BMC Cancer

RESEARCH Open Access

Pan-cancer evaluation of gene expression and somatic alteration data for cancer prognosis prediction

Xingyu Zheng[1], Christopher I. Amos[1,2*] and H. Robert Frost[1*]

Abstract

Background: Over the past decades, approaches for diagnosing and treating cancer have seen significant improvement. However, the variability of patient and tumor characteristics has limited progress on methods for prognosis prediction. The development of high-throughput omics technologies now provides multiple approaches for characterizing tumors. Although a large number of published studies have focused on integration of multi-omics data and use of pathway-level models for cancer prognosis prediction, there still exists a gap of knowledge regarding the prognostic landscape across multi-omics data for multiple cancer types using both gene-level and pathway-level predictors.

Methods: In this study, we systematically evaluated three often available types of omics data (gene expression, copy number variation and somatic point mutation) covering both DNA-level and RNA-level features. We evaluated the landscape of predictive performance of these three omics modalities for 33 cancer types in the TCGA using a Lasso or Group Lasso-penalized Cox model and either gene or pathway level predictors.

Results: We constructed the prognostic landscape using three types of omics data for 33 cancer types on both the gene and pathway levels. Based on this landscape, we found that predictive performance is cancer type dependent and we also highlighted the cancer types and omics modalities that support the most accurate prognostic models. In general, models estimated on gene expression data provide the best predictive performance on either gene or pathway level and adding copy number variation or somatic point mutation data to gene expression data does not improve predictive performance, with some exceptional cohorts including low grade glioma and thyroid cancer. In general, pathway-level models have better interpretative performance, higher stability and smaller model size across multiple cancer types and omics data types relative to gene-level models.

Conclusions: Based on this landscape and comprehensively comparison, models estimated on gene expression data provide the best predictive performance on either gene or pathway level. Pathway-level models have better interpretative performance, higher stability and smaller model size relative to gene-level models.

Keywords: Cancer prognosis prediction, Multi-omics data, Pathway analysis, L1 penalized regression model

* Correspondence: chris.amos@bcm.edu; hildreth.r.frost@dartmouth.edu
[1]Department of Biomedical Data Science, Geisel School of Medicine, Dartmouth College, Hanover, NH, USA
Full list of author information is available at the end of the article

Background

Over the past decades, considerable progress has been achieved in diagnosing and treating cancer, with the overall cancer death rates between 1999 and 2015 decreasing by 1.8% per year for men and 1.4% per year for women [1]. However, the variability of patient and tumor characteristics has limited progress on methods for prognosis prediction, despite significant efforts by members of the cancer research community [2, 3]. Several prognostic models for cancer patients using clinical and pathological variables have been developed and widely used in clinical oncology practice [4–6]. With the development of microarrays to detect molecular profiles of patients, some multi-gene assays have been designed and successfully applied in clinical care, such as the assays for prediction of breast cancer recurrence [7, 8]. The development of high-throughput technologies now enables the integration of large-scale molecular profiling data for developing cancer prognostic tools, e.g., RNA profiling through arrays or sequencing enables the measurement of gene-level expression [9], DNA sequencing enables the calling of somatic mutations [10] and application of SNP arrays enable the detection of copy number variation [11]. Many gene-level prognostic models based on gene expression data have been published [12–15], copy number variation has provided insights for cancer prognosis prediction [16, 17], and somatic mutations are often reliably associated with cancer prognosis [18–21]. Given the high level of stochastic variation found in the measures of individual genes, various studies have focused on developing pathway-level models for cancer prognosis prediction [22–25]. A limitation of some single-omics prognostic models is that a single type of genomic measurement may be insufficient to characterize fully the features that lead to cancer progression.

Over the past decade, several large repositories, such as The Cancer Genome Atlas (TCGA) [26] and The International Cancer Genome Consortium (ICGC) [27], have been developed to collect comprehensive multi-omics data on a large group of cancer patients spanning the most common types of human cancer. In TCGA, tumor and normal samples from over 6000 patients have been profiled, covering 37 types of genomic and clinical data for 33 cancer types. Studies based on the analysis of TCGA data range from the comprehensive analysis of specific cancers to more comprehensive landscapes across the most common cancer types. The development of these repositories offers extraordinary opportunities to integrate multi-omics data and researchers have noted that accurate modeling of cancer biology requires multi-dimensional genomic measurements [28–31]. Several recent studies have focused on integration of multi-omics data, especially for survival analysis. For example, studies such as [32–35] have integrated copy number variation and gene expression, and [36–39] have integrated somatic mutation and gene expression. Although a large number of studies have explored the integration of multi-omics data for cancer prognosis prediction [25, 28, 29, 34, 36–40], there still exists a gap of knowledge regarding the prognostic landscape across multi-omics data for multiple cancer types and both gene-level and pathway-level models. In this study, we systematically evaluate three types of omics data (gene expression, copy number variation and somatic point mutation) covering both DNA-level and RNA-level features. We construct the landscape of predictive performance using these three types of omics data for 33 cancer types on both the gene and pathway levels. Based on this landscape, we highlight the cancer types and omics modalities that support the most accurate prognostic models.

Methods

Data sources

TCGA data were accessed via the UCSC Xena data hub [41]. In all, 33 cohorts listed in Supplementary Table X1 were retained for analysis, which included 30 different cancer types and 3 combinations of cancer subtypes; 4 cancer cohorts were excluded because of an insufficient number of samples (Bile Duct Cancer cohort, Formalin Fixed Paraffin-Embedded Pilot Phase II cohort, Large B-cell Lymphoma cohort and Uterine Carcinosarcoma cohort). We downloaded and analyzed gene expression (GE) RNA-seq data, gene-level copy number variation (CNV) data, gene-level non-silent somatic point mutation (SPM) data and survival data for these 33 cancer type cohorts. We focused on the overall survival (OS) end point as the prognostic outcome. Overall survival (OS) is the gold standard primary end point since OS is universally recognized as being unambiguous, unbiased and clinical relevant [42].

For the pathway definitions, we adopted the Hallmark pathway collection from the Molecular Signatures Database (MSigDB) version 6.2 [43]. The Hallmark pathways were generated by a hybrid approach combining computation with manual expert curation and can reduce redundancy and produce more robust enrichment analysis results. The Hallmark pathway collection of MSigDB consists of 50 gene sets derived by aggregating and clustering all other MSigDB gene sets, followed by assignment of well-defined biological states or processes and refinement of genes relevant to the corresponding biological theme [44].

Prognostic models

In this study, we used penalized Cox proportional hazards models with either gene-level or pathway-level predictors as the prognostic models. Our workflow for both the gene-level and pathway-level models is illustrated in Fig. 1.

