wessels, L., Weigelt, B., Nuyten, D. S., Nobel, A. B., . . . Perou, C. M. (2006). Concordance among gene-expression–based predictors for breast cancer. *New England Journal of Medicine, 355*(6), 560-569.

Fan, C., Prat, A., Parker, J. S., Liu, Y., Carey, L. A., Troester, M. A., & Perou, C. M. (2011). Building prognostic models for breast cancer patients using clinical variables and hundreds of gene expression signatures. *BMC Med Genomics, 4*, 3. doi:10.1186/1755-8794-4-3

10.1186/1755-8794-4-3.

Fan, P., Agboke, F. A., McDaniel, R. E., Sweeney, E. E., Zou, X., Creswell, K., & Jordan, V. C. (2014). Inhibition of c-Src blocks oestrogen-induced apoptosis and restores oestrogen-stimulated growth in long-term oestrogen-deprived breast cancer cells. *Eur J Cancer, 50*(2), 457-468. doi:10.1016/j.ejca.2013.10.001

Fan, Y., Murphy, T. B., Byrne, J. C., Brennan, L., Fitzpatrick, J. M., & Watson, R. W. (2011). Applying random forests to identify biomarker panels in serum 2D-DIGE data for the detection and staging of prostate cancer. *J Proteome Res, 10*(3), 1361-1373. doi:10.1021/pr1011069

Fiehn, O. (2002). Metabolomics--the link between genotypes and phenotypes. *Plant Mol Biol, 48*(1-2), 155-171.

Fong, M. Y., McDunn, J., & Kakar, S. S. (2011). Identification of metabolites in the normal ovary and their transformation in primary and metastatic ovarian cancer. *PLoS One, 6*(5), e19963. doi:10.1371/journal.pone.0019963

Fu, M., Maresh, E. L., Helguera, G., Kiyohara, M., Qin, Y., Ashki, N., . . . Wadehra, M. (2014). Rationale and pre-clinical efficacy of a novel anti-EMP2 antibody for the treatment of invasive breast cancer. *Mol Cancer Ther*. doi:10.1158/1535-7163.MCT-13-0199

Garcia, E., Andrews, C., Hua, J., Kim, H. L., Sukumaran, D. K., Szyperski, T., & Odunsi, K. (2011). Diagnosis of early stage ovarian cancer by 1H NMR metabonomics of serum explored by use of a microflow NMR probe. *J Proteome Res, 10*(4), 1765-1771. doi:10.1021/pr101050d

Garcia, M., Jemal, A., Ward, E., Center, M., Hao, Y., Siegel, R., & Thun, M. (2016). Global cancer facts & figures 2016. *Atlanta, GA: American cancer society, 1*(3), 52.

Gill, R. D. (1992). Multistate life-tables and regression models. *Math Popul Stud, 3*(4), 259-276. doi:10.1080/08898489209525345

Goeman, J. J. (2009). L1 penalized estimation in the Cox proportional hazards model. *Biom J, 52*, 70-84. doi:10.1002/bimj.200900028

10.1002/bimj.200900028.

Goeman, J. J., & Buhlmann, P. (2007). Analyzing gene expression data in terms of gene sets: methodological issues. *Bioinformatics, 23*(8), 980-987. doi:10.1093/bioinformatics/btm051

Goeman, J. J., & Bühlmann, P. (2007). Analyzing gene expression data in terms of gene sets: methodological issues. *Bioinformatics, 23*(8), 980-987.

Goodison, S., Chang, M., Dai, Y., Urquidi, V., & Rosser, C. J. (2012). A multi-analyte assay for the non-invasive detection of bladder cancer. *PloS one, 7*(10), e47469.

Gossai, D., & Lau-Cam, C. A. (2009). The effects of taurine, taurine homologs and hypotaurine on cell and membrane antioxidative system alterations caused by type 2 diabetes in rat erythrocytes. *Adv Exp Med Biol, 643*, 359-368. doi:10.1007/978-0-387-75681-3_37

Griffiths, T. L., & Ghahramani, Z. (2005). *Infinite latent feature models and the Indian buffet process.* Paper presented at the NIPS.

Guille, A., Chaffanet, M., & Birnbaum, D. (2013). Signaling pathway switch in breast cancer. *Cancer cell international, 13*. doi:10.1186/1475-2867-13-66

Guth, U., Huang, D. J., Huber, M., Schotzau, A., Wruk, D., Holzgreve, W., . . . Zanetti-Dallenbach, R. (2008). Tumor size and detection in breast cancer: Self-examination and clinical breast examination are at their limit. *Cancer Detect Prev, 32*(3), 224-228. doi:10.1016/j.cdp.2008.04.002

Hagan, S., Al-Mulla, F., Mallon, E., Oien, K., Ferrier, R., Gusterson, B., . . . Kolch, W. (2005). Reduction of Raf-1 kinase inhibitor protein expression correlates with breast cancer metastasis. *Clinical cancer research: an official journal of the American Association for Cancer Research, 11*, 7392-7397. doi:10.1158/1078-0432.CCR-05-0283

Hagerty, R., Butow, P., Ellis, P., Dimitry, S., & Tattersall, M. (2005). Communicating prognosis in cancer care: a systematic review of the literature. *Annals of Oncology, 16*(7), 1005-1053.

Hall, M. A. (1999). *Correlation-based feature selection for machine learning.* The University of Waikato.

Hanke, M., Kausch, I., Dahmen, G., Jocham, D., & Warnecke, J. M. (2007). Detailed technical analysis of urine RNA-based tumor diagnostics reveals ETS2/urokinase plasminogen activator to be a novel marker for bladder cancer. *Clinical chemistry, 53*(12), 2070-2077.

Hanley, J. A., & McNeil, B. J. (1982). The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology, 143*(1), 29-36.

Haque, R., Ahmed, S. A., Inzhakova, G., Shi, J., Avila, C., Polikoff, J., . . . Press, M. F. (2012). Impact of breast cancer subtypes and treatment on survival: an analysis spanning two decades. *Cancer Epidemiol Biomarkers Prev, 21*(10), 1848-1855. doi:10.1158/1055-9965.EPI-12-0474

Harrell, F. E., Lee, K. L., & Mark, D. B. (1996). Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Statistics in medicine, 15*(4), 361-387.

