Modules of Effector T cell Cytolytic Activity are Prognostic in Multiple Solid Tumors and are Predictive of Response to Immune Checkpoint Blockade Therapy.

Chaitanya R. Acharya Ph.D.1, Casey Shuptrine Ph.D.1, Zack Hartman, Ph.D.1 and Herbert Kim Lyerly M.D.1\*

1 Center for Applied therapeutics, Department of Surgery, Duke University Medical Center, Durham NC 27705

\*Corresponding author:

Herbert K. Lyerly,

Center for Applied Therapeutics,

Department of Surgery,

Duke University Medical Center, Durham NC 27705 USA

[kim.lyerly@duke.edu](mailto:kim.lyerly@duke.edu)

Phone: 919-684-6408 Fax: 919-681-7970

**Abstract**

The prognostic and predictive value of tumor infiltrating lymphocytes for immune checkpoint blockade has been recognized in a variety of tumor types, including triple negative breast cancer (TNBC). We hypothesize that a specific effector phenotype of T cell cytolytic activity (ECA) is a consistent feature of epithelial tumors, possibly varying by tumor types with a range of inflammatory features. We interrogated 6,311 CD3+ single T cells previously isolated from human primary TNBC samples to derive functional gene expression modules of ECA, and trained Random Forest classifiers using each such module. We demonstrate prognostic value of each such classifier in more than 6,000 tumor samples encompassing 15 tumor types. Classifier developed from one of our ECA modules was significantly associated with improved patient survival and predicts response to immune checkpoint inhibitors in fifty-one advanced melanoma patients. Further evaluation of these ECA modules is important to understand their promising role in individualizing immunotherapy.

**Main**

The primary objective of cancer immunotherapy is to enable our immune system to efficiently identify and attack tumor cells. Currently, there are two FDA approved strategies to accomplish this – 1) adoptive transfer of tumor-specific autologous T cells, engineered to recognize antigens on tumor cells using chimeric antigen receptors (CAR), a process also known as CAR T-cell therapy, and 2) blocking key inhibitory receptors of T cells enabling anti-tumor immune response activation and effector function. While the former approach involves modifying autologous T cells *ex vivo* for production of CAR T cells [1](#_ENREF_1), the latter approach primarily targets immune-inhibitory checkpoints such as programmed death-ligand 1(PDL1), programmed death 1 (PD1) and cytotoxic T-lymphocyte-associated protein 4 (CTLA4) in order to enable and expand adaptive anti-tumor immune response [2](#_ENREF_2),[3](#_ENREF_3). A growing body of research has indicated that these tumor infiltrating lymphocytes (TILs) have both prognostic and predictive value in many solid tumors, which may be useful in selecting for individuals who are most likely to respond to immune checkpoint blockade (ICB). In fact, a recent study of 3,771 breast cancer patients suggested that an increased TIL concentration also predicted response to neo-adjuvant chemotherapy in TNBC and HER2+ patients [4](#_ENREF_4).

Given the clinical utility of existing gene expression biomarkers such as OncotypeDx [5](#_ENREF_5), MammaPrint [6](#_ENREF_6), Prosignia [7](#_ENREF_7), and some accumulating evidence supporting the use of ‘TILs scoring’ as a prognostic biomarker [8](#_ENREF_8), it is essential to develop a reliable immune-based diagnostic assay to predict the efficacy of immunotherapies. In order to develop a reliable molecular assay, sufficient samples are necessary to develop and validate the assay. Currently, publicly available gene expression data are available only for a limited set of immunotherapy pre-treatment samples. A biomarker of immunotherapy derived from such data will likely need extensive evaluation for clinical utility in large randomized clinical trials, which demand many years, patients and resources [9](#_ENREF_9). Consequently, there has been a proliferation of proposed surrogate biomarker studies that focus on biologically plausible markers, including immune evasion [10](#_ENREF_10), cytokines in the immune microenvironment [2](#_ENREF_2) and tissue-resident memory T cell differentiation [11](#_ENREF_11),[12](#_ENREF_12) as a guide to prioritize potential prognostic and predictive biomarkers of immunotherapeutic agents.

To develop more precision in quantifying the changes in T cells infiltrating tumors, we sought to exploit the heterogeneity present in a population of T cells by using single-cell RNA-seq (scRNA-seq) analysis to characterize functional modules of ECA. To overcome the challenges posed by signal sparsity in scRNA-seq data (due to ‘dropout events’) and strong correlation structure among genes, we used Random Forests (RFs) [13](#_ENREF_13) to train classifiers of ECA activity, which were subsequently validated in multiple solid tumor types. RF classification models are non-parametric machine learning tree-based ensemble learning approaches that are increasingly frequently used in addressing “large p, small n” problems [13](#_ENREF_13). They are often a collection of many trees, where each tree is grown as a bootstrap sample of the original data. Such methods not only help eliminate any model selection bias but also adeptly deal with the correlation and interaction among genes.

In this study, we used recently published scRNA-seq data of human T cells from TNBC [11](#_ENREF_11),[12](#_ENREF_12) to develop RF classification models from, and assess their association with disease progression in multiple solid tumors and in predicting response to ICB. We choose TNBC data for not only their moderate success in response to ICB [14-16](#_ENREF_14) but also due to the availability of specific transcriptional profiles of T cells in the public domain required for our study.

