Manuscript Title:

A novel visual analysis interface enables integrative analyses of cancer transcriptomics data and identifies potential markers of immunotherapy response via machine learning.

Authors:

M. Emre Kus[1], Cagatay Sahin[1,†], Emre Kilic[1,†], Arda Askin[1,†], M. Mert Ozgur[2], Gokhan Karahanogulları[3], Ahmet Aksit[4], Ryan M. O'Connell[5], H. Atakan Ekiz[1,*]

Affiliations:

[1] The Department of Molecular Biology and Genetics, Izmir Institute of Technology, Gulbahce, Izmir 35430, TURKEY

[2] The Department of Molecular Biology and Genetics, Bilkent University, Cankaya, Ankara 06800 TURKEY

[3] The Department of Mathematics, Izmir Institute of Technology, Gulbahce, Izmir 35430, TURKEY

[4] The Department of Information Technologies, Izmir Institute of Technology, Gulbahce, Izmir 35430, TURKEY

[5] The Department of Pathology, University of Utah, Salt Lake City, Utah 84112, USA

[†] These authors contributed equally

* To whom correspondence should be addressed. Phone: +90-232-750-7363, Email: atakanekiz@iyte.edu.tr

Abstract:

The Cancer Genome Atlas (TCGA) initiative has been essential for revealing key mechanisms in human cancer leading to the development of novel therapeutics. Analysis of the cancer transcriptomics data in the TCGA and other public repositories require coding skills that are associated with a steep learning curve for most non-specialists. To enable a wider utilization of these data, we introduce The Cancer Genome Explorer (TCGEx), a web-based visual data analysis interface that can perform a number of sophisticated analyses ranging from survival modeling and gene set enrichment analysis to unsupervised clustering and linear regression-based machine learning. In addition to providing access to preprocessed data from TCGA and immune checkpoint inhibition studies on cBioportal and CRI-iAtlas, the TCGEx platform allows users to upload and investigate their own data. Using this tool, we investigated the molecular subsets of human melanoma and identified microRNAs associated with intratumoral interferon signaling. We validated these findings using independent data from clinical trials involving immune checkpoint inhibitors for melanoma and other cancers. Moreover, our analyses unveiled a subset of cytokines predictive of positive responses to diverse immune checkpoint inhibitors prior to treatment initiation. Built on the R/Shiny framework, TCGEx modules offer customizable features to tailor the analysis to different study contexts and help generate publication-ready plots. TCGEx is freely available at https://tcgex.iyte.edu.tr, and it provides an interactive solution to extract meaningful insights from cancer transcriptomics data and guide scientific inquiry.

Keywords: TCGA, transcriptomics, immunotherapy, integrative analysis, interactive data analysis, survival modeling, machine learning, Shiny

Introduction

The Cancer Genome Atlas (TCGA) launched in 2006 is among the most ambitious projects to better understand and tackle human cancer (1). As of this writing, more than 11,000 tumor samples and matched healthy tissues have been characterized at the molecular level across 33 different cancer types as part of this initiative. TCGA studies employ a variety of experimental approaches including RNA and microRNA (miRNA) sequencing, whole exome sequencing, and genotyping and methylation arrays to generate 2.5+ petabytes of publicly available data. Landmark studies from the TCGA network described common genetic aberrations in cancer such as TP53 loss (2) as well as other cancer-specific changes including BRCA1/2 mutations in breast and ovarian tumors (3, 4) and APC mutations in colorectal carcinoma (5). Furthermore, transcriptomic profiling has led to the characterization of immune signatures within the tumor microenvironment (TME). Follow-up studies have revealed that the genomic and transcriptomic features of tumors determine whether they will respond to targeted therapies and immunotherapeutics (6–9), suggesting that a better understanding of the heterogeneity within the TME is essential for improved therapeutics and clinical outcomes.

Analyzing genomic data in the TCGA and other public repositories effectively requires programming skills and it is often challenging for the non-specialists. Multiple web-based applications have been developed to facilitate TCGA data analysis including cBioPortal, Web-TCGA, UCSC Xena, and GEPIA2 (10–13). Although these visual analysis interfaces reduced the barriers to data utilization for many investigators in the field, there is still a need for a centralized platform that integrates multiple methods for customized analyses. In particular, no interactive tool to date is equipped with sophisticated functional genomic analysis capabilities that allow examining transcriptomics data from various sources including TCGA, landmark immunotherapy studies, and one's own research. To address this need and democratize the use of high-throughput data in the cancer research field, we introduce The Cancer Genome Explorer (TCGEx), an open-source web-based analysis platform written in R/Shiny that offers several modules for sophisticated custom analyses. Using this novel interface, we investigated the TCGA skin cutaneous melanoma (SKCM) data and identified gene expression patterns and mutational subtypes associated with differential survival. We also examined intratumoral immune signatures using TCGEx machine learning modules and revealed miRNAs associated with elevated interferon-gamma (IFNγ) signaling. We validated these findings on independent data sets from immune checkpoint inhibition (ICI) trials and further identified a cytokine signature associated with positive responses to ICI across multiple types of cancers.

TCGEx features 10 analysis modules that can be flexibly adapted to various study contexts. Integrative analyses by aggregating data from multiple projects are also possible to investigate commonalities

and differences among cancers. TCGEx's comprehensive computational pipeline utilizes pre-processed data from the TCGA repository that was prepared by combining transcriptomics data with clinical metadata and intratumoral immune signatures described by the leading studies in the field (14). Additionally, TCGEx features transcriptomic data from multiple landmark ICI studies (6, 15–26) and it allows users to upload and analyze their own data sets. We provide step-by-step in-app tutorials to increase user-friendliness and help researchers implement TCGEx into their workflows and generate publication-ready plots. TCGEx is publicly available without requiring user registration at https://tcgex.iyte.edu.tr web server as well as through a docker image to facilitate local execution and open-source program development (https://hub.docker.com/r/atakanekiz/tcgex). The source code of TCGEx is accessible at https://github.com/atakanekiz/TCGEx.

Materials and Methods

Summary of the pipeline

TCGEx was created using the R programming language and Shiny framework and it runs on a publicly accessible GNU/Linux server. TCGEx is designed to work with pre-processed harmonized data allowing rapid and interactive analyses of TCGA projects, landmark ICI studies, and custom data sets. Pre-processed TCGA data for the pipeline were prepared by integrating RNA sequencing (RNAseq), microRNA sequencing (miRNAseq) data, and clinical metadata. The latter data type contains dozens of features ranging from patient demographics to tumor subtypes and genetic aberrations. We further expanded the sample-level metadata by incorporating intratumoral immune cell signatures described by Thorsson et al (14). The ICI data sets were downloaded and prepared from cBioPortal and CRI-iAtlas repositories (27, 28). TCGEx analysis starts by selecting one or more cancer projects in the data selection module. After data selection, relevant data objects are loaded into the session and passed to individual analysis modules **(Fig.1)**. TCGEx offers the following user-friendly modules with graphical interfaces: i) Kaplan-Meier survival analysis; ii) Cox proportional hazards survival analysis; iii) feature-to-metadata exploratory graphing; iv) feature-to-feature correlation graphing; v) correlative features analysis; vi) heatmap and hierarchical clustering; vii) gene set enrichment analysis (GSEA); viii) receiving operating characteristics (ROC) analysis; ix) principal component analysis (PCA); and x) lasso/ridge/elastic net machine learning analysis. Each module allows users to examine specific data subsets such as metastatic and/or primary tumor biopsies as well as healthy control samples. Furthermore, flexible options enable users to tailor the pipeline to their specific needs and generate informative publication-ready plots. The TCGEx Shiny app consists of modularized R scripts facilitating troubleshooting and incorporation of the code into other analysis pipelines. Thus, TCGEx aims to reduce barriers to high-throughput data analysis in cancer research by providing a scalable and user-friendly solution.

Data Selection Module

This module provides the interface for selecting from 33 TCGA and 14 ICI data sets to be analyzed in TCGEx. Upon data selection, pre-processed data objects are loaded, and descriptive statistics are visualized to help users examine the age and gender distribution of the selected patients as well as the tissue origin of the biopsy samples. TCGA data objects were prepared by the sample-level merging of the log(cpm+1) transformed RNAseq/miRNAseq data and clinical metadata. To facilitate downstream analyses, we removed genes that are not expressed in 25% or more of the samples in each TCGA project. ICI studies obtained from CRI-iAtlas repository were quartile-normalized while the data downloaded from cBioPortal were normalized through RPKM, FPKM, or TPM methods. For convenience and more control over the analyses, we added options for normalizing user-uploaded data sets and filtering lowly-expressed genes. Unlike some of the previously available tools for TCGA data analysis (12, 13), TCGEx allows users to select single or multiple data sets for integrative analyses. If multiple TCGA projects are selected, pre-processed data sets are aggregated keeping only the matched genes and clinical metadata variables. In the end, the data selection module creates a tidy data frame containing normalized RNAseq/miRNAseq data and clinical metadata; and feeds it to the other analysis modules.

Survival Analysis Modules

Studying the association between gene expression and survival outcome is important for understanding key mechanisms of cancer progression (29). TCGEx features two distinct survival analysis methods: Kaplan-Meier (KM) and Cox proportional hazards (CoxPH) modeling (30, 31). In both modules, categorical and continuous variables can be examined. Categorical meta-variables can differ between data sets, but it can include tumor grade, mutation types, and patient demographic information, among other features. Continuous variables such as gene expression and intratumoral immune cell signatures (14) can be directly modeled in CoxPH analysis or they can be categorized as “high”, “mid”, and “low” at user-defined thresholds for KM analysis. If more than one cancer type is selected by the user, gene expression can be categorized across all the projects at a common threshold, or within each cancer type separately. In KM analysis involving categorical variables, the user can further determine data subsets to include in the analysis and further customize the pipeline. Both KM and CoxPH modules allow including covariates in the model to examine the association of multiple variables with the survival outcome simultaneously. Furthermore, the CoxPH module is designed to accept user-defined interaction terms to accommodate arbitrarily complex multivariable study designs. Survival analysis pipelines of TCGEx generate customizable plots and show various statistical details including log-rank p-value for KM and CoxPH, and likelihood ratio test and Wald test p-values along with hazard ratios for CoxPH. Additionally, the proportionality assumption of the CoxPH model is checked using Schoenfeld residuals against time, and the user is

prompted a message regarding the validity of the model. Taken together, TCGEx survival analysis modules provide flexible options for customizing analysis to specific needs and studying the relationship between gene expression and survival outcomes in cancer.

Boxplot module

In this module, users can explore the relationship between gene expression and categorical clinical meta-variables. Categorical variables vary according to tumor type, and they include disease pathological stage, patient gender and race, tumor molecular subtype (32, 33), genome-wide methylation state (33), oncogenic signaling (OncoSign) cluster type (34), genetic aberrations (2), genomic instability levels (5), intratumoral lymphocyte scores (33), and numerous other metadata features **(Supplementary Table-1)**. This module allows selecting specific categories to generate gene expression plots to emphasize differences in data subsets. To offer more flexibility in the analyses, we also added an optional faceting variable which can further break data into sub-categories. This way, for instance, it becomes possible to plot the expression of a given gene in various mutation subsets in male and female patients separately. To facilitate publication-ready plotting, the program allows the user to select journal-specific color palettes, toggle individual data points on/off, and define specific statistical comparisons to highlight. Here, users can select the t-test or Wilcoxon test for pairwise comparisons between data sub-groups. Thus, this module helps researchers explore gene expression in clinically meaningful data subsets.

Correlation analysis modules

Examining correlations among genes and continuous metadata features can help study co-regulation mechanisms (6) and miRNA-mRNA interactions (35). TCGEx offers two modules for correlation analysis which can generate sample-level scatter plots or correlation matrices and tables. Sample-level plots are generated by plotting two continuous variables against each other, where each point denotes a patient sample. Continuous variables for this analysis can include transcript levels from RNAseq and miRNAseq data and numerical metadata such as mutation load or intratumoral immune cell scores (14). The correlation scatter plot is designed to be responsive, and it can show patient barcode, gender, and race upon hovering over individual data points. Furthermore, an optional faceting variable can be used to break data points into various subcategories and generate informative plots, as described in the previous modules. Users can visualize the best-fitting line and the linear regression equation on the graph to demonstrate the relationship between variables. In addition to these functionalities, we anticipated that it could be of interest to examine the top positive and negative correlators of a specific gene and plot a gene-to-gene correlation matrix. The second correlation analysis module of TCGEx enables these analyses by calculating pairwise Pearson and Spearman correlation coefficients and creating tabular and graphical

outputs. Therefore, these pipelines facilitate examining the linear relationships between variables and help investigate potential genetic interactions.