Hastie, T., & Stuetzle, W. (1989). Principal Curves. *Journal of the American Statistical Association, 84*(406), 502-516. doi:Doi 10.2307/2289936

Hofmann, T., Schölkopf, B., & Smola, A. J. (2008). Kernel methods in machine learning. *The annals of statistics*, 1171-1220.

Holyoake, A., O'Sullivan, P., Pollock, R., Best, T., Watanabe, J., Kajita, Y., . . . Kerr, N. (2008). Development of a multiplex RNA urine test for the detection and stratification of transitional cell carcinoma of the bladder. *Clinical Cancer Research, 14*(3), 742-749.

Hornik, K., Buchta, C., & Zeileis, A. (2009). Open-source machine learning: R meets Weka. *Computational Statistics, 24*(2), 225-232. doi:DOI 10.1007/s00180-008-0119-7

Huang, S., Chaudhary, K., & Garmire, L. X. (2017). More Is Better: Recent Progress in Multi-Omics Data Integration Methods. *Front Genet, 8*, 84.

Huang, S., Chong, N., Lewis, N. E., Jia, W., Xie, G., & Garmire, L. X. (2016). Novel personalized pathway-based metabolomics models reveal key metabolic pathways for breast cancer diagnosis. *Genome medicine, 8*(1), 34.

Huang, S., Yee, C., Ching, T., Yu, H., & Garmire, L. X. (2014). A novel model to combine clinical and pathway-based transcriptomic information for the prognosis prediction of breast cancer. *PLoS Comput Biol, 10*(9), e1003851. doi:10.1371/journal.pcbi.1003851

Huang, S., Yee, C., Ching, T., Yu, H., & Garmire, L. X. (2014). A novel model to combine clinical and pathway-based transcriptomic information for the prognosis prediction of breast cancer. *PLoS Computational Biology, 10*(9), e1003851.

Hudson, T. J., Anderson, W., Aretz, A., Barker, A. D., Bell, C., Bernabé, R. R., . . . Gerhard, D. S. (2010). International network of cancer genome projects. *Nature, 464*(7291), 993-998.

Ideker, T., Ozier, O., Schwikowski, B., & Siegel, A. F. (2002). Discovering regulatory and signalling circuits in molecular interaction networks. *Bioinformatics, 18*(suppl 1), S233-S240.

Imoto, S., Higuchi, T., Goto, T., Tashiro, K., Kuhara, S., & Miyano, S. (2004). Combining microarrays and biological knowledge for estimating gene networks via Bayesian networks. *Journal of bioinformatics and computational biology, 2*(01), 77-98.

Ivshina, A. V., George, J., Senko, O., Mow, B., Putti, T. C., Smeds, J., . . . Nordgren, H. (2006). Genetic reclassification of histologic grade delineates new clinical subtypes of breast cancer. *Cancer Res, 66*(21), 10292-10301.

Ivshina, A. V., George, J., Senko, O., Mow, B., Putti, T. C., Smeds, J., . . . Miller, L. D. (2006). Genetic reclassification of histologic grade delineates new clinical subtypes of breast cancer. *Cancer Res, 66*, 10292-10301. doi:10.1158/0008-5472.CAN-05-4414

Jaraj, S. J., Augsten, M., Häggarth, L., Wester, K., Pontén, F., Östman, A., & Egevad, L. (2011). GAD1 is a biomarker for benign and malignant prostatic tissue. *Scandinavian journal of urology and nephrology, 45*(1), 39-45.

Jemal, A., Bray, F., Center, M. M., Ferlay, J., Ward, E., & Forman, D. (2011). Global cancer statistics. *CA: a cancer journal for clinicians, 61*(2), 69-90.

Jemal, A., Siegel, R., & Ward, E. (2009). Cancer facts and figures 2009. *American Cancer Society, Technical Report*.

Jennings, E. M., Morris, J. S., Carroll, R. J., Manyam, G. C., & Baladandayuthapani, V. (2013). Bayesian methods for expression-based integration of various types of genomics data. *EURASIP Journal on Bioinformatics and Systems Biology, 2013*(1), 13.

Jewison, T., Su, Y., Disfany, F. M., Liang, Y., Knox, C., Maciejewski, A., . . . Wishart, D. S. (2014). SMPDB 2.0: big improvements to the Small Molecule Pathway Database. *Nucleic Acids Res, 42*(Database issue), D478-484. doi:10.1093/nar/gkt1067

Jobard, E., Pontoizeau, C., Blaise, B. J., Bachelot, T., Elena-Herrmann, B., & Tredan, O. (2014). A serum nuclear magnetic resonance-based metabolomic signature of advanced metastatic human breast cancer. *Cancer Lett, 343*(1), 33-41. doi:10.1016/j.canlet.2013.09.011

Johnson, S. R., & Lange, B. M. (2015). Open-access metabolomics databases for natural product research: present capabilities and future potential. *Front Bioeng Biotechnol, 3*, 22. doi:10.3389/fbioe.2015.00022

Kanehisa, M., & Goto, S. (2000). KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res, 28*(1), 27-30.

Kanehisa, M., & Goto, S. (2000). KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Research, 28*(1), 27-30.

Kasprzyk, A. (2011). BioMart: driving a paradigm change in biological data management. *Database (Oxford), 2011*, bar049. doi:10.1093/database/bar049

Kattan, M. W., Eastham, J. A., Stapleton, A. M., Wheeler, T. M., & Scardino, P. T. (1998). A preoperative nomogram for disease recurrence following radical prostatectomy for prostate cancer. *JNCI: Journal of the National Cancer Institute, 90*(10), 766-771.

Katz, R. (2004). Biomarkers and surrogate markers: an FDA perspective. *NeuroRx, 1*(2), 189-195. doi:10.1602/neurorx.1.2.189

Khadra, M., Pickard, R., Charlton, M., Powell, P., & Neal, D. (2000). A prospective analysis of 1,930 patients with hematuria to evaluate current diagnostic practice. *The Journal of urology, 163*(2), 524-527.