**Results**

The following are the questions addressed in this study: 1) Is it possible to isolate effective T cell cytolytic activity (ECA) profile from single cells (scRNA-seq) and validate it in a population-based analysis of cells (bulk RNA-seq)? 2) How similar is the ECA profile (with respect to biomarkers) across multiple epithelial tumors? 3) Are these biomarkers associated with clinical prognosis? 4) Can the ECA profile be a predictive biomarker of response to immunotherapy?

**Discovering gene expression modules from single-cell RNA-seq (scRNA-seq) data.**

We begin with publicly available scRNA-seq data from GEO repository (GSE110686) comprising of 6,311 purified CD3+ single T cells from two patients diagnosed with triple negative breast cancer (TNBC). Raw data from 3’ transcriptional profiling of 5,174 CD3+ T cells from the first patient and 1,137 CD3+ T cells from the second was made available in 10x Genomics™ matrix format. Unique molecular identifier (UMI) count matrix was extracted from this format, and the UMI counts for a given gene were aggregated across all the representative transcripts by computing median values in order to generate gene-level UMI count matrix. This data was appropriately normalized (Supplementary Fig. 1; Online Methods) and used in downstream analyses (Figure 1a). In order to exploit the heterogeneity within the T cell infiltrate as captured by the scRNA-seq data, we made use of twenty published immune metagenes [17](#_ENREF_17) (Supplementary table 1). These includes genes describing immune checkpoint inhibition and stimulation, T cell activation and cytolytic activity, macrophages, Natural Killer (NK) cells, lymphocyte-specific kinase (LCK), dendritic cells, chemokines, MHC class I and II, B cells, and other published gene expression signatures such as T helper cells (Th1) signature [17](#_ENREF_17), Regulatory T cells (Treg) signature [17](#_ENREF_17), Follicular B helper T cells (Tfh) signature [17](#_ENREF_17), STAT1 signature [17](#_ENREF_17), Interferon-γ signature [2](#_ENREF_2), Immune1 signature [18](#_ENREF_18) and Immune2 signature [19](#_ENREF_19) (Table 1).

We then computed enrichment scores for each pairing of sample and immune metagene using single sample gene set enrichment analysis (ssGSEA), thus transforming data to a higher-level space, leading to a more biologically meaningful interpretation.

This transformed enrichment score data was clustered to identify 14 unique clusters (Supplementary fig. 2a), and the clusters were visualized using t-distributed stochastic neighbor embedding (t-SNE). We focused on two clusters with the highest and lowest median enrichment score of T cell cytolytic activity (CYT) defined by granzyme A (GZMA) and perforin 1 (PRF1), which are secreted by effector cytotoxic T cells and NK cells (Supplementary fig. 2b). Cluster 1 (N = 426) has highest median CYT enrichment score whereas, cluster 11 (N = 495) has the lowest median CYT enrichment score. Cluster 1 is henceforth identified as ECA ‘High’ and cluster 11, as ECA ‘Low’.

Gene expression data for samples in ECA ‘High’ and ‘Low’ groups were used for downstream analysis including dimension reduction and module discovery (Figure 1b). Following spectral decomposition (principal component analysis) and jackstraw analysis (Supplementary fig. 3a) of data subsetted on highly variable genes, we identified eight gene expression-based modules (Figure 1c). Each module is defined as a linear combination of multiple genes, robust to any drop-out events in any individual gene. Module 1 (MOD1) has 1,438 genes and is enriched for cell cycle progression and DNA repair pathways. Module 2 (MOD2) has 1,018 genes and enriched for IL12 pathway, cytokine signaling, interferon signaling, NK T cell signaling pathway, IL2/STAT5 pathway, Th1/Th2 signaling pathway, CD40 pathway, and p53 signaling pathway. Interestingly, genes that describe pathway indicative of allograft rejection are also enriched in this module. Modules 3 (MOD3), 4 (MOD4) and 5 (MOD5) have 382, 374 and 301 genes, respectively, and are enriched for interferon and interferon stimulated signaling pathways along with forkhead box M1 (FOXM1) signaling pathway, and CD40 pathway. They are also enriched for activating transcription factor 2 (ATF2) pathway and IL12 pathway. Module 6 (MOD6) is comprised of 339 genes, and is dominated by genes enriched for interferon signaling pathways. Module 7 (MOD7) has 226 genes and is enriched for calcium-nuclear factor of activated T cells (NFAT) pathway while module 8 is comprised of 150 genes and is enriched for calcium signaling in CD4+ T cell receptor signaling (TCR) pathway.

We measured relative expression of some key immune-related genes in ECA ‘High’ and ‘Low’ phenotypes. We observed that ECA ‘High’ group has a relative high expression of PRF1 and GZMA, an expected result since patient groupings are based on the expression of these two genes (Figure 1d). In addition, we also observed that patients characterized as ECA ‘Low’ have a relative higher expression of forkhead box P3 (FOXP3), a biomarker for T regulatory cells (Treg cells). ECA ‘High’ group is CD8+, and enriched for T cell immunoglobulin and mucin-domain containing-3 (TIM3) along with programmed cell death 1 (PDCD1) gene. On the other hand, ECA ‘Low’ group is CD4+, and enriched for cytotoxic T-lymphocyte-associated protein 4 (CTLA4).

