scientific reports

Check for updates

OPEN

Dissecting transcriptome signals of anti-PD-1 response in lung adenocarcinoma

Kyeongmi Lee[1,8], Honghui Cha[2,8], Jaewon Kim[3], Yeongjun Jang[3], Yelin Son[3], Cheol Yong Joe[2,7], Jaesang Kim[4,5], Jhingook Kim[6], Se-Hoon Lee[2,7]✉ & Sanghyuk Lee[1,3,4]✉

Immune checkpoint blockades are actively adopted in diverse cancer types including metastatic melanoma and lung cancer. Despite of durable response in 20–30% of patients, we still lack molecular markers that could predict the patient responses reliably before treatment. Here we present a composite model for predicting anti-PD-1 response based on tumor mutation burden (TMB) and transcriptome sequencing data of 85 lung adenocarcinoma (LUAD) patients who received anti-PD-(L)1 treatment. We found that TMB was a good predictor (AUC = 0.81) for PD-L1 negative patients (n = 20). For PD-L1 positive patients (n = 65), we built an ensemble model of 100 XGBoost learning machines where gene expression, gene set activities and cell type composition were used as input features. The transcriptome-based models showed excellent accuracy (AUC > 0.9) and highlighted the contribution of T cell activities. Importantly, nonresponder patients with high prediction score turned out to have high CTLA4 expression, which suggested that neoadjuvant CTLA4 combination therapy might be effective for these patients. Our data and analysis results provide valuable insights into developing biomarkers and strategies for treating LUAD patients using immune checkpoint inhibitors.

Keywords Lung adenocarcinoma, Immunotherapy, Machine learning, Transcriptomics

Immune checkpoint blockades (ICBs) have brought a paradigm shift for the treatment of diverse types of cancers with remarkably durable clinical gains[1,2]. In non-small-cell lung cancer (NSCLC), several ICBs that target the programmed death 1 (PD-1) and its major ligand PD-L1 were approved by FDA as a first-line treatment for PD-L1 positive metastatic patients. However, the response rate is less than 50% even in PD-L1 positive tumors and larger than 10% in PD-L1 negative patients[3,4]. Tumor mutation burden (TMB) has been established as an independent marker to predict the ICB response in many tumors[5–8], but its prediction accuracy is much lower to meet the standards for clinical applications. Thus, we still lack reliable molecular markers of clinical utility for immunotherapy of lung cancer.

Transcriptome signatures have been regarded as a valuable source of deducing immune activities that could be used as a proxy variable of ICB responses. For example, Hugo et al.[9] reported the innate anti-PD-1 resistance-related signature (IPRES) in metastatic melanoma that included mesenchymal transition, cell adhesion, ECM remodeling, angiogenesis, and wound healing. The IPRES signature was significantly down-regulated in anti-PD-1 responding patients. Ayers et al.[10] identified the T-cell inflamed gene expression profiles (GEPs) that consisted of diverse immune-related signatures encompassing IFN-γ–responsive genes related to antigen presentation, chemokine expression, cytotoxic activity, and adaptive immune resistance. A pan-fibroblast TGFβ response signature (F-TBRS) was also developed to measure the TGFβ pathway activity that played roles in immune-excluded subtype of the metastatic urothelial cancer[11]. Jiang et al.[12] constructed a computation model for tumor immune dysfunction and exclusion (TIDE) that integrated diverse expression signatures for T cell

[1]Department of Bio-Information Science, Ewha Womans University, Seoul 03760, South Korea. [2]Department of Health Sciences and Technology, Samsung Advanced Institute of Health Science and Technology, Sungkyunkwan University, Seoul 06351, South Korea. [3]Ewha Research Center for Systems Biology (ERCSB), Ewha Womans University, Seoul 03760, South Korea. [4]Department of Life Sciences, Ewha Womans University, Seoul 03760, South Korea. [5]Ewha-JAX Cancer Immunotherapy Research Center, Ewha Womans University, Seoul 03760, South Korea. [6]Department of Lung Surgery, Samsung Medical Center, Sungkyunkwan University School of Medicine, Seoul 06351, South Korea. [7]Division of Hematology-Oncology, Department of Medicine, Samsung Medical Center, Sungkyunkwan University School of Medicine, Seoul 06351, South Korea. [8]These authors contributed equally: Kyeongmi Lee and Honghui Cha. ✉email: shlee119@skku.edu; sanghyuk@ewha.ac.kr

dysfunction and exclusion mechanisms. However, none of these expression signatures showed reasonable performance in predicting the anti-PD-1 response in lung adenocarcinoma patients in our preliminary analysis.

In this study, we report a machine learning model for predicting anti-PD-(L)1 treatment response in lung adenocarcinoma (LUAD), utilizing gene expression as well as the molecular features computationally inferred from the transcriptome sequencing data. We also dissect the molecular signatures of immune regulation contributing to the response of ICBs.

Results

Study overview

The overall study design is summarized in Fig. 1. Our patient cohort, a subset of LUAD immunotherapy cohort in the European Nucleotide Archive under accession number EGAS00001006461[13], consists of 85 LUAD patients who received PD-(L)1 mAb monotherapy. We had 23 complete or partial responders (CR or PR), 15 stable disease (SD) patients, and 47 progressive disease (PD) patients (Supplementary Table S1). SD patients were regarded as non-responders (NRs) in this study. Molecular features for predicting ICB response were derived from exome sequencing data (TMB and mutations) and transcriptome sequencing data (cell type abundance and gene set activities). Patients were divided into the PD-L1 positive and negative groups mainly because TMB turned out to be a good classifier for the PD-L1 negative patients. For the PD-L1 positive group, we built transcriptome-based prediction models that are essentially an ensemble of 100 XGBoost (XGB) machines where each machine was trained with a pseudo-randomly selected subset (80%) of patients. Lastly, we evaluated the performance of our predictors and deduced features that provided insights into understanding molecular mechanisms and developing biomarkers for immunotherapy of LUAD. Patient characteristics are summarized in the Supplementary Figure S1.

Feature space for predicting the ICB response

Tumor mutation burden (TMB) was estimated from exome sequencing data for each patient. Transcriptome data was used as input to the machine learning model in two ways – gene expression values themselves or inferred properties such cell type composition and pathway activities (Fig. 2a). Cell type composition was deduced by MCP-counter for 8 immune cell types (CD3+ T cells, CD8+ T cells, cytotoxic lymphocytes, NK cells, B lymphocytes, monocytic lineage cells, myeloid dendritic cells, and neutrophils) as well as endothelial cells and fibroblasts. ESTIMATE was used to calculate the stromal and immune cell portions. Patient-specific pathway activity was obtained by ssGSEA for a number of gene sets in MSigDB (v7.2). The 'HCP' gene sets were defined to include 50 hallmark gene sets (H) and 2922 curated canonical pathways (CP). The 'emerging' gene sets, defined to include 3368 gene sets from chemical and genetic perturbations (CGP), 189 oncogenic signature gene sets, and 5219 immunologic signature gene sets, represent the collection of differentially expressed genes in various experimental environments. Although being poorly characterized and annotated, they can serve as a gold mine to identify novel targets or relationships.