Kim, D., Li, R., Dudek, S. M., & Ritchie, M. D. (2013). ATHENA: Identifying interactions between different levels of genomic data associated with cancer clinical outcomes using grammatical evolution neural network. *BioData mining, 6*(1), 23.

Kim, D., Li, R., Lucas, A., Verma, S. S., Dudek, S. M., & Ritchie, M. D. (2016). Using knowledge-driven genomic interactions for multi-omics data analysis: metadimensional models for predicting clinical outcomes in ovarian carcinoma. *Journal of the American Medical Informatics Association*, ocw165.

Kim, D., & Ritchie, M. D. (2014). Data Integration for Cancer Clinical Outcome Prediction. *J Health Med Informat, 5*, e122.

Kim, D., Shin, H., Song, Y. S., & Kim, J. H. (2012). Synergistic effect of different levels of genomic data for cancer clinical outcome prediction. *Journal of biomedical informatics, 45*(6), 1191-1198.

Kirk, P., Griffin, J. E., Savage, R. S., Ghahramani, Z., & Wild, D. L. (2012). Bayesian correlated clustering to integrate multiple datasets. *Bioinformatics, 28*(24), 3290-3297.

Krumsiek, J., Suhre, K., Illig, T., Adamski, J., & Theis, F. J. (2011). Gaussian graphical modeling reconstructs pathway reactions from high-throughput metabolomics data. *BMC Syst Biol, 5*, 21. doi:10.1186/1752-0509-5-21

Krumsiek, J., Suhre, K., Illig, T., Adamski, J., & Theis, F. J. (2012). Bayesian independent component analysis recovers pathway signatures from blood metabolomics data. *J Proteome Res, 11*(8), 4120-4131. doi:10.1021/pr300231n

Lanckriet, G. R., De Bie, T., Cristianini, N., Jordan, M. I., & Noble, W. S. (2004). A statistical framework for genomic data fusion. *Bioinformatics, 20*(16), 2626-2635.

Lê Cao, K.-A., Martin, P. G., Robert-Granié, C., & Besse, P. (2009). Sparse canonical methods for biological data integration: application to a cross-platform study. *BMC bioinformatics, 10*(1), 34.

Lee, D. D., & Seung, H. S. (2001). *Algorithms for non-negative matrix factorization.* Paper presented at the Advances in neural information processing systems.

Lee, E., Chuang, H.-Y., Kim, J.-W., Ideker, T., & Lee, D. (2008). Inferring pathway activity toward precise disease classification. *PLoS Computational Biology, 4*(11), e1000217.

Lee, E., Chuang, H. Y., Kim, J. W., Ideker, T., & Lee, D. (2008). Inferring pathway activity toward precise disease classification. *PLoS Comput Biol, 4*(11), e1000217. doi:10.1371/journal.pcbi.1000217

Lee, K. H., Ho, W. Y., Wu, S. J., Omar, H. A., Huang, P. J., Wang, C. C., & Hung, J. H. (2014). Modulation of Cyclins, p53 and Mitogen-Activated Protein Kinases Signaling in Breast Cancer Cell Lines by 4-(3,4,5-Trimethoxyphenoxy)benzoic Acid. *Int J Mol Sci, 15*(1), 743-757. doi:10.3390/ijms15010743

Li, J. C. A. (2003). Modeling survival data: Extending the Cox model. *Sociological Methods & Research, 32*(1), 117-120. doi:Doi 10.1177/0049124103031004005

Lin, D., Zhang, J., Li, J., Calhoun, V. D., Deng, H.-W., & Wang, Y.-P. (2013). Group sparse canonical correlation analysis for genomic data integration. *BMC bioinformatics, 14*(1), 245.

Lock, E. F., & Dunson, D. B. (2013). Bayesian consensus clustering. *Bioinformatics*, btt425.

Lock, E. F., Hoadley, K. A., Marron, J. S., & Nobel, A. B. (2013). Joint and individual variation explained (JIVE) for integrated analysis of multiple data types. *The annals of applied statistics, 7*(1), 523.

Lokeshwar, V. B., Habuchi, T., Grossman, H. B., Murphy, W. M., Hautmann, S. H., Hemstreet, G. P., . . . Schmitz-Dräger, B. J. (2005). Bladder tumor markers beyond cytology: International Consensus Panel on bladder tumor markers. *Urology, 66*(6), 35-63.

Louhimo, R., & Hautaniemi, S. (2011). CNAmet: an R package for integrating copy number, methylation and expression data. *Bioinformatics, 27*(6), 887-888.

Ma, S., Kosorok, M. R., Huang, J., & Dai, Y. (2011). Incorporating higher-order representative features improves prediction in network-based cancer prognosis analysis. *BMC Med Genomics, 4*, 5. doi:10.1186/1755-8794-4-5

10.1186/1755-8794-4-5.

Ma, S., Kosorok, M. R., Huang, J., & Dai, Y. (2011). Incorporating higher-order representative features improves prediction in network-based cancer prognosis analysis. *BMC medical genomics, 4*(1), 5.

Ma, X.-J., Wang, Z., Ryan, P. D., Isakoff, S. J., Barmettler, A., Fuller, A., . . . Tuggle, J. T. (2004). A two-gene expression ratio predicts clinical outcome in breast cancer patients treated with tamoxifen. *Cancer Cell, 5*(6), 607-616.

Madeb, R., & Messing, E. M. (2008). Long-term outcome of home dipstick testing for hematuria. *World journal of urology, 26*(1), 19-24.

Mankoo, P. K., Shen, R., Schultz, N., Levine, D. A., & Sander, C. (2011). Time to recurrence and survival in serous ovarian tumors predicted from integrated genomic profiles. *PloS one, 6*(11), e24709.

Marshall, K. C. (1965). The role of beta-alanine in the biosynthesis of nitrate by Aspergillus flavus. *Antonie Van Leeuwenhoek, 31*(4), 386-394.