**Random Forest prediction models derived from the modules are prognostic in multiple solid malignancies.**

In order to demonstrate the usefulness of our modules, we developed classifiers or prediction models of ECA activity using Random Forests (RFs). Each module was used as a training set for RF prediction model. After optimizing for the number of genes and the number of trees, each RF prediction model was validated in multiple datasets. We demonstrated the validity of our prediction models by comparing them with previously published gene expression-based signatures/biomarkers such as Tfh signature (8 genes), Th1 signature (12 genes), Treg signature (7 genes), STAT1 signature (4 genes), Immune1 signature (7 genes), Immune2 signature (95 genes) and Interferon-γ signature (27 genes). Of note, there is no strong overlap of genes comprised in the aforementioned signatures and genes within our modules, with only STAT1 signature having a strong overlap with modules MOD3 and MOD4, both sharing 3 out of 4 genes from the signature (Figure 2a). Univariable cox regression analysis was performed with the best RF prediction model (defined by the strength of association between predicted ECA activity and progression-free survival) for a given tumor type and compared its performance with the aforementioned signatures. As seen in Table 2, our prediction models are consistently found to be prognostic across all the tumors as opposed to other signatures. Within a given tumor type, our prediction model’s association with disease-free survival is either better or comparable with other signatures. In addition, many of the prognostic models developed from these modules have favorable association with good prognosis (Figure 2b). For example, in renal cell carcinoma, our prediction models are both negatively and positively associated with progression-free survival i.e. an enrichment of a set of genes within some modules are associated with either poor prognosis or good prognosis, respectively. Surprisingly, prediction models contained within the modules in stomach adenocarcinoma, glioblastomas, low-grade gliomas and prostate adenocarcinomas are negatively associated with progression-free survival, a relationship that was previously described [20-23](#_ENREF_20). The role of TILs in stomach adenocarcinoma is still unclear [24](#_ENREF_24). Majority of the modules are prognostic in case of renal cell carcinoma and low-grade gliomas reflecting on a potential role ECA profile plays in determining response to ICB. Interestingly, RF model developed from MOD7 is prognostic in non-small cell lung cancer, both adeno- and squamous cell carcinoma.

As proof-of-concept, since we developed these modules from TNBC single T cells, we identified prediction models developed from MOD5 and MOD6 to be prognostic TNBC. Our best prediction model comprising of 50 genes (developed from MOD5) was prognostic in TNBC samples obtained from both TCGA (N = 150; left and center panels of Figure 3a) and METABRIC (N = 320; left and center panels of Figure 3b) datasets. Patient samples from TCGA dataset predicted to be in group ECA ‘High’ have better progression-free survival (log-rank test p-value: 0.0098l; HR: 0.30; 95% CI: 0.14-0.64) and better overall survival (log-rank test p-value: 0.0066; HR: 0.17; 95% CI: 0.06-0.42). Similarly, patient samples from METABRIC dataset predicted to be in group ECA ‘High’ have better progression-free survival (log-rank test p-value: 0.015; HR: 0.63; 95% CI: 0.45-0.89) and better overall survival (log-rank test p-value: 0.0078; HR: 0.63; 95% CI: 0.46-0.87). In an effort to characterize the genes within this prediction model, we verified their expression levels in all of TCGA (N = 877) and METABRIC (N = 1965) samples within each subtype. In both cases, these 50 genes are relatively under-expressed in ER+ tumors (right panels of Figures 3a and 3b) as opposed to HER2+ and TNBC (in both cases, ANOVA p-value<2.2e-16), suggesting that TNBC tumors have a higher potential likelihood of response to ICB.

Out of the 50 genes in the prediction model, granzymes B (GZMB) and K (GZMK), interferon-γ (IFNG), vascular cell adhesion molecule 1 (VCAM1) are some of the genes that are up-regulated while, lymphotoxin beta (LTB), baculoviral IAP repeat containing 3 (BIRC3), L-selectin (SELL), forkhead box P3 (FOXP3), nuclear factor kappa B inhibitor alpha (NFKBIA) and C-C motif chemokine receptor 7 (CCR7) are some of the genes that are down-regulated in predicted ECA ‘High’ group (Figure 3c). A quick look at the training set from MOD5 (N = 259) samples, shows the same genes to be up- and down-regulated (Figure 3d). In an effort to distinguish the immunogenicity of normal breast tissue from pure ductal carcinoma *in situ* (DCIS) and mixed DCIS/invasive ductal carcinoma (IDC), we assessed the expression levels of the same genes across all the three tissue types in GSE26304 (N = 114). Pure DCIS samples have a relative lower expression of the top genes in the predictive model as opposed to IDC samples [25](#_ENREF_25) (Figure 3e). The significantly lower expression of ECA profile in DCIS tumors may reflect their muted response to ICB.