Heatmap and hierarchical clustering module

Heatmaps are commonly used for visualizing high-throughput gene expression data. Especially when combined with hierarchical clustering, heatmaps can reveal interesting patterns in the data and help categorize samples with different characteristics (36). We created an interactive heatmap module equipped with flexible options to facilitate visualization and analysis of transcriptomics data. Here, users can manually enter genes of interest, or select curated gene sets described in the Molecular Signatures Database (MSigDB) (37). Extensively utilized in both cancer and non-cancer literature, MSigDB contains thousands of annotated gene sets that define numerous biological states and processes. We anticipated that heatmaps created with these pre-defined gene sets could be especially informative when paired with user-selected sample-level annotations. Thus, we added functionality to show custom annotation bars above the heatmaps. In these annotation bars, both categorical and continuous features can be specified to generate complex heatmaps. As described previously, numerous cancer-specific categorical features are available for selection including tumor stage, mutational subtype, and oncogenic copy number alterations (38). When continuous annotation variables are selected, we designed the pipeline to categorize the samples at the median value to create two color-coded groups. The pipeline also allows selecting multiple continuous variables, in which case the categorization is done for each variable separately or after averaging the variables, thus making it possible to create meta-features on the fly. In the heatmap module, users can filter out genes with low variance and change the hierarchical clustering parameters to tailor the analysis to specific needs. Therefore, this module can help researchers create informative heatmaps to highlight meaningful patterns in gene expression data.

Gene set enrichment analysis module

In transcriptome analysis, expression patterns of genes in specific pathways can be examined to study the biological states of the samples. Gene set enrichment analysis (GSEA) is a powerful approach for comparing two groups of samples by focusing on a list of genes that share common biological characteristics (39). We developed an efficient GSEA pipeline that allows comparisons between user-defined data subsets. Similar to previous modules, users can define these data subsets by selecting specific values for categorical variables (such as mutation subtypes, tumor stage, and patient gender), or by binarizing continuous variables (such as gene expression, nonsilent mutation rate, and intratumoral immune scores) at custom thresholds. After defining two data subsets, the pipeline ranks genes based on their expression levels and variance using the previously defined signal-to-noise approach (39). The

GSEA module utilizes thousands of readily available gene sets from the MSigDB repository and facilitates the investigation of various biological pathways in iterative analyses (37). We further added gene signatures predictive of immunotherapy response in numerous studies that are curated in a recent report (40). Importantly, to further increase the usability of this module, we gave the user an option to provide custom gene sets and examine their expression patterns in user-defined data subsets. This module can create enrichment plots for individual gene sets and print a table of leading-edge genes that mostly drive the enrichment scores (39), or it can show the most highly enriched gene sets among others provided to the pipeline. Taken together, the GSEA module provides a flexible and customizable platform for examining functional associations in cancer gene expression data.

Receiving operator characteristics analysis module

Receiving operator characteristics (ROC) analysis is a method for evaluating the discriminatory performance of variables in a binary classification system (41). ROC curves are generated by plotting the true positive rate (i.e. sensitivity) against the false positive rate (i.e. 1-specificity) across all possible thresholds for the variable of interest. The area under the ROC curve (AUC or AUROC) can be examined to assess the power and accuracy of the classifier, a commonly used approach in the field for investigating diagnostic and prognostic biomarkers (42–44). TCGEx features a powerful and flexible module for performing ROC analyses using gene expression data and clinical metadata. Users can binarize features of interest at custom thresholds to create two groups needed for the ROC analysis and specify custom predictor variables to assess their classifier potential. Using this pipeline one can examine, for instance, whether the expression of a specific gene is associated with certain tumor characteristics such as mutation rate, and intratumoral immune signatures. As in the previously described modules, users can work with both categorical variables and numerical variables through flexible input options and tailor the analysis to specific needs. We also provided an option to add ROC curves to the graph generated using MSigDB gene sets. This design facilitates comparing user-selected custom predictors and other functionally annotated predictors. Thus, we anticipate that the ROC module can be conveniently used in various study contexts to assess novel predictors in cancer data.

Principal component analysis module

Dimensionality reduction techniques such as principal component analysis (PCA) are commonly employed in transcriptomics (45). PCA involves constructing a new coordinate system for the data using linear combinations of its variables (i.e., genes in RNAseq). The axes of this new coordinate system, called principal components, represent the variation in the experiment, and they help visualize multidimensional data sets on a 2D graph. TCGEx program features a user-friendly PCA module with flexible

227 inputs to be utilized in various study contexts. Similar to other modules, one can specify sample types
228 and select a custom list of genes to prepare the gene expression matrix for PCA. Here, the module
229 conveniently allows selecting all genes in the RNAseq/miRNAseq data or specifying genes from MSigDB
230 pathways. Furthermore, convenient options allow selecting miRNAs or lncRNAs exclusively to tailor anal-
231 yses to specific contexts. Although PCA accounts for both highly and lowly variable genes, we provided
232 a pre-filtering option to speed up the analysis by removing genes with low variance. In PCA plots, it is
233 usually helpful to annotate individual points based on specific characteristics. For instance, one may want
234 to see whether samples with certain genetic features form a separate cluster in the PCA space (46). To
235 address this need, this module allows color-coding data points based on categorical or continuous vari-
236 ables, the latter of which is categorized at the median value to form “high” and “low” groups. PCA module
237 generates customizable and interactive graphics and allows exporting publication-ready plots.

238 **Machine learning module**

239 Machine learning (ML) refers to a range of applications and algorithms that aim to extract relevant and
240 useful information from data. In the context of bioinformatics, ML is utilized to perform classification, pre-
241 diction, and feature selection from biological data (47, 48). Regularized regression, a type of supervised
242 machine learning technique, is a derivative of linear regression that allows one to simultaneously create
243 a model and perform feature selection in high dimensional data. It is optimized to minimize the sum of
244 squared residuals while penalizing the generated model coefficient estimates. The penalty term is applied
245 to the model equation to reduce model complexity and make the prediction with the limited mean squared
246 error. Ridge, lasso, and elastic net regression are all types of regularized regression methods with varying
247 strengths depending on the underlying structure of the data. Ridge regression shrinks the estimated
248 coefficients without making them zero and assigns correlating parameters with similar coefficients (49).
249 Lasso regression on the other hand can shrink model coefficients to zero and therefore performs feature
250 selection (50). Elastic net regression can refer to a middle ground in between ridge and lasso regression
251 and it is ideal for data sets in which the number of predictor variables significantly exceeds the number
252 of samples and/or there is a high number of correlating variables. The optimal degree of penalization for
253 each approach can be selected using cross-validation across the data set. The ML module in TCGEx
254 provides a user-friendly interface for applying these regularized linear regression methods on the tran-
255 scriptomics data. Users choose response and predictor variables for the model and specify the ML algo-
256 rithm by setting the penalization parameter alpha. The response variable (i.e., dependent variable) in
257 these analyses can be constructed from a custom list of genes or MSigDB gene sets by taking the aver-
258 age of the expression values. Alternatively, users can select categorical metadata features as the re-
259 sponse variable and binarize data subsets in a custom way to adapt the pipeline to various study

260 contexts. This module also provides convenient options to specify different types of predictor variables
261 (i.e. independent variables) including genes belonging to specific biological pathways or encoding miR-
262 NAs and lncRNAs. This way, one can easily examine, for instance, miRNA/lncRNAs that are most closely
263 associated with a pathway of interest (51). The ML module can use the entire data to train a model or
264 split it into training and testing subsets to evaluate the overall model accuracy (48). The findings of the
265 analyses are displayed through interactive graphs which show penalized coefficients and the mean
266 squared error of the cross-validated model across various levels of regularization. Taken together, the
267 ML module provides a customizable platform for sophisticated analysis of the TCGA transcriptomics data.

268 **Results and Discussion:**

269 In this study, we have used the TCGEx pipeline to investigate the transcriptomic landscape of human
270 melanoma and other cancers with the goal of identifying gene expression patterns associated with anti-
271 tumor immunity and immunotherapy response. Melanoma is an aggressive form of cancer that develops
272 in pigment-producing cells of the skin, although it can rarely be observed in other organs including the
273 eye and gastrointestinal tract as well. More than 320,000 new cutaneous melanoma cases are reported
274 globally each year and 55,000+ lives are lost because of this disease (52). UV exposure is a significant
275 risk factor for the development of melanoma, and most cases are diagnosed at an advanced stage where
276 the cancer cells have already metastasized to nearby lymph nodes or distant organs. Accordingly, the
277 TCGA-SKCM study is composed of mainly metastatic tumor biopsies, as shown in the TCGEx data se-
278 lection module **(Fig.2a)**. These samples originated from comparable numbers of male and female pa-
279 tients in various age groups **(Fig.2b)**. The landmark TCGA paper describing the genetic features of SKCM
280 classified tumors into subtypes based on mutations in BRAF, NF1, and RAS genes (33). These subtypes
281 are characterized by distinct sets of accompanying genomic aberrations including PTEN loss in BRAF-
282 mutant tumors, KIT amplification in triple wild-type (WT) tumors, Akt3 amplification in RAS-mutant and
283 wild-type tumors which are readily observed in transcriptomics data **(Fig.2c)**. When these graphs are
284 faceted by patient gender, some of these genes showed sex-dependent differences in their expression
285 patterns **(Fig.S1a, b)**. Interestingly, BRAF-mutant melanomas expressed higher levels of transferrin (TF)
286 which was recently shown to protect circulating tumor cells from oxidative stress leading to BRAF inhibitor
287 resistance **(Fig.S1c)** (53). On the other hand, NF1-mutant melanoma was distinctly marked by the high
288 expression of hepatocyte-specific transcription factor ONECUT1 which was shown to be associated with
289 tumor progression **(Fig.S1c)** (54). Our analyses also revealed that ZC3H13, a component of the RNA
290 methyltransferase complex, is enriched in RAS-mutant melanoma; and triple-WT tumors selectively ex-
291 pressed FOXF2, a transcription factor that plays a complex role in regulating oncogenic signaling

292 pathways **(Fig.S1c)** (55). These findings suggest that distinct molecular mechanisms may be responsible
293 for driving different tumor subtypes in melanoma and indicate possible intervention points.

294 BRAF-mutant melanoma constitutes about 50% of the cases in the TCGA cohort followed by RAS-
295 mutant (~30%), NF1-mutant, and triple-WT tumors (~10% each). Of these 4 groups, BRAF-mutant mel-
296 anoma has the most favorable prognosis **(Fig.2d)**, although this may be confounded by the younger age
297 of the patients in this group (33). While melanoma incidence is comparable between both sexes, females
298 were reported to have a reduced risk of mortality (56). Examination of the TCGA-SKCM cohort as a whole
299 did not show a significant survival difference between the two groups; however, a slightly increased risk
300 was observed for the males **(Fig.S2a, b)**. Notably, when we examined only BRAF and NRAS-mutant
301 tumors, males were associated with poorer survival outcomes, especially in the NRAS-mutant subgroup
302 **(Fig.2e)**. These findings suggest that sex-specific factors may differentially influence the prognosis in
303 different mutation subtypes.