Mengual, L., Burset, M., Ribal, M. J., Ars, E., Marín-Aguilera, M., Fernández, M., . . . Alcaraz, A. (2010). Gene expression signature in urine for diagnosing and assessing aggressiveness of bladder urothelial carcinoma. *Clinical Cancer Research, 16*(9), 2624-2633.

Messing, E. (2007). *Markers of detection.* Paper presented at the Urologic Oncology: Seminars and Original Investigations.

Miller, J. A., Pappan, K., Thompson, P. A., Want, E. J., Siskos, A. P., Keun, H. C., . . . Chow, H. H. (2015). Plasma metabolomic profiles of breast cancer patients after short-term limonene intervention. *Cancer Prev Res (Phila), 8*(1), 86-93. doi:10.1158/1940-6207.CAPR-14-0100

Miller, L. D., Smeds, J., George, J., Vega, V. B., Vergara, L., Ploner, A., . . . Liu, E. T. (2005). An expression signature for p53 status in human breast cancer predicts mutation status, transcriptional effects, and patient survival. *Proc Natl Acad Sci U S A, 102*(38), 13550-13555.

Miller, L. D., Smeds, J., George, J., Vega, V. B., Vergara, L., Ploner, A., . . . Bergh, J. (2005). An expression signature for p53 status in human breast cancer predicts mutation status, transcriptional effects, and patient survival. *Proc Natl Acad Sci U S A, 102*, 13550-13555. doi:10.1073/pnas.0506230102

Miyagi, Y., Higashiyama, M., Gochi, A., Akaike, M., Ishikawa, T., Miura, T., . . . Okamoto, N. (2011). Plasma free amino acid profiling of five types of cancer patients and its application for early detection. *PLoS One, 6*(9), e24143. doi:10.1371/journal.pone.0024143

Mo, Q., Wang, S., Seshan, V. E., Olshen, A. B., Schultz, N., Sander, C., . . . Shen, R. (2013). Pattern discovery and cancer gene identification in integrated cancer genomic data. *Proceedings of the National Academy of Sciences, 110*(11), 4245-4250.

Montironi, R., & Lopez-Beltran, A. (2005). The 2004 WHO classification of bladder tumors: a summary and commentary. *International journal of surgical pathology, 13*(2), 143-153.

Mosca, E., Alfieri, R., Merelli, I., Viti, F., Calabria, A., & Milanesi, L. (2010). A multilevel data integration resource for breast cancer study. *BMC systems biology, 4*(1), 76.

Nakamura, K., Kasraeian, A., Iczkowski, K. A., Chang, M., Pendleton, J., Anai, S., & Rosser, C. J. (2009). Utility of serial urinary cytology in the initial evaluation of the patient with microscopic hematuria. *BMC urology, 9*(1), 12.

Nam, H., Chung, B. C., Kim, Y., Lee, K., & Lee, D. (2009). Combining tissue transcriptomics and urine metabolomics for breast cancer biomarker identification. *Bioinformatics, 25*(23), 3151-3157. doi:10.1093/bioinformatics/btp558

Nishimura, D. (2001a). BioCarta. *Biotech Software & Internet Report, 2*(3), 4.

Nishimura, D. (2001b). BioCarta. *Biotech Software & Internet Report: The Computer Software Journal for Scient, 2*(3), 117-120.

O'Brien, K. M., Cole, S. R., Tse, C. K., Perou, C. M., Carey, L. A., Foulkes, W. D., . . . Millikan, R. C. (2010). Intrinsic breast tumor subtypes, race, and long-term survival in the Carolina Breast Cancer Study. *Clin Cancer Res, 16*(24), 6100-6110. doi:10.1158/1078-0432.CCR-10-1533

Oakman, C., Tenori, L., Claudino, W. M., Cappadona, S., Nepi, S., Battaglia, A., . . . Di Leo, A. (2011). Identification of a serum-detectable metabolomic fingerprint potentially correlated with the presence of micrometastatic disease in early breast cancer patients at varying risks of disease relapse by traditional prognostic methods. *Ann Oncol, 22*(6), 1295-1301. doi:10.1093/annonc/mdq606

Ovaska, K., Laakso, M., Haapa-Paananen, S., Louhimo, R., Chen, P., Aittomäki, V., . . . Karinen, S. (2010). Large-scale data integration framework provides a comprehensive view on glioblastoma multiforme. *Genome medicine, 2*(9), 65.

Paik, S., Shak, S., Tang, G., Kim, C., Baker, J., Cronin, M., . . . Park, T. (2004). A multigene assay to predict recurrence of tamoxifen-treated, node-negative breast cancer. *New England Journal of Medicine, 351*(27), 2817-2826.

Parker, J. S., Mullins, M., Cheang, M. C., Leung, S., Voduc, D., Vickery, T., . . . Hu, Z. (2009). Supervised risk predictor of breast cancer based on intrinsic subtypes. *Journal of clinical oncology, 27*(8), 1160-1167.

Parkhomenko, E., Tritchler, D., & Beyene, J. (2009). Sparse canonical correlation analysis with application to genomic data integration. *Statistical Applications in Genetics and Molecular Biology, 8*(1), 1-34.

Pasikanti, K. K., Esuvaranathan, K., Ho, P. C., Mahendran, R., Kamaraj, R., Wu, Q. H., . . . Chan, E. C. (2010). Noninvasive urinary metabonomic diagnosis of human bladder cancer. *J Proteome Res, 9*(6), 2988-2995. doi:10.1021/pr901173v

Pawitan, Y., Bjöhle, J., Amler, L., Borg, A.-L., Egyhazi, S., Hall, P., . . . Klaar, S. (2005). Gene expression profiling spares early breast cancer patients from adjuvant therapy: derived and validated in two population-based cohorts. *Breast cancer research, 7*(6), R953.