**ECA activity as a predictive and prognostic surrogate biomarker of response to immune checkpoint inhibitors.**

We validated our RF classifiers of ECA activity in sixty-five patients (fifty-one pre-treatment and fifty-eight on treatment) with advanced stage melanoma patients exposed to immune checkpoint inhibitor, Nivolumab (anti-PD-1 agent). Of these fifty-one pre-treatment samples, twenty-six patients had previously progressed on ipilimumab (Ipi) therapy (NIV-PROG cohort); twenty-five patients were Ipi-naïve (NIV-NAÏVE cohort). Ten patients responded to immunotherapy and the rest did not. Our primary objective was to identify a RF classifier of ECA activity that was predictive of response to Nivolumab and prognostic i.e. any correlation with improved progression-free survival. We hypothesized that there is a strong correlation between ECA activity and response to immunotherapy i.e. ECA ‘High’ phenotypes is positively associated with patient response to immunotherapy. RF predictive model developed from MOD1 is not only predictive of response to Nivolumab but also prognostic (Figure 4b). Our ECA module was able to classify responders and non-responders with 77% accuracy (Fisher test p-value = 0.02) and was associated with progression-free survival (log-rank p-value = 0.03; HR: 0.28 [0.05-1.5]). Average gene expression of biomarkers within the RF model show a statistically significant (t-test p-value < 2.2e-16) enrichment in responders as opposed to non-responders (Figure 4c). Of the top genes in the RF model, macrophage inflammatory proteins or chemokines such as C-C motif chemokine ligand 3 (CCL3), C-C motif chemokine ligand 3 (CCL4), B cell attracting chemokine such as C-X-C motif chemokine ligand 13 (CXCL13), natural killer cell granule protein 7 (NKG7), IFNG, HLA type 2 molecules such as HLA-DRB1, HLA-DRA and HLA-DQA1 are some of them that are upregulated. On the other hand, NFKBIA, LTB, interleukin 1 receptor type 2 (IL1R2), FOS and JUNB are some of the genes that are downregulated in ECA ‘High’ phenotype (Figure 4a). A receiver-operator-characteristic (ROC) curve analysis (top panel of Figure 4d) of our model predictions has an area under the curve of 0.73 [95% CI: 0.56-0.9]. Out of the 41 non-responders, 33 of them are positively correlated with ECA ‘Low’ phenotype (80.5% specificity). Similarly, out of 10 responders, 6 of them are positively correlated with ECA ‘High’ phenotype (60% sensitivity). Additionally, when focused on NIV-PROG cohort (N = 26), the same model classified responders and non-responders at 81% accuracy (Fisher test p-value = 0.008), with all the responders (N = 4) correlated with ECA ‘High’ phenotype and 17 non-responders (81% specificity) associated with ECA ‘Low’ phenotype. ROC plot (bottom panel of Figure 4d) has an AUC of 0.88 [95% CI: 0.73-1]. These results indicate that ECA activity profile can be a potential surrogate biomarker of response to ICB.

An independent analysis of patients in NIV3-NAÏVE cohort before and after Nivolumab therapy have a higher average expression of the top genes from the RF model after therapy in both responders (t-test p-value = 1.6e-03) and non-responders (t-test p-value = 2.2e-05). Interestingly, responders in NIV3-PROG cohort did not have any meaningful biological changes before and after therapy (t-test p-value = 0.71) as opposed to non-responders (t-test p-value = 0.02). Of note, the average expression of the model genes is still low after therapy in non-responders from NIV3-PROG cohort indicating that the gene expression in majority of the non-responders that did not respond to Ipi therapy remained the same post-therapy (Figure 4f). This can also be visualized in the heatmap of all the genes pre- and post-Nivolumab therapy (Figure 4e).

RF models developed from MOD4, MOD6 and MOD8 are also predictive of patient response to Nivolumab but not prognostic (Figure 2b). Similarly, RF model developed from MOD5 is not predictive of patient response to Nivolumab but is prognostic. Of note, RF models developed from modules MOD1 and MOD6 are also prognostic in primary melanoma samples (N = 102).

**Discussion**

Single-cell RNA sequencing has emerged as a powerful tool to interrogate heterogeneity in an otherwise homogenous population of cells and extract signal from noise. The fine precision of thus extracted signal (as a set of biomarkers) can be lost in bulk RNA-seq samples. Using machine learning approaches, we identified a defined set of biomarkers from a population of CD3+ single T cells extracted from triple negative breast tumors, which is a recognized molecular subset of breast cancer which have an active immune microenvironment that clinically responds to immune checkpoint. We evaluated the validity of these biomarkers in bulk RNA-seq datasets from a variety of tumor types, including those known to reponse to immune checkpoint blockade. While a few previous studies have shown that biomarker identified in the scRNA-seq data can be interrogated in bulk RNA-seq data, this is one of the first attempts to identify both prognostic and predictive biomarkers, and validate them in multiple additional solid tumor types.

There are some caveats to our approach, including the implicit assumption that the final common anti-tumor pathway, specifically the cytotoxic activity profile of the effector T cell is preserved across multiple tumor types. Nonetheless, our approach does not rely or focus on the mechanistic aspects of ECA activity and instead, identifies different sets of prognostic biomarkers that define an ECA profile in multiple solid malignancies. Second, our approach exploits the richness of data of immune cells within one of the most well-defined molecular subtypes of cancer (TNBC), thus restricting the heterogeneity of the initial T cell profile to a single tissue type, which appears to respond to ICB. These ECA pathways defined by this well characterized sets of genes, were found in T cells infiltrating multiple tumor types, including tumor types well known to respond to ICB. This suggests that indeed, the activated effector T cell pathways identified are biologically relevant to humans, being employed in tumor associated T cells across the range of human tumors responding to immune therapy. As more scRNA-seq is made available, it is important to develop statistical approaches that efficiently model scRNA-seq data from multiple tumor types that are currently treated using FDA-approved immunotherapy to confirm the predictive value of this biomarker in different tumor types.

Our study highlights one important application of single-cell genomics in our understanding of immune microenvironment and potentially identify new immunotherapy targets. Further *in vitro* and *in vivo* studies are warranted to study the validity of these biomarkers.