304 In addition to defining potential genetic drivers in melanoma, the TCGA-SKCM manuscript also iden-
305 tified three distinct tumor subsets based on gene expression profiles as immune, keratin, and MITF-low
306 groups (33). The immune subset was characterized by the higher expression of immune cell-specific
307 genes **(Fig.3a)**, while the keratin and MITF-low subsets were characterized by the high and low expres-
308 sion of epithelial genes, respectively, in the absence of a clear immune infiltration signature. Of these
309 three groups, the immune subset exhibits the most favorable survival outcome **(Fig.3b, S3)**, which is
310 consistent with the positive role of inflammation in SKCM reported previously (14). IFNγ is a key inflam-
311 matory cytokine released from T and NK cells to eliminate tumor cells. IFNγ exerts its multifaceted func-
312 tions within the TME by signaling through IFNγ-receptors (IFNGR) and regulating dozens of downstream
313 genes in tumor cells and the neighboring immune cells (57). We examined IFNγ response gene expres-
314 sion signature in SKCM subsets and observed higher levels of most of these transcripts in the immune
315 subset **(Fig.3c)** (58), and as expected, these trends were paralleled by the elevated IFNγ transcript levels
316 and CD8+ T cell signatures within the TME. We next wanted to investigate the relationship between the
317 intratumoral CD8+ T cell score and the T cell receptor (TCR) diversity (14). This is of interest because
318 multiple studies have suggested that the clonally expanded T cells within melanoma are associated with
319 a spectrum of dysfunctional phenotypes (59, 60). Increasing CD8+ T cell signature within the TCGA-
320 SKCM tumors positively correlated with the overall TCR diversity, suggesting a polyclonal T cell infiltrate
321 and the presence of T cell clones with different antigenic specificities **(Fig.3d)** (61). Notably, the increased
322 TCR diversity correlated with higher levels of cytotoxic effector molecule perforin (PRF1) and decreasing
323 genetic signatures of M2 macrophages that are known to exert pro-tumorigenic functions **(Fig.3e)** (14).

Taken together, TCGEx can be utilized to explore multifaceted data types in TCGA including the tumor-specific subtypes and the intratumoral immune landscape.

MicroRNAs (miRNAs), 20-22 nucleotide noncoding RNAs responsible for post-transcriptional modulation, are key regulators of antitumor immunity and tumor immunoevasion (62). Given the importance of IFNγ signaling in cancer, we next focused on TCGA-SKCM miRNAseq data and utilized lasso regression to select top miRNAs associated with the IFNγ response. The patient cohort was divided into training (70%) and test (30%) subsets and the expression of miRNAs (529 genes after low-expression filtering) were modeled as the predictor variables against the average expression of IFNγ signature genes (MSigDB Hallmark IFNγ response gene set) as the response variable. 26 positive and 18 negative coefficients remained in the optimally regularized model (log lambda + 1 standard error) **(Fig.4a, S4a, b)**. 5 miRNAs, miRs-155-5p, -150-3p, -150-5p, -142-3p and -7702 were the last five predictors remaining in the model when the model penalty term lambda (λ) is increased further at the expense of minimized mean-squared error. As expected from prior work, these miRNAs showed a strong positive correlation with T cell effector molecules including IFNγ, PRF1, and granzyme B (GZMB) **(Fig.4b)**. Of note, top positive correlators of miR-155-5p were immune-specific genes such as markers of T cells and costimulatory receptors; and top negative correlators included regulators of nucleotide biosynthesis, metabolism, and signal transduction **(Fig.S4c-e)**. Since a single miRNA can control dozens of targets and miR-155-5p correlates with immunity in SKCM (63), we next examined whether the expression levels of miR-155-5p can distinguish the immune subset of SKCM from the keratin and MITF-low subsets. We performed ROC analysis and found that miR-155-5p predicted the immune-enriched SKCM subset with an AUC of 0.819, although the average expression of IFNγ response genes was a slightly better predictor with the AUC of 0.891 **(Fig.4c)**. However, we noted that the predictive power of the averaged expression of 5 miRNAs that remained in the highly regularized lasso model was a better predictor of the immune subset of SKCM **(Fig.4d)**. To further examine gene expression patterns in miR-155-5p-high melanoma, we performed GSEA using annotated Hallmark gene sets from the MSigDB database (58) and noted that the expression of genes associated with adaptive immunity was enriched in the miR-155-5p-high subset **(Fig.4e, f)**. Lastly, we performed dimensionality reduction using 685 genes belonging to the “adaptive immune response” gene ontology class and color-coded the individual samples based on miR-155-5p expression to see whether miR-155 levels can differentially mark samples in the PCA space. Interestingly, the first principal component distinguished a portion of the samples suggesting that miR-155 is a marker of adaptive immunity signature within the TME which is supported by the work from our group and others **(Fig.4g)** (63–66).

356 Next, we wanted to validate the associations between miRNAs and intratumoral IFNγ signature in
357 independent data sets obtained from melanoma biopsies prior to ICI therapy (6, 15–17, 21). No miRNA
358 sequencing has been done in these clinical trials yet, but the miRNA host gene expression is available
359 in the RNAseq data. We observed that the expression of the mature miRNAs and their host genes show
360 a strong positive correlation in the TCGA data **(Fig.S5)**, and thus we focused on the miRNA host genes
361 in the ICI data sets as a surrogate for mature miRNA expression. Using quartile-normalized integrated
362 data from 5 different melanoma ICI studies (313 total samples prior to treatment with anti-CTLA4 or anti-
363 PD1), we constructed lasso regression against the averaged IFNγ signature genes as the response var-
364 iable and 45 miRNA genes that are expressed in at least 25% of the samples as the predictor variables
365 **(Fig.5a)**. This model predicted 11 miRNAs with positive coefficients, including MIR155HG, and 2 miRNAs
366 with negative coefficients at the maximum penalization within 1 standard error of the minimum cross-
367 validation error **(Fig.5b, c)**. We showed that MIR155HG expression is positively correlated with the cyto-
368 toxic lymphocyte markers CD8A, IFNG, and PRF1 across melanoma ICI data sets **(Fig.5d)**. As expected,
369 the GSEA analysis revealed an enrichment of IFNγ signaling in MIR155HG-high samples **(Fig.5e)**. After
370 validating MIR155HG's association with intratumoral immunity in melanoma ICI data, we performed sim-
371 ilar analyses on pre-treatment Pan-ICI data prepared by integrating 10 quartile-normalized ICI studies (5
372 melanoma, 3 kidney, 1 stomach, and 1 bladder cancer studies adding up to 938 samples prior to treat-
373 ment with anti-CTLA4, anti-PD1 or anti-PD-L1 inhibitors) (6, 15–17, 20, 22, 24–26, 67). Lasso modeling
374 against averaged IFNγ-responsive genes revealed 15 (including MIR155HG), and 7 miRNA genes with
375 positive and negative coefficients, respectively **(Fig.5f-h)** which were mostly consistent with the mela-
376 noma-specific analyses. When the genome-wide expression patterns of the pan-ICI pre-treatment data
377 were examined, we noted that MIR155HG-high subset was associated with an enrichment of IFNγ-re-
378 sponsive gene set **(Fig.5i)** as well as other pathways associated with the immune response **(Fig.5j)**.
379 Taken together, machine learning-based analysis of melanoma and other tumor biopsies revealed the
380 miRNAs associated with antitumor immunity and further supported the connection between intratumoral
381 IFNγ signaling and MIR155HG. As these data originate from bulk methods, it is not possible to ascertain
382 the gene expression to specific cell types within the TME, however there is literature supporting the im-
383 munoregulatory roles of MIR155HG in both tumor cells, and the infiltrating immune cells (68, 69).

384 Pre-existing immune infiltration and particularly IFNγ signaling have been linked to better clinical out-
385 comes in cancer and improved responses to immunotherapies (6, 8). After observing the association of
386 MIR155HG with immune parameters within the TME, we next asked whether MIR155HG expression
387 prior to ICI could stratify melanoma survival and therapy response. We separately categorized patients
388 at the median MIR155HG for each of the 5 melanoma ICI data sets and, similar to what was observed
389 for TCGA-SKCM, noted MIR155HG-high tumors showed an improved overall survival **(Fig.6a)**. When

390 the patients were also segregated by immunotherapy response, we noted that the higher MIR155HG still
391 marked better prognosis in non-responding patients, but not in the responding cohort **(Fig.S6a)**. We next
392 wanted to examine the baseline differences at MIR155HG levels between responding and non-respond-
393 ing patients and noted that MIR155HG expression was elevated in one of the 5 melanoma ICI data sets
394 prior to treatment with anti-PD1 or a combination of anti-CTLA4/anti-PD1 ICI **(Fig.6b)** (15). Interestingly,
395 ICI-responsive tumors tended to express higher levels of MIR155HG in this data set both prior to and
396 during the immunotherapy although this was not statistically significant **(Fig.S6b)**. To understand whether
397 miRNA expression can predict positive responses to ICI, we binarized therapy response (complete and
398 partial responders vs progressive disease cases) and utilized binomial lasso regression in the pre-treat-
399 ment combined melanoma data set. Our modeling revealed 11 and 17 miRNAs with positive and negative
400 coefficients, respectively at the highest penalization level within 1 standard error of the maximum model
401 AUC **(Fig.S6c-e)**. Interestingly, some of the miRNAs associated with ICI-response in this analysis (in-
402 cluding MIR155HG, MIR23AHG and MIR181A1HG) were also observed when using lasso regression
403 against IFNγ response as the response variable. We then expanded the scope of the lasso regression
404 and utilized the aggregated Pan-ICI data set as described above to investigate miRNAs as potential
405 predictors of ICI response. This analysis returned only two miRNA genes with negative coefficients at the
406 maximal penalization level (value of lambda within 1 standard deviation for the maximum cross-validated
407 area under the roc curve [AUC]), however, relaxing the penalization level (value of lambda at the maxi-
408 mum cross-validated AUC) resulted in a total of 12 non-zero coefficients in the model **(Fig.6c-e)**.
409 MIR155HG was among the 3 miRNAs with positive coefficients in the model, although the overall model
410 performance was not too high when only using miRNAs as predictors. Although MIR155HG was not
411 found to be a powerful predictor of response in these analyses, it was consistently expressed across
412 many data sets examined for this study. Thus, we used a recently reported online portal, Immuno-Oncol-
413 ogy Signatures Explorer (IOSig) (40), to assess the association of MIR155HG with ICI responsiveness
414 in a larger cohort of ICI studies in comparison to other previously published response signatures **(Table-**
415 **S2,3)**. We noted that the single-gene MIR155HG signature was variably associated with ICI response,
416 although an AUC between 0.7-0.8 was seen for 8 out of 30 pre-ICI treatment data sets where MIR155HG
417 expression was detected **(Fig.6f, Table-S4)**. Notably, these studies involved various cancer types includ-
418 ing melanoma, kidney, breast, and gastrointestinal cancers. We validated IOSig platform's findings using
419 Gide et al. data set in the ROC module of TCGEx and showed that MIR155HG expression could distin-
420 guish responders (complete and partial response) from non-responders (stable or progressive disease),
421 albeit more poorly compared to a previously published 6-gene IFNγ response signature shown to predict
422 immunotherapy response **(Fig.6g)** (8). Of note, MIR155HG expression performed better in distinguishing
423 responders to the anti-PD1/anti-CTLA4 combination therapy in this data suggesting differences in

424 molecular pathways that are involved in the mechanisms of ICI therapeutics. Thus, our results suggest
425 that, while MIR155HG rises among other miRNAs to differentiate some ICI responders, it does not con-
426 stitute a strong signature to be generalized across data sets.