Computational model for predicting patient response to ICBs

In building the transcriptome-based prediction models with thousands of genes or gene sets, we adopted the XGBoost algorithm for machine learning because of its speed, accuracy, and support for automatic feature

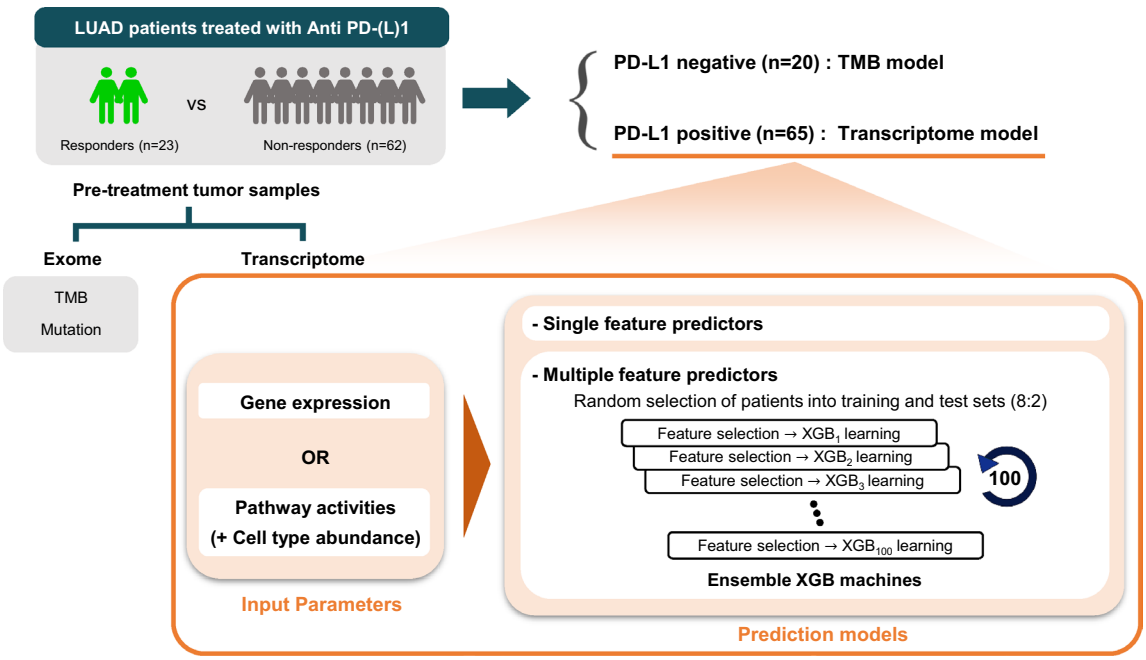


Fig. 1. Overview of predicting response to anti-PD(L)1 in 85 lung adenocarcinoma patients.

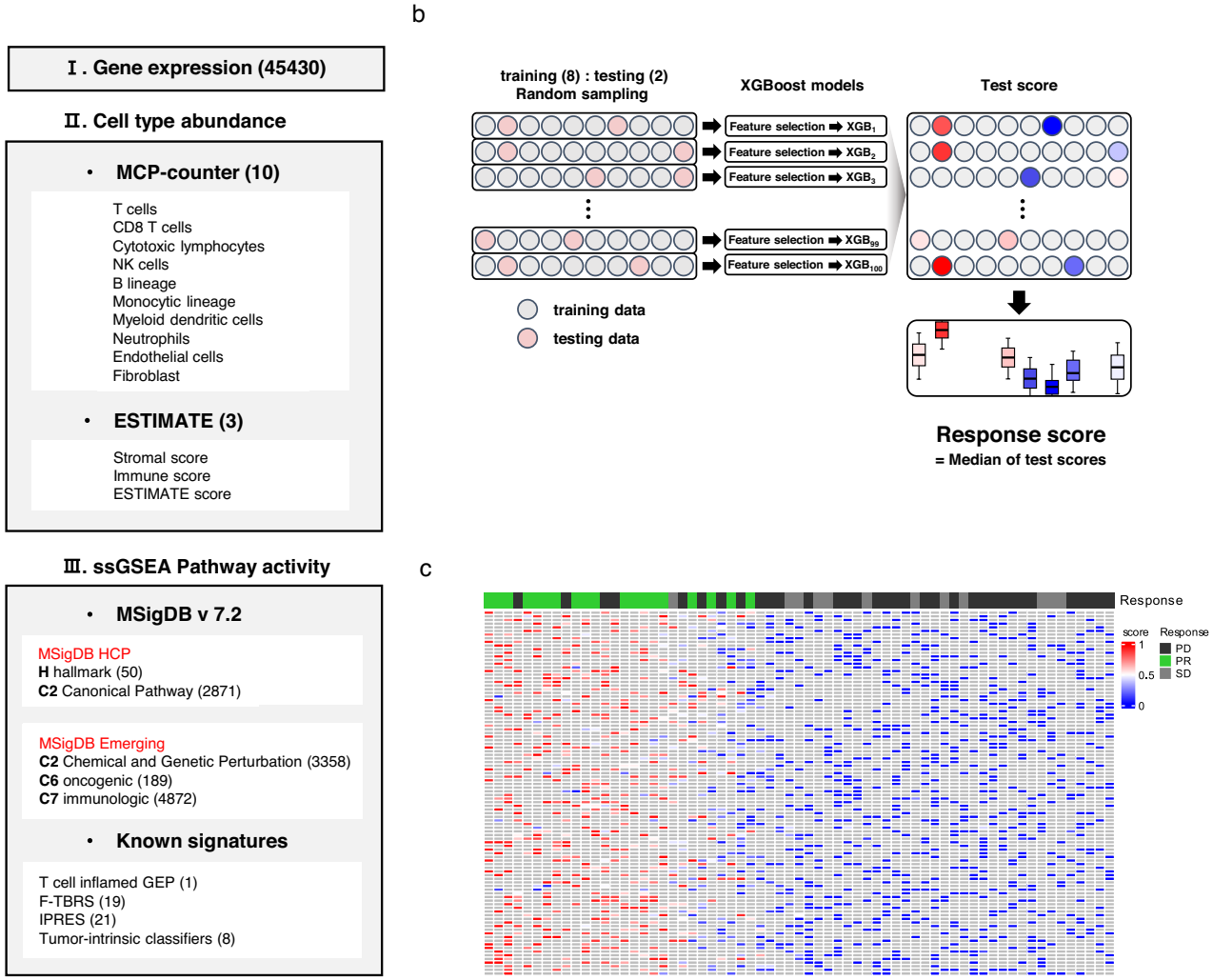


Fig. 2. Ensemble machine learning model for analyzing transcriptome data. (**a**) Three types of input variables. Number of input variables for each type is indicated in the parenthesis. (**b**) Pseudo-random patient selection of training and testing data in 8:2. The responder: non-responder ratio of each XGB machine was fixed to be the same as that of total patients. Test scores were evaluated for each XGB machine and the median value for each patient was used as the final prediction score. (**c**) Heat map of test scores. Each row and column represent XGB machine and patient, respectively. Gray rectangles indicate the training data for each XGB machine. Note that the test scores show considerable variation across 100 XGB machines, manifesting that each XGB machine reflects the heterogeneous nature of patient subgroups adequately.

selection and feature importance analysis. Considering the complexity and heterogeneity of tumors, patient number of 85 is far short to build robust and accurate prediction models for ICB response. To alleviate the instability of a single XGB machine and to increase the prediction accuracy, we built an ensemble of 100 XGBoost machines (Fig. 2b). For each XGB machine, patients were randomly divided into 80% training and 20% test sets, keeping the responder and non-responder ratio the same as the total patients. Input variables were prefiltered according to the single-variable AUC score to feed manageable number of informative features into the XGB training process. Once the training step was finished for all XGB machines, we calculated the response score of each patient by amassing XGB machines that did not use the patient of interest in the training process and taking the median score of those XGB machines as the final response score. The heatmap of test scores showed that our ensemble model of XGB machines reflected properly the heterogeneous and diverse characteristics of patient subgroups used in the training process (Fig. 2c).