**Acknowledgements**

This research was supported by Clinical Translational Research Award (Grant No. BC113107) from the Department of Defense. We would like to thank Pankaj Agarwal for his IT support, and Dr. Erica Crosby of the Lyerly lab for useful discussions.

**Author Contributions**

C.A wrote the original draft with author contributions from H.K.L and Z.H.; Writing – Review & editing, C. A, H.K.L, C.S and Z.H.; Conceptualization, C.A.; Supervision: H.K.L.; Study Design, C.A and C.S.; Funding Acquisition, H.K.L.; Formal Analysis, C.A.;

**Declarations of Interest**

No potential conflicts of interest were disclosed by the other authors.

**Figure Legends**

**Figure 1. An overview of the strategy to identify modules of effector T cell cytolytic activity (ECA). a,** Work flow of our strategy to develop ECA modules**. b,** Gene expression data subset derived from human primary triple negative breast cancer single-cell RNA-seq dataset based on T cell cytolytic enrichment scores. Two groups, ECA ‘High’ (N = 495) and ECA ‘Low’ (N = 426), were identified based on the average enrichment scores of T cell cytolytic activity (as defined by GZMA and PRF1 genes). Spectral decomposition followed by Jackstraw analyses helped identify eight functional modules. **c**, An illustration of gene overlap among all eight modules. **d**, 2-D scattered plot showing ECA ‘High’ and ‘Low’ groups, visualized here as a t-SNE map. Samples in red are in ECA ‘High’ group while samples in blue are in ECA ‘Low’ group. Feature plots demonstrating expression of some key immune-related genes. Dark blue points represent a relative over-expression of that specific gene.

**Figure 2. ECA module characterization and their prognostic value in multiple solid tumors. a**,An illustration of gene overlap between various immune-related gene expression-based signatures and our functional ECA modules. **b**, Plot showing prognostic significance of each module in various solid tumors. Each module was used to develop a Random Forest classifier and each such classifier was then prognosticated in several tumor types. A blue bar indicates a positive association between ECA module and favorable prognosis (usually, progression-free survival). It means that an enriched ECA module may result in a favorable prognostic outcome. On the other hand, a red bar indicated negative association between ECA module and favorable prognosis i.e. an enriched ECA module may not result in a favorable prognostic outcome. A white bar indicates no association between ECA module and patient survival. In the case of dataset GSE91061 (CA209-038 study), our modules are predictive of response to immunotherapeutic agent (white box with lines) or associated with patient progression-free survival (blue box) or both (blue box with lines).

**Figure 3.** **An example of** **superior prognostic ability of a prognostic ECA module (MOD5) in triple negative breast adenocarcinomas.** **a,b,** Left panels show Kaplan-Meier survival curves for progression-free survival in TCGA and (N=150) and METABRIC (N=320) datasets, respectively. Middle panels show Kaplan-Meier survival curves for overall survival in the same datasets. Right panels are barplots (with standard error) of log2 gene expression of the top genes in the Random Forest (RF) prediction model in each subtype of breast cancer. Statistical analysis was performed using one-way analysis of variance. **c,** Diverging barplot indicating relative gene expression levels of top genes in the RF model. The length of each bar denotes log2 fold-change values. **d,** Volcano-plot of the differentially expressed genes between predicted ECA ‘High’ and ECA ‘Low’ groups. Vertical axis represents -log10 p-values and the horizontal axis represents log2 fold change in gene expression. **e,** Barplots (with standard error) of log2 gene expression of the top genes in the Random Forest (RF) prediction model in GSE26304 (N = 114) composed of patients with ductal carcinoma *in situ* (DCIS; N = 31), mixed DCIS and invasive breast adenocarcinoma (IDC; N = 77). Some normal samples were also identified in this dataset. Statistical analysis was performed using one-way analysis of variance.

**Figure 4. Prognostic and predictive module of cancer immunotherapy in patients with advanced melanoma. a,** Diverging barplot indicating relative gene expression levels of top genes in the RF model generated from ECA MOD1 within predicted ECA ‘High’ and ‘Low’ groups of patients from GSE91066 dataset. The length of each bar denotes log2 fold-change values. **b,** Progression-free survival for cases that were stratified into ECA ‘High’ and ‘Low’ with a log-rank test p-value of 0.029. **c,** Barplots showing pre-therapy baseline gene expression levels of top 50 genes from the RF model in responders (N = 10) and non-responders (N = 41). Statistical analysis was performed with t-test. **d,** Top panel shows ROC curve for all the patients (N = 51) using their predicted ECA values. Non-responders are expected to have ECA ‘Low’ enrichment while responders to be enriched for ECA ‘High’. Area under the curve (AUC) = 0.73 (95% CI – 0.56-0.9). Bottom panel shows ROC curve of patients previously exposed to ipilimumab (N = 26). Of these patients, all responders (N = 4) are enriched for ECA ‘High’. AUC = 0.88 (95% CI – 0.73-1). **e,** Heatmap showing average expression levels of top 50 genes from the RF model in patients before and after treatment with Nivolumab. Two cohorts of patients are shown separately – one set of patients are ipilimumab naïve patients (NIV3-NAÏVE; N = 25) and the second set of patients were previously treated with ipilimumab (NIV3-PROG; N = 26). Red indicates high expression and blue indicates low expression of genes. **f,** Barplots showing gene expression levels of top 50 genes from the RF model in responders and non-responders by treatment cohort. Statistical analysis was done using t-test.