Pawitan, Y., Bjohle, J., Amler, L., Borg, A. L., Egyhazi, S., Hall, P., . . . Bergh, J. (2005). Gene expression profiling spares early breast cancer patients from adjuvant therapy: derived and validated in two population-based cohorts. *Breast Cancer Res, 7*, R953-964. doi:10.1186/bcr1325

Perou, C. M., Sørlie, T., Eisen, M. B., van de Rijn, M., Jeffrey, S. S., Rees, C. A., . . . Akslen, L. A. (2000). Molecular portraits of human breast tumours. *Nature, 406*(6797), 747-752.

Pham, T. T., Angus, S. P., & Johnson, G. L. (2013). MAP3K1: Genomic Alterations in Cancer and Function in Promoting Cell Survival or Apoptosis. *Genes & cancer, 4*, 419-426. doi:10.1177/1947601913513950

Polyak, K. (2011). Heterogeneity in breast cancer. *The Journal of clinical investigation, 121*(10), 3786-3788.

Poschke, I., Mao, Y., Kiessling, R., & de Boniface, J. (2013). Tumor-dependent increase of serum amino acid levels in breast cancer patients has diagnostic potential and correlates with molecular tumor subtypes. *J Transl Med, 11*, 290. doi:10.1186/1479-5876-11-290

Pradhan, M. P., Desai, A., & Palakal, M. J. (2013). Systems biology approach to stage-wise characterization of epigenetic genes in lung adenocarcinoma. *BMC Syst Biol, 7*, 141. doi:10.1186/1752-0509-7-141

Puig, T., Vázquez-Martín, A., Relat, J., Pétriz, J., Menéndez, J. A., Porta, R., . . . Colomer, R. (2008). Fatty acid metabolism in breast cancer cells: differential inhibitory effects of epigallocatechin gallate (EGCG) and C75. *Breast cancer research and treatment, 109*, 471-479. doi:10.1007/s10549-007-9678-5

Qiu, Y., Cai, G., Su, M., Chen, T., Liu, Y., Xu, Y., . . . Jia, W. (2010). Urinary metabonomic study on colorectal cancer. *J Proteome Res, 9*(3), 1627-1634. doi:10.1021/pr901081y

Rabin, B. A., Gaglio, B., Sanders, T., Nekhlyudov, L., Dearing, J. W., Bull, S., . . . Marcus, A. (2013). Predicting cancer prognosis using interactive online tools: a systematic review and implications for cancer care providers. *Cancer Epidemiology and Prevention Biomarkers*, cebp. 0513.2013.

Ramskold, D., Luo, S., Wang, Y. C., Li, R., Deng, Q., & Faridani, O. R. (2012). Full-length mRNA-seq from single-cell levels of RNA and individual circulating tumor cells. *Nat Biotechnol, 30*. doi:10.1038/nbt.2282

Rand, W. M. (1971). Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical Association, 66*(336), 846-850.

Ray, P., Zheng, L., Lucas, J., & Carin, L. (2014). Bayesian joint analysis of heterogeneous genomics data. *Bioinformatics, 30*(10), 1370-1376.

Reis-Filho, J. S., & Pusztai, L. (2011). Gene expression profiling in breast cancer: classification, prognostication, and prediction. *The Lancet, 378*(9805), 1812-1823.

Reyal, F., Van Vliet, M. H., Armstrong, N. J., Horlings, H. M., de Visser, K. E., Kok, M., . . . Caldas, C. (2008). A comprehensive analysis of prognostic signatures reveals the high predictive capacity of the proliferation, immune response and RNA splicing modules in breast cancer. *Breast cancer research, 10*(6), R93.

Reyal, F., van Vliet, M. H., Armstrong, N. J., Horlings, H. M., de Visser, K. E., Kok, M., . . . Wessels, L. F. (2008). A comprehensive analysis of prognostic signatures reveals the high predictive capacity of the proliferation, immune response and RNA splicing modules in breast cancer. *Breast Cancer Res, 10*(6), R93. doi:10.1186/bcr2192

Ritchie, M. D., Holzinger, E. R., Li, R., Pendergrass, S. A., & Kim, D. (2015). Methods of integrating data to uncover genotype-phenotype interactions. *Nature Reviews Genetics, 16*(2), 85-97.

Rodriguez, C. I., & Setaluri, V. (2014). Cyclic AMP (cAMP) signaling in melanocytes and melanoma. *Arch Biochem Biophys, 563*, 22-27. doi:10.1016/j.abb.2014.07.003

Rosser, C. J., Chang, M., Dai, Y., Ross, S., Mengual, L., Alcaraz, A., & Goodison, S. (2014). Urinary protein biomarker panel for the detection of recurrent bladder cancer. *Cancer Epidemiology and Prevention Biomarkers*.

Rosser, C. J., Ross, S., Chang, M., Dai, Y., Mengual, L., Zhang, G., . . . Goodison, S. (2013). Multiplex protein signature for the detection of bladder cancer in voided urine samples. *The Journal of urology, 190*(6), 2257-2262.

Roy, D., Mondal, S., Wang, C., He, X., Khurana, A., Giri, S., . . . Shridhar, V. (2014). Loss of HSulf-1 promotes altered lipid metabolism in ovarian cancer. *Cancer Metab, 2*, 13. doi:10.1186/2049-3002-2-13

Rubino, D., Driggers, P., Arbit, D., Kemp, L., Miller, B., Coso, O., . . . Segars, J. (1998). Characterization of Brx, a novel Dbl family member that modulates estrogen receptor action. *Oncogene, 16*(19), 2513-2526. doi:10.1038/sj.onc.1201783

Rudy, J., & Valafar, F. (2011). Empirical comparison of cross-platform normalization methods for gene expression data. *Bmc Bioinformatics, 12*. doi:Artn 467

Doi 10.1186/1471-2105-12-467

Ruffalo, M., Koyutürk, M., & Sharan, R. (2015). Network-based integration of disparate omic data to identify" silent players" in cancer. *PLoS Comput Biol, 11*(12), e1004595.

Schaefer, M. H., Fontaine, J.-F., Vinayagam, A., Porras, P., Wanker, E. E., & Andrade-Navarro, M. A. (2012). HIPPIE: Integrating protein interaction networks with experiment based quality scores. *PloS one, 7*(2), e31826.