Patient subgroup analysis by PD-L1 expression

We examined the performance of our composite model for diverse subsets of patients (PD-L1 positives vs. PD-L1 negatives, smokers vs. nonsmokers, and TMB-high vs. TMB-low patients), using various features that included TMB, gene expression, cell type composition, and gene set activities of known expression signatures as well as thousands of gene sets in the MSigDB collection (Supplementary Table 2). The best performance was obtained when patients were divided into the PD-L1 positive (n = 65) and negative groups (n = 20). The responder ratio was slightly larger in the PD-L1 positive subgroup, consistent with the previous knowledge[14] (Fig. 3a). Other classifications by smoking status or TMB-high/low groups did not show satisfactory performance.

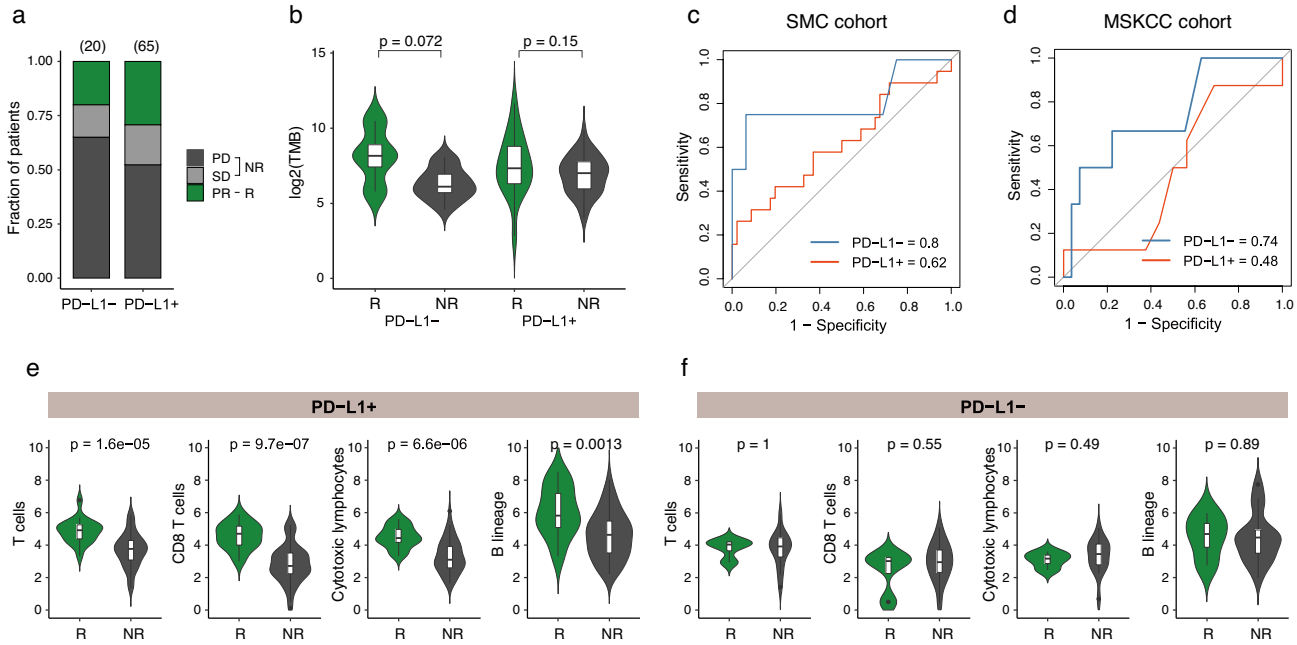


Fig. 3. Patient characteristics of TMB and immune cells. (**a**) Response rate in the PD-L1 positive and negative patients. (**b**) Tumor mutation burden (TMB) comparison of responders (R) and non-responders (NR) in the PD-L1 positive and negative patients. (**c–d**) Receiver Operating Characteristic (ROC) curve for response prediction of TMB in the Samsung Medical Center (SMC) (**c**) and MSKCC (**d**) cohorts. Numbers for PD-L1 positive and negative patients indicate the Area Under Curve (AUC) values. (**e–f**) Abundance of immune cell types inferred from RNA-seq data using MCP-counter in the PD-L1 positive (**e**) and negative (**f**) patients.

TMB was higher for the responder group in the PD-L1 negative patients ($p = 0.072$), whereas it showed no difference between the responder and nonresponder (NR) groups in the PD-L1 positive patients (Fig. 3b). In our predictive modeling, TMB turned out to be the best classifier for the PD-L1 negative patients with the area under the ROC curve (AUC) = 0.8, whereas TMB showed marginal predictive power (AUC = 0.62) for the PD-L1 positive patients (Fig. 3c). This observation was validated using an independent cohort of the MSKCC LUAD patients (n = 57) (Fig. 3d). Thus, we concluded that TMB is a decent classifier of ICB response for the PD-L1 negative LUAD patients.

We also examined the abundance of immune cell types, inferred from transcriptome data by MCP-counter, for the PD-L1 positive and negative patients. T cells, CD8 T cells, cytotoxic lymphocytes, and B lineage cells were significantly more abundant in the PD-L1 positive patients (Fig. 3e), whereas their amount were rather similar in the PD-L1 negative patients (Fig. 3f). Other immune cell types showed a similar but weaker trend (Supplementary Fig. 2). This strongly suggested that the transcriptome-based model, from which the immune signatures were derived, would be successful only for the PD-L1 positive patients. In fact, we achieved the best performance of AUC = 0.93 for the PD-L1 positive patients whose details are given in the following sections.

Transcriptome-based ensemble prediction models for the PD-L1 positive patients

In an effort to deduce key features (i.e. genes, gene sets or pathways) of ICB response, we tested the performance of our ensemble prediction models for diverse subset of transcriptome signatures. Among the known gene signatures, only the IMPRES[15] signature showed a better predictive power (AUC = 0.76) than TMB (AUC = 0.62), and all others (TIDE[12], IPRES[9], and F-TBRS[11]) showed poor performance similar to random prediction (Fig. 4a). This implies that LUAD is much different, in the immune perspective at least, from other tumor types such as melanoma or urothelial cancer where the known signatures were derived previously.

We then explored the performance of the transcriptome-based XGB ensemble models in the two prediction modes of gene expression and signatures. In the gene expression mode where the expression values from RNA-seq were used as input features, the ensemble model showed variable performance of $0.85 < AUC < 0.93$ depending on the prefiltering stringency (Fig. 4b). We obtained the best AUC = 0.93 when three genes for each XGB machine were used after prefiltering step, where 36 genes in total were used in 100 XGB machines. This test demonstrated that our ensemble approach was accurate as well as robust and that the prefiltering step enhanced the performance substantially.

