BMC Research Notes

RESEARCH NOTE

Open Access

Strategies for improving the performance of prediction models for response to immune checkpoint blockade therapy in cancer

Tiantian Zeng[1*], Jason Z. Zhang[2], Arnold Stromberg[1], Jin Chen[3] and Chi Wang[4*]

Abstract

Immune checkpoint blockade (ICB) therapy holds promise for bringing long-lasting clinical gains for the treatment of cancer. However, studies show that only a fraction of patients respond to the treatment. In this regard, it is valuable to develop gene expression signatures based on RNA sequencing (RNAseq) data and machine learning methods to predict a patient's response to the ICB therapy, which contributes to more personalized treatment strategy and better management of cancer patients. However, due to the limited sample size of ICB trials with RNAseq data available and the vast number of candidate gene expression features, it is challenging to develop well-performed gene expression signatures. In this study, we used several published melanoma datasets and investigated approaches that can improve the construction of gene expression-based prediction models. We found that merging datasets from multiple studies and incorporating prior biological knowledge yielded prediction models with higher predictive accuracies. Our finding suggests that these two strategies are of high value to identify ICB response biomarkers in future studies.

Keywords Immune checkpoint blockade (ICB) therapy, RNA sequencing, Predictive model, Machine learning

Introduction

Immunotherapy has emerged recently as a promising and viable treatment option for many cancer patients [1]. Among multiple types of immunotherapy, the immune checkpoint blockade (ICB) therapy, which aims at blocking the interaction of inhibitory receptors expressed on the surface of immune cells [2], has been proved applicable in helping the immune system target and attack cancer cells [3, 4]. Particularly, ICB can provide exceptional clinical gains in the treatment of a handful cancer, melanoma, mostly because the spontaneous regression of melanoma is closely related to the immune response [5, 6]. Despite the success of ICB therapy in the treatment of melanoma, however, recent studies showed that only around one-third of patients would respond to the ICB therapy [7]. The reason for the heterogeneous response still remains unclear and requires further investigations [1, 8, 9].

It is therefore desired to develop biomarkers that can predict patient's response to ICB therapy, which will contribute to better stratification of patients to maximize therapeutic benefit. Previous studies showed that tumor mutational burden, microsatellite instability are predictive biomarkers [10–12]. Gene expression signatures have also been demonstrated as valuable for predicting ICB

*Correspondence:
Tiantian Zeng
zengtiantian009@gmail.com
Chi Wang
chi.wang@uky.edu
[1] Department of Statistics, University of Kentucky, 725 Rose St, Lexington, KY 40536, USA
[2] Wake Forest University, Winston-Salem, NC 27109, USA
[3] Department of Medicine - Nephrology, University of Alabama at Birmingham, 703 19th St S, Birmingham, AL 35294, USA
[4] Department of Internal Medicine, University of Kentucky, 800 Rose St, Lexington, KY 40536, USA

treatment response in melanoma patients [13–17]. However, the sample sizes of ICB clinical studies that have gene expression profiling data available are very limited [13]. Table 1 listed three published studies, each of which had less than 60 patients. The lack of large scale datasets makes it challenging to construct reliable prediction models. Although an alternative approach of using data from patients without ICB treatment to develop immune response signatures and transfering the results to predict ICB treatment response has been proposed [13, 14], it is still highly desired if the signatures could be directly built on patients with ICB treatment. Further, gene expression profiling technologies, such as RNA-sequencing (RNA-seq), are powerful to simultaneously quantify more than 10,000 genes' expression levels. It is challenging to identify informative gene features and their complex relationships to build accurate prediction models, especially when the sample size is small.

In this paper, we investigated the potential of the following two strategies to enhance the development of gene expression signatures for ICB treatment effect prediction in melanoma patients. The first strategy is to merge data from different ICB clinical studies. Merging datasets has been shown as a viable approach to increase the sample size and thus improve the power of biomarker development in various biomedical applications [18, 19]. We explored the potential benefit of merging three published datasets [15–17] for the prediction of ICB treatment response. The second strategy is to leverage prior biological knowledge to use more informative and biologically relevant features for the construction of prediction models. It has been suggested that expressions of immune checkpoint genes and their interactions are relevant to tumor response to ICB therapy [13, 14]. We explored whether focusing on pairwise relation features among these immune checkpoint genes, as suggested by [14], could improve feature selection and prediction performance of the models.

Methods

Study design

Figure 1 presents an overview of our study design. Firstly, using each individual RNA-seq dataset, we leveraged prior biological knowledge and focused on immune checkpoint genes, where the pairwise relation of those genes were considered as candidate features for subsequent prediction model development. Secondly, we merged individual datasets to increase the sample size, where the presence of batch effect was assessed. Thirdly, based on the merged data, we built prediction models based on three commonly used machine learning algorithms, i.e. Random forest [20], Least absolute shrinkage and selection operation (LASSO) [21], and XGBoost [22]. Fourthly, we evaluated the performance of prediction models based on the receiver operating characteristic (ROC) curve and area under the curve (AUC) using cross-validation. To investigate the power of merging multiple datasets, we compared AUCs from models based on the merged dataset versus those based on each individual dataset. Besides, in order to investigate the benefit of incorporating prior information in feature selection, we compared AUCs from models built based on features characterizing pairwise relation of immune checkpoint genes versus those based on the original expression features of all genes.