Seigel, R., Naishadham, D., & Jemal, A. (2012). Cancer statistics 2012. *CA Cancer J Clin, 62*(1), 10-29.

Seoane, J. A., Day, I. N., Gaunt, T. R., & Campbell, C. (2014). A pathway-based data integration framework for prediction of disease progression. *Bioinformatics, 30*(6), 838-845.

Shen, J., Yan, L., Liu, S., Ambrosone, C. B., & Zhao, H. (2013). Plasma metabolomic profiles in breast cancer patients and healthy controls: by race and tumor receptor subtypes. *Transl Oncol, 6*(6), 757-765.

Shen, R., Mo, Q., Schultz, N., Seshan, V. E., Olshen, A. B., Huse, J., . . . Sander, C. (2012). Integrative subtype discovery in glioblastoma using iCluster. *PloS one, 7*(4), e35236.

Shen, R., Olshen, A. B., & Ladanyi, M. (2009). Integrative clustering of multiple genomic data types using a joint latent variable model with application to breast and lung cancer subtype analysis. *Bioinformatics, 25*(22), 2906-2912.

Silverman, D. T., Hartge, P., Morrison, A., & Devesa, S. (1992). Epidemiology of bladder cancer. *Hematology/oncology clinics of North America, 6*(1), 1-30.

Singletary, S. E., Allred, C., Ashley, P., Bassett, L. W., Berry, D., Bland, K. I., . . . Greene, F. L. (2002). Revision of the American Joint Committee on Cancer staging system for breast cancer. *J Clin Oncol, 20*(17), 3628-3636.

Smyth, G. K. (2004). Linear models and empirical bayes methods for assessing differential expression in microarray experiments. *Stat Appl Genet Mol Biol, 3*, Article3. doi:10.2202/1544-6115.1027

Society, A. C. (2008). Breast cancer facts & figures 2007-2008: American Cancer Society Atlanta.

Society, A. C. (2013). *Cancer Facts & Figures 2013*. Retrieved from Atlanta: http://www.cancer.org/acs/groups/content/@epidemiologysurveilance/documents/document/acspc-036845.pdf

Society, A. C. (2015). *Cancer facts & figures*: The Society.

Society, A. C. (2017). Cancer facts and figures 2017: American Cancer Society Atlanta.

Sørlie, T., Perou, C. M., Tibshirani, R., Aas, T., Geisler, S., Johnsen, H., . . . Jeffrey, S. S. (2001). Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. *Proceedings of the National Academy of Sciences, 98*(19), 10869-10874.

Sørlie, T., Tibshirani, R., Parker, J., Hastie, T., Marron, J., Nobel, A., . . . Geisler, S. (2003). Repeated observation of breast tumor subtypes in independent gene expression data sets. *Proceedings of the National Academy of Sciences, 100*(14), 8418-8423.

Sotiriou, C., & Pusztai, L. (2009). Gene-expression signatures in breast cancer. *New England Journal of Medicine, 360*(8), 790-800.

Sotiriou, C., Wirapati, P., Loi, S., Harris, A., Fox, S., Smeds, J., . . . Haibe-Kains, B. (2006). Gene expression profiling in breast cancer: understanding the molecular basis of histologic grade to improve prognosis. *Journal of the National Cancer Institute, 98*(4), 262-272.

Speicher, N. K., & Pfeifer, N. (2015). Integrating different data types by regularized unsupervised multiple kernel learning with application to cancer subtype discovery. *Bioinformatics, 31*(12), i268-i275.

Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S., Ebert, B. L., Gillette, M. A., . . . Mesirov, J. P. (2005). Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci U S A, 102*(43), 15545-15550. doi:10.1073/pnas.0506580102

Sveen, A., Agesen, T. H., Nesbakken, A., Meling, G. I., Rognum, T. O., Liestol, K., . . . Lothe, R. A. (2012). ColoGuidePro: a prognostic 7-gene expression signature for stage III colorectal cancer patients. *Clin Cancer Res, 18*, 6001-6010. doi:10.1158/1078-0432.CCR-11-3302

10.1158/1078-0432.CCR-11-3302. Epub 2012 Sep 18.

Tao, M. H., Mason, J. B., Marian, C., McCann, S. E., Platek, M. E., Millen, A., . . . Freudenheim, J. L. (2011). Promoter methylation of E-cadherin, p16, and RAR-beta(2) genes in breast tumors and dietary intake of nutrients important in one-carbon metabolism. *Nutr Cancer, 63*(7), 1143-1150. doi:10.1080/01635581.2011.605982

Tenori, L., Oakman, C., Morris, P. G., Gralka, E., Turner, N., Cappadona, S., . . . Di Leo, A. (2015). Serum metabolomic profiles evaluated after surgery may identify patients with oestrogen receptor negative early breast cancer at increased risk of disease recurrence. Results from a retrospective study. *Mol Oncol, 9*(1), 128-139. doi:10.1016/j.molonc.2014.07.012

Teschendorff, A. E., Gomez, S., Arenas, A., El-Ashry, D., Schmidt, M., Gehrmann, M., & Caldas, C. (2010). Improved prognostic classification of breast cancer defined by antagonistic activation patterns of immune response pathway modules. *BMC cancer, 10*(1), 604.

Teschendorff, A. E., Gomez, S., Arenas, A., El-Ashry, D., Schmidt, M., Gehrmann, M., & Caldas, C. (2010). Improved prognostic classification of breast cancer defined by antagonistic activation patterns of immune response pathway modules. *BMC Cancer, 10*, 604. doi:10.1186/1471-2407-10-604

Têtu, B. (2009). Diagnosis of urothelial carcinoma from urine. *Modern Pathology, 22*, S53-S59.

Therneau, T. M., & Grambsch, P. (2000). Extending the Cox model. *Edited by P. Bickel, P. Diggle, S. Fienberg, K. Krickeberg*, 51.

Thibaux, R., & Jordan, M. I. (2007). *Hierarchical Beta Processes and the Indian Buffet Process.* Paper presented at the AISTATS.