**Methods**

**Processing single-cell RNA-seq dataset**

Quantified unique molecular identifier (UMI) count data was accessed from Gene Expression Omnibus (GEO)[26](#_ENREF_26) repository under the accession number GSE110686. The Cell Ranger™ R kit (version 2.0.0) provided by 10x Genomics™ was used to load processed data files. Transcript-level abundances of single-cell RNA-seq (scRNA-seq) data were aggregated to gene-level counts using median count values resulting in a gene expression matrix with 33,275 genes and 6,311 samples. Gene-level UMI count matrix was further processed using the R package ‘Monocle’ (version 2.6.1) [27](#_ENREF_27) by assuming negative binomial distribution (with fixed variance). As a quality control (QC) step, genes and cells were filtered with a global expression threshold of 0.1 (at least one count of gene expression) in at least ten expressing cells (Supplementary Fig. 1). The resulting gene expression matrix (12,786 genes and 6,311 samples) was library size normalized and then variance stabilized transformed. The variance stabilized data was used to compute single sample gene-set enrichment analysis scores (ssGSEA scores) [28](#_ENREF_28) of all immune metagenes using R package ‘GSVA’ (version 1.22.4)[29](#_ENREF_29). ssGSEA scores are computed as difference in empirical cumulative distribution functions of gene expression ranks inside and outside the immune metagene gene-sets. These ssGSEA scores are further normalized throughout all the gene-sets and samples by the absolute difference between the minimum and maximum values and then scaled. Truncated principal component analysis of the resulting normalized and scaled ssGSEA score matrix was performed using Lanczos bidiagonalization algorithm [30](#_ENREF_30) as implemented in R package ‘irlba’ (version 2.3.2) followed by dimension reduction using t-distributed stochastic neighbor embedding (t-SNE) method [31](#_ENREF_31) implemented in R package ‘Rtsne’ (version 0.13). We then performed clustering by fast search and find of density peaks using a clustering algorithm [32](#_ENREF_32) using R package ‘densityClust’ (version 0.3). This resulted in fourteen distinct clusters. We identified clusters with least and highest average ssGSEA score for T-cell cytotoxic activity and called them ECA ‘Low’ (N=426) and ECA ‘High’ (N=495), respectively.

**Processing bulk RNA-seq datasets**

***The Cancer Genome Atlas (TCGA) data:*** RNA-seq data from the TCGA analysis pipeline (RNASeqV2) was downloaded from Broad Institute GDAC Firehose server as raw count data using R package TCGA2STAT [33](#_ENREF_33) (version 1.2). The resulting data consisted of RSEM [34](#_ENREF_34) count values for 20,501 genes, which usually range from 0 to 106. The raw counts were normalized using trimmed mean of M-values (TMM) [35](#_ENREF_35) followed by a subsequent transformation to log-CPM (counts per million) values using voom [36](#_ENREF_36), which are assumed to be approximately normally distributed. Additionally, these transformed values were quantile normalized. Patient clinical data was accessed from the Genomic Data Commons (GDC) data portal after obtaining appropriate permissions.

***METABRIC data:*** Previouslynormalized [37](#_ENREF_37)gene expression and clinical data were obtained from the European Genome-Phenome Archive (EGA) under the accession id EGAS00000000098 after appropriate permissions. The discovery dataset was composed of 997 primary breast tumors and a second validation set was composed of 995 primary breast tumors. The expression data were arrayed on Illumina HT12 Bead Chip composed of 48,803 transcripts. Transcript-level gene expression values were aggregated to gene-level counts using median expression values resulting in a gene expression matrix of 28,503 genes.

***Ductal carcinoma in situ (DCIS) and invasive ductal carcinoma (IDC) data:*** Gene expression data from 31 patients with pure DCIS, 36 patients with IDC and 42 mixed DCIS/IDC along with 6 normal breast samples were downloaded from GEO under the accession id GSE26304. Transcript-level data was aggregated to gene-level using maximum values for that gene across all the transcripts. Data were arrayed on Agilent-012391 Whole Human Genome Oligo Microarray G4112A platform.

***Immune checkpoint blockade data from CA209-038 study:*** RNA-seq data from 109 samples (58 on-treatment with Nivolumab and 51 pre-treatment) over 65 patients with advanced melanoma were downloaded from Riaz *et al* [*38*](#_ENREF_38) at https:// github.com/riazn/bms038\_analysis/). Thirty-five patients had previously progressed on ipilimumab therapy, and the other thirty-three patients were ipilimumab-naïve. The raw counts were normalized using trimmed mean of M-values (TMM) followed by a subsequent transformation to log-CPM (counts per million) values using voom, which are assumed to be approximately normally distributed.