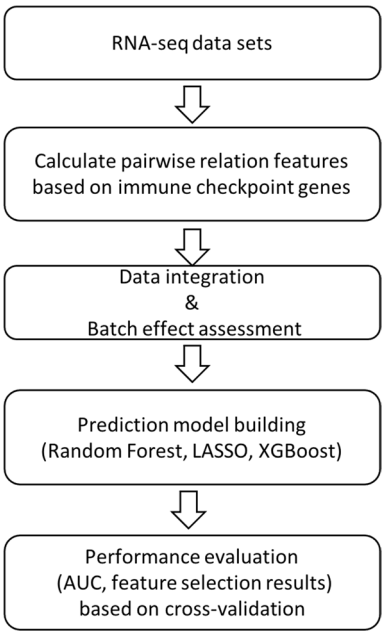

Fig. 1 The flowchart of this study

Datasets

We considered three published melanoma datasets [15–17] as listed in Table 1. All the datasets we utilized were sourced from studies concentrating on the immune treatment of melanoma. Consequently, any additional datasets incorporated should also originate from melanoma research. For Van Allen et al. and Hugo et al, gene expression data were provided in the units of fragments per kilobase of transcript per million mapped reads (FPKM). For Riaz et al., the data were provided in counts. We uniformly transformed all the data to the unit of transcripts per million (TPM) before our analysis. There were 18,878 common gene features across the three datasets. The total sample size was 125. The detailed information of the three datasets is summarized in Table 1.

The response variable was defined to be a binary variable, 'response' or 'non-response' to the treatment.

Table 1 Datasets used in the study

	Van Allen et al. [15]	Hugo et al. [16]	Riaz et al. [17]
Accession ID	dbGaP phs000452.v2.p1.	GEO GSE78220	GEO GSE91061
Number of subjects	42	27	56
Number of genes	57731	25268	20771

Since the response annotations for each dataset were not exactly the same, a standard classification used in this study was given following the definition from Auslander paper [14], where 'complete response', and 'partial response' were classified as 'response', and 'nonresponse', 'progressive disease', etc. were classified as 'non-response'. A patient characteristics table is provided in Additional file 1: Appendix C.

Features characterizing pairwise relations of immune checkpoint genes

Due to the large number of gene features in one dataset, we focused on immune checkpoint genes and considered a similar set of candidate features as in [14]. The authors proposed to use the pairwise relations between the expressions of immune checkpoint genes as features to develop prediction models for immune checkpoint blockade therapy. They formed a comprehensive list of 28 immune checkpoint genes, known for their costimulatory or co-inhibitory functions, as documented in previous studies [23–26]. It is expected that essential immune interactions are encapsulated through specific pairwise relations of immune checkpoint genes. Among those immune checkpoint genes, six genes, i.e. PD-1, PD-L1, CTLA-4, CD28, CD80 and CD86, are directly associated with anti-CTLA-4 and anti-PD-1 blockade therapies [14, 23–27], which are two major types of immune checkpoint blockade therapy. Auslander et al. focused on pairwise relations containing at least one of the six genes. In this paper, we considered a similar set of pairwise relations. The only difference is that we only included 26 out of the 28 gene considered in Auslander et al. [14] because the other two genes do not have expression data available across all RNA-seq datasets analyzed in our study.

For a gene pair x and y, we define the following expression function was used:

$f_{x,y}(i) = 1$, if $exp_x(i) > exp_y(i)$;

$f_{x,y}(i) = 0$, otherwise,

where $exp_x(i)$ and $exp_y(i)$ denote expressions of x and y in sample i. Since we further focused on the gene pairs containing at least one of the six genes, i.e. PD-1, PD-L1, CTLA-4, CD28, CD80 and CD86, that are directly associated with anti-CTLA-4 and anti-PD-1 blockade therapy as stated above, we obtained a total of 135 pairs forming candidate features for building prediction models.

Data integration

To integrate datasets from different sources, we applied the following procedure to integrate datasets from different sources. First, we ensured to merge by the common genes, and uniformly transformed all the RNA-seq data to the unit of TPM. The response variable was also uniformly defined across datasets as stated in the above section. Next, we calculated the pairwise relations between the expressions of immune checkpoint genes. Note that the pairwise relation features only consider the order of expressions between genes, but not the quantitative expression levels. Therefore, it can reduce the impact of non-biological experimental variations, i.e. batch effect, in the analysis. Finally, we visually assessed the possible batch effect and outliers in data integration by using heatmaps [28, 29]with hierarchical clustering as well as UMAP plots [30, 31]

Prediction model building

We considered the following three frequently used statistical/machine learning methods to build models for predicting response to immune checkpoint blockade therapy based on features of the pairwise relations between immune checkpoint genes.

Random Forest Random forest is a well established ensemble learning algorithm that can be applied for classification. It is formed by a large amount of individual decision trees, and then operates as an ensemble. Random forest applies a widespread technique of bagging, or called bootstrap aggregating while training the algorithm, but it includes implementing an essential modification of bagging in order to obtain an ensemble of de-correlated trees [20]. The feature selection is reflected in the "Gini importance" metric, which serves as an indicator of feature relevance, offering a comparative ranking of features derived as a secondary outcome during the classifier's training process [32, 33]. In our random forest model, features are ranked according to their Gini index, with the model selecting the highest-ranked features for use. Package 'randomForest' in R (Version: 4.6-14) was used in this study.