As shown in Fig. 1, we first conducted filtering on the gene list. To make a fair comparison between gene-level and pathway-level models, we restricted the genes to include only the genes that are present in the Hallmark pathway collection. In addition to this filtering, we also evaluated the prognostic accuracy of models after further

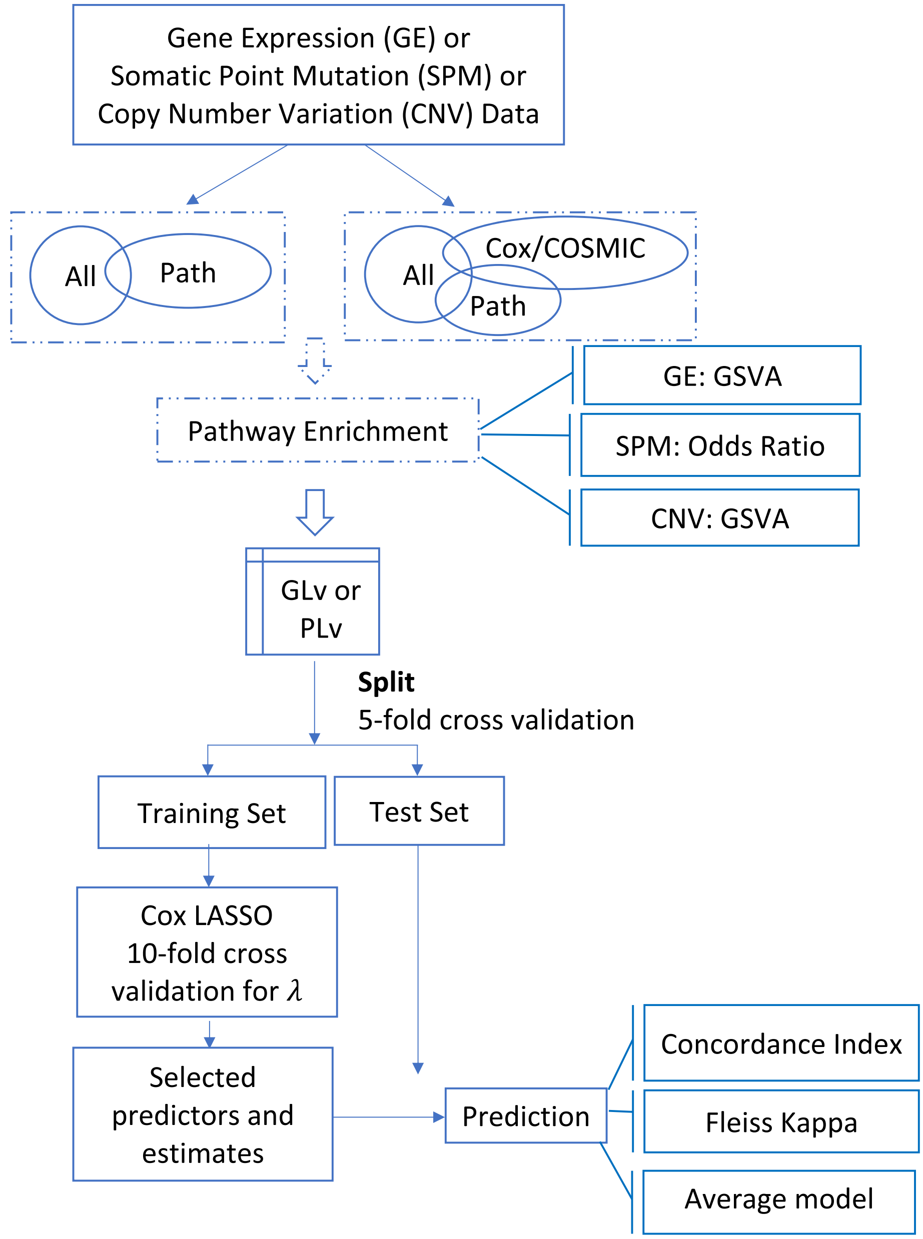

Fig. 1 Workflow of gene-level and pathway-level models. Gene-level data matrix of GE/SPM/CNV is input into the workflow. Genes are pre-filtered either by intersecting with the pathway collection (shown as 'Path') or further filtering the genes by intersecting with COSMIC genes (shown as 'COSMIC') or significant genes (*p*-value less than 0.05) in univariable Cox models (shown as 'Cox'). Then, for the pathway-level models, gene set enrichment is conducted to transform the gene-level matrix into a pathway-level matrix. For GE and CNV data, GSVA is applied and for SPM, odds ratio is applied to conduct gene set enrichment. While for the gene-level models, this step is skipped. With the filtered gene-level data matrix or the transformed pathway-level data matrix as the predictor matrix, we conducted nested cross validation to test the predictive performance of gene-level and pathway-level models. A 5-fold cross validation separates the data into training and test sets. In the training set, a Lasso (least absolute shrinkage and selection operator) or L1-penalized Cox model is fit with the shrinkage parameter chosen by a nested 10-fold cross validation. With the selected predictors and coefficient estimates, the estimated model is applied to the test set and three metrics are adopted to measure the prediction: i) the predictive performance is measured by the concordance index, ii) the model robustness is measured by Fleiss Kappa, iii) the model parsimony is measured by average model size

filtering the genes by either intersecting with COSMIC (The Catalogue Of Somatic Mutations In Cancer) genes [45] or significant genes (*p*-value less than 0.05) in univariable Cox models. To avoid an overfitting bias when filtering according to univariable Cox models, the models were estimated on the training set.

After the filtering step, for the pathway-level model, we conducted single sample gene set enrichment to calculate the sample-level pathway scores and transform the sample-by-gene data matrix into a sample-by-pathway data matrix. For the GE data, we adopted the GSVA (Gene set variation analysis) method [46]. GSVA is an unsupervised and sample-wise gene set enrichment method designed for gene expression data, which calculates a score indicating pathway activity for each sample and pathway. GSVA generates probability density estimates for each gene, which protects it against systematic gene-specific biases and brings distinct profiles to a common scale. Considering the rationale of GSVA and the similar structure of GE and TCGA level 3 CNV data (both are gene-level continuous data), we directly applied GSVA to the CNV data. Since the SPM data is binary, we computed sample-level pathway scores using a log-odds ratio method. Specifically, for each pathway and sample, we created a two-by-two table counting the number of genes according to the presence of somatic point mutations and pathway membership. To avoid the 0 count in the two-by-two table, we added 0.5 to each of the cells (known as Haldane-Anscombe correction [47, 48]). Haldane-Anscombe correction is a common practice, which also removes some bias from the estimator. Using this table, an odds ratio is calculated to indicate the association between pathway membership and mutation status and the log of this odds ratio is used as the sample-level pathway score.

Then, with the filtered gene-level data matrix or the transformed pathway-level data matrix as the predictor matrix, we conducted cross validation to test and compare the predictive performance of gene-level and pathway-level models. Specifically, we conducted 5-fold cross validation of a Lasso-penalized [49] Cox model with the shrinkage parameter chosen by a nested 10-fold cross validation. The Lasso-penalized Cox model was implemented using the functions 'cv.glmnet()' and 'glmnet()' in the R package 'glmnet' [50] with default parameter values.