427 Finally, we set out to identify a gene signature that can perform well across a broader range of ICI
428 studies using the TCGEx analysis pipeline. As cytokine signaling is a core determinant of the immuno-
429 logical outcome in cancer, we utilized the 238-gene KEGG cytokine-cytokine receptor interaction gene
430 set curated in the MSigDB repository (37). We binarized ICI response as described above and con-
431 structed a binomial elastic net regression model (with elastic mixing parameter of alpha=0.5) using the
432 Pan-ICI pre-treatment data set described above. Out of the 238 genes, the calculated model retained 5
433 cytokines with positive coefficients, IFNG, CXCL9, CXCL13, IL21, and CCL5, at the highest regularization
434 level **(Fig.7a-c)**. We next used IOSig platform to assess the association of this signature (Cytokine-5)
435 with the immunotherapy response across a diverse set of cancer types and treatments. In these anal-
436 yses, Cytokine-5 signature attained a higher overall AUC value compared to the previously published
437 response signatures including 6- or 18-gene IFNγ response signature (8) and Impres signature (9), how-
438 ever, a degree of variation existed among datasets **(Fig.7d, S7a-b, Table-S5)**. Out of 34 pre-treatment
439 data sets expressing all 5 genes of the signature, 4 showed AUC values of 0.9 or more and, interestingly,
440 these studies were conducted on 4 different cancer types (thymic, colorectal, non-small cell lung cancer
441 and melanoma) and involved different treatment agents (anti-PD1, anti-CTLA4, or combination). Addi-
442 tionally, the Cytokine-5 signature corresponded to an AUC between 0.7-0.8 in 8 other data sets, and an
443 AUC between 0.6-0-7 for yet another 8 data sets across cancer types and treatment modalities. Our
444 analyses revealed that, among the published signatures available in the IOSig, Cytokine-5 has the high-
445 est average AUC score for the smallest number of genes in the signature for predicting immunotherapy
446 responsiveness **(Fig.7e)**. Notably, the 6-gene IFNg signature previously reported to predict anti-PD-1
447 response (8) showed positive correlation with the Cytokine-5 signature, although the Cytokine-5 score
448 performed better in predicting responses in multiple cancer types **(Fig.7f, S7b)**. It is important to note
449 that a previously reported 12-gene chemokine signature (marked as Chemokines in the heatmap) con-
450 taining CCL5 and CXCL9 had a lower albeit comparable overall performance in predicting ICI response.
451 Similarly, the 6- or 18-gene IFNγ signature described by Ayers et al. (8) (IFNG6 or IFNG18) demonstrated
452 a strong positive correlation with the Cytokine-5 signature, but it had a lower prediction power overall.
453 Two genes from Cytokine-5 signature, IFNG and CXCL9, overlapped with IFNG6 gene set, while CCL5
454 and CXC13 genes were common with IFNG18, suggesting that the unique composition of the Cytokine-
455 5 signature may provide a higher prediction performance across a broader range of ICI studies. To vali-
456 date these findings, we next examined the average expression of Cytokine-5 signature in the pre-ICI
457 datasets available on the TCGEx platform. As expected, the Cytokine-5 signature was found significantly

458 higher in responders at baseline across melanoma, kidney, stomach, and bladder cancers **(Fig.7g, S7c)**.
459 Interestingly, the higher levels of this signature in therapy-responsive melanoma samples were evident
460 both prior to and on ICI treatment **(Fig.7h)**. We also noted that the expression of this signature increased
461 during ICI treatment regardless of the response classification suggesting that the Cytokine-5 signature is
462 induced upon immunotherapy, albeit more strongly in responding tumors. Lastly, we examined how Cy-
463 tokine-5 signature performs in specific cancers and treatment types. Although the number of studies was
464 limiting for some cancers, Cytokine-5 score corresponded to an AUC of >0.75 for multiple cancer types
465 including thymic, colorectal, pancreatic, gastrointestinal, and non-small cell lung cancers **(Fig S8a)**.
466 When we categorized studies based on treatment type, we observed that Cytokine-5 signature has an
467 AUC of >0.6 for single agent anti-CTLA4 or anti-PD1 as well as their combinations, although these pat-
468 terns may be confounded by cancer type **(FigS8b)**. Collectively, these results highlight the utility of the
469 TCGEx web server to investigate cancer transcriptomics data and identify clinically relevant predictors
470 such as the Cytokine-5 signature in the context of immunotherapy.

471 TGCEx platform provides a user-friendly solution for high throughput data analysis and can benefit
472 researchers beyond the scope of this work. Although we developed this tool with RNAseq/miRNAseq in
473 mind and not tested elsewhere, the TCGEx modules are also compatible with other data types, such as
474 microarray, Nanostring, and CHIP-seq data. While this platform allows users to upload and explore their
475 own data through flexible interfaces, it is important to consider the effects of data normalization and
476 filtering on the analyses. We provide convenient input options to help with data processing, but more
477 customized data manipulations may be needed in specific study contexts. We openly provide the TCGEx
478 source code to help researchers who wish to further control the analysis pipeline and tailor the code to
479 their needs.

480 **Conclusions:**

481 Taken together, TCGEx provides a powerful and flexible interface for analyzing cancer transcriptomics
482 data originating from various sources. The TCGEx pipeline can be adapted to study various aspects of
483 cancer including oncogenic drivers, immune associations, prognostic markers in specific subtypes,
484 among others. Furthermore, TCGEx allows integrative analyses of multiple data sets to study the molec-
485 ular underpinnings of human cancer and enable users to interact with their own data. Visual analysis
486 interfaces increase the accessibility and reusability of high throughput data and facilitate research by
487 serving as a bridge between scientific disciplines. TCGEx, publicly and freely accessible on the web, not
488 only reduces barriers to data analysis for researchers with no programming experience but also can be
489 implemented into larger analysis pipelines to facilitate hypothesis-driven research.

Acknowledgments:

We thank the interim director of the IzTech IT Department Dr. Ozgur Orun for helpful discussions in hosting the TCGEx app and facilitating the establishment of the Shiny server at IzTech. We also extend our gratitude to Emin Bayindirli for his involvement during the initial development of the app. We are also grateful to the TCGA initiative as a whole and the patients who participated in the TCGA projects and clinical trials. Funding: A.A. received Abdi Ibrahim Foundation undergraduate scholarship. C.S., E.K., and M.M.O. received Turkish Health Institutes Directorate (TUSEB) undergraduate research project funding (TUSEB-A1-28154). C.S. and E.K. also received Scientific and Technological Council of Turkey (TUBITAK) 2247C-STAR scholarships. M.E.K. was supported by a TUBITAK-2210-A graduate scholarship. H.A.E. is supported by TUBITAK (2232-121C115, 1001-122S337), Turkish Academy of Sciences (TUBA-GEBIP-2022), and institutional grants (2022IYTE-2-0060, 2023IYTE-1-0053, 2023IYTE-1-0054).

Author Contributions:

M.E.K. prepared the input data, developed the code connecting different modules of TCGEx, participated in developing the ML module, and performed app deployment and maintenance. C.S. developed survival analysis and heatmap modules. E.K. developed the ROC and heatmap modules. A. Askin developed the PCA, correlation, and GSEA modules. M.M.O developed the ML module and participated in preparing and cleaning up the input data. G.K. contributed to the correlation modules and website analytics. A. Aksit was instrumental in setting up the Shiny server within the institutional infrastructure. R.O.C. provided critical insight during the development of the TCGEx app. H.A.E. developed the initial prototype for most of the modules, oversaw the TCGEx app development, and set up the web-accessible server. All authors contributed to preparing and reviewing the manuscript.

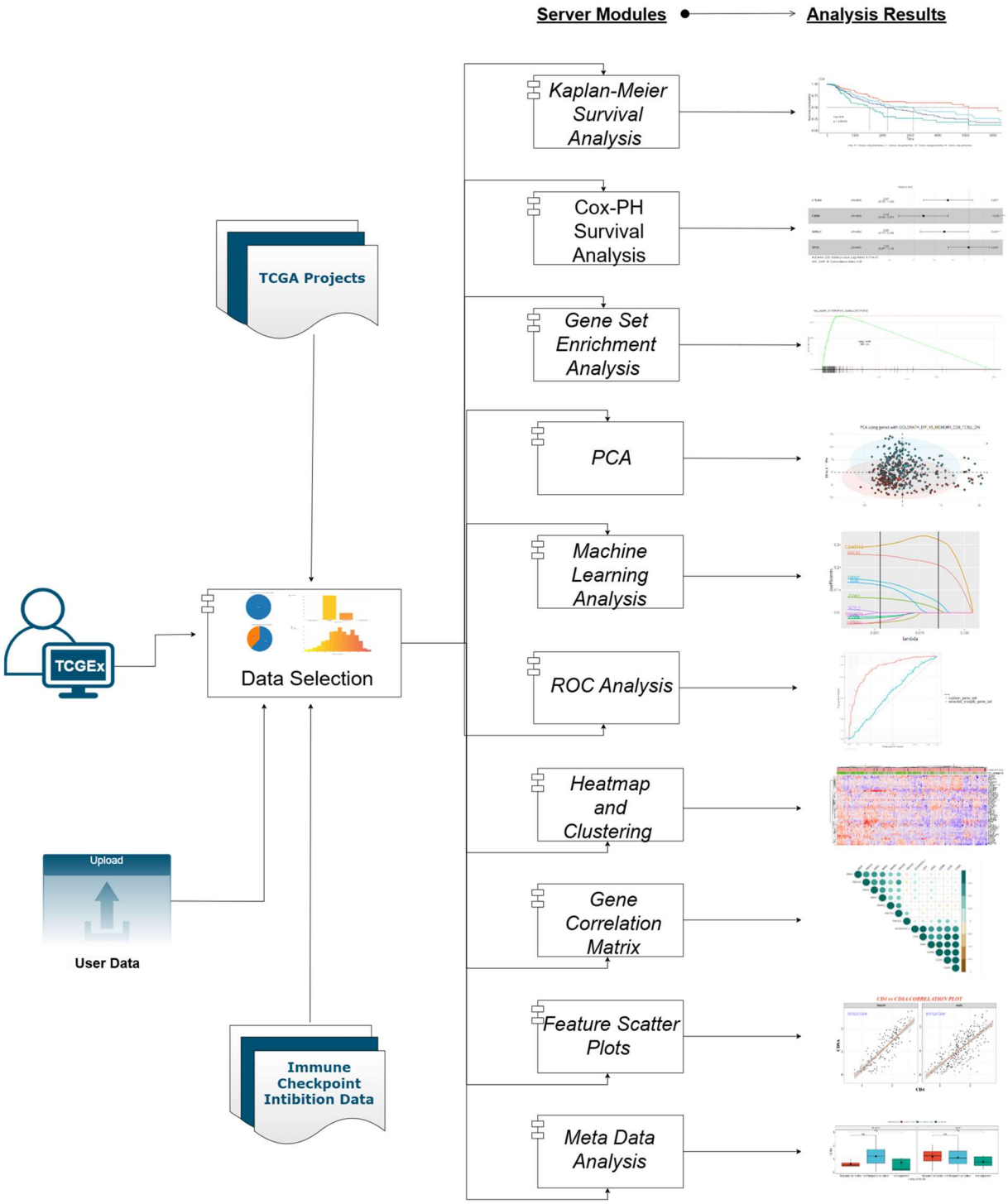
Data availability:

The TCGEx webserver is accessible at https://www.tcgex.iyte.edu.tr. All code for the TCGEx app, data download, and pre-processing is publicly accessible at https://github.com/atakanekiz/TCGEx. To facilitate local execution and code development, TCGEx docker image can be accessed at https://hub.docker.com/repository/docker/atakanekiz/tcgex/. The data sets used in TCGEx were derived from the TCGA data repository in the public domain (https://portal.gdc.cancer.gov/). Processed data files and TCGEx source code are accessible on Figshare (doi:10.6084/m9.figshare.23912532) (https://figshare.com/s/22f21b780acd57fb5dfb).

Conflict of interest:

The authors declare no conflicts of interest.