Lasso Least absolute shrinkage and selection operation (LASSO) is a well-known method in machine learning, especially for the datasets that have more number of features than number of observations. In regression

analysis, LASSO can perform feature selection and regularization at the same time, so as to improve the accuracy of model prediction performance as well as strengthen the interpretability of the obtained model [21]. Lasso could also be applied for classification problem. Lasso model employs regularization to penalize regression coefficients, reducing some to zero. Variables with non-zero coefficients after this process are chosen for the model, aiming to minimize prediction error [21, 34]. The function 'glmnet' in R (Version: 4.1-1) was used for the LASSO model building, while setting family to 'binomial' could build classifiers for the binary outcome.

XGBoost XGBoost is an implementation of the gradient boosted decision tree algorithm. Boosting is also an ensemble algorithm that can combine the output of many weak classifiers into a strong one. The algorithm enables to work on both classification and regression problems. XGBoost, which is defined as a scalable end-to-end tree boosting system, is a very strong boosting method that can be used to build a classifier, and exhibits outstanding prediction performance according to recent studies [22]. It conducts feature selection by assigning importance scores to features based on their contribution to node purity and model performance. This process is integrated into its training, where an ensemble of decision trees prioritizes more significant features, inherently filtering out less relevant ones [22]. The 'xgboost' package in R (Version: 1.4.1.1) allows for applying XGBoost in the classification problem in R.

Performance evaluation

The prediction performance of models based on random forest [20], LASSO [21] and XGBoost [22] were evaluated and compared by using tenfold cross validation. The merged dataset was randomly divided into ten folds. Each time, the model was built on a combination of the nine folds, and then evaluated on the leave-out fold. The ROC and AUC were calculated to measure the predictive accuracy of a prediction model [35]. The whole cross validation procedure was replicated 10 times and averaged results were reported. In an ROC curve, the Y-axis indicates sensitivity, while the X-axis indicates 1-specificity. Each point on the ROC curve represents a sensitivity/specificity pair under a given threshold. Therefore, the ROC curve provides a comprehensive comparison of sensitivity versus specificity over various thresholds for predicting binary outcomes. Furthermore, the area under the ROC curve serves as an additional metric for evaluating the overall performance of a prediction model.

The feature selection results of these machine learning methods were assessed by calculating the probability of each feature being selected, i.e. the proportion of times that the feature was included in the prediction model based on the tenfold cross validation. The Spearman's correlation coefficient was calculated to measure similarity in feature selection between every two methods [36]. The Heatmap and correlation plot were generated to visualize and compare the results.

Results

Candidate features

Given the extensive number of gene features in RNA-seq datasets, our analysis concentrated on immune checkpoint genes, aligning with the candidate features outlined in the research paper from Auslander et al. [14]. We selected 26 immune checkpoint genes, previously identified in the literature, and present within the RNA-seq datasets. Recognizing the co-stimulatory or co-inhibitory nature of these genes, we examined the pairwise interactions between their expression levels [14, 23–26]. To do this, we utilized indicator functions to compare the expression levels of pairs of immune checkpoint genes, capturing their intricate relationships. A total of 135 pairwise relations features were considered as candidate features in our analysis.

Data integration

To enhance statistical power, we merged three published datasets, i.e. Van Allen et al. [15], Hugo et al. [16], and Riaz et al. [17] (Table 1), and obtained a combined dataset with 135 subjects. We first assessed the batch effect for data from different sources. Figure 2A shows the hierarchical clustering of samples from those sources based on the original gene expression data. It is clear that samples from the same source were clustered together, suggesting the presence of batch effect. In contrast, Fig. 2B shows the hierarchical clustering of samples based on the 135 features characterizing the pairwise relations of immune checkpoint genes. Samples from different sources were intermixed. Thus, by focusing on those pairwise relations features, the batch effect was minimized. This is likely due to the fact that pairwise-relation features focus on the relative orders between expressions of pairs of genes. Such information is more robust against batch effect compared to the original gene expression levels. We also generated UMAP plots on the original features as well as pairwise relation features, see Additional file 1: Appendix B. Those UMAP plots show that samples from different batches came much closer when considering the pairwise relation features, as compared to the original gene expression features. In addition, the UMAP based on the pairwise relation features did not indicate the presence of outliers because there was no subject that was far away from all other subjects.