Integrative models

In this study, we also evaluated the integration of multi-omics data for cancer prognosis prediction. For the integrated analysis, we evaluated two integration methods. In the first method, we combined the data matrices for each omics modality into a single predictor matrix and then performed cross validation as detailed above. The predictor standardization was implemented by default in glmnet to bring different types of variables to the same scale. In the second method, we explored the use of Group Lasso [51] to integrate multi-omics data. Group Lasso is an extension of Lasso for data with a group structure. The principle of Group Lasso is that the variables in the same group should be either all included or all discarded. In this study, for each gene or pathway, we have scores for GE, CNV and SPM separately. Each gene or pathway can function as a group in the Group Lasso with its GE, CNV and SPM variables as group members, which indicate three dimensions of each gene or pathway and may capture a similar biological association with cancer prognosis. After model estimation using a Group Lasso penalty using the function 'cv.grpsurv()' in the R package 'grpreg' [52], all the GE, CNV and SPM variables in the remaining non-zero groups were included into the prognostic models.

Model evaluation metrics

The concordance index (CI), or c-index, is one of the most widely used metrics for survival models and can be interpreted as the measurement of concordance between the predicted and true survival outcomes with a value of 1 indicating perfect prediction and a value of 0.5 indicating random prediction [53]. In our study, we used the average concordance index across cross validation replications to quantify the predictive performance of each model. Inter-rater reliability represents the ability of a model to assign the same score to the same variable for different repeated raters [54]. The Fleiss kappa statistic is widely used to test inter-rater reliability and can be interpreted as the measurement of agreement among different replications with a value of 1 indicating perfect agreement and a value equal to or less than 0 indicating no agreement. In our study, we used the Fleiss kappa statistic [55] to evaluate the repeatability and inter-rater reliability among replications. Specifically, each trained model is a rater that is assigning each variable (gene or pathway) to either being included or excluded in the model. Finally, we used the average number of predictors retained in the trained models to measure model parsimony.

Results

Figure 2 displays the results for both gene-level and pathway-level prognostic models estimated on GE, SPM and CNV data from 33 TCGA cancer types. Rows a and b show that models estimated on GE data most often provide the best predictive performance using either gene-level or pathway-level predictors. The comparison of pathway-level and gene-level GE models in row c shows that these models have similar predictive power.

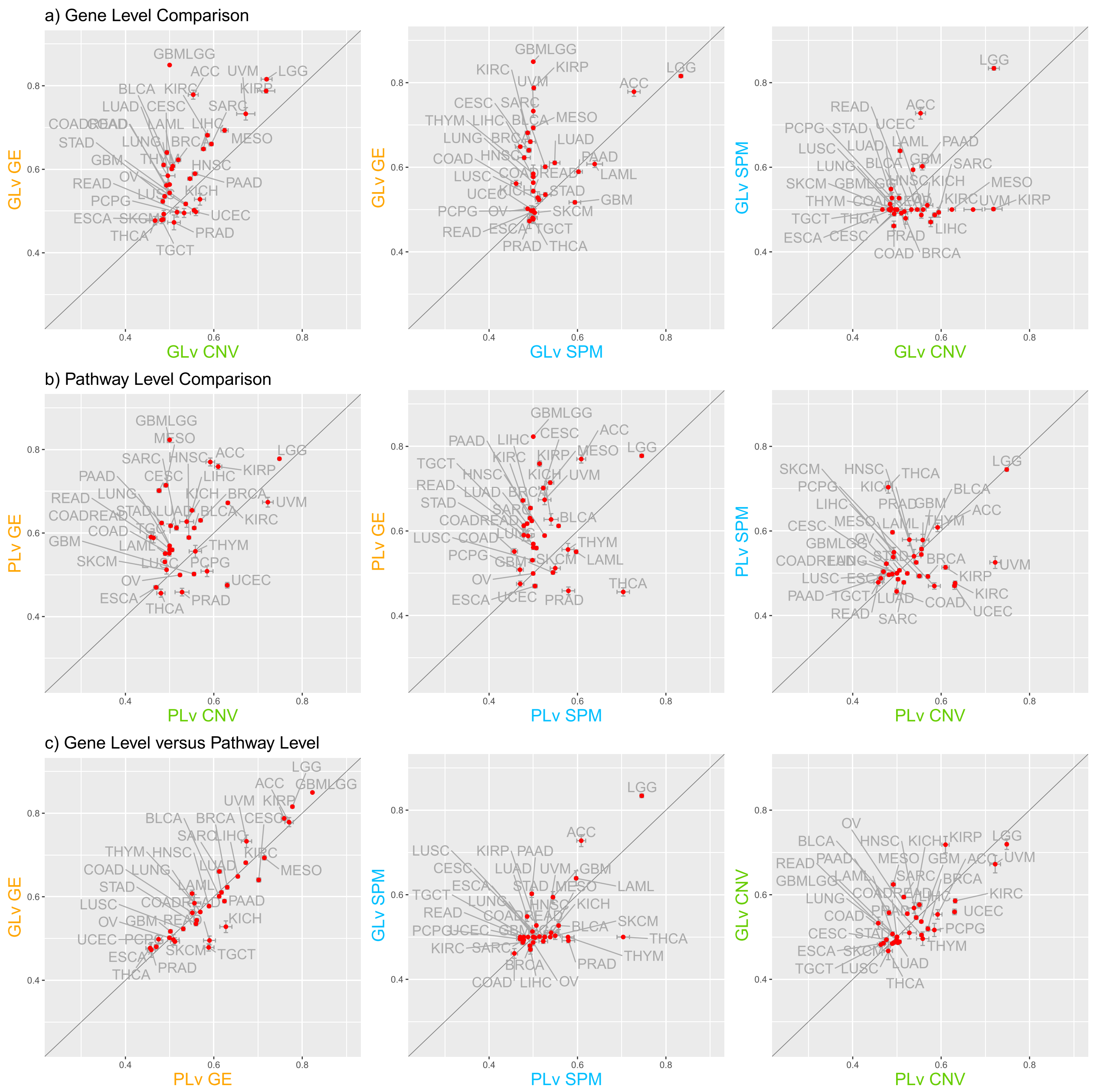

Fig. 2 The comparative results for both gene-level and pathway-level prognostic models estimated using GE, SPM and CNV data from multiple cancer types. 'PLv' represents 'pathway-level' and 'GLv' represents 'gene-level'. The dots represent the values of the concordance index and the bars represent the standard error

For CNV and SPM data, the gene-level models perform slightly better than the pathway-level models.