Next, we tested various features derived from the transcriptome data where the number of input features after prefiltering was fixed to 5 after trying a few alternatives (Fig. 4c). Cell type abundance achieved the lowest AUC = 0.78 likely due to the limited number of cell types and accuracy in deducing their abundance from the transcriptome data. Predictive models using 49 known ICB-related signatures from the previous studies or the HCP gene sets yielded the AUC = 0.87. Using ~ 8400 'emerging' gene sets from the MSigDB (designated as MSigDB_emerging) gave AUC = 0.92. The best performance (AUC = 0.93) was obtained when we used all gene sets (i.e. the cell type abundance and the known ICB-related signatures added to the emerging gene sets).

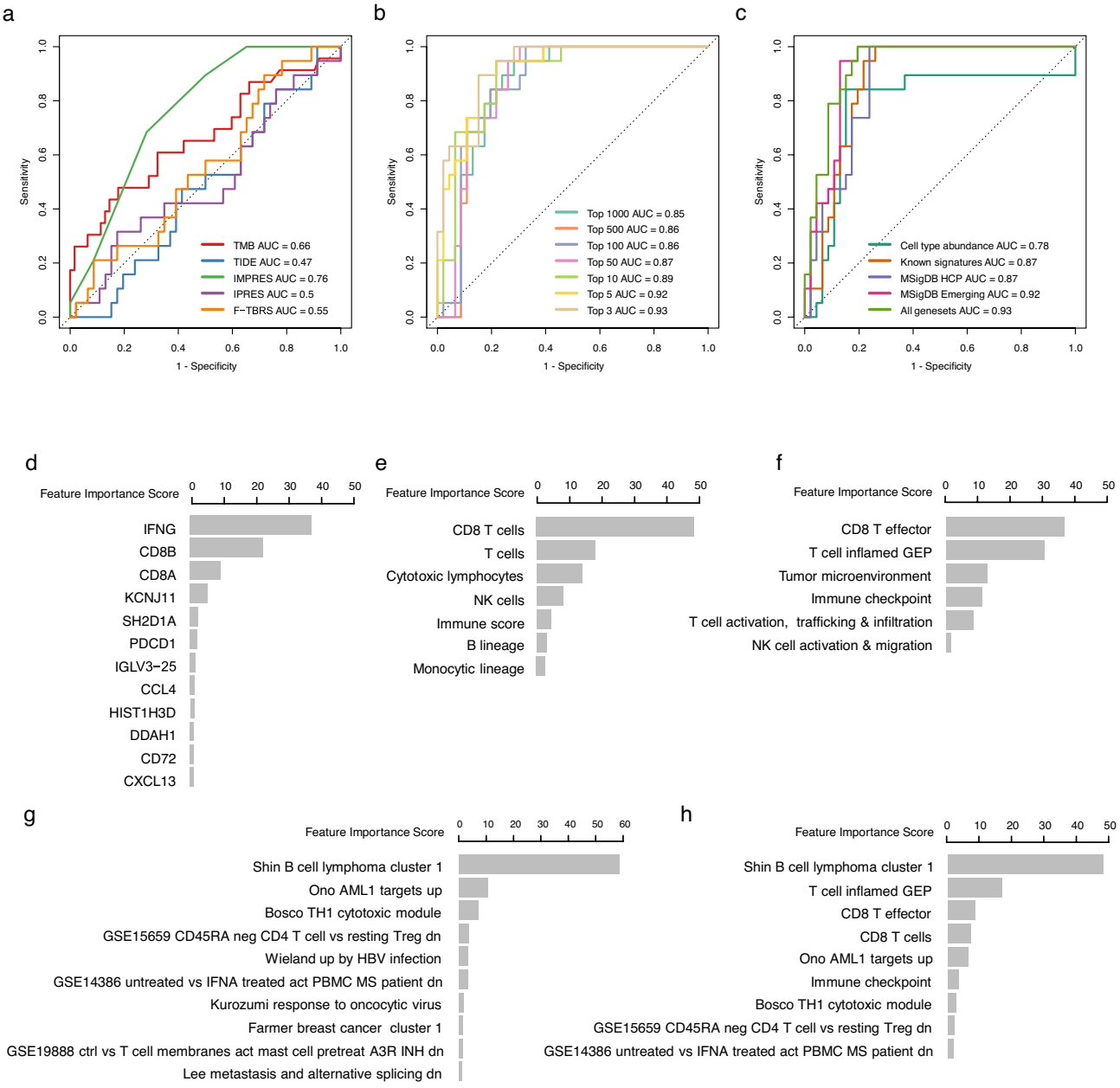


Fig. 4. ROC curves and important features from transcriptome-based models. **a**-**c**) ROC curves for individual known signatures (**a**), gene expression (**b**), and cell type abundance and gene sets (**c**). In the gene expression, Top numbers indicate the number of features used after prefiltering. (**d**–**h**) Importance scores from our ensemble XGBoost machines using individual gene expression (**d**), cell type abundance (**e**), known signatures from literature (**f**), emerging gene sets (**g**), and all transcriptome-derived variables (*i.e.* **e**+**f**+**g**) (**h**).

One of the major advantages of XGB machine is its automatic feature selection and feature importance estimation. Each XGB machine reported 4–5 important features and their contribution scores. The importance scores were highly variable across 100 XGB machines, implying that our ensemble approach sampled the heterogeneous characteristics of different patients groups adequately (Supplementary Fig. 3). In order to deduce key features of predicting the ICB response, the cumulative importance scores from 100 XGB machines were shown in the bar plots (Fig. 4d–g). Among the gene expression input features, *IFNG* and *CD8B* were the top 2 genes followed by *KCNJ11*, *CD8A*, *CXCL9*, *SLC29A2*, *GZMA*, and so on (Fig. 4d). The cell type abundance features showed the importance score in the order of CD8 T cells, T cells, Cytotoxic lymphocytes, NK cells, B lineage, and monocytic lineage (Fig. 4e). Among the known ICB-related signatures, CD8 T effector and 'T cell inflamed GEP' signatures were the top 2 features, followed by tumor microenvironment and immune checkpoint signatures (Fig. 4f). All these signatures were closely related to the immune activity including T cells and tumor microenvironments. Among the novel 'emerging' gene sets, we obtained 10 important features, most of which were poorly characterized (Fig. 4g). Dominant contribution came from the 'SHIN B cell lymphoma cluster 1' that consisted of 13 genes including *IFNG*, *CXCL9*, *IL10*, and *CCL3*. When we analyzed the full sets, the 'SHIN B cell lymphoma cluster 1'

was still the dominant contributor followed by the 'T cell inflamed GEP' and 'CD8 T effector' signatures from the known signatures and 'CD8 T cells' from the cell type abundance (Fig. 4h). All others except the immune checkpoint gene set were from the MSigDB_emerging gene sets.

Prediction outcome for the PD-L1 positive patients

The prediction score from the full set predictor (AUC = 0.93) is shown as the waterfall plot for the PD-L1 positive patients (Fig. 5a). Our ensemble model predicted all 19 responders correctly, achieving the perfect sensitivity in predicting responders. All patients predicted to be non-responders were correct except nine case (32 out of 41). In the clinical point-of-view, our prediction model has extremely desirable characteristics – rescuing all responders and filtering out majority of non-responders successfully. Of note, we suggest an alternative treatment method for unexpected nonresponders (see below).