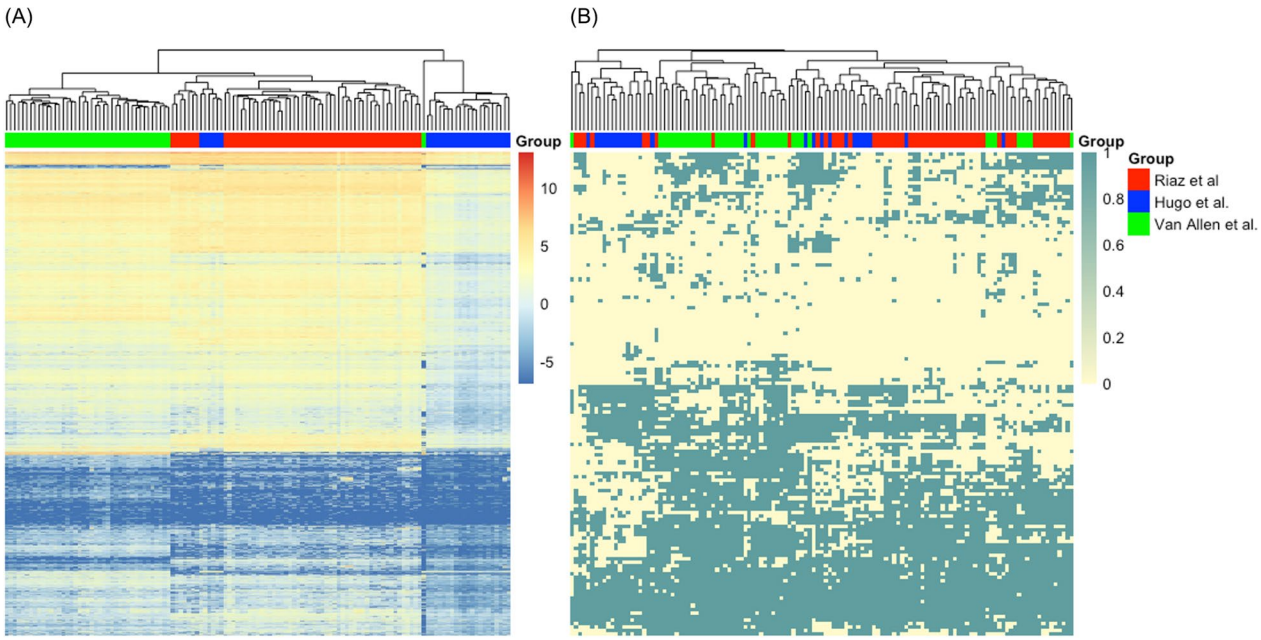

Fig. 2 Heatmaps and sample clustering of the merged datasets based on **A** the original 18,878 features and **B** the 135 features charaterizing the pairwise relations of immune checkpoint genes

Model prediction results

We applied three frequently used statistical/machine learning methods, including random forest [20], least absolute shrinkage and selection operation (LASSO) [21], and XGBoost [22], to build prediction models. Figure 3 presents the ROC curves and AUCs of the prediction models built by the above-mentioned machine learning methods based on tenfold cross validation. Random forest, and LASSO both had AUCs above 0.7, providing good predictions of the immune response. The XGBoost had a lower AUC of 0.667. The difference in prediction performance across methods is likely due to different feature selection or model building strategies of these methods.

We next investigated the impact on model's predictive accuracy by using the combined dataset versus using a single dataset. We applied the same three machine learning methods to build prediction models based on each

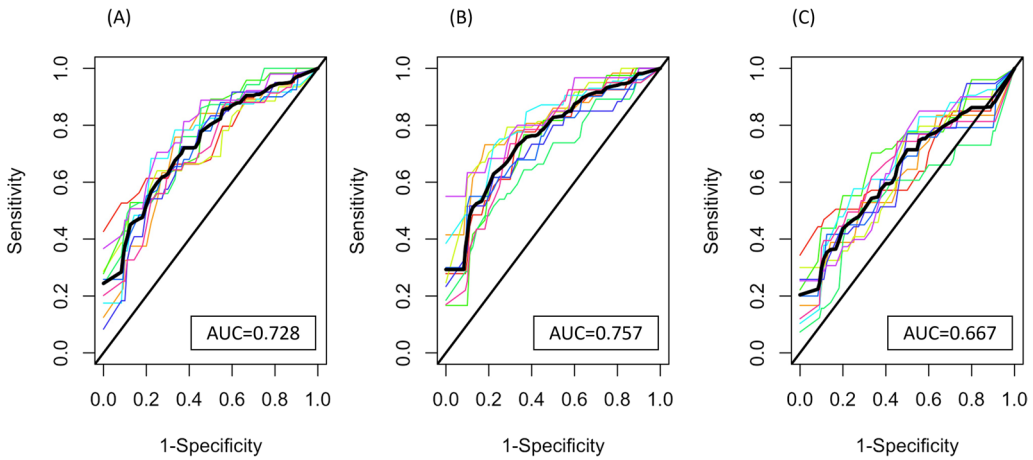

Fig. 3 ROC curves of the combined dataset for all algorithms Comparison of ROC curves among Random Forest (**A**), LASSO (**B**), and XGBoost (**C**) based on the combined dataset. Each colored dashed curve indicates one tenfold cross validation replicate. The solid black curve indicates the average curve across ten replicates. Results were averaged across ten tenfold cross-validations. The average area under the curve (AUC) were calculated over the ten replicates

of the three individual datasets. As shown in Table 2, the AUC from a single dataset was lower than that from the combined dataset. For example, the LASSO AUCs from the single datasets were 0.547, 0.464, and 0.729, respectively, while that from the combined dataset was 0.757. In addition, the ROC curves from different cross-validation replicates had much larger variations (Additional file 1: Appendix D), suggesting that the prediction performance of models based on a single dataset is less stable than that based on the combined dataset. Further, the AUCs from the Hugo et al. dataset tended to be lower than those from the other two datasets for a given machine learning method. This is likely due to the fact that the Hugo et al. dataset had a smaller sample size compared to the other two datasets, which further demonstrates the importance of sample size in prediction model development. In addition, we also considered combining the two datasets from Van Allen et al. and Riaz et al. given the small sample size of Hugo et al. We employed the same three methodologies to construct the predictive models. The results are shown in Additional file 1: Appendix E, where AUCs from the three models were 0.743, 0.704, and 0.689, respectively. The results indicate that combining two datasets yielded a higher AUC than using a single dataset, which is consistent with the findings from combining three datasets.