Row c in Fig. 2 shows the predictive power of single-omics data on both the gene and pathway level. Focusing on cohorts with concordance index values larger than 0.7, which indicates good model performance and is widely used as a standard in the literature [56, 57], GE-based models can predict well for LGG, GBMLGG, KIRP, ACC, MESO on both levels, CESC on the pathway level and UVM on the gene level; SPM-based models can predict well for LGG and ACC on the gene level and THCA on the pathway level; CNV-based models can predict well for LGG and UVM on the pathway level and KIRP on the gene level.

Based on the results shown in Fig. 2, survival models estimated using GE data most often have the best predictive performance among all single omics models. Given this, we next investigated whether the integration of SPM or CNV data with GE data could improve performance over models based on just GE data. Figure 3

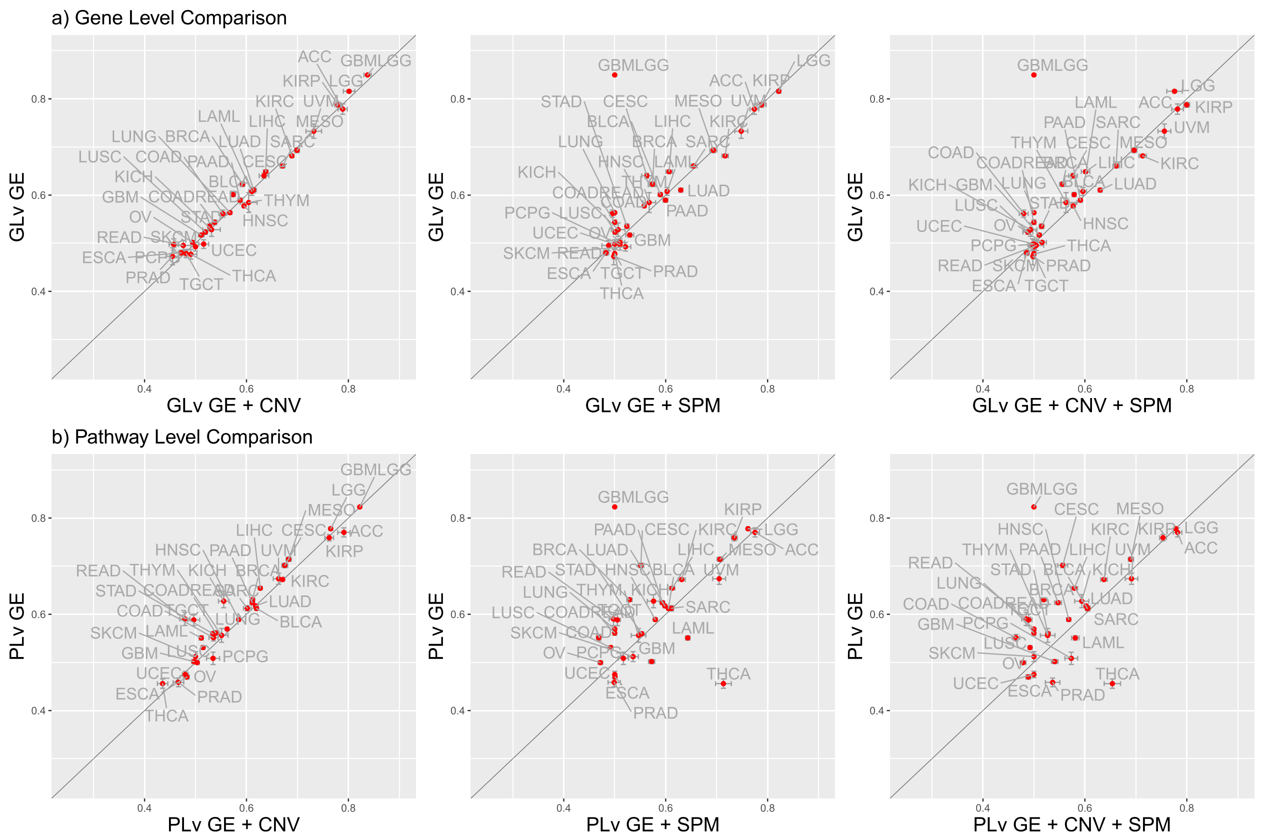

Fig. 3 Comparative results of adding SPM or CNV data to GE data. 'PLv' represents 'pathway-level' and 'GLv' represents 'gene-level'. The dots represent the values of the concordance index and the bars represent the standard error

displays the comparative results of GE data alone versus integration of GE with SPM or CNV data on both the gene and pathway levels. These results show that, in general, adding CNV or SPM data to GE data does not improve predictive performance. This is consistent with findings from Zhao et al. [28]. Adding CNV data to GE data neither increases nor decreases the prediction relative to GE data alone, on both the gene and pathway levels. A similar result is obtained by adding SPM data to GE data on the gene level. The additional SPM data to GE data on the pathway level increases predictive power for cohorts such as THCA but decreases performance for cohorts such as CESC. It is worth noting that, for THCA cohort, the integration of pathway-level SPM and GE data does not perform better than the SPM-only model. Therefore, for THCA cohort, the SPM-only models are optimal. We then investigated the prognostic predictors used in the THCA pathway-level SPM model and found that the MSigDB Hallmark Glycolysis and Spermatogenesis pathways were included as predictors in more than 95% of the estimated models. The biological association of these two pathways to thyroid cancer has been detailed by other researchers. The thyroid gland, previously assumed to not have an impact on spermatogenesis and male fertility, is now recognized to have an important role in male reproductive functions [58, 59]. A considerable amount of data shows that thyroid hormone influences steroidogenesis as well as spermatogenesis [60]. And it is reported that glycolysis-related proteins, such as LDHA, are associated with invasiveness and prognosis of thyroid cancer [61].

In addition to predictive performance, other features such as model robustness and parsimony are also important metrics for the evaluation cancer prognosis prediction models. We utilized the concordance index (CI) to evaluate prediction, Fleiss Kappa statistics to evaluate model robustness and average model size to evaluate parsimony. Figure 4 includes heatmaps that illustrate the pattern of these three metrics across all cohorts and models.