**Identifying modules and developing Random Forest classifiers or prediction models**

In order to develop modules of ECA activity, patient samples in ECA ‘Low’ and ECA ‘High’ groups were identified and the corresponding UMI count data was processed using the R package ‘Seurat’ (version 2.0.1) [39](#_ENREF_39). In order to focus more on biologically meaningful variation, we focused on the highly variable genes while controlling for the strong relationship between variability and average expression. We chose genes whose log-mean values were between 0.015 and 3.2, and whose dispersion was above 0.5, resulting in 4,690 highly variable genes. We next performed spectral decomposition or principal component analysis of this matrix to compute scores (pca scores) with each principal component representing a metagene or module that combines information across a correlated gene set. In order to determine the number of significant modules for any downstream analysis, we ran a resampling test [40](#_ENREF_40) inspired by jackstraw procedure [41](#_ENREF_41), which involves computing an empirical null distribution of the pca scores by repeated resampling of 1% of the data. Jackstraw analysis of 1,000 data replicates resulted in eight modules at p-value less than or equal to 0.05. Functional enrichment analysis of each module was implemented by R package ‘clusterProfiler’ (version 3.8.1) [42](#_ENREF_42).

Each module was then used to train Random Forest (RF) classifiers implemented in R package ‘caret’ (version 6.0.81) [43](#_ENREF_43). For every RF classifier, we optimized the following three parameters until prognostic significance was achieved by the final predictions – 1) mtry: the number of genes randomly sampled at every tree split, 2) ntree: the number of trees to grow in a RF model, and 3) data partitioning into train and test sets. In order to optimize the ‘mtry’ parameter, we tried a minimum of 50 genes and a maximum number not exceeding the number of genes in the RF model; we tried 10 – 1000 trees at increments of 50, and tried train/test partition sizes of 40%, 45%, 50%, 60%, 75% and 100%. To avoid overfitting, each classifier was validated using 10-fold cross-validation procedure.

**Survival analyses with the RF classifiers developed from the modules**

In order to assess the prognostic significance of the RF classifiers developed from each module, we generated Kaplan-Meier survival curves on the predicted groups (ECA ‘Low’ or ECA ‘High’) using R package ‘survminer’ (version 0.4.3). Distributions of Kaplan-Meir survival curves for progression-free and overall survival were compared using log-rank test, and a log-rank test p*-*value ≤ 0.05 is considered to be statistically significant. For the univariable forest plots implemented using R package ‘survutils’ (version 1.0.1), hazard ratios were derived using Cox proportional hazards survival models with the end points of progression-free survival.

**Statistical analyses**

All the statistical analyses were performed in R (version 3.5.1). Unpaired t-test was used for group comparisons and one-way ANOVA for was used for comaprisons between more than two groups. Statistical significance for all the analyses was accepted at p-value < 0.05. For all differential expression analyses, R package ‘limma’ was used and the p-values were adjusted for multiple comparisons using Benjamini-Hochberg method.

**Data availability**

All the codes required to reproduce the analysis are located at <https://github.com/cramanuj/ECA_modules_paper>

Further information and requests for resources should be made to the corresponding author, Herbert K. Lyerly ([kim.lyerly@duke.edu](mailto:kim.lyerly@duke.edu)).

**References**

1. June, C.H. Adoptive T cell therapy for cancer in the clinic. *J Clin Invest* **117**, 1466-1476 (2007).

2. Ayers, M.*, et al.* IFN-gamma-related mRNA profile predicts clinical response to PD-1 blockade. *J Clin Invest* **127**, 2930-2940 (2017).

3. Burugu, S., Dancsok, A.R. & Nielsen, T.O. Emerging targets in cancer immunotherapy. *Semin Cancer Biol* **52**, 39-52 (2018).

4. Denkert, C.*, et al.* Tumour-infiltrating lymphocytes and prognosis in different subtypes of breast cancer: a pooled analysis of 3771 patients treated with neoadjuvant therapy. *Lancet Oncol* **19**, 40-50 (2018).

5. Paik, S.*, et al.* A multigene assay to predict recurrence of tamoxifen-treated, node-negative breast cancer. *N Engl J Med* **351**, 2817-2826 (2004).

6. van 't Veer, L.J.*, et al.* Gene expression profiling predicts clinical outcome of breast cancer. *Nature* **415**, 530-536 (2002).

7. Parker, J.S.*, et al.* Supervised risk predictor of breast cancer based on intrinsic subtypes. *J Clin Oncol* **27**, 1160-1167 (2009).

8. Hendry, S.*, et al.* Assessing Tumor-infiltrating Lymphocytes in Solid Tumors: A Practical Review for Pathologists and Proposal for a Standardized Method From the International Immunooncology Biomarkers Working Group: Part 1: Assessing the Host Immune Response, TILs in Invasive Breast Carcinoma and Ductal Carcinoma In Situ, Metastatic Tumor Deposits and Areas for Further Research. *Adv Anat Pathol* **24**, 235-251 (2017).

9. Morse, M.A., Clay, T.M., Hobeika, A.C., Mosca, P.J. & Lyerly, H.K. Surrogate markers of response to cancer immunotherapy. *Expert Opin Biol Ther* **1**, 153-158 (2001).

10. Jiang, P.*, et al.* Signatures of T cell dysfunction and exclusion predict cancer immunotherapy response. *Nat Med* **24**, 1550-1558 (2018).

11. Savas, P.*, et al.* Publisher Correction: Single-cell profiling of breast cancer T cells reveals a tissue-resident memory subset associated with improved prognosis. *Nat Med* **24**, 1941 (2018).

12. Savas, P.*, et al.* Single-cell profiling of breast cancer T cells reveals a tissue-resident memory subset associated with improved prognosis. *Nat Med* **24**, 986-993 (2018).