In addition, we assessed the value of incorporating prior biological information and focusing on immune checkpoint genes. As a comparison, we applied the three machine learning methods and feature selection/model building procedure to the original 18,878 gene expression features based on either a single dataset or the combined dataset. For the combined dataset, the Combat normalization method had been applied to remove batch effect [37, 38]. Table 2 shows that the AUCs of those resulting models were only around 0.5, even for using the combined dataset. Thus, those models based on original gene expression features without incorporating prior biological knowledge had much poorer prediction performance compared to models using pairwise relationships of immune checkpoint genes. The result indicates that by suggesting biologically relevant features and their combinations, prior biological knowledge can contribute to building better performed prediction models.

Feature selection results

We also compared the features selected by the three statistical/machine learning methods. For each method, the probability of a feature being selected was calculated based on the cross-validation procedure. Features with selection probabilities greater or equal to 0.2 from at least one of the three methods are presented in Fig. 4A. Some features, such as 'PD-1 > PDL-1', 'PD-1 > CTLA4', 'PD-1 > CD200R1', 'PD-1 > TNFRSF18', 'PD-1 > CD137L', 'PDL-1 > CTLA4', 'CTLA4 > CD200R1', 'CD80 > CD137L', 'CD86 > IL2RB', tended to be selected by all the methods with similar probabilities, while some other features had very different selection probabilities for different methods. A table with detailed records of probabilities is provided in Additional file 1: Appendix A.

We further quantified the consistency in feature selection between each pair of methods based on the Spearman correlation coefficient. Results are presented in Fig. 4B. The two tree-based methods, random forest and XGBoost, had high consistency in feature selection. The Spearman's correlation coefficients was around 0.82. In contrast, the features selected by LASSO were very different from the tree-based methods with the Spearman's correlation coefficients less than 0.1. This is likely due to the fact that tree-based methods focus on non-linear combinations while LASSO focuses on linear combinations across features.

Table 2 Summary of perdiction performance

	Build models based on the 135 pairwise relations features of immune checkpoint genes			Build models based on the original 18,878 gene expression features		
	Random forest	Lasso	XGBoost	Random forest	Lasso	XGBoost
	AUC	AUC	AUC	AUC	AUC	AUC
Combined	0.728	0.757	0.667	0.595*	0.552*	0.582*
Van Allen et al.	0.723	0.547	0.607	0.410	0.444	0.446
Hugo et al.	0.559	0.464	0.445	0.671	0.559	0.322
Riaz et al.	0.711	0.729	0.622	0.568	0.441	0.447

Models were built using Random forest, Lasso, XGBoost based on the combined dataset or an individual dataset. AUCs were calculated based on tenfold cross validation except for the case of using the Hugo et al. dataset alone, where a fivefold cross validation was performed because the sample size of the dataset was so small that the tenfold validation did not yield robust result

*Data had been normalized based on the Combat method [37, 38] when merging the three datasets

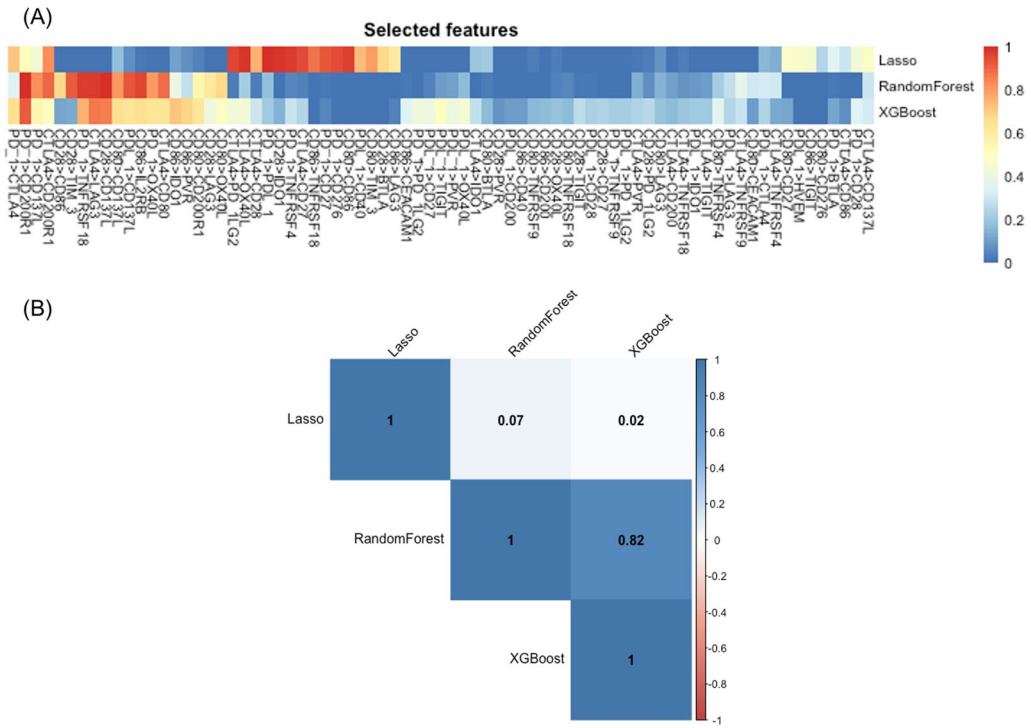

Fig. 4 Feature selection results. **A** Comparison of feature selection probabilities across LASSO, random forest and XGBoost. Only features with probability of selection larger or equal to 0.2 from at least one of the three methods are presented. **B** Spearman correlation of feature selection probabilities between each pair of methods