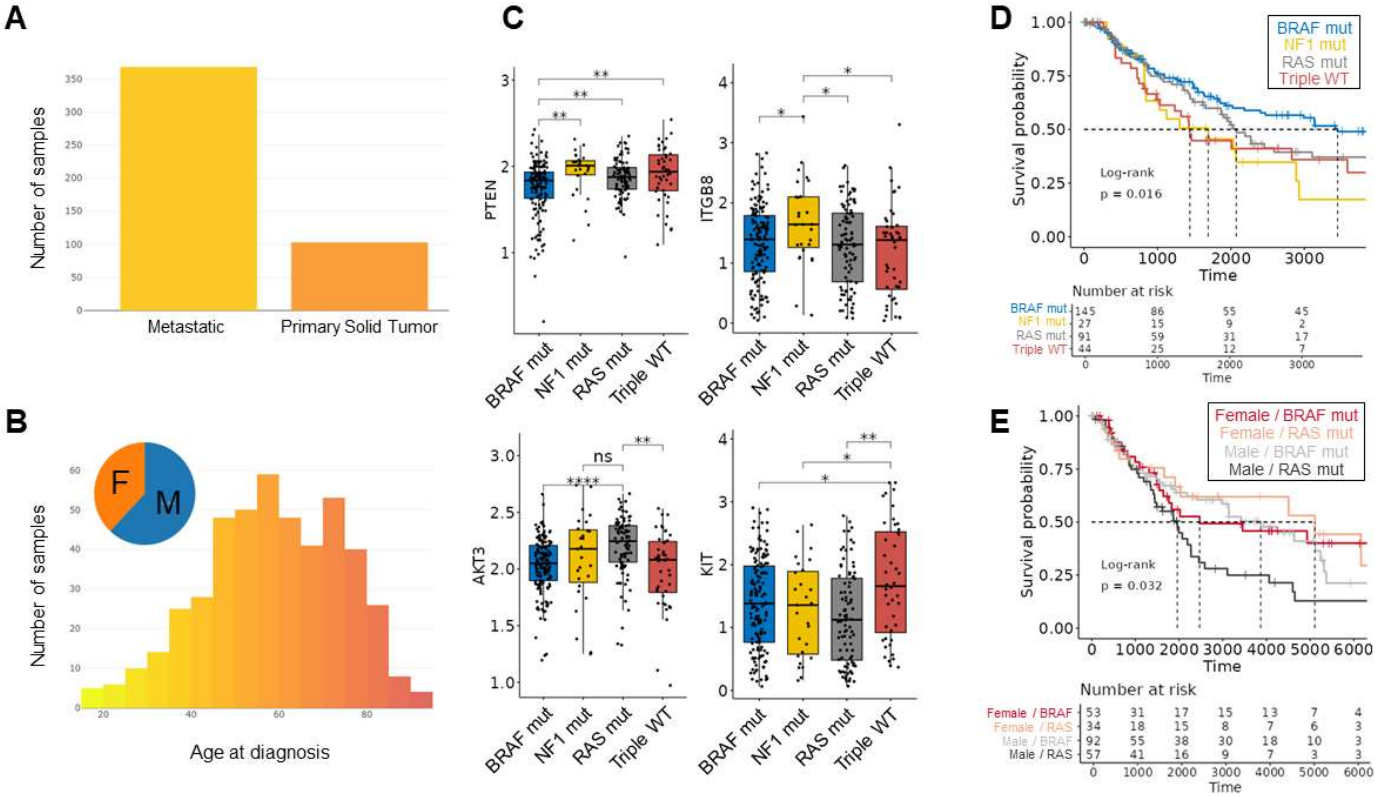
521 **Figure-1**



522

523 **Figure 1. TCGEx workflow overview.** TCGEx pipeline begins with the user selection of pre-pro-
524 cessed transcriptomics data from a single TCGA project or multiple projects for integrated analyses. The
525 data selection module shows descriptive statistics including the sample types, gender, and age distribu-
526 tion of the patients in the selected project(s). The data is subsequently passed onto 10 distinct modules
527 for specialized analyses including survival modeling, gene set enrichment analysis (GSEA), principal
528 component analysis (PCA), correlation calculation, and exploratory graphing in the form of scatter plots,
529 boxplots, and heatmaps. In each module, the user can further focus on specific sample types such as
530 primary and/or metastatic tumor samples, and some modules allow further subsetting in the analyses.
531 Algorithmic and graphical parameters are easily changed through visual and responsive user interfaces.
532 TCGEx modules are supported by interactive tutorials and provide options to export analysis results and
533 graphical outputs for publication purposes. TCGEx is freely available on the web as well as through a
534 docker image. All the subsequent figures in this manuscript are generated using the TCGEx modules.

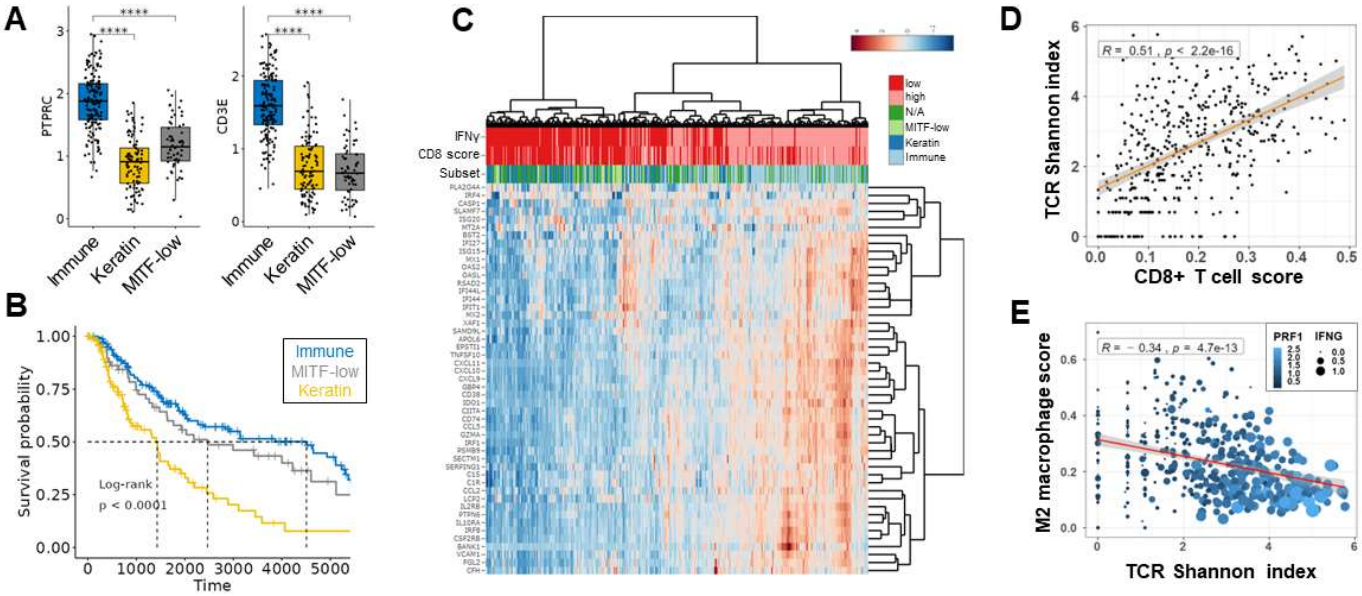
535 **Figure-2**


A
Number of samples
Metastatic
Primary Solid Tumor
B
F
M
Number of samples
Age at diagnosis
C
PTEN
ITGB8
AKT3
KIT
BRAF mut
NF1 mut
RAS mut
Triple WT
ns
D
BRAF mut
NF1 mut
RAS mut
Triple WT
Survival probability
Log-rank
p = 0.016
Time
Number at risk
E
Female / BRAF mut
Female / RAS mut
Male / BRAF mut
Male / RAS mut
Survival probability
Log-rank
p = 0.032
Time
Number at risk

536 **Figure 2. Analysis of mutational subsets in skin cutaneous melanoma. A.** The TCGA-SKCM pro-
537 ject consists of mostly metastatic samples. Analyses described herein are performed on both metastatic
538 and primary tumors. **B.** Comparable numbers of male and female individuals from a range of age groups
539 are included in the SKCM patient cohort. **C.** Expression profiles of PTEN, ITGB8, AKT3, and KIT can
540 distinguish mutational subtypes of SKCM, suggesting different molecular signaling landscapes exist in
541 disease subsets (33). Statistical comparisons are performed using the t-test although users have the
542 option to select non-parametric Wilcoxon test as well. The p-values are adjusted for multiple comparisons
543 ($p<0.0001$, "****"; $p<0.001$, "***"; $p<0.01$, "**"; $p<0.05$, "*"; $p>0.05$, "ns") **D.** Kaplan-Meier survival analysis
544 revealed that BRAF-mutant SKCM has the most favorable clinical outcome. 10-year survival curves are
545 shown where dashed lines indicate the median survival time points in each group. The risk table provided
546 below the survival curve denotes the number of surviving patients (i.e. at-risk patients) at different follow-
547 up times. The log-rank p-value shown on the graph tests the null hypothesis that none of the groups are
548 different from one another. **E.** Survival modeling using only BRAF- and RAS-mutant melanoma patients
549 suggests that males in the RAS-mutant group have poorer survival outcomes compared to BRAF-mutant
550 males and females in general. TCGEx provides users with flexible options to change the color scheme
551 and the range of time axis in the graphs.

552

Figure-3

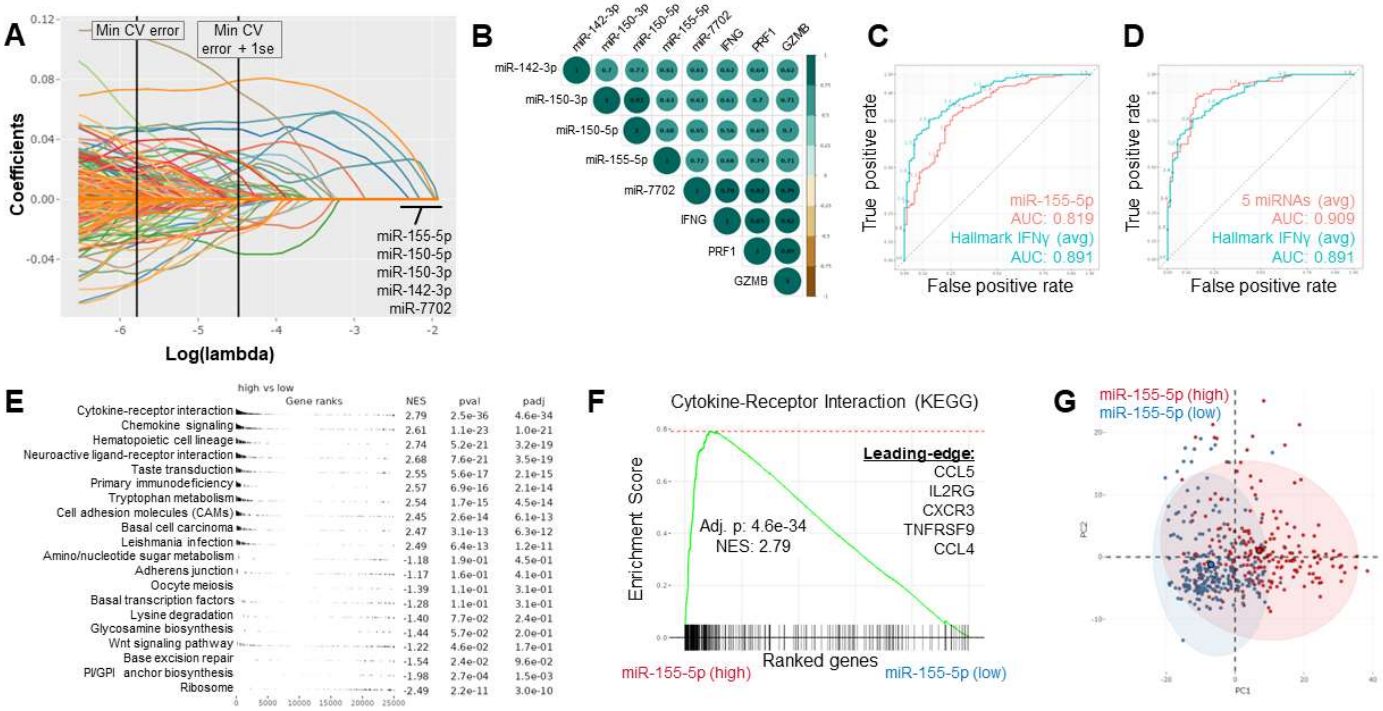

A
PTPRC
CD3E
Immune
Keratin
MITF-low
B
Survival probability
Log-rank
p < 0.0001
Time
C
IFNγ
CD8 score
Subset
low
high
N/A
MITF-low
Keratin
Immune
D
TCR Shannon index
CD8+ T cell score
E
M2 macrophage score
TCR Shannon index
PRF1
IFNG

555 **Figure 3. Analysis of the melanoma intratumoral immune landscape. A.** The immune-enriched
556 melanoma subset previously defined based on multi-gene expression profiles (33) is characterized by
557 higher levels of PTPRC (CD45, a pan-leukocyte marker) and CD3E (a T cell marker) transcripts. T-test
558 p-values are adjusted for multiple comparisons ($p<0.0001$, "****"; $p<0.001$, "***"; $p<0.01$, "**"; $p<0.05$, "*";
559 $p>0.05$, "ns"). **B.** Immune subset of the SKCM is associated with an improved survival outcome compared
560 to keratin and MITF-low subsets. **C.** Heatmap visualization of the top 25% variably expressed genes from
561 the MSigDB Hallmark IFNγ response gene set (58) in primary and metastatic melanoma is shown (50
562 out of 200 genes are plotted). The annotation rows on top of the heatmap were prepared by categorizing
563 IFNγ transcript level and CIBERSORT CD8+ T cell score (70) at their median values. Melanoma subsets
564 defined previously based on genome-wide expression patterns are also shown. CPM-normalized and
565 log-transformed transcript counts are scaled by row to better show expression differences in data sub-
566 sets. The default TCGEx parameters were used for hierarchical clustering (Euclidean distance and
567 Ward.D2 clustering algorithm for both genes and samples). **D.** Increasing intratumoral CD8+ T cell sig-
568 nature is associated with a higher TCR diversity suggesting a polyclonal T cell infiltrate. **E.** The higher
569 TCR diversity in melanoma tumors correlates with increased levels of T cell effector genes including
570 perforin (PRF1) and IFNγ and decreased pro-tumorigenic M2 macrophage signatures (14).