The CI heatmap in Fig. 4 shows that the predictive performance is cancer dependent. This result is concordant with findings reported in Jardillier et al. [62]. For the cohorts located in the left area of the CI heatmap, all evaluated models have poor prognostic power. These models are also associated with lower Kappa values and smaller model sizes as shown in the other two heatmaps, indicating that these models tend to

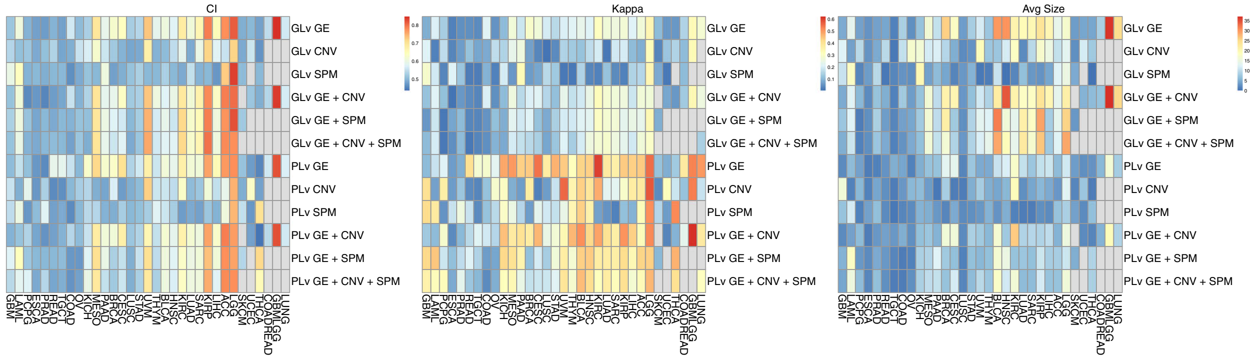

Fig. 4 Heatmap of concordance index, Fleiss Kappa statistics and average model size across cohorts and models. 'PLv' represents 'pathway-level' and 'GLv' represents 'gene-level'. The cells in grey represent models that cannot converge and in this case, no predictors could be selected to predict prognosis

select a small set of random predictors thus the prediction is poor and the models are unstable. For the cohorts located in the right portion of the CI heatmap, the evaluated models had relatively good predictive performance. For these cohorts, the Kappa heatmap indicates that the pathway-level models have higher Kappa values, which indicates better robustness across multiple cross validation splits and replications. As shown in the average model size heatmap, the pathway-level models for these cohorts are also more parsimonious. Overall, these results demonstrate that the pathway-level models have the advantages of better interpretation, higher stability and smaller model size across multiple cancer types and omics data types.

In addition to the models above, we also investigated two different approaches for filtering genes before model estimation. The first filtering approach we evaluated retained only those genes with a significant *p*-value in a univariable Cox model fit on the training set during cross validation. Supplementary Figure S1 displays the predictive performance achieved by this filtering approach relative to models estimated without gene filtering. As shown in this figure, filtering genes with a univariable Cox model failed to improve predictive performance for gene-level models but did improve performance for pathway-level GE and CNV models. For pathway-level model estimated using SPM data, however, filtering resulted in a model without any pathway-level predictors at the optimal Lasso penalization threshold (the relative performance for this model is therefore not included in Supplementary Figure S1). The failure of the filtered SPM pathway-level model to retain predictors after Lasso penalization may be due to the fact that the SPM data itself is sparse and binary and that, after filtering, too few genes are retained to accurately estimate single sample pathway scores. In this case, it is likely that the pathway-level variables contain insufficient information to predict cancer prognosis. Surprisingly, filtering genes based on the results from univariable Cox models did not improve predictive performance for either gene-level or pathway-level multi-omic models. Supplementary Figure S2 row a displays the comparative results of gene filtering for these integration models. The second type of filtering we investigated was limited to the SPM-based models and it filtered the genes according to the COSMIC database. Specifically, we removed any genes without a known cancer association according to COSMIC. As shown in Supplementary Figure S2 row b, COSMIC-based filtering failed to improve predictive performance for either the gene-level or pathway-level models.

In addition to gene filtering, we also investigated the use of a Group Lasso penalty for multi-omics models and the incorporation of clinical stage as a predictor. As illustrated in Supplementary Figure S2 row c the use of a Group Lasso penalty did not improve the predictive performance for the multi-omics models. Supplementary Figure S3 illustrates the impact of adding clinical stage to the models. Surprisingly, adjusting for clinical stage failed to improve the predictive performance for expression data and only weakly improved predictive accuracy for selected CNV or SPM prognostic models. This finding suggests that gene expression levels and clinical staging are correlated, so that little is gained by adding stage information to models for expression data. Other factors that may be driving this result include: i) insufficient samples for many TCGA cohorts to achieve good results via stage-based stratification and five-fold cross validation, and ii) the fact that some cancer types in TCGA represent stage-specific subtypes, such as the LGG and GBM cohorts.

Discussion

In this study, we construct the prognostic landscape using three types of omics data for 33 cancer types on both the gene and pathway levels. Based on this landscape, we found that predictive performance is cancer type dependent and that, relative to gene-level models, pathway-level models have better interpretative performance, higher stability and smaller model size across multiple cancer types and omics data types. We also highlight the cancer types and omics modalities that support the most accurate prognostic models. Beyond this landscape, we evaluated the impact of other modeling parameters including gene filtering, integrative methods and adjustment of clinical stage. In general, models estimated on GE data provide the best predictive performance on either gene or pathway level and adding CNV or SPM data to GE data does not improve predictive performance. Although adding CNV or SPM data into the GE models did not on average improve the predictive power significantly on the pathway level, as shown in the Supplementary Figure S4, the pathway level variables of CNV and SPM still contributed to risk prediction for some models. In the pathway-level integrative model of GE and CNV, the average proportion of CNV variables across all cohorts is 0.49 and for ESCA and UCEC, the proportions are even larger than 0.70. In the pathway-level integrative model of GE and SPM, the average proportion of SPM variables across all cohorts is 0.42 and for ESCA, GBM, PRAD and LUSC, the proportions are even larger than 0.70. Compared with the average proportions of 0.18 and 0.25 respectively in the gene-level integrative models, this finding implies that pathway-level models may exploit more information from CNV and SPM data than gene-level models.

Among the cohorts with concordance index values above 0.7, LGG, ACC and THCA are noteworthy. The LGG cohort performs better than all other cohorts with strong predictive power, robustness across replications and relatively parsimonious models. For the LGG cohort, all 6 models have high concordance index values above 0.7. As shown in the CI heatmap in Fig. 4, the LGG cohort performed remarkably well for all models with the gene-level CNV model having the worst predictive performance (CI is 0.72) and gene-level SPM having the best predictive performance (CI is 0.83). This implies that effective prognostic performance for this cohort can be achieved without gene expression data. Specifically, the LGG SPM models have equivalent performance as the LGG GE models on both the pathway-level and gene-level. Equivalent predictive performance results have also been reported in Zheng et al. [63]. While the gene-level LGG CNV model is slightly worse than the GE and SPM models, the pathway-level LGG CNV model works as well as the GE and SPM models. For the ACC cohort, only the GE and SPM-based models work well using gene-level predictors. For models estimated using pathway-level predictors, only those based on GE data work well for the ACC cohort. Gene expression data is therefore not required to generate effective predictive models for ACC. For the THCA cohort, it is surprising that among all 6 models, only the pathway-level SPM model can predict well. This result may be due to the fact that thyroid cancer has a very low death rate (0.03 in the TCGA data), which makes estimation of survival models challenging.

The underlying factors leading to the heterogenous predictive performance for different cohorts are unknown. These cohorts, located in the left portion in the CI heatmap in Fig. 4, have variable death rates, ranging from 0.02 to 0.75, and variable sample sizes, ranging from 80 to 500. For these cohorts, these three omics data types could not predict prognosis and other characteristics beyond the scope of this study may dominate prognosis, such as clinical variables specific to each cancer type, more accurate characteristics of each sample and even more accurate measurement of each tumor cell with the high-speed development of single-cell sequencing technology.