Discussion

We considered three frequently used statistical/machine learning methods to construct prediction models. For all three methods, models based on merged data had higher predictive accuracy than those based on individual datasets. This result suggested that the improved prediction performance is not sensitive to the choice of model construction method. We also noticed some difference in predictive accuracy between models from different machine learning methods. Further investigation of methods' predictive accuracy under different sample size settings will be needed to more comprehensively evaluate and compare the performance of different methods for predicting ICB treatment response in melanoma patients.

We focused on interactions between immune checkpoint genes as candidate features and followed Auslander et. al to use logical relations between the expression levels of pairs of immune checkpoint genes as candidate features to characterize interactions between those genes [14]. One can consider other function forms, e.g. products of expression levels of pairs of genes, to describe the co-stimulatory and co-inhibitory effects. An interesting topic for future research is to compare different function forms and identify more informative function forms to enhance prediction.

There are other factors, such as tumor heterogeneity, comorbidities, genetic variations, mutational burden, and immune cell infiltration, that could affect the response to ICB therapy. However, the current sample size from a clinical study is inadequate for creating an all-encompassing model for all the potential important factors. In fact, a primary objective of our study is to investigate the feasibility of combining data from multiple studies to increase the sample size for constructing predictive models. Nonetheless, the resultant sample size remains inadequate for encompassing all types of features. Therefore, we focused on gene expression features in this study. To further narrow down the numbers of features we need to include in the analysis, we leverage prior biological knowledge to only consider immune checkpoint genes and their interactions, which have been shown to be relevant to tumor response to ICB therapy. We hope that our study will provide a viable approach for predictive model development under the practical situation where the sample size is limited. But we acknowledge that ignoring other factors, such as tumor heterogeneity, comorbidities, genetic variations, mutational burden, and immune cell infiltration, is a limitation of our current method. With the accumulation of more clinical studies on ICB therapy and the feasible combination across datasets as demonstrated in this paper, we believe that those other

factors could be incorporated in the future to improve the model performance.

Removing batch effect is an important task when merging different datasets. We showed that features on pairwise relations between immune checkpoint genes were less affected by batches compared to the original features. This is because the pairwise relation features only consider the order of expressions between genes but not the quantitative expression levels. Therefore, focusing on pairwise relation features reduces the impact of batch effect in our analysis. We also tried a more traditional approach, ComBat [37, 38], for removing batch effect. However, we noticed that the ComBat adjusted expression values did not pertain the order of expressions between genes. Therefore, such batch effect removal may cause disturbance of the useful information contained in the original data. In addition, another problem with ComBat is that it requires all datasets to be analyzed together, where the batch effect removal modifies gene expression values in all the datasets. A model developed based on batch effect corrected training datasets cannot be directly applied to a new independent dataset since the new dataset needs to be batch corrected first. But ComBat would require jointly analyzing the new dataset and the training datasets, where the expression values of not only the new dataset but also the training datasets will be modified. As a result, the prediction model will have to be re-built based on the modified expression values in the training datasets. Therefore, the generalizability of the prediction model based on ComBat is limited.

The ICB therapy for melanoma primarily targets cytotoxic T-lymphocyte-associated protein 4 (CTLA-4) and programmed cell death protein 1 (PD-1) [27]. Both CTLA-4 and PD-1 binding have similar negative regulatory effects on the activity of T-cells. Anti-PD-1 and Anti-CTLA-4 immunotherapies inhibit these targets and prevent melanoma cells from evading the immune system [27]. In the literature, it has been shown that gene expression signatures can be generally applicable to predict treatment effects of both anti-CTLA-4 and anti-PD-1 drugs [13, 14]. Therefore, in this paper, we combined datasets for drugs targeting CTLA-4 and datasets for drugs targeting PD-1 to ensure an adequate sample size for model building [15–17]. In the future, as clinical trial data acumulate, one can possibly focus on trials for drugs targeting one of the two targets to develop target-specific gene expression signatures. It would be interesting to investigate whether those gene signatures could further improve the predictive accuracy of treatment effect.

Conclusion

In summary, we have demonstrated that merging datasets and incorporating prior biological knowledge are useful strategies to improve the prediction performance of ICB treatment using gene expression signatures. The batch effect could be minimized by capturing pairwise-relation features. Classical machine learning algorithms were applied to the integrated datasets with features of pairwise relations, and demonstrated satisfactory classification performance, with AUC around 0.70. When compared with the model built on the single dataset, the result showed that the model with dataset merging improved and stabilized the prediction performance. In addition, the prediction performance of models based on the pairwise relations of immune checkpoint genes was higher than models built on the original dataset without incorporating prior biological knowledge. Overall, our finding demonstrated that merging datasets from multiple studies and incorporating prior biological knowledge are of high value to identify ICB response biomarkers in future studies.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s13104-024-06760-5.