13. Breiman, L. Random Forests. *Machine Learning* **45**, 5-32 (2001).

14. Emens, L.A.*, et al.* Long-term Clinical Outcomes and Biomarker Analyses of Atezolizumab Therapy for Patients With Metastatic Triple-Negative Breast Cancer: A Phase 1 Study. *JAMA Oncol* (2018).

15. Adams, S.*, et al.* Pembrolizumab Monotherapy for Previously Treated Metastatic Triple-Negative Breast Cancer: Cohort A of the Phase 2 KEYNOTE-086 Study. *Ann Oncol* (2018).

16. Adams, S.*, et al.* Title: Pembrolizumab Monotherapy for Previously Untreated, PD-L1-Positive, Metastatic Triple-Negative Breast Cancer: Cohort B of the Phase 2 KEYNOTE-086 Study. *Ann Oncol* (2018).

17. Bianchini, G.*, et al.* Immune modulation of pathologic complete response after neoadjuvant HER2-directed therapies in the NeoSphere trial. *Ann Oncol* **26**, 2429-2436 (2015).

18. Teschendorff, A.E., Miremadi, A., Pinder, S.E., Ellis, I.O. & Caldas, C. An immune response gene expression module identifies a good prognosis subtype in estrogen receptor negative breast cancer. *Genome Biol* **8**, R157 (2007).

19. Desmedt, C.*, et al.* Biological processes associated with breast cancer clinical outcome depend on the molecular subtypes. *Clin Cancer Res* **14**, 5158-5165 (2008).

20. Strasner, A. & Karin, M. Immune Infiltration and Prostate Cancer. *Front Oncol* **5**, 128 (2015).

21. Geissler, K.*, et al.* Immune signature of tumor infiltrating immune cells in renal cancer. *Oncoimmunology* **4**, e985082 (2015).

22. Yue, Q.*, et al.* The prognostic value of Foxp3+ tumor-infiltrating lymphocytes in patients with glioblastoma. *J Neurooncol* **116**, 251-259 (2014).

23. Han, S.*, et al.* Tumour-infiltrating CD4(+) and CD8(+) lymphocytes as predictors of clinical outcome in glioma. *Br J Cancer* **110**, 2560-2568 (2014).

24. Kang, B.W., Kim, J.G., Lee, I.H., Bae, H.I. & Seo, A.N. Clinical significance of tumor-infiltrating lymphocytes for gastric cancer in the era of immunology. *World J Gastrointest Oncol* **9**, 293-299 (2017).

25. Toss, M.S.*, et al.* Prognostic significance of tumor-infiltrating lymphocytes in ductal carcinoma in situ of the breast. *Mod Pathol* **31**, 1226-1236 (2018).

26. Edgar, R., Domrachev, M. & Lash, A.E. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Res* **30**, 207-210 (2002).

27. Trapnell, C.*, et al.* The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. *Nat Biotechnol* **32**, 381-386 (2014).

28. Barbie, D.A.*, et al.* Systematic RNA interference reveals that oncogenic KRAS-driven cancers require TBK1. *Nature* **462**, 108-112 (2009).

29. Hanzelmann, S., Castelo, R. & Guinney, J. GSVA: gene set variation analysis for microarray and RNA-seq data. *BMC Bioinformatics* **14**, 7 (2013).

30. Kokiopoulou, E., Bekas, C. and Gallopoulos, E. Computing smallest singular triplets with implicitly restarted Lanczos bidiagonalization. *Applied Numerical Mathematics* **49**, 39-61 (2004).

31. Hinton, L.J.P.v.d.M.a.G.E. Visualizing High-Dimensional Data Using t-SNE. *Journal of Machine Learning Research* **9**, 2579-2605 (2008).

32. Rodriguez, A. & Laio, A. Machine learning. Clustering by fast search and find of density peaks. *Science* **344**, 1492-1496 (2014).

33. Wan, Y.W., Allen, G.I. & Liu, Z. TCGA2STAT: simple TCGA data access for integrated statistical analysis in R. *Bioinformatics* **32**, 952-954 (2016).

34. Li, B. & Dewey, C.N. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics* **12**, 323 (2011).

35. Robinson, M.D. & Oshlack, A. A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biol* **11**, R25 (2010).

36. Law, C.W., Chen, Y., Shi, W. & Smyth, G.K. voom: Precision weights unlock linear model analysis tools for RNA-seq read counts. *Genome Biol* **15**, R29 (2014).

37. Margolin, A.A.*, et al.* Systematic analysis of challenge-driven improvements in molecular prognostic models for breast cancer. *Sci Transl Med* **5**, 181re181 (2013).

38. Riaz, N.*, et al.* Tumor and Microenvironment Evolution during Immunotherapy with Nivolumab. *Cell* **171**, 934-949 e916 (2017).

39. Butler, A., Hoffman, P., Smibert, P., Papalexi, E. & Satija, R. Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nat Biotechnol* **36**, 411-420 (2018).

40. Macosko, E.Z.*, et al.* Highly Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets. *Cell* **161**, 1202-1214 (2015).

41. Chung, N.C. & Storey, J.D. Statistical significance of variables driving systematic variation in high-dimensional data. *Bioinformatics* **31**, 545-554 (2015).

42. Yu, G., Wang, L.G., Han, Y. & He, Q.Y. clusterProfiler: an R package for comparing biological themes among gene clusters. *OMICS* **16**, 284-287 (2012).

43. Kuhn, M. Building predictive models in R using the caret package. *Journal of Statistical Software* **28**(2008).