We also examined the performance of previously known markers for each patient (Fig. 5b). PD-L1 expression was significantly higher ($p = 0.0002$) in the responder group (Supplementary Fig. 4a) and TMB was slightly higher in the responder group with no statistical significance (Fig. 3b). Progression-free survival (PFS) was better in the responder group as well (Supplementary Fig. 4b). Driver mutations, sex, smoking showed no difference between the responder and non-responder groups (Supplementary Fig. 1). Other expression-based signatures (TIDE, IPRES, IMPRES, and F-TBRS) showed virtually no predictive power for our patients.

The heatmap of individual activity scores (Fig. 5c) or gene expression (Fig. 5d) of important features showed that those features could be classified into two groups. Most features showing elevated activity or expression in the responder group were closely related to T cell activation and their pairwise correlations were extremely high. Of note, we observed three features that showed the opposite trend—'GSE15659 CD45RA negative CD4 T cell vs. Resting Treg DN' activity and *KCNJ11* and *DDAH1* expression. Function of these features are poorly characterized, but they might play roles in negative regulation of immune activity, thus impeding successful treatment of ICBs.

Importantly, we noticed that several patients with high immune activities according to mRNA expression data were falsely predicted to respond. We systematically searched for differential factors between the expected responders (true positives) and unexpected nonresponders (false positives). Differential analysis of gene expression and gene set activities yielded 27 gene sets $p < 0.01$ and 51 genes at $p < 0.005$ (Supplementary Fig. 5), most of which were difficult to interpret their biological meanings. Alternatively, we examined gene expression of immune checkpoint genes known to play roles in NSCLC[16]. Among 7 receptors and 14 ligands, only the CTLA4 expression showed a noticeable difference ($p = 0.054$) (Fig. 5e). This elevated CTLA4 expression might be the cause of anti-PD-1 treatment failure, which suggested that combination treatment of anti-PD-1 and anti-CTLA4 could be successful for those patients. Another factor that showed a similar trend with the CTLA4 expression was the gene set activity of 'GO:2000318_Positive regulation of T-helper 17 type immune response', which included *PRKCQ* (protein kinase C theta) with significant difference among our patient subgroups (Fig. 5f,g).

Validation of prediction models using independent datasets

Performance of machine learning models usually needs to be validated with independent datasets due to the overfitting possibility. We identified an NSCLC study of 16 patients who received nivolumab (anti-PD-1) treatment at the Yonsei Medical School (Seoul, Korea) and whose transcriptome sequencing data were available in public (GSE126044). Our prediction model showed an excellent performance of AUC = 0.96 when the emerging or all gene sets were used (Supplementary Fig. 6). Our ensemble prediction models with gene expression, cell type abundance, or known gene signatures were equally good for this data set. However, only the TIDE score among the known prediction signatures showed a decent performance (AUC = 0.84).

Another relevant datasets are from two randomized clinical trials for atezolizumab (anti-PD-L1) as a second-line treatment of NSCLC[17]. The OAK and POPLAR datasets contained RNA-seq data from 241 and 55 nonsquamous tumor samples, respectively. Our ensemble models yielded a decent but limited performance with the best AUC = 0.68 and 0.74 in two datasets (Supplementary Table 3 & Supplementary Fig. 7). We reasoned that the cause of limited performance could be the inclusion of PD-L1 negative patients. Since the OAK dataset provided PD-L1 expression level as well, we examined the PD-L1 + subgroup (n = 56), which did not improve the performance. We further defined PD-L1 + patients into strong plosive (n = 21) and weak positive (n = 35) subgroups. Our models achieved the best performance in the PD-L1 strong positive patient subgroup with AUC = 0.87. Of note, the performance with gene set activities was better and more robust than gene expression. Reasons for the poor performance of PD-L1 weak positive patients are not clear, but the immune activities were clearly suppressed in this subgroup in spite of PD-L1 expression albeit low.

Overall, our validation test with independent datasets confirmed that transcriptome-based ensemble models work best for the PD-L1 positive patients and that the geneset-based predictors are more robust than expression-based models. It also showed that the prediction accuracy of our predictive models was much larger in the Yonsei NSCLC cohort than in the OAK/POPLAR cohorts. The worse accuracy in the OAK/POPLAR datasets can be ascribed to the difference in ethnicity and ICBs (nivolumab vs atezolizumab; anti-PD-1 vs. anti-PD-L1).

Discussion

Identifying patients responding to ICB treatment is of uttermost importance in cancer therapy. We developed a machine learning model that reflected diverse signals of gene expression, cell type abundances and pathway activities based on transcriptome sequencing data. Our computational model is accurate as well as robust by coupling the ensemble aggregation method with a state-of-the-art machine learning algorithm XGBoost. Two major strengths in the algorithmic aspect were the ensemble method and XGBoost, where the former successfully