Additional file 1: Table S1. Feature selection with different algorithms. **Figure S1**. UMAP for the merged datasets based on (a) the original 18878 features and (b) the 135 features characterizing the pairwise relations of immune checkpoint genes. **Table S2**. Clinical Characteristics. **Figure S2**. Random forest - ROC plot for thecomparison of the result. Comparison of ROC curves with applying Random Forest on each single dataset. (A) Cross validation using single dataset Van Allen et al.; (B) Cross validation using single dataset Hugo et al.; (C) Cross validation using single dataset Riaz et al. **Figure S3**. Lasso - ROC plot for the comparison of the result. Comparison of ROC curves with applying Lasso on each single dataset. (A) Cross validation using single dataset Van Allen et al.; (B) Cross validation using single dataset Hugo et al.; (C) Cross validation using single dataset Riaz et al. **Figure S4**. XGBoost - ROC plot for the comparison of the result. Comparison of ROC curves with applying XGBoost on each single dataset. (A) Cross validation using single dataset Van Allen et al.; (B) Cross validation using single dataset Hugo et al.; (C) Cross validation using single dataset Riaz et al. **Figure S5**. ROC curves of the combined two datasets (Van Allen et al. and Riaz et al.) for all algorithms.

Acknowledgements

The authors acknowledge Daheng He for the help in RNA-seq data processing.

Author contributions

The authors confirm contribution to the paper as follows: study conception and design: CW, TZ; method exploration: TZ, JZ, JC, CW; analysis and interpretation of results: TZ, CW; draft manuscript preparation: TZ, JZ, AS, JC, CW.

Funding

This work is supported by the Biostatistics and Bioinformatics Shared Resource Facility of the University of Kentucky Markey Cancer Center [P30CA177558]. The Van Allen et al. dataset downloaded from dbGaP was supported by the National Human Genome Research Institute (NHGRI) Large Scale Sequencing Program, Grant U54 HG003067 to the Broad Institute (PI, Lander).

Availability of data and materials
The supporting data can be available to public.

Code availability
The code can be available to public.

Declarations

Ethics approval and consent to participate
This study is approved by the University of Kentucky Institutional Review Board.

Consent for publication
Not applicable

Competing interests
The authors have no competing of interest to declare.

Received: 22 December 2023 Accepted: 27 March 2024
Published online: 09 April 2024