Although our study was conducted with the Hallmark pathway collection and OS end point as justified in the Data sources section, our method can be extended to other pathway collections in the MSigDB database and other end points including Disease-Specific Survival (DSS), Disease-Free Interval (DFI) and Progression-Free Interval (PFI) in UCSC Xena datahub. We have conducted the basic analysis of an alternative survival outcome, disease free interval (DFI). We explored GE, CNV and SPM data for DFI prediction on both the gene and pathway level. Supplementary Figure S5 shows the predictive power of single-omics data on both the gene and pathway level. The conclusion is consistent with the prediction of overall survival, that predictive performance is cancer type dependent and in general GE data provides the best predictive performance on either gene or pathway level relative to CNV and SPM data. One limitation of our study is that the analysis results were generated on only TCGA data and the conclusions have not been validated in non-TCGA datasets. The reason why we focused on TCGA in this study is that TCGA is the largest and richest collection of multi-omics and clinical data on a large group of cancer patients spanning the most common types of human cancer. It is difficult to find a large-scale database besides TCGA to conduct a comprehensive validation for this pan-cancer and multi-omics study. Some validation for specific cohorts and specific omics data types could be conducted through an analysis of curated datasets from individual research studies. Considering that the goal of this study is to comprehensively compare multi-omics data and pathway predictors relative to gene predictors for cancer prognosis prediction, this specific validation is beyond the scope of this study. Although the validation on other datasets besides TCGA is beyond the scope of our study, it is an important consideration and something we hope to explore in future work. One limitation of our study is that not all genes are included in our analysis because our analysis restricted the genes to include only the genes that are present in the Hallmark pathway collection to make a fair comparison between gene-level and pathway-level models. The models may fail to explore some genes that have a true association with patient survival. Subsequently, the definition of pathways in existing databases could also affect the performance. When gene level predictors are the only focus of interest, this restriction could be released to retain the full performance of prediction. Another limitation of our study is that we only fully explored three types of omics data (GE, CNV, SPM) and there are many omics data, which are not explored in our study, such as methylation data, miRNA data and proteomics data. For example, there are studies reporting the more stable prognostic power of methylation data relative to GE data on the univariable gene level [64]. We have conducted the basic analysis of using methylation data to predict the overall survival on both the gene and pathway level. Supplementary Figure S6 displays the comparison of concordance index values using single omics data on both the pathway and gene level. It shows that in general the methylation data provides similar predictive performance with gene expression data, better than CNV and SPM data. This is biological meaningful since methylation could regulate gene expression. Figure S7 displays the comparison of Fleiss Kappa values using single omics data on both the pathway and gene level. It is consistent with Fig. 4 that the pathway-level models have higher Kappa values than gene-level models, which indicates better robustness across multiple cross validation splits and replications. The methylation data is slightly less stable than gene expression data in our multivariable model on both the pathway and gene level. The detailed exploration on methylation data is beyond the scope of this study and may be explored in the future work.

Conclusion

Based on this study, we found that predictive performance is cancer type dependent and, for the cohorts including GBM, LAML, PCPG, ESCA, PRAD, READ, TGCT, COAD and OV, all evaluated models have poor prognostic power. This finding implies that for these cancer types, more cancer specific clinical information should be used for model estimation in addition to multi-omics data to achieve significant predictive performance. For all other cohorts, we demonstrated that the pathway-level models have the advantages of better interpretation, higher stability and smaller model size, and in general GE data provides the best predictive performance on either gene or pathway level relative to CNV and SPM data. We also highlighted omics modalities that support the most accurate prognostic models. Beyond this, we showed that based on our results, the LGG, ACC and THCA cohorts are noteworthy. For the LGG cohort, all models have good predictive power with the SPM- having the best predictive performance (CI is 0.83). This implies that effective prognostic performance for this cohort can be achieved without gene expression data. For the THCA cohort, it is surprising that among all models, only the pathway-level SPM model can predict well.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12885-021-08796-3.

Additional file 1.

Acknowledgements
We would like to acknowledge the supportive research environments at the Geisel School of Medicine at Dartmouth College and Department of Medicine at Baylor College of Medicine.

Authors' contributions

XZ conducted the analysis and drafted the manuscript. Both HRF and CA guided the analysis and revised the manuscript. All authors read and approved the final manuscript.

Funding
The authors are grateful for the financial support from research grants K01LM012426 NIH/NLM, R21CA253408 NIH/NCI, P20GM130454 NIH/NIGMS, U01CA196386 NIH/NCI and P30CA023108 NIH/NCI. Dr. Amos is an Established Research Scholar supported by the Cancer Research Institute of Texas grant RR170048 and by NIH grant U19CA203654.

Availability of data and materials

The datasets analyzed during the current study are available in the UCSC TCGA repository, http://xena.ucsc.edu/, and MSigDB database, http://software.broadinstitute.org/gsea/msigdb/genesets.jsp?collection=H

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Author details
[1]Department of Biomedical Data Science, Geisel School of Medicine, Dartmouth College, Hanover, NH, USA. [2]Department of Medicine, Institute for Clinical and Translational Research, Baylor College of Medicine, Houston, TX, USA.

Received: 9 February 2021 Accepted: 16 August 2021
Published online: 25 September 2021

References

1. Cronin KA, Lake AJ, Scott S, Sherman RL, Noone AM, Howlader N, et al. Annual report to the nation on the status of Cancer, part I: national cancer statistics. Cancer. 2018;124(13):2785–800. https://doi.org/10.1002/cncr.31551.
2. Lee VC. Cancer immunotherapy, part 3: challenges and future trends. Pharm Ther. 2017;42(8):514–21.