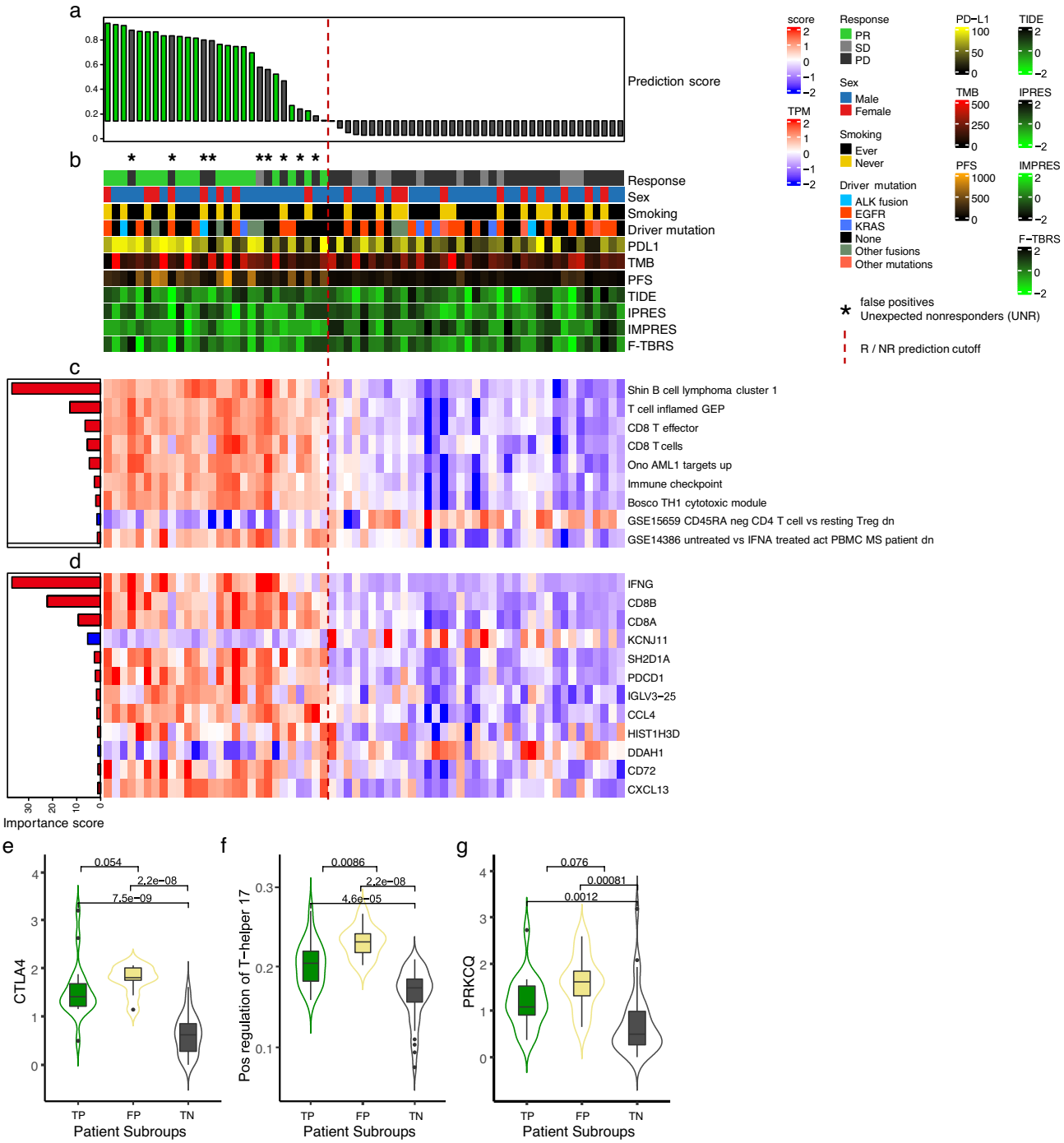


Fig. 5. Prediction outcome from the transcriptome-based model using all gene sets. **a**) Waterfall plot of the response score. The responder/non-responder cutoff was selected to maximize the prediction accuracy. (**b**) Clinical information and prediction scores from four previously known methods (TIDE, IPRES, IMPRES, and F-TBRS). **c**) Gene set activities of each patient in the order of the importance scores on the left bar graph. (**d**) Expression level of important genes obtained from the expression-based analysis. Blue color in the importance score indicate the features showing negative correlation with typical immune features. (**e**–**g**) Comparison of true positives (TP), false positives (FP indicated with asterisks), and true negatives (TN). *CTLA4* expression (**e**), Th-17 related gene set activity (**f**), and *PRKCQ* expression (**g**) are shown in violin plots.

overcame the small number of patient cases and the latter provided the importance scores that could be utilized to identify key elements of ICB response.

Our computational model achieved the accuracy over AUC > 0.9 using either gene expression or gene set activity as input variables. Each mode has unique merits. The gene expression mode is more straightforward and

easier to understand the result by investigating the gene function one by one. Developing biomarkers should be possible as well since our model showed a good performance with as small number of genes as five. Even though the gene selection and measurement method need to be further tuned and optimized for clinical usage, it is likely that an expression panel with less than 10 genes offer a satisfactory performance.

On the other hand, the prediction models based on the gene set activity showed similar performance to the expression-based models, but they were more robust as can be seen in validation with independent datasets. Critical disadvantage seems that biological meaning of gene sets is rather unclear especially for the emerging gene sets. To decipher the biological meaning of these novel gene sets, we examined the pairwise correlation of gene expression and gene set activities used in this study—the ICP-related gene sets (IPRES, F-TBRS, T-cell inflamed GEP), the emerging gene sets from MSigDB, and the immune cell-type scores (ESTIMATE, MCP-counter). Hierarchical clustering of Pearson correlation coefficients yielded four distinct clusters (Supplementary Fig. 8). The largest cluster represented the immune function as can be seen in the immune-related genes (*IFNG*, *CXCL9*, *CD8A*, *CD8B*, *GZMA*, and *IL2RB*) and various immune cells (T, B, neutrophil, and NK cells), and made the largest contribution to the predictive power of ICB response. The second largest cluster consisted of signatures related to the tumor microenvironment (TME) such as the stromal score, TGF-β signaling, fibroblasts, angiogenesis, and Pan F-TBRS. Although importance of TME in cancer immunotherapy was emphasized in many recent literature, the predictive power of these signatures was rather limited based on the importance score from the XGB machines. The third cluster was associated with the DNA damage repair (DDR) process mostly originated from the F-TBRS signatures. Their contribution to the predictive models was negligible as well. Lastly, observed a small cluster of gene sets and genes whose activities were the opposite of the immune signatures (Fig. 5c,d). The importance score of these signatures was comparable to the immune signatures. Genes associated with these signatures might play roles as negative regulators of immune activities, which can be utilized as a target of combination therapies with anti-PD-1 agents.

LUAD is a heterogeneous cancer in the viewpoint of immunotherapy with no single prediction model showing satisfactory performance. Our best solution was dividing patients into two groups according to the PD-L1 expression level where TMB was a good predictor in the PD-L1 negative patients and the transcriptome-derived model showed a good performance in the PD-L1 positive patients. This is partly explained by the fact that immune-hot tumors are more frequent in the PD-L1+ patients.

Identifying novel negative regulators of immune activity is critical to develop combination treatment methods with anti-PD-1/PD-L1 agents. We identified several genes and gene set signatures that showed the opposite trend to the immune activity. It is not clear whether these are causal factors or just the result of immune activation, but they could be targets or biomarkers of successful ICTs. In this sense, we found that nonresponders with high predictive score had elevated expression of *CTLA4* and *PRKCQ*, which might have clinical utilities.

CTLA4 is thought to regulate T-cell proliferation early in an immune response, primarily in lymph nodes, whereas PD-1 suppresses T cells later in an immune response, primarily in peripheral tissues. In patients with high *CTLA4* expression, it is possible that the tumor may be utilizing *CTLA4*'s immune evasion mechanisms rather than PD-1. In such cases, a combination therapy of anti-PD-1 and anti-CTLA4 could be suggested. For the metastatic melanoma patients, neoadjuvant ipilimumab + nivolumab therapy showed much improved rate of complete responders[18] which might be applicable to LUAD patients with high *PD-L1* and *CTLA4* expression. *PRKCQ* encodes PKCθ, a crucial effector molecule in T cell receptor (TCR)[19]. PKCθ plays a crucial role in T cell activation and is known to reduce the suppressive function of regulatory T cells (Tregs)[20]. However, Tregs treated with a PKCθ-specific antibody exhibited increased PD-1 expression in a humanized mouse model[21]. Additionally, in patients with cutaneous T cell lymphoma (CTCL) who experienced rapid progression after anti-PD-1 therapy, it was found that the PD-1 inhibitor failed to suppress tumor growth in malignant T cells with *PRKCQ* amplification, and instead, it potentially promoted cancer cell proliferation[22]. These results suggest that *PRKCQ* inhibition may help overcome resistance to anti-PD-(L)1 therapies.

Recently, a number of predictive models for immunotherapy response have been introduced. Chang et al.[23] introduced the LORIS (Logistic Regression-based Immunotherapy-Response Score) model, which utilizes six clinical and pathological characteristics as input features, including TMB, history of systemic therapy, blood albumin, blood NLR, age, and cancer type. The average predictive performance of LORIS in lung cancer yielded an AUC of 0.72, while TMB and PD-L1 alone showed AUCs of 0.65 and 0.66, respectively. Notably, in our lung cancer dataset, TMB exhibited a similar predictive performance of AUC 0.62. In addition, Xie et al.[24] developed the EPL-GNN (End-to-end Part Learning with Graph Neural Networks) model based on deep learning interpretation of H&E slide image data. This model demonstrated an AUC of 0.69 in lung cancer, whereas PD-L1 and TMB showed AUCs of 0.68 and 0.62, respectively. Farina et al.[25] constructed a predictive model for NSCLC patients using a Random Forest approach that integrated CT scan image data with various clinical data. This model achieved an AUC of 0.75 in predicting treatment response at nine months. While these models, leveraging clinical, pathological, CT scan, and H&E slide data, struggle to surpass an AUC of 0.8, our model distinguishes itself by achieving an AUC greater than 0.9 using solely gene expression data. It should be pointed out that such a good performance may have been achieved by focusing our prediction on a very specific patient population, possibly Asian PD-L1(+) LUAD patients, which requires further studies and critical evaluation in more diverse populations. Nevertheless, our model and computational methods have provided valuable insights into patient stratification for precision medicine, biomarker and target gene discovery, and potential therapeutic development.