References

1. Mahoney KM, Rennert PD, Freeman GJ. Combination cancer immunotherapy and new immunomodulatory targets. Nat Rev Drug Discov. 2015;14(8):561–84.
2. Petitprez F, Meylan M, de Reyniès A, Sautès-Fridman C, Fridman WH. The tumor microenvironment in the response to immune checkpoint blockade therapies. Front Immunol. 2020;11:533888.
3. Hargadon KM, Johnson CE, Williams CJ. Immune checkpoint blockade therapy for cancer: an overview of FDA-approved immune checkpoint inhibitors. Int Immunopharmacol. 2018;62:29–39.
4. Pardoll DM. The blockade of immune checkpoints in cancer immunotherapy. Nat Rev Cancer. 2012;12(4):252–64.
5. Aung PP, Nagarajan P, Prieto VG. Regression in primary cutaneous melanoma: etiopathogenesis and clinical significance. Lab Invest. 2017;97(6):657–68.
6. Bramhall RJ, Mahady K, Peach AHS. Spontaneous regression of metastatic melanoma-clinical evidence of the abscopal effect. Eur J Surg Oncol (EJSO). 2014;40(1):34–41.
7. Sharma P, Hu-Lieskovan S, Wargo JA, Ribas A. Primary, adaptive, and acquired resistance to cancer immunotherapy. Cell. 2017;168(4):707–23.
8. Ayers M, Lunceford J, Nebozhyn M, Murphy E, Loboda A, Kaufman DR, Albright A, Cheng JD, Peter Kang S, Shankaran V, et al. IFN-γ-related mRNA profile predicts clinical response to PD-1 blockade. J Clin Investig. 2017;127(8):2930–40.
9. Gao J, Aksoy BA, Dogrusoz U, Dresdner G, Gross B, Sumer SO, Sun Y, Jacobsen A, Sinha R, Larsson E, et al. Integrative analysis of complex cancer genomics and clinical profiles using the cbioportal. Sci Signal. 2013;6(269):pl1.
10. Klempner SJ, Fabrizio D, Bane S, Reinhart M, Peoples T, Ali SM, Sokol ES, Frampton G, Schrock AB, Anhorn R, et al. Tumor mutational burden as a predictive biomarker for response to immune checkpoint inhibitors: a review of current evidence. Oncologist. 2020;25(1): e147.
11. Vanderwalde A, Spetzler D, Xiao N, Gatalica Z, Marshall J. Microsatellite instability status determined by next-generation sequencing and compared with PD-l1 and tumor mutational burden in 11,348 patients. Cancer Med. 2018;7(3):746–56.
12. Davis AA, Patel VG. The role of PD-l1 expression as a predictive biomarker: an analysis of all us food and drug administration (FDA) approvals of immune checkpoint inhibitors. J Immunother Cancer. 2019;7(1):1–8.
13. Jiang P, Shengqing G, Pan D, Jingxin F, Sahu A, Xihao H, Li Z, Traugh N, Xia B, Li B, et al. Signatures of t cell dysfunction and exclusion predict cancer immunotherapy response. Nat Med. 2018;24(10):1550–8.
14. Auslander N, Zhang G, Lee JS, Frederick DT, Miao B, Moll T, Tian T, Wei Z, Madan S, Sullivan RJ, et al. Robust prediction of response to immune checkpoint blockade therapy in metastatic melanoma. Nat Med. 2018;24(10):1545–9.
15. Van Allen EM, Miao D, Schilling B, Shukla SA, Blank C, Zimmer L, Sucker A, Hillen U, Foppen MHG, Goldinger SM, et al. Genomic correlates of response to CTLA-4 blockade in metastatic melanoma. Science. 2015;350(6257):207–11.
16. Hugo W, Zaretsky JM, Sun L, Song C, Moreno BH, Hu-Lieskovan S, Berent-Maoz B, Pang J, Chmielowski B, Cherry G, et al. Genomic and transcriptomic features of response to anti-PD-1 therapy in metastatic melanoma. Cell. 2016;165(1):35–44.
17. Riaz N, Havel JJ, Makarov V, Desrichard A, Urba WJ, Sims JS, Stephen Hodi F, Martín-Algarra S, Mandal R, Sharfman WH, et al. Tumor and microenvironment evolution during immunotherapy with nivolumab. Cell. 2017;171(4):934–49.
18. Yasrebi H, Sperisen P, Praz V, Bucher P. Can survival prediction be improved by merging gene expression data sets? PLoS ONE. 2009;4(10):e7431.
19. McDermott JE, Wang J, Mitchell H, Webb-Robertson B-J, Hafen R, Ramey J, Rodland KD. Challenges in biomarker discovery: combining expert insights with statistical analysis of complex omics data. Expert Opin Med Diagn. 2013;7(1):37–51.
20. Hastie T, Tibshirani R, Friedman J. Random forests, the elements of statistical learning. Data Min Inference Pred. 2009;2:587–604.
21. Tibshirani R. Regression shrinkage and selection via the lasso. J R Stat Soc: Ser B (Methodol). 1996;58(1):267–88.
22. Chen T, Guestrin C. Xgboost: a scalable tree boosting system. In: Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining; 2016. pp. 785–94.
23. Chen L, Flies DB. Molecular mechanisms of T cell co-stimulation and co-inhibition. Nat Rev Immunol. 2013;13(4):227–42.
24. Zhang Q, Vignali DAA. Co-stimulatory and co-inhibitory pathways in autoimmunity. Immunity. 2016;44(5):1034–51.
25. Fuertes Marraco SA, Neubert NJ, Grégory V, Speiser DE. Inhibitory receptors beyond T cell exhaustion. Front Immunol. 2015;6:310.
26. Ramsay AG. Immune checkpoint blockade immunotherapy to activate anti-tumour T-cell immunity. Br J Haematol. 2013;162(3):313–25.
27. Buchbinder EI, Desai A. CTLA-4 and PD-1 pathways: similarities, differences, and implications of their inhibition. Am J Clin Oncol. 2016;39(1):98.
28. Scherer A. Batch effects and noise in microarray experiments: sources and solutions, vol. 868. Hoboken: Wiley; 2009.
29. Zuguang G, Eils R, Schlesner M. Complex heatmaps reveal patterns and correlations in multidimensional genomic data. Bioinformatics. 2016;32(18):2847–9.
30. McInnes L, Healy J, James M. Umap: uniform manifold approximation and projection for dimension reduction. arXiv preprint arXiv:1802.03426 (2018).
31. Becht E, McInnes L, Healy J, Dutertre C-A, Kwok IWH, Ng LG, Ginhoux F, Newell EW. Dimensionality reduction for visualizing single-cell data using UMAP. Nat Biotechnol. 2019;37(1):38–44.
32. Breiman L. Random forests. Mach Learn. 2001;45:5–32.
33. Menze BH, Michael Kelm B, Masuch R, Himmelreich U, Bachert P, Petrich W, Hamprecht FA. A comparison of random forest and its gini importance with standard chemometric methods for the feature selection and classification of spectral data. BMC Bioinform. 2009;10:1–16.
34. Muthukrishnan R, Rohini R. Lasso: a feature selection technique in predictive modeling for machine learning. In: 2016 IEEE international conference on advances in computer applications (ICACA); 2016. IEEE. pp. 18–20.
35. Walter SD. The partial area under the summary ROC curve. Stat Med. 2005;24(13):2025–40.
36. de Winter JCF, Gosling SD, Potter J. Comparing the Pearson and spearman correlation coefficients across distributions and sample sizes: a tutorial using simulations and empirical data. Psychol Methods. 2016;21(3):273.
37. Evan Johnson W, Li C, Rabinovic A. Adjusting batch effects in microarray expression data using empirical Bayes methods. Biostatistics. 2007;8(1):118–27.
38. Leek JT, Evan Johnson W, Parker HS, Jaffe AE, Storey JD. The SVA package for removing batch effects and other unwanted variation in high-throughput experiments. Bioinformatics. 2012;28(6):882–3.

Publisher's Note