571

572 **Figure-4**

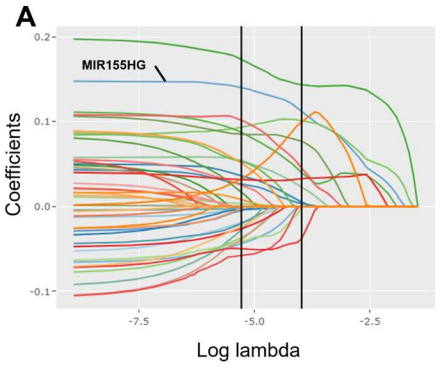
573

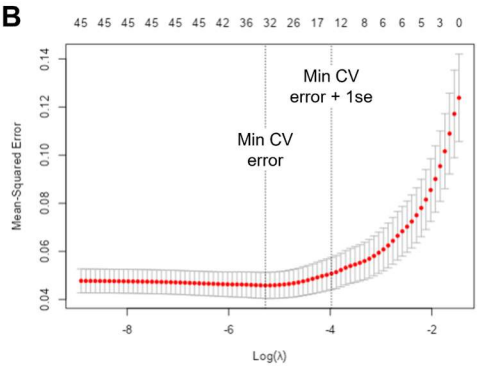


574 **Figure 4. Machine-learning analysis of the associations between miRNAs and antitumor im-**
575 **munity in TCGA-SKCM data. A.** Lasso regression was performed by selecting IFNγ response score as
576 the response variable and mature miRNA transcripts as predictor variables. 529 miRNA predictors in the
577 pre-processed SKCM data entered the lasso model after splitting the data set into training (70%) and
578 test (30%) subsets for cross-validation. The graph shows the coefficient regularization paths for individual
579 predictors at increasing levels of penalty term lambda (λ). Vertical lines indicate the cutoffs where the
580 cross-validation (CV) error was at its minimum value and within one standard error. 5 miRNAs that con-
581 tinued to have non-zero coefficients at arbitrarily high levels of regularization are shown. **B.** 5 miRNAs
582 highlighted in panel (A) show positive correlations with each other and immune effector transcripts IFNγ,
583 PRF1, and GZMB. **C.** ROC analysis with miR-155-5p and averaged Hallmark IFNγ response gene set
584 was performed to test their power in predicting the immune subset of SKCM (binarized as 1 for correct
585 grouping) over keratin and MITF-low groups (both binarized as 0 for incorrect grouping). AUC values
586 indicate both predictors are strongly associated with the immune subset, although the IFNγ response
587 gene set performed better compared to miR-155-5p. **D.** ROC analysis was performed as described in
588 (C), with the averaged expression of 5 miRNAs (and IFNγ response) as predictors. In this analysis, the
589 5 miRNA-expression signature performed slightly better than the IFNγ response score for predicting the
590 immune subset in melanoma. **E.** Gene set enrichment analysis was performed between miR-155-5p-
591 high and -low subsets of melanoma (categorized at the median expression value) using genome-wide
592 signal-to-noise ranked (39) transcripts and the KEGG gene sets (37). The top 10 positively and negatively
593 enriched gene sets are shown in the table where NES>0 indicates enrichment in the miR-155-5p-high
594 subset, and "padj" shows the Benjamini-Hochberg adjusted p-value to account for multiple comparisons
595 across the gene set. **F.** Enrichment plot showing the most significantly enriched KEGG gene set in the
596 miR-155-5p-high melanoma samples where leading-edge genes contributing strongly to the score are
597 also shown. **G.** Principal component analysis of melanoma samples was performed using 685 genes
598 associated with the adaptive immune response GO annotation. Samples are categorized at the median
599 expression of miR-155-5p and color-coded to demonstrate miR-155-5p-high and -low samples in the
600 PCA space.

601

602 **Figure-5**

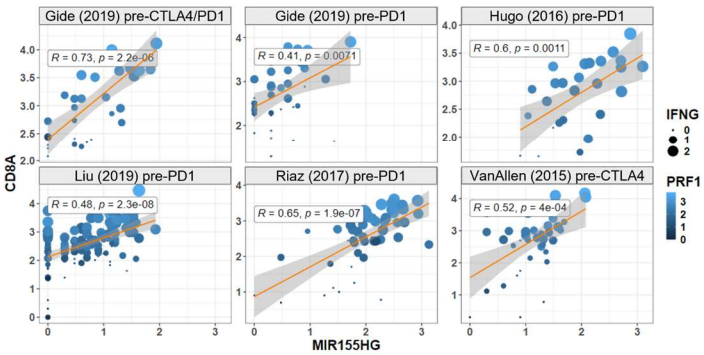


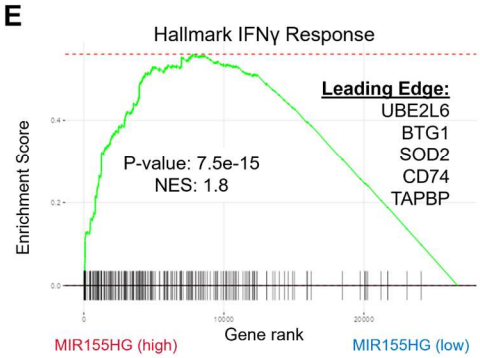


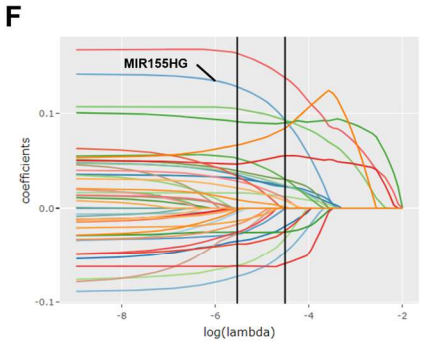
C

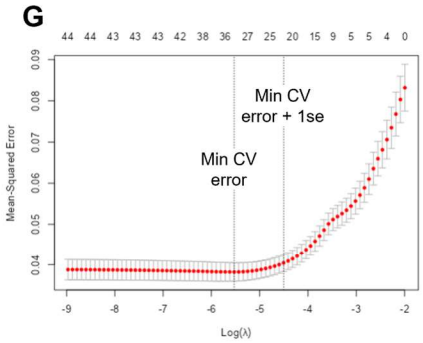
Gene	Lasso Coef.
MIR29B2CHG	0.14378
MIR155HG	0.11164
MIR223HG	0.10085
MIRLET7BHG	0.09946
MIR320B2	0.07686
MIR4435.2HG	0.04535
MIR23AHG	0.03945
MIR497HG	0.03296
MIR2052HG	0.02294
MIR1915HG	0.00506
MIR181A1HG	0.00334

Gene	Lasso Coef.
MIR4458HG	-0.03821
MIR210HG	-0.00113





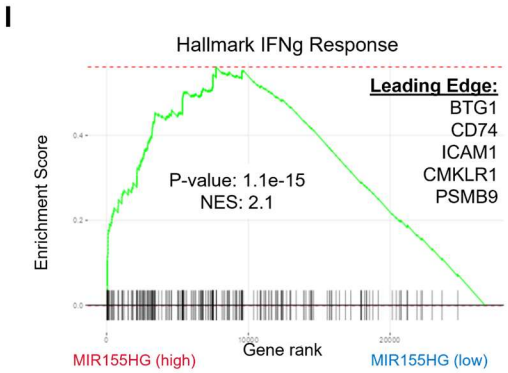


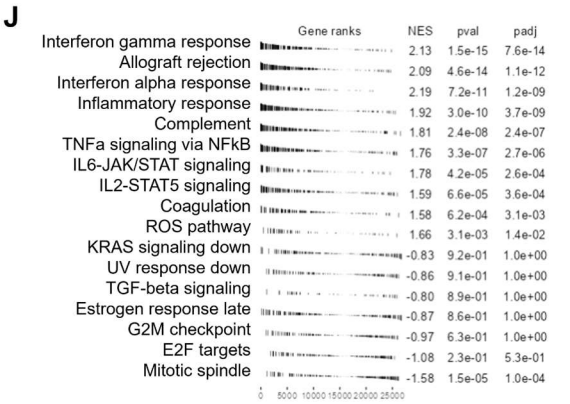


H

Gene	Lasso Coef.
MIR4435.2HG	0.13809
MIR155HG	0.09316
MIR223HG	0.09308
MIR29B2CHG	0.08977
MIRLET7BHG	0.08435
MIR497HG	0.05536
MIR23AHG	0.0307
MIR320B2	0.03042
MIR2052HG	0.02408
MIR1915HG	0.02312
MIR181A1HG	0.02287
MIR205HG	0.01246
MIR3936HG	0.01048
MIR9.1HG	0.00794
MIR99AHG	0.00119

Gene	Lasso Coef.
MIR4458HG	-0.05944
MIR133A1HG	-0.04703
MIR210HG	-0.03219
MIR3142HG	-0.02534
MIR503HG	-0.02408
MIR181A2HG	-0.01931
MIR17HG	-0.00012