Methods

Public patient cohort

Our study uses a LUAD treated with PD-(L)1 mAb monotherapy cohort data from Korean patients at Samsung Medical Center. RNA-seq data and clinical data are available in the European Genome-phenome Archive (EGA)

(https://ega-archive.org/) with restricted access (EGA: EGAS00001006461[13]). Exome sequencing data (TMB and mutations) not included in EGA data are obtained from lead contact upon reasonable request. Patients without TMB and PD-L1 IHC information were excluded from our study. We defined the PD-L1 negative patients as the tumor proportion score (TPS) = 0% and the PD-L1 positive as TPS > 0%. Detailed information of the patient cohort is presented in Supplementary Table 1, and Supplementary Fig. 1.

Data processing

FASTQ files from transcriptome sequencing were initially checked for quality control using FastQC[26] (version 0.11.7). Adapter sequences were trimmed using Sickle[27] (version 1.33), and low-quality reads below quality score of 20 and short reads < 80 bp were eliminated. For read mapping and subsequent quantification of transcript abundance, we used the STAR-RSEM (STAR[28] version 2.6.0 and RSEM[29] version 1.3.1) pipeline with the parameters as in the ENCODE Phases 3. Hg19 human genome from Ensembl was used as the reference genome.

Since we used two different kits in RNA-Seq library preparation, we checked for any systematic biases between the two kits and found substantial batch effect (Supplementary Fig. 9a). ComBat[30] in sva R package (version 3.34.0) was used to remove the batch effect in the form of the log2-transformed expression values (Supplementary Fig. 9b). The batch-corrected expression values were used in all downstream analyses.

Calculation of transcriptome-based scores

Transcriptome data were used to infer two types of scores—cell type composition (cell types & their abundance) and gene set activities for each sample. For the gene set activity scores, we used GSVA[31] (version 1.34.0) with 50 hallmark (H), 2871curated (C2), 189 oncogenic signature (C6), and 4871 immunology signature (C7) gene sets in MSigDB (version 7.2) as well as 49 known signatures related to ICB responses, which included the T cell inflamed GEP, TGF-β related signatures, IPRES signatures, and tumor-intrinsic classifier signatures (Fig. 2). To estimate the cell type abundance, we used the MCP-counter[32] (version 1.1.0) and ESTIMATE[33] (version 1.0.13) methods using the batch-corrected read counts and TPM values, respectively.

We also examined the performance of other classifiers such as IPRES[9], F-TBRS[11], TIDE[12], and IMPRES[15] methods using batch-corrected expression values. The IPRES score was obtained from the average of geneset-wise normalized Z scores calculated from single-sample GSEA (ssGSEA) for 21 IPRES gene sets. Similarly, the ssGSEA score of the PanF-TBRS signature was used as the F-TBRS score. For the TIDE score, expression values were normalized by each gene and used as input in the TIDE web application. The IMPRES score was calculated by comparing the TPM expression of two checkpoint genes constituting 15 gene pairs of IMPRES prediction features, and the logical relationships from 15 gene pairs were added to give the final result.

Building the ensemble machine-learning model

To avoid dominance of any specific features, we performed the feature scaling by converting all input features into the z-scores. We used the ensemble learning method that combined the results from 100 individual predictors of the response to ICB treatment. Patients were divided into the training and testing data sets randomly in 8:2 ratio, respectively, keeping the responder and non-responder ratio the same as the total patients. This pseudo-random patient selection was repeated to make 100 sets of training and testing data.

Since the feature space included so many numbers of genes or gene sets, we applied an extensive prefiltering procedure to select informative features. From each training set, we produced 100 new patient subgroups by bootstrapping method and calculated the performance of each feature based on the number of AUC scores > 0.9. We selected the top 5 features after trying several alternatives. For inputs including the MSigDB gene sets, we chose 2 features from the MSigDB gene sets and 3 features from other input types for feature balancing.

For each training and test data sets, we adopted the gradient boosted tree-based xgboost architecture, known as the XGBoost (eXtreme Gradient Boosting) algorithm. The xgboost package in R was used with the dart booster option at the learning rate of 0.5 to prevent overfitting. The maximum number of boosting iterations was set to 200 and the maximum depth was 30. Default values were used for all other parameters. The learning task is a logistic regression for binary classification (responder or nonresponder) of objective function, and the output comes with the probability value.

After finishing the learning process for 100 different training datasets, we applied each XGBoost machine to patients in the test data set to obtain the probability score of responding. The final prediction score for each patient was obtained by amassing XGBoost machines that did not use the patient of interest in the training process and taking the median value of those prediction scores.

Analysis of publicly available data

We identified three datasets of NSCLC patients treated with anti-PD-1 or anti-PD-L1. TMB and PD-L1 expression values were available in the clinical information of the MSKCC LUAD dataset[34] (n = 57). These patients were divided into the PD-L1 positive and negative groups, where patients with nonzero PD-L1 expression were regarded as the PD-L1 positive patients. Then the predictive power of mutation rate (TMB) was estimated in the receiver operating characteristic (ROC) analysis (Fig. 3d).

Two studies provided the full transcriptome data as well and were used for independent validation of our ensemble prediction models. The Yonsei dataset (n = 16) were downloaded from GEO accession GSE126044 as the raw counts[35]. The OAK (n = 241) and POPLAR (n = 55) datasets were downloaded from EGA study ID EGAS00001005013 as TPM[17]. The raw expression counts were converted into the activity scores for the MSigDB as well as 49 known signature gene sets using ssGSEA. Cell type abundance was calculated by MCP-Counter and ESTIMATE in the same way as our data. The resulting values were used as input to the prediction model. Output responding scores were subject to the ROC analysis to obtain the AUC value.

Data availability

The data used to support the findings of this study are available in the European Genome-phenome Archive (EGA) (https://ega-archive.org/) with restricted access (EGA : EGAS00001006461[13]).