Thiele, I., Swainston, N., Fleming, R. M., Hoppe, A., Sahoo, S., Aurich, M. K., . . . Palsson, B. O. (2013). A community-driven global reconstruction of human metabolism. *Nat Biotechnol, 31*(5), 419-425. doi:10.1038/nbt.2488

Tian, S., Simon, I., Moreno, V., Roepman, P., Tabernero, J., Snel, M., . . . Capella, G. (2013). A combined oncogenic pathway signature of BRAF, KRAS and PI3KCA mutation improves colorectal cancer classification and cetuximab treatment prediction. *Gut, 62*(4), 540-549. doi:10.1136/gutjnl-2012-302423

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 267-288.

Tibshirani, R. (1997). The lasso method for variable selection in the Cox model. *Stat Med, 16*, 385-395.

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

Tipping, M. E. (2001). Sparse Bayesian learning and the relevance vector machine. *Journal of machine learning research, 1*(Jun), 211-244.

Tiruppathi, C., Brandsch, M., Miyamoto, Y., Ganapathy, V., & Leibach, F. H. (1992). Constitutive expression of the taurine transporter in a human colon carcinoma cell line. *Am J Physiol, 263*(5 Pt 1), G625-631.

Toussaint, J., Sieuwerts, A. M., Haibe-Kains, B., Desmedt, C., Rouas, G., Harris, A. L., . . . Durbecq, V. (2009). Improvement of the clinical applicability of the Genomic Grade Index through a qRT-PCR test performed on frozen and formalin-fixed paraffin-embedded tissues. *BMC genomics, 10*(1), 424.

Triantaphyllou, E. (2000). Multi-criteria decision making methods *Multi-criteria decision making methods: A comparative study* (pp. 5-21): Springer.

Trivedi, D., & Messing, E. M. (2009). Commentary: the role of cytologic analysis of voided urine in the work-up of asymptomatic microhematuria. *BMC urology, 9*(1), 13.

Urquidi, V., Goodison, S., Cai, Y., Sun, Y., & Rosser, C. J. (2012). A candidate molecular biomarker panel for the detection of bladder cancer. *Cancer Epidemiology and Prevention Biomarkers, 21*(12), 2149-2158.

Valdehita, A., Bajo, A. M., Fernández-Martínez, A. B., Arenas, M. I., Vacas, E., Valenzuela, P., . . . Carmena, M. J. (2010). Nuclear localization of vasoactive intestinal peptide (VIP) receptors in human breast cancer. *Peptides, 31*, 2035-2045. doi:10.1016/j.peptides.2010.07.024

Van't Veer, L. J., Dai, H., Van De Vijver, M. J., He, Y. D., Hart, A. A., Mao, M., . . . Witteveen, A. T. (2002). Gene expression profiling predicts clinical outcome of breast cancer. *Nature, 415*(6871), 530-536.

van 't Veer, L. J., Dai, H., van de Vijver, M. J., He, Y. D., Hart, A. A., Mao, M., . . . Friend, S. H. (2002). Gene expression profiling predicts clinical outcome of breast cancer. *Nature, 415*(6871), 530-536. doi:10.1038/415530a

van de Vijver, M. J., He, Y. D., van't Veer, L. J., Dai, H., Hart, A. A., Voskuil, D. W., . . . Bernards, R. (2002). A gene-expression signature as a predictor of survival in breast cancer. *N Engl J Med, 347*(25), 1999-2009. doi:10.1056/NEJMoa021967

van den Akker, E. B., Passtoors, W. M., Jansen, R., van Zwet, E. W., Goeman, J. J., Hulsman, M., . . . Beekman, M. (2013). Meta-analysis on blood transcriptomic studies identifies consistently coexpressed protein-protein interaction modules as robust markers of human aging. *Aging Cell*. doi:10.1111/acel.12160

van Iterson, M., van de Wiel, M. A., Boer, J. M., & de Menezes, R. X. (2013). General power and sample size calculations for high-dimensional genomic data. *Stat Appl Genet Mol Biol, 12*(4), 449-467. doi:10.1515/sagmb-2012-0046

Van Rhijn, B. W., Van der Poel, H. G., & van Der Kwast, T. H. (2005). Urine markers for bladder cancer surveillance: a systematic review. *European urology, 47*(6), 736-748.

van Vliet, M. H., Horlings, H. M., van de Vijver, M. J., Reinders, M. J., & Wessels, L. F. (2012). Integration of clinical and gene expression data has a synergetic effect on predicting breast cancer outcome. *PLoS One, 7*(7), e40358. doi:10.1371/journal.pone.0040358

Vaske, C. J., Benz, S. C., Sanborn, J. Z., Earl, D., Szeto, C., Zhu, J., . . . Stuart, J. M. (2010). Inference of patient-specific pathway activities from multi-dimensional cancer genomics data using PARADIGM. *Bioinformatics, 26*(12), i237-i245.

Vickers, A. J., & Elkin, E. B. (2006). Decision curve analysis: a novel method for evaluating prediction models. *Medical Decision Making, 26*(6), 565-574.

Volinia, S., & Croce, C. M. (2013). Prognostic microRNA/mRNA signature from the integrated analysis of patients with invasive breast cancer. *Proc Natl Acad Sci U S A, 110*, 7413-7417. doi:10.1073/pnas.1304977110

10.1073/pnas.1304977110. Epub 2013 Apr 15.

Wang, B., Mezlini, A. M., Demir, F., Fiume, M., Tu, Z., Brudno, M., . . . Goldenberg, A. (2014). Similarity network fusion for aggregating data types on a genomic scale. *Nat Methods, 11*(3), 333-337.

Wang, Y., Klijn, J. G., Zhang, Y., Sieuwerts, A. M., Look, M. P., Yang, F., . . . Yu, J. (2005). Gene-expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer. *The Lancet, 365*(9460), 671-679.