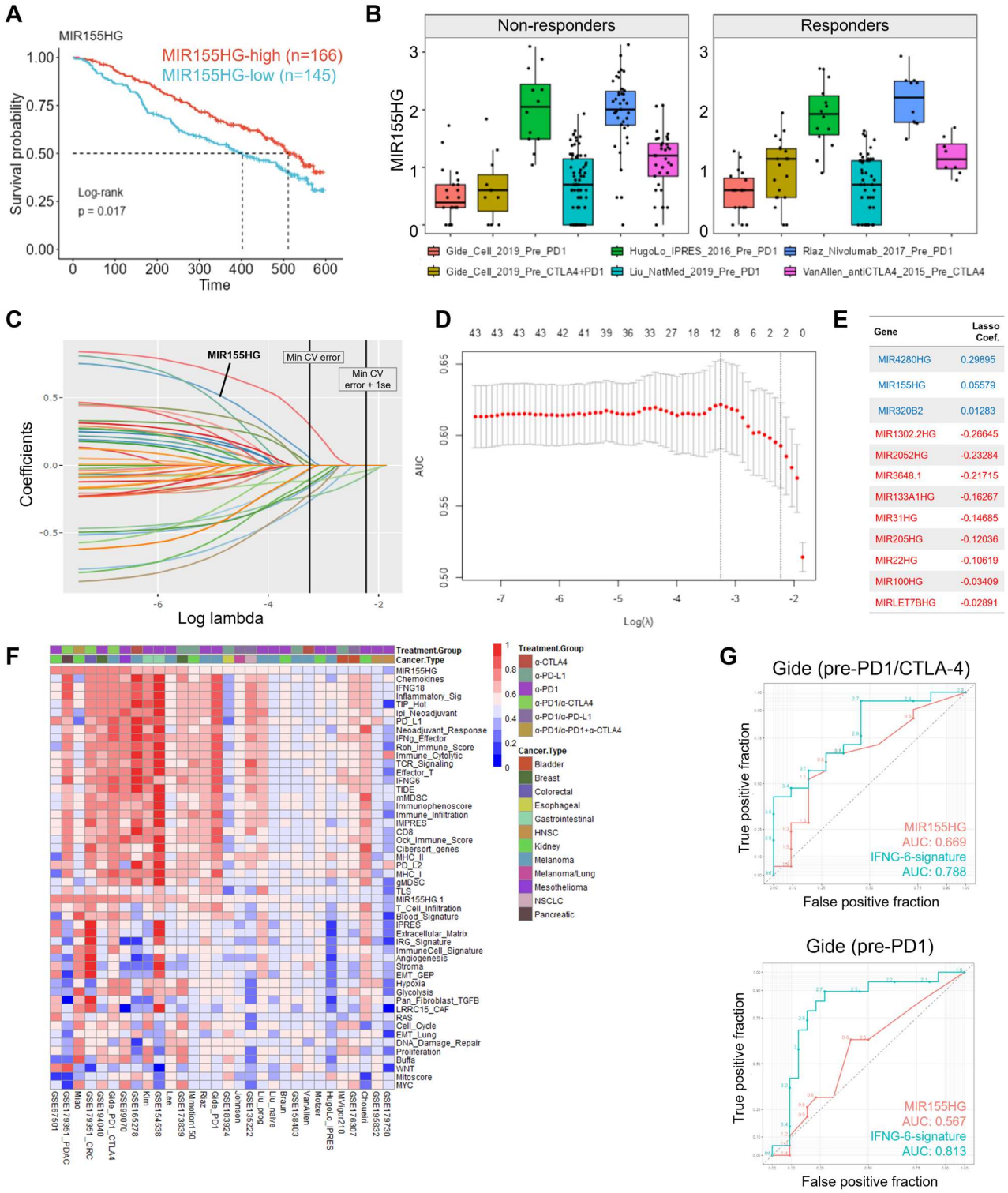




603 **Figure 5. Investigating the associations between miRNAs and immunity markers in cancer im-**
604 **munotherapy studies. A-C.** Lasso regression was performed using the average expression of MSigDB
605 Hallmark IFNg response gene set as the response variable and miRNA host genes as predictors in mel-
606 anoma biopsies prior to immunotherapy. The penalized coefficients (A) and the mean-squared cross-
607 validation error of penalized models are shown (B). miRNA predictors with positive and negative coeffi-
608 cients at the lambda value corresponding to within 1 standard error of the minimum cross-validation error
609 are listed (C). **D.** Scatter plots show the relationship between MIR155HG and cytotoxic T lymphocyte-
610 specific gene expression in melanoma ICI clinical trial data sets (6, 16, 17, 67). **E.** Gene set enrichment
611 analysis reveals higher expression of IFNγ-responsive genes in MIR155HG-high melanoma biopsies
612 prior to ICI therapy. **F-H.** Lasso regression was performed using miRNA host genes as predictors and
613 the average expression of MSigDB Hallmark IFNγ responsive genes as response variable across pan-
614 ICI data prepared by integrating 10 different ICI studies comprising of 938 samples prior to treatment.
615 Feature coefficients (F), mean squared error of the penalized models (G) and the genes with non-zero
616 coefficients (at minimum lambda + 1 standard error) are shown. **I.** MIR155HG-high biopsies demonstrate
617 a significant enrichment of IFNγ-response gene signature in the pan-ICI dataset. **J.** The top differentially
618 enriched pathways are shown for the MIR155HG-high and MIR155HG-low biopsies prior to ICI treatment.

619

Figure-6


A
MIR155HG
MIR155HG-high (n=166)
MIR155HG-low (n=145)
Survival probability
Log-rank
p = 0.017
Time
B
Non-responders
Responders
MIR155HG
Gide_Cell_2019_Pre_PD1
HugoLo_IPRES_2016_Pre_PD1
Riaz_Nivolumab_2017_Pre_PD1
Gide_Cell_2019_Pre_CTLA4+PD1
Liu_NatMed_2019_Pre_PD1
VanAllen_antiCTLA4_2015_Pre_CTLA4
C
MIR155HG
Min CV error
Min CV error + 1se
Coefficients
Log lambda
D
AUC
Log(λ)
E
Gene | Lasso Coef.
MIR4280HG | 0.29895
MIR155HG | 0.05579
MIR320B2 | 0.01283
MIR1302.2HG | -0.26645
MIR2052HG | -0.23284
MIR3648.1 | -0.21715
MIR133A1HG | -0.16267
MIR31HG | -0.14685
MIR205HG | -0.12036
MIR22HG | -0.10619
MIR100HG | -0.03409
MIRLET7BHG | -0.02891
F
Treatment.Group
Cancer.Type
α-CTLA4
α-PD-L1
α-PD1
α-PD1/α-CTLA4
α-PD1/α-PD-L1
α-PD1/α-PD1+α-CTLA4
Bladder
Breast
Colorectal
Esophageal
Gastrointestinal
HNSC
Kidney
Melanoma
Melanoma/Lung
Mesothelioma
NSCLC
Pancreatic
G
Gide (pre-PD1/CTLA-4)
True positive fraction
False positive fraction
MIR155HG
AUC: 0.669
IFNG-6-signature
AUC: 0.788
Gide (pre-PD1)
MIR155HG
AUC: 0.567
IFNG-6-signature
AUC: 0.813

622 **Figure 6. Investigating the associations between miRNAs and immunotherapy response.**
623 **A.** The higher MIR155HG expression in tumors prior to immune checkpoint inhibition (ICI) corre-
624 sponds to improved survival in metastatic melanoma. **B.** Boxplot shows MIR155HG expression in
625 tumors responding or refractive to ICI agents. MIR155HG expression is higher in responders in
626 some but not all studies examined. **C-E.** Binomial lasso regression was applied to pre-treatment
627 SKCM immunotherapy studies. In these analyses, a positive response (partial or complete re-
628 sponse) is binarized to "1" and no response (progressive disease) is binarized to "0" as the re-
629 sponse variable, and miRNA host gene transcripts in the RNAseq data were used as predictors.
630 Regularized coefficients at various levels of lambda (C) and the AUC of the model (D) are shown.
631 Non-zero coefficients were extracted at the lambda value with the highest cross-validated AUC
632 (E). **F.** Performance of MIR155HG and various other gene signatures for predicting ICI response
633 is assessed by the IOSig platform and summarized via the heatmap. The color-coded AUC esti-
634 mates between 0 and 1 from 12 different cancer clinical trials are shown. **G.** The ROC module of
635 TCGEx was used to assess the potential of MIR155HG and 6-gene IFNγ signatures in predicting
636 response to immunotherapy (complete/partial response vs progressive disease) in data set gen-
637 erated by Gide et al.

Figure-7

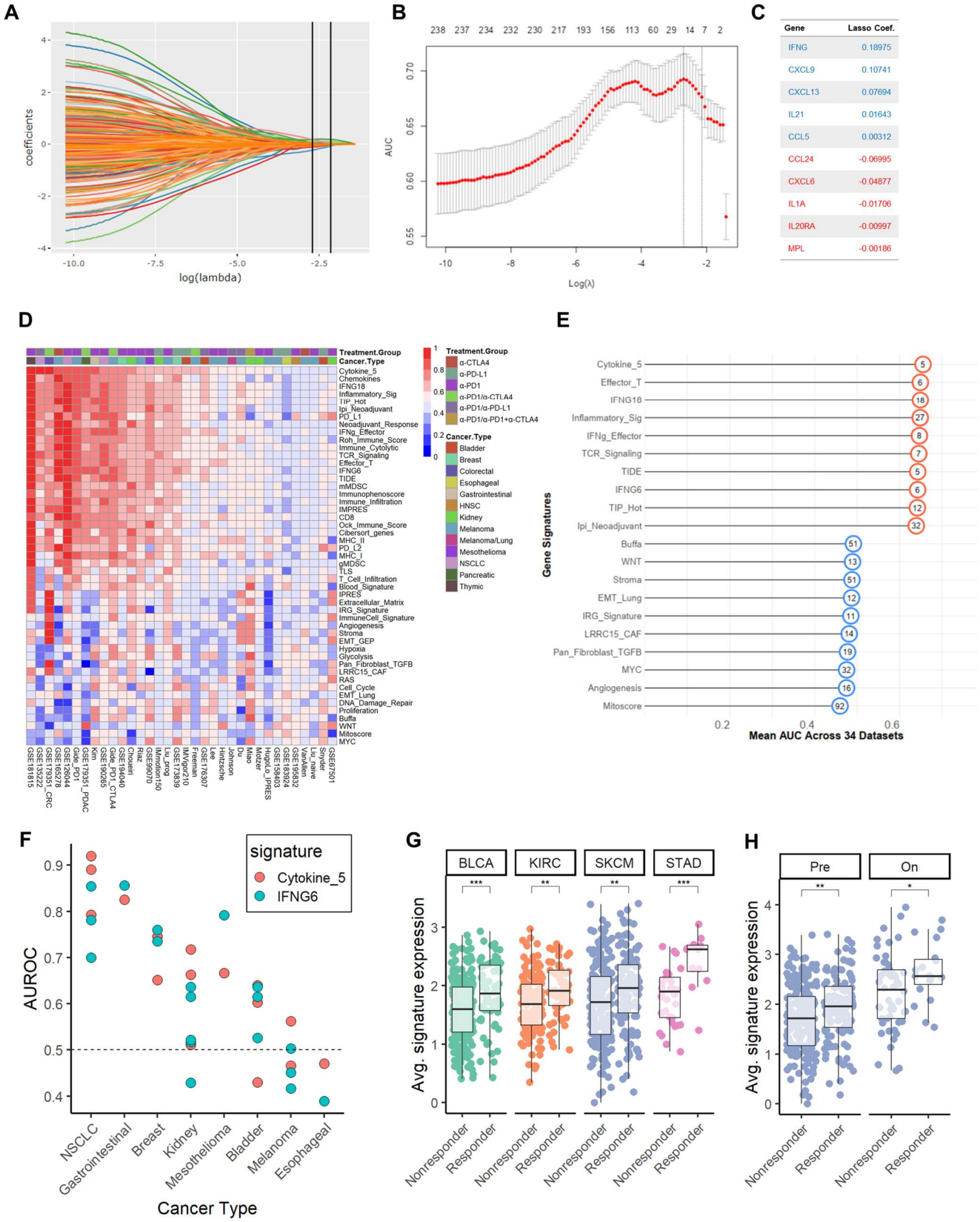

A
coefficients
log(lambda)
B
238 237 234 232 230 217 193 156 113 60 29 14 7 2
AUC
Log(λ)
C
Gene | Lasso Coef.
IFNG | 0.18975
CXCL9 | 0.10741
CXCL13 | 0.07694
IL21 | 0.01643
CCL5 | 0.00312
CCL24 | -0.06995
CXCL6 | -0.04877
IL1A | -0.01706
IL20RA | -0.00997
MPL | -0.00186
D
Treatment.Group
Cancer.Type
α-CTLA4
α-PD-L1
α-PD1
α-PD1/α-CTLA4
α-PD1/α-PD-L1
α-PD1/α-PD1+α-CTLA4
Bladder
Breast
Colorectal
Esophageal
Gastrointestinal
HNSC
Kidney
Melanoma
Melanoma/Lung
Mesothelioma
NSCLC
Pancreatic
Thymic
E
Gene Signatures
Mean AUC Across 34 Datasets
F
signature
Cytokine_5
IFNG6
AUROC
Cancer Type
NSCLC
Gastrointestinal
Breast
Kidney
Mesothelioma
Bladder
Melanoma
Esophageal
G
BLCA
KIRC
SKCM
STAD
Avg. signature expression
Nonresponder
Responder
H
Pre
On
Avg. signature expression
Nonresponder
Responder

Figure 7. TCGEx identifies a new cytokine signature predicting immunotherapy response in a broad range of cancers. A-C. Binomial elastic net regression (alpha=0.5) was performed using the Pan-ICI data consisting of 938 samples across 10 clinical trials and 4 different tumor types. In these analyses binarized ICI response (CR/PR vs PD) was used as the independent variable, and 238 transcripts associated with cytokine-cytokine receptor interaction (KEGG) were used as independent (i.e. predictor) variables. Regularized coefficients (A), and the AUC of the cross-validated model (B) are shown. Predictors with non-zero coefficients at the smallest value of lambda within the 1 standard error of the maximum AUC of the cross validated model are listed (C). “Cytokine-5” signature is derived from the transcripts with positive coefficients. **D.** The IOSig platform was used to assess the potential of Cytokine-5 and other previously published signatures for predicting immunotherapy response. AUC values ranging between 0 and 1 from various studies are depicted with the color gradient. **E.** The average AUC of the top and bottom signatures are summarized in the graph where the numbers in the circles indicate the number of genes in the signature. **F.** The AUC values of Cytokine-5 and IFNG6 signatures in individual clinical trials are shown across various cancer types. Each point indicates a distinct data set (data were obtained from the IOSig platform). **G.** Boxplot module of TCGEx was used to plot the average expression of Cytokine-5 signature in tumors responding or refractory to ICI agents. Individual samples collected prior to ICI treatment are depicted by points. **H.** Boxplots using the data from melanoma ICI studies demonstrate responder tumors express higher levels of Cytokine-5 signature prior to and during the treatment.