Received: 16 April 2024; Accepted: 3 September 2024
Published online: 10 September 2024

References

1. Postow, M. A., Callahan, M. K. & Wolchok, J. D. Immune checkpoint blockade in cancer therapy. *J. Clin. Oncol.* **33**, 1974–1982. https://doi.org/10.1200/JCO.2014.59.4358 (2015).
2. Wei, S. C., Duffy, C. R. & Allison, J. P. Fundamental mechanisms of immune checkpoint blockade therapy. *Cancer Discov.* **8**, 1069–1086. https://doi.org/10.1158/2159-8290.CD-18-0367 (2018).
3. Herbst, R. S. *et al.* Pembrolizumab versus docetaxel for previously treated, PD-L1-positive, advanced non-small-cell lung cancer (KEYNOTE-010): A randomised controlled trial. *Lancet* **387**, 1540–1550 (2016).
4. Garon, E. B. *et al.* Pembrolizumab for the treatment of non–small-cell lung cancer. *New Engl. J. Med.* **372**, 2018–2028 (2015).
5. Goodman, A. M. *et al.* Tumor mutational burden as an independent predictor of response to immunotherapy in diverse cancers. *Mol. Cancer Ther.* **16**, 2598–2608 (2017).
6. Yarchoan, M., Hopkins, A. & Jaffee, E. M. Tumor mutational burden and response rate to PD-1 inhibition. *New Engl. J. Med.* **377**, 2500–2501 (2017).
7. Hellmann, M. D. *et al.* Tumor mutational burden and efficacy of nivolumab monotherapy and in combination with ipilimumab in small-cell lung cancer. *Cancer Cell* **33**, 853-861.e4 (2018).
8. Kowanetz, M. *et al.* OA20.01 tumor mutation burden (TMB) is associated with improved efficacy of atezolizumab in 1L and 2L+ NSCLC patients. *J. Thorac. Oncol.* **12**, S321–S322 (2017).
9. Hugo, W. *et al.* Genomic and transcriptomic features of response to anti-PD-1 therapy in metastatic melanoma. *Cell* **165**, 35–44 (2016).
10. Ayers, M. *et al.* IFN-γ-related mRNA profile predicts clinical response to PD-1 blockade. *J. Clin. Invest.* **127**, 2930–2940 (2017).
11. Mariathasan, S. *et al.* TGFβ attenuates tumour response to PD-L1 blockade by contributing to exclusion of T cells. *Nature* **554**, 544–548 (2018).
12. Jiang, P. *et al.* Signatures of T cell dysfunction and exclusion predict cancer immunotherapy response. *Nat. Med.* **24**, 1550–1558 (2018).
13. Roh, W. *et al.* High-resolution profiling of lung adenocarcinoma identifies expression subtypes with specific biomarkers and clinically relevant vulnerabilities. *Cancer Res.* **82**, 3917–3931 (2022).
14. Herbst, R. S. *et al.* Predictive correlates of response to the anti-PD-L1 antibody MPDL3280A in cancer patients. *Nature* **515**, 563–567 (2014).
15. Auslander, N. *et al.* Robust prediction of response to immune checkpoint blockade therapy in metastatic melanoma. *Nat. Med.* **24**, 1545–1549 (2018).
16. Lehtiö, J. *et al.* Proteogenomics of non-small cell lung cancer reveals molecular subtypes associated with specific therapeutic targets and immune-evasion mechanisms. *Nat. Cancer* **2**, 1224–1242 (2021).
17. Patil, N. S. *et al.* Intratumoral plasma cells predict outcomes to PD-L1 blockade in non-small cell lung cancer. *Cancer Cell* **40**, 289-300.e4 (2022).
18. Burton, E. M. *et al.* neoadjuvant systemic therapy in melanoma: recommendations of the international neoadjuvant melanoma consortium. *Policy Rev. Lancet Oncol.* **20**, e378–e389 (2019).
19. Brezar, V., Tu, W. J. & Seddiki, N. PKC-theta in regulatory and effector T-cell functions. *Front. Immunol.* **6**, 530. https://doi.org/10.3389/fimmu.2015.00530 (2015).
20. Zanin-Zhorov, A. *et al.* Protein kinase C-θ mediates negative feedback on regulatory T cell function. *Science* **1979**(328), 372–376 (2010).
21. Ozay, E. I. *et al.* Cell-penetrating anti-protein kinase C theta antibodies act intracellularly to generate stable, highly suppressive regulatory T cells. *Mol. Ther.* **28**, 1987–2006 (2020).
22. Gao, Y. *et al.* Hyperprogression of cutaneous T cell lymphoma after anti-PD-1 treatment. https://doi.org/10.1172/jci (2023).
23. Chang, T. G. *et al.* LORIS robustly predicts patient outcomes with immune checkpoint blockade therapy using common clinical, pathologic and genomic features. *Nat. Cancer* https://doi.org/10.1038/s43018-024-00772-7 (2024).
24. Xie, C. Computational biomarker predicts lung ICI response via deep learning-driven hierarchical spatial modelling from H&E. https://doi.org/10.21203/rs.3.rs-1251762/v1.
25. Farina, B. *et al.* Integration of longitudinal deep-radiomics and clinical data improves the prediction of durable benefits to anti-PD-1/PD-L1 immunotherapy in advanced NSCLC patients. *J. Transl. Med.* **21**, 174 (2023).
26. Andrews, S. FastQC: A Quality Control Tool for High Throughput Sequence Data [Online] (2010). Available online at: http://www.bioinformatics.babraham.ac.uk/projects/fastqc/.
27. Joshi, N. A., Fass JN. Sickle: A sliding-window, adaptive, quality-based trimming tool for FastQ files (Version 1.33) [Software] (2011). Available at https://github.com/najoshi/sickle.
28. Dobin, A. *et al.* STAR: Ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).
29. Li, B. & Dewey, C. N. RSEM: Accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinform.* **12**, 1–16 (2011).
30. Johnson, W. E., Li, C. & Rabinovic, A. Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics* **8**, 118–127 (2007).
31. Hänzelmann, S., Castelo, R. & Guinney, J. GSVA: gene set variation analysis for microarray and RNA-Seq data. *BMC Bioinform.* **14**, 1–15 (2013).
32. Becht, E. *et al.* Estimating the population abundance of tissue-infiltrating immune and stromal cell populations using gene expression. *Genome Biol.* **17**, 1–20 (2016).
33. Yoshihara, K. *et al.* Inferring tumour purity and stromal and immune cell admixture from expression data. *Nat. Commun.* **4**, 2612 (2013).
34. Rizvi, H. *et al.* Molecular determinants of response to anti-programmed cell death (PD)-1 and anti-programmed death-ligand 1 (PD-L1) blockade in patients with non-small-cell lung cancer profiled with targeted next-generation sequencing. *J. Clin. Oncol.* **36**, 633–641 (2018).
35. Cho, J. W. *et al.* Genome-wide identification of differentially methylated promoters and enhancers associated with response to anti-PD-1 therapy in non-small cell lung cancer. *Exp. Mol. Med.* **52**, 1550–1563 (2020).

Author contributions

S.L., S.H.L., Jhingook Kim, and Jaesang Kim.: Conceptualization. S.H.L., Jhingook Kim: Data collection. H.C, C.Y.J: Data curation. K.L., Jaewon Kim, Y.J., and Y.S.: performed the analysis. S.L., S.H.L, K.L.: wrote the manuscript.

Funding

This work was supported by the grants from the National Research Foundation of Korea (No. NRF-2020R1A2C2099719 to SL; 2020R1A2C3006535 to SHL; RS-2023-00225255 to SHL). The authors acknowledge the Korea Research Environment Open Network (KREONET) service and the usage of the Global Science Experimental Data Hub Center (GSDC) provided by Korea Institute of Science and Technology Information (KISTI). This study was also supported by Future Medicine 2030 Project of the Samsung Medical Center [#SMX1240011] and a grant of the Korea Health Technology R&D Project through the Korea Health Industry Development Institute (KHIDI) (No. HR20C0025).

Competing interests

The authors declare no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-024-72108-5.

Correspondence and requests for materials should be addressed to S.-H.L. or S.L.

Reprints and permissions information is available at www.nature.com/reprints.

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

Open Access This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by-nc-nd/4.0/.

© The Author(s) 2024