Wang, Y., Klijn, J. G., Zhang, Y., Sieuwerts, A. M., Look, M. P., Yang, F., . . . Foekens, J. A. (2005). Gene-expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer. *Lancet, 365*(9460), 671-679. doi:10.1016/S0140-6736(05)17947-1

Wei, J., Xie, G., Zhou, Z., Shi, P., Qiu, Y., Zheng, X., . . . Jia, W. (2011). Salivary metabolite signatures of oral cancer and leukoplakia. *Int J Cancer, 129*(9), 2207-2217. doi:10.1002/ijc.25881

Wei, R., De Vivo, I., Huang, S., Zhu, X., Risch, H., Moore, J. H., . . . Garmire, L. X. (2016). Meta-dimensional data integration identifies critical pathways for susceptibility, tumorigenesis and progression of endometrial cancer. *Oncotarget, 7*(34), 55249.

Weigel, M. T., & Dowsett, M. (2010). Current and emerging biomarkers in breast cancer: prognosis and prediction. *Endocrine-related cancer, 17*(4), R245-R262.

Weinstein, J. N., Collisson, E. A., Mills, G. B., Shaw, K. R. M., Ozenberger, B. A., Ellrott, K., . . . Network, C. G. A. R. (2013). The cancer genome atlas pan-cancer analysis project. *Nature genetics, 45*(10), 1113-1120.

Wishart, D. S., Jewison, T., Guo, A. C., Wilson, M., Knox, C., Liu, Y., . . . Scalbert, A. (2013). HMDB 3.0--The Human Metabolome Database in 2013. *Nucleic Acids Res, 41*(Database issue), D801-807. doi:10.1093/nar/gks1065

Witten, D. M., & Tibshirani, R. J. (2009). Extensions of sparse canonical correlation analysis with applications to genomic data. *Statistical Applications in Genetics and Molecular Biology, 8*(1), 1-27.

Wu, T.-F., Lin, C.-J., & Weng, R. C. (2004). Probability estimates for multi-class classification by pairwise coupling. *Journal of Machine Learning Research, 5*(Aug), 975-1005.

Xia, J., Sinelnikov, I. V., Han, B., & Wishart, D. S. (2015). MetaboAnalyst 3.0--making metabolomics more meaningful. *Nucleic Acids Res, 43*(W1), W251-257. doi:10.1093/nar/gkv380

Xia, J., & Wishart, D. S. (2010). MSEA: a web-based tool to identify biologically meaningful patterns in quantitative metabolomic data. *Nucleic Acids Res, 38*(Web Server issue), W71-77. doi:10.1093/nar/gkq329

Xie, G., Zhou, B., Zhao, A., Qiu, Y., Zhao, X., Garmire, L., . . . Jia, W. (2015). Lowered circulating aspartate is a metabolic feature of human breast cancer. *Oncotarget, 6*(32), 33369-33381. doi:10.18632/oncotarget.5409

Yang, C., Richardson, A. D., Smith, J. W., & Osterman, A. (2007). Comparative metabolomics of breast cancer. *Pac Symp Biocomput*, 181-192.

Yang, N., Feng, S., Shedden, K., Xie, X., Liu, Y., Rosser, C. J., . . . Goodison, S. (2011). Urinary glycoprotein biomarker discovery for bladder cancer detection using LC/MS-MS and label-free quantification. *Clinical Cancer Research, 17*(10), 3349-3359.

Yang, W., Yoshigoe, K., Qin, X., Liu, J. S., Yang, J. Y., Niemierko, A., . . . Yang, M. (2014). Identification of genes and pathways involved in kidney renal clear cell carcinoma. *BMC Bioinformatics, 15 Suppl 17*, S2. doi:10.1186/1471-2105-15-S17-S2

Yizhak, K., Gaude, E., Le Devedec, S., Waldman, Y. Y., Stein, G. Y., van de Water, B., . . . Ruppin, E. (2014). Phenotype-based cell-specific metabolic modeling reveals metabolic liabilities of cancer. *Elife, 3*. doi:10.7554/eLife.03641

Yong, H.-Y., Hwang, J.-S., Son, H., Park, H.-I., Oh, E.-S., Kim, H.-H., . . . Moon, A. (2011). Identification of H-Ras-Specific Motif for the Activation of Invasive Signaling Program in Human Breast Epithelial Cells. *Neoplasia (New York, N.Y.), 13*(2), 98-107.

You, Z.-H., Yin, Z., Han, K., Huang, D.-S., & Zhou, X. (2010). A semi-supervised learning approach to predict synthetic genetic interactions by combining functional and topological properties of functional gene network. *BMC bioinformatics, 11*(1), 343.

Yuan, Y., Savage, R. S., & Markowetz, F. (2011). Patient-specific data fusion defines prognostic cancer subtypes. *PLoS Comput Biol, 7*(10), e1002227.

Yuan, Y., Van Allen, E. M., Omberg, L., Wagle, N., Amin-Mansour, A., Sokolov, A., . . . Diao, L. (2014). Assessing the clinical utility of cancer genomic and proteomic data across tumor types. *Nature biotechnology, 32*(7), 644-652.

Zhang, F., & Du, G. (2012). Dysregulated lipid metabolism in cancer. *World J Biol Chem, 3*(8), 167-174. doi:10.4331/wjbc.v3.i8.167

Zhang, S., Li, Q., Liu, J., & Zhou, X. J. (2011). A novel computational framework for simultaneous integration of multiple types of genomic data to identify microRNA-gene regulatory modules. *Bioinformatics, 27*(13), i401-i409.

Zhang, S., Liu, C.-C., Li, W., Shen, H., Laird, P. W., & Zhou, X. J. (2012). Discovery of multi-dimensional modules by integrative analysis of cancer genomic data. *Nucleic acids research*, gks725.

Zhao, B., Rubinstein, B. I., Gemmell, J., & Han, J. (2012). A bayesian approach to discovering truth from conflicting sources for data integration. *Proceedings of the VLDB Endowment, 5*(6), 550-561.

Zhu, Y., Qiu, P., & Ji, Y. (2014). TCGA-assembler: open-source software for retrieving and processing TCGA data. *Nat Methods, 11*(6), 599-600.