661 **References:**

662 1. Blum,A., Wang,P. and Zenklusen,J.C. (2018) SnapShot: TCGA-Analyzed Tumors. *Cell*, **173**, 530.

663 2. Campbell,P.J., Getz,G., Korbel,J.O., Stuart,J.M., Jennings,J.L., Stein,L.D., Perry,M.D., Nahal-Bose,H.K., Ouel-
664 lette,B.F.F., Li,C.H., *et al.* (2020) Pan-cancer analysis of whole genomes. *Nature 2020 578:7793*, **578**, 82–93.

665 3. Bell,D., Berchuck,A., Birrer,M., Chien,J., Cramer,D.W., Dao,F., Dhir,R., Disaia,P., Gabra,H., Glenn,P., *et al.* (2011)
666 Integrated genomic analyses of ovarian carcinoma. *Nature 2011 474:7353*, **474**, 609–615.

667 4. Koboldt,D.C., Fulton,R.S., McLellan,M.D., Schmidt,H., Kalicki-Veizer,J., McMichael,J.F., Fulton,L.L., Dooling,D.J.,
668 Ding,L., Mardis,E.R., *et al.* (2012) Comprehensive molecular portraits of human breast tumours. *Nature*
669 *2012 490:7418*, **490**, 61–70.

670 5. Muzny,D.M., Bainbridge,M.N., Chang,K., Dinh,H.H., Drummond,J.A., Fowler,G., Kovar,C.L., Lewis,L.R., Mor-
671 gan,M.B., Newsham,I.F., *et al.* (2012) Comprehensive molecular characterization of human colon and rectal
672 cancer. *Nature 2012 487:7407*, **487**, 330–337.

673 6. Van Allen,E.M., Miao,D., Schilling,B., Shukla,S.A., Blank,C., Zimmer,L., Sucker,A., Hillen,U., Foppen,M.H.G.,
674 Goldinger,S.M., *et al.* (2015) Genomic correlates of response to CTLA-4 blockade in metastatic melanoma.
675 *Science (1979)*, **350**, 207–211.

676 7. Hugo,W., Shi,H., Sun,L., Piva,M., Song,C., Kong,X., Moriceau,G., Hong,A., Dahlman,K.B., Johnson,D.B., *et al.*
677 (2015) Non-genomic and Immune Evolution of Melanoma Acquiring MAPKi Resistance. *Cell*, **162**, 1271–
678 1285.

679 8. Ayers,M., Ribas,A., Mcclanahan,T.K., Ayers,M., Lunceford,J., Nebozhyn,M., Murphy,E., Loboda,A., Kauf-
680 man,D.R., Albright,A., *et al.* (2017) IFN-γ–related mRNA profile predicts clinical response to PD-1 blockade. *J*
681 *Clin Invest*, **127**, 2930–2940.

682 9. Auslander,N., Zhang,G., Lee,J.S., Frederick,D.T., Miao,B., Moll,T., Tian,T., Wei,Z., Madan,S., Sullivan,R.J., *et al.*
683 (2018) Robust prediction of response to immune checkpoint blockade therapy in metastatic melanoma.
684 *Nat Med*, **24**, 1545–1549.

685 10. Cerami,E., Gao,J., Dogrusoz,U., Gross,B.E., Sumer,S.O., Aksoy,B.A., Jacobsen,A., Byrne,C.J., Heuer,M.L., Lars-
686 son,E., *et al.* (2012) The cBio Cancer Genomics Portal: An Open Platform for Exploring Multidimensional
687 Cancer Genomics Data. *Cancer Discov*, **2**, 401–404.

688 11. Deng,M., Brägelmann,J., Schultze,J.L. and Perner,S. (2016) Web-TCGA: An online platform for integrated
689 analysis of molecular cancer data sets. *BMC Bioinformatics*, **17**, 1–7.

690 12. Goldman,M.J., Craft,B., Hastie,M., Repečka,K., McDade,F., Kamath,A., Banerjee,A., Luo,Y., Rogers,D.,
691 Brooks,A.N., *et al.* (2020) Visualizing and interpreting cancer genomics data via the Xena platform. *Nature*
692 *Biotechnology 2020 38:6*, **38**, 675–678.

693 13. Tang,Z., Kang,B., Li,C., Chen,T. and Zhang,Z. (2019) GEPIA2: an enhanced web server for large-scale expres-
694 sion profiling and interactive analysis. *Nucleic Acids Res*, **47**, W556–W560.

695 14. Thorsson,V., Gibbs,D.L., Brown,S.D., Wolf,D., Bortone,D.S., Ou Yang,T.H., Porta-Pardo,E., Gao,G.F., Plaisier,C.L.,
696 Eddy,J.A., *et al.* (2018) The Immune Landscape of Cancer. *Immunity*, **48**, 812-830.e14.

697 15. Gide,T.N., Quek,C., Menzies,A.M., Tasker,A.T., Shang,P., Holst,J., Madore,J., Lim,S.Y., Velickovic,R.,
698 Wongchenko,M., *et al.* (2019) Distinct Immune Cell Populations Define Response to Anti-PD-1 Monother-
699 apy and Anti-PD-1/Anti-CTLA-4 Combined Therapy. *Cancer Cell*, **35**, 238-255.e6.

700 16. Riaz,N., Havel,J.J., Makarov,V., Desrichard,A., Urba,W.J., Sims,J.S., Hodi,F.S., Martín-Algarra,S., Mandal,R.,
701 Sharfman,W.H., *et al.* (2017) Tumor and Microenvironment Evolution during Immunotherapy with
702 Nivolumab. *Cell*, **171**, 934-949.e15.

703 17. Hugo,W., Zaretsky,J.M., Sun,L., Song,C., Moreno,B.H., Hu-Lieskovan,S., Berent-Maoz,B., Pang,J., Chmie-
704 lowski,B., Cherry,G., *et al.* (2016) Genomic and Transcriptomic Features of Response to Anti-PD-1 Therapy
705 in Metastatic Melanoma. *Cell*, **165**, 35–44.

706 18. Zhao,J., Chen,A.X., Gartrell,R.D., Silverman,A.M., Aparicio,L., Chu,T., Bordbar,D., Shan,D., Samanamud,J., Ma-
707 hajan,A., *et al.* (2019) Immune and genomic correlates of response to anti-PD-1 immunotherapy in glioblas-
708 toma. *Nature Medicine 2019 25:3*, **25**, 462–469.

709 19. Cloughesy,T.F., Mochizuki,A.Y., Orpilla,J.R., Hugo,W., Lee,A.H., Davidson,T.B., Wang,A.C., Ellingson,B.M.,
710 Rytlewski,J.A., Sanders,C.M., *et al.* (2019) Neoadjuvant anti-PD-1 immunotherapy promotes a survival ben-
711 efit with intratumoral and systemic immune responses in recurrent glioblastoma. *Nature Medicine 2019*
712 *25:3*, **25**, 477–486.

713 20. Miao,D., Margolis,C.A., Gao,W., Voss,M.H., Li,W., Martini,D.J., Norton,C., Bossé,D., Wankowicz,S.M., Cul-
714 len,D., *et al.* (2018) Genomic correlates of response to immune checkpoint therapies in clear cell renal cell
715 carcinoma. *Science*, **359**, 801–806.

716 21. Liu,D., Schilling,B., Liu,D., Sucker,A., Livingstone,E., Jerby-Amon,L., Zimmer,L., Gutzmer,R., Satzger,I., Lo-
717 quai,C., *et al.* (2019) Integrative molecular and clinical modeling of clinical outcomes to PD1 blockade in pa-
718 tients with metastatic melanoma. *Nat Med*, **25**, 1916–1927.

719 22. Kim,S.T., Cristescu,R., Bass,A.J., Kim,K.M., Odegaard,J.I., Kim,K., Liu,X.Q., Sher,X., Jung,H., Lee,M., *et al.*
720 (2018) Comprehensive molecular characterization of clinical responses to PD-1 inhibition in metastatic gas-
721 tric cancer. *Nat Med*, **24**, 1449–1458.

722 23. Rosenberg,J.E., Hoffman-Censits,J., Powles,T., Van Der Heijden,M.S., Balar,A. V., Necchi,A., Dawson,N.,
723 O'Donnell,P.H., Balmanoukian,A., Loriot,Y., *et al.* (2016) Atezolizumab in patients with locally advanced and
724 metastatic urothelial carcinoma who have progressed following treatment with platinum-based chemo-
725 therapy: a single-arm, multicentre, phase 2 trial. *Lancet*, **387**, 1909–1920.

726 24. Balar,A. V., Galsky,M.D., Rosenberg,J.E., Powles,T., Petrylak,D.P., Bellmunt,J., Loriot,Y., Necchi,A., Hoffman-
727 Censits,J., Perez-Gracia,J.L., *et al.* (2017) Atezolizumab as first-line treatment in cisplatin-ineligible patients
728 with locally advanced and metastatic urothelial carcinoma: a single-arm, multicentre, phase 2 trial. *Lancet*,
729 **389**, 67–76.

730 25. McDermott,D.F., Huseni,M.A., Atkins,M.B., Motzer,R.J., Rini,B.I., Escudier,B., Fong,L., Joseph,R.W., Pal,S.K.,
731 Reeves,J.A., *et al.* (2018) Clinical activity and molecular correlates of response to atezolizumab alone or in
732 combination with bevacizumab versus sunitinib in renal cell carcinoma. *Nat Med*, **24**, 749–757.

733 26. Choueiri,T.K., Fishman,M.N., Escudier,B., McDermott,D.F., Drake,C.G., Kluger,H., Stadler,W.M., Perez-Gra-
734 cia,J.L., McNeel,D.G., Curti,B., *et al.* (2016) Immunomodulatory Activity of Nivolumab in Metastatic Renal
735 Cell Carcinoma. *Clin Cancer Res*, **22**, 5461–5471.

736 27. Gao,J., Aksoy,B.A., Dogrusoz,U., Dresdner,G., Gross,B., Sumer,S.O., Sun,Y., Jacobsen,A., Sinha,R., Larsson,E.,
737 *et al.* (2013) Integrative Analysis of Complex Cancer Genomics and Clinical Profiles Using the cBioPortal. *Sci*
738 *Signal*, **6**, pl1.

739 28. Eddy,J.A., Shmulevich,I., Guinney,J., Thorsson,V., Lamb,A.E., Gibbs,D.L., Heimann,C., Yu,J.X., Chung,V., Chae,Y.,
740 *et al.* (2020) CRI iAtlas: an interactive portal for immuno-oncology research. *F1000Res*, **9**.

741 29. Liu,J., Lichtenberg,T., Hoadley,K.A., Poisson,L.M., Lazar,A.J., Cherniack,A.D., Kovatich,A.J., Benz,C.C., Lev-
742 ine,D.A., Lee,A. V., *et al.* (2018) An Integrated TCGA Pan-Cancer Clinical Data Resource to Drive High-Quality
743 Survival Outcome Analytics. *