3. Dalton WS, Friend SH. Cancer biomarkers - An invitation to the table. Science. 2006;312(5777):1165–8.
4. Gaspar L, Scott C, Rotman M, Asbell S, Phillips T, Wasserman T, et al. Recursive partitioning analysis (RPA) of prognostic factors in three radiation therapy oncology group (RTOG) brain metastases trials. Int J Radiat Oncol Biol Phys. 1997;37(4):745–51. https://doi.org/10.1016/S0360-3016(96)00619-0.
5. Sperduto PW, Berkey B, Gaspar LE, Mehta M, Curran W. A new prognostic index and comparison to three other indices for patients with brain metastases: an analysis of 1,960 patients in the RTOG database. Int J Radiat Oncol Biol Phys. 2008;70(2):510–4. https://doi.org/10.1016/j.ijrobp.2007.06.074.
6. Sperduto PW, Kased N, Roberge D, Xu Z, Shanley R, Luo X, et al. Effect of tumor subtype on survival and the graded prognostic assessment for patients with breast cancer and brain metastases. Int J Radiat Oncol Biol Phys. 2012;82(5):2111–7. https://doi.org/10.1016/j.ijrobp.2011.02.027.
7. Mook S, Van't Veer LJ, Rutgers EJT, Piccart-Gebhart MJ, Cardoso F. Individualization of therapy using mammaprint®™: from development to the MINDACT trial. Cancer Genomics Proteomics. 2007;4(3):147–55.
8. Sparano JA, Paik S. Development of the 21-gene assay and its application in clinical practice and clinical trials. J Clin Oncol. 2008;26(5):721–8. https://doi.org/10.1200/JCO.2007.15.1068.
9. Wang Z, Gerstein M, Snyder M. RNA-Seq: a revolutionary tool for transcriptomics. Nat Rev Genet. 2009;10(1):57–63. https://doi.org/10.1038/nrg2484.
10. Heather JM, Chain B. The sequence of sequencers: the history of sequencing DNA. Genomics. 2016;107(1):1–8. https://doi.org/10.1016/j.ygeno.2015.11.003.
11. LaFramboise T. Single nucleotide polymorphism arrays: a decade of biological, computational and technological advances. Nucleic Acids Res. 2009;37(13):4181–93. https://doi.org/10.1093/nar/gkp552.
12. Bøvelstad HM, Nygård S, Borgan Ø. Survival prediction from clinico-genomic models - a comparative study. BMC Bioinformatics. 2009;413:1–9.
13. Van De Vijver MJ, He YD, Van't Veer LJ, Dai H, Hart AAM, Voskuil DW, et al. a gene-expression signature as a predictor of survival in breast cancer. N Engl J Med 2002;347(25):1999–2009, DOI: https://doi.org/10.1056/NEJMoa021967.
14. Feng Y, Sun B, Li X, Zhang L, Niu Y, Xiao C, et al. Differentially expressed genes between primary cancer and paired lymph node metastases predict clinical outcome of node-positive breast cancer patients. Breast Cancer Res Treat. 2007;103(3):319–29. https://doi.org/10.1007/s10549-006-9385-7.
15. Mook S, Schmidt MK, Viale G, Pruneri G, Eekhout I, Floore A, et al. The 70-gene prognosis-signature predicts disease outcome in breast cancer patients with 1-3 positive lymph nodes in an independent validation study. Breast Cancer Res Treat. 2009;116(2):295–302. https://doi.org/10.1007/s10549-008-0130-2.
16. Kawano O, Sasaki H, Okuda K, Yukiue H, Yokoyama T, Yano M, et al. PIK3CA gene amplification in Japanese non-small cell lung cancer. Lung Cancer. 2007;58(1):159–60. https://doi.org/10.1016/j.lungcan.2007.06.020.
17. Go H, Jeon YK, Park HJ, Sung SW, Seo JW, Chung DH. High MET gene copy number leads to shorter survival in patients with non-small cell lung cancer. J Thorac Oncol. 2010;5(3):305–13. https://doi.org/10.1097/JTO.0b013e3181ce3d1d.
18. Walker BA, Wardell CP, Murison A, Boyle EM, Begum DB, Dahir NM, et al. APOBEC family mutational signatures are associated with poor prognosis translocations in multiple myeloma. Nat Commun. 2015;6(1):6997. https://doi.org/10.1038/ncomms7997.
19. Walker BA, Boyle EM, Wardell CP, Murison A, Begum DB, Dahir NM, et al. Mutational spectrum, copy number changes, and outcome: results of a sequencing study of patients with newly diagnosed myeloma. J Clin Oncol. 2015;33(33):3911–20. https://doi.org/10.1200/JCO.2014.59.1503.
20. Haricharan S, Bainbridge MN, Scheet P, Brown PH. Somatic mutation load of estrogen receptor-positive breast tumors predicts overall survival: an analysis of genome sequence data. Breast Cancer Res Treat. 2014;146(1):211–20. https://doi.org/10.1007/s10549-014-2991-x.
21. Miller A, Asmann Y, Cattaneo L, Braggio E, Keats J, Auclair D, et al. High somatic mutation and neoantigen burden are correlated with decreased progression-free survival in multiple myeloma. Blood Cancer J. 2017;7:e612.
22. Jones S, Zhang X, Parsons DW, Lin JCH, Leary RJ, Angenendt P, et al. Core signaling pathways in human pancreatic cancers revealed by global genomic analyses. Science. 2008;321(5897):1801–6.
23. Zhang X, Li Y, Akinyemiju T, Ojesina AI, Buckhaults P, Liu N, et al. Pathway-structured predictive model for cancer survival prediction: a two-stage approach. Genetics. 2017;205(1):89–100. https://doi.org/10.1534/genetics.116.189191.
24. Eng KH, Wang S, Bradley WH, Rader JS, Kendziorski C. Pathway index models for construction of patient-specific risk profiles. Stat Med. 2013;32(9):1524–35. https://doi.org/10.1002/sim.5641.
25. Bennett BD, Xiong Q, Mukherjee S, Furey TS. A predictive framework for integrating disparate genomic data types using sample-specific gene set enrichment analysis and multi-task learning. PLoS One. 2012;7(9):e44635. https://doi.org/10.1371/journal.pone.0044635.
26. The Cancer Genome Atlas Database. Available from: https://www.cancer.gov/tcga. Accessed 16 July 2019.
27. Hudson TJ, Anderson W, Aretz A, Barker AD, Bell C, Bernabé RR, et al. International network of cancer genome projects. Nature. 2010;464(7291):993–8. https://doi.org/10.1038/nature08987.
28. Zhao Q, Shi X, Xie Y, Huang J, BenShia C, Ma S. Combining multidimensional genomic measurements for predicting cancer prognosis: observations from TCGA. Brief Bioinform. 2015;16(2):291–303. https://doi.org/10.1093/bib/bbu003.
29. Kim D, Shin H, Song YS, Kim JH. Synergistic effect of different levels of genomic data for cancer clinical outcome prediction. J Biomed Inform. 2012;45(6):1191–8. https://doi.org/10.1016/j.jbi.2012.07.008.
30. Li W, Zhang S, Liu CC, Zhou XJ. Identifying multi-layer gene regulatory modules from multi-dimensional genomic data. Bioinformatics. 2012;28(19):2458–66. https://doi.org/10.1093/bioinformatics/bts476.
31. Wang W, Baladandayuthapani V, Morris JS, Broom BM, Manyam G, Do KA. IBAG: integrative Bayesian analysis of high-dimensional multiplatform genomics data. Bioinformatics. 2013;29(2):149–59. https://doi.org/10.1093/bioinformatics/bts655.
32. Menezes RX, Boetzer M, Sieswerda M, van Ommen GJB, Boer JM. Integrated analysis of DNA copy number and gene expression microarray data using gene sets. BMC Bioinformatics. 2009;10(1):203. https://doi.org/10.1186/1471-2105-10-203.
33. Soneson C, Lilljebjörn H, Fioretos T, Fontes M. Integrative analysis of gene expression and copy number alterations using canonical correlation analysis. BMC Bioinformatics. 2010;11(1):191. https://doi.org/10.1186/1471-2105-11-191.
34. Xu C, Liu Y, Wang P, Fan W, Rue TC, Upton MP, et al. Integrative analysis of DNA copy number and gene expression in metastatic oral squamous cell carcinoma identifies genes associated with poor survival. Mol Cancer. 2010;9(1):143. https://doi.org/10.1186/1476-4598-9-143.
35. Lu TP, Lai LC, Tsai MH, Chen PC, Hsu CP, Lee JM, et al. Integrated analyses of copy number variations and gene expression in lung adenocarcinoma. PLoS One. 2011;6(9):e24829. https://doi.org/10.1371/journal.pone.0024829.
36. Gerstung M, Pellagatti A, Malcovati L, Giagounidis A, Della Porta MG, Jädersten M, et al. Combining gene mutation with gene expression data improves outcome prediction in myelodysplastic syndromes. Nat Commun. 2015;6(1):5901. https://doi.org/10.1038/ncomms6901.
37. Yang Q, Xiong Y, Jiang N, Zeng F, Huang C, Li X. Integrating genomic data with transcriptomic data for improved survival prediction for adult diffuse glioma. J Cancer. 2020;11(13):3794–802. https://doi.org/10.7150/jca.44032.
38. Song Y, Chen D, Zhang X, Luo Y, Li S. Integrating genetic mutations and expression profiles for survival prediction of lung adenocarcinoma. Thorac Cancer. 2019;10(5):1220–8. https://doi.org/10.1111/1759-7714.13072.
39. Zhang Y, Yang W, Li D, Yang JY, Guan R, Yang MQ. Toward the precision breast cancer survival prediction utilizing combined whole genome-wide expression and somatic mutation analysis. BMC Med Genet. 2018;11(S5):104. https://doi.org/10.1186/s12920-018-0419-x.
40. Kim YW, Koul D, Kim SH, Lucio-Eterovic AK, Freire PR, Yao J, et al. Identification of prognostic gene signatures of glioblastoma: a study based on TCGA data analysis. Neuro-Oncology. 2013;15(7):829–39. https://doi.org/10.1093/neuonc/not024.
41. The UCSC Xena Datahub. Available from: http://xena.ucsc.edu/. Accessed 16 July 2019.
42. Driscoll JJ, Rixe O. Overall survival: still the gold standard: why overall survival remains the definitive end point in cancer clinical trials. Cancer J. 2009;15(5):401–5.
43. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci U S A. 2005;102(43):15545–50. https://doi.org/10.1073/pnas.0506580102.
44. Liberzon A, Birger C, Thorvaldsdóttir H, Ghandi M, Mesirov JP, Tamayo P. The molecular signatures database Hallmark gene set collection. Cell Syst. 2015;1(6):417–25. https://doi.org/10.1016/j.cels.2015.12.004.

45. Tate JG, Bamford S, Jubb HC, Sondka Z, Beare DM, Bindal N, et al. COSMIC: the catalogue of somatic mutations in Cancer. Nucleic Acids Res. 2019; 47(D1):D941–7. https://doi.org/10.1093/nar/gky1015.
46. Hänzelmann S, Castelo R, Guinney J. GSVA: gene set variation analysis for microarray and RNA-seq data. BMC Bioinformatics. 2013;14(7):1–5.
47. HALDANE BJBS. The estimation and significance of the logarithm of a ratio of frequencies. Ann Hum Genet. 1956;20(4):309–11. https://doi.org/10.1111/j.1469-1809.1955.tb01285.x.
48. Anscombe FJ. On estimating binomial response relations. Biometrika. 1956; 43(3):s461–4.
49. Tibshirani R. Regression Selection and Shrinkage via the Lasso. J Royal Stat Soc B. 1996;58:267–88.
50. Friedman J, Hastie T, Tibshirani R. Regularization paths for generalized linear models via coordinate descent. J Stat Softw. 2010;33(1):1–22. https://doi.org/10.18637/jss.v033.i01.
51. Yuan M, Lin Y. Model selection and estimation in regression with grouped variables. J R Stat Soc Ser B Stat Methodol. 2006;68(1):49–67. https://doi.org/10.1111/j.1467-9868.2005.00532.x.
52. Breheny P, Huang J. Group descent algorithms for nonconvex penalized linear and logistic regression models with grouped predictors. Stat Comput. 2013;25(2):173–87. https://doi.org/10.1007/s11222-013-9424-2.
53. Harrell FE. Evaluating the yield of medical tests. JAMA. 1982;247(18):2543–6. https://doi.org/10.1001/jama.1982.03320430047030.
54. McHugh ML. Interrater reliability: the kappa statistic. Biochem Med. 2012; 22(3):276–82. https://doi.org/10.11613/BM.2012.031.
55. Fleiss JL. Measuring nominal scale agreement among many raters. Psychol Bull. 1971;76(5):378–82. https://doi.org/10.1037/h0031619.
56. Huang Z, Zhan X, Xiang S, Johnson TS, Helm B, Yu CY, et al. Salmon: Survival analysis learning with multi-omics neural networks on breast cancer. Front Genet. 2019;10:166.
57. Signorell A. DescTools: Tools for descriptive statistics. R Packag version 09938; 2020.
58. Krassas GE, Poppe K, Glinoer D. Thyroid function and human reproductive health. Endocr Rev. 2010;31(5):702–55. https://doi.org/10.1210/er.2009-0041.
59. Krajewska-Kulak E, Sengupta P. Thyroid function in male infertility. Front Endocrinol. 2013;4:174.
60. Wagner MS, Wajner SM, Maia AL. The role of thyroid hormone in testicular development and function. J Endocrinol. 2008;199(3):351–65. https://doi.org/10.1677/JOE-08-0218.
61. Wen SS, Zhang TT, Xue DX, Wu WL, Wang YL, Wang Y, et al. Metabolic reprogramming and its clinical application in thyroid cancer (review). Oncol Lett. 2019;18(2):1579–84. https://doi.org/10.3892/ol.2019.10485.
62. Jardillier R, Guyon L. Benchmark of lasso-like penalties in the Cox model for TCGA datasets reveal improved performance with pre-filtering and wide differences between cancers. bioRxiv Bioinforma. 2020. https://doi.org/10.1101/2020.03.09.984070.
63. Zheng X, Amos CI, Frost HR. Comparison of pathway and gene-level models for cancer prognosis prediction. BMC Bioinformatics. 2020;21(76):1–7.
64. Hu WL, Zhou XH. Identification of prognostic signature in cancer based on DNA methylation interaction network. BMC Med Genet. 2017;10(4):63. https://doi.org/10.1186/s12920-017-0307-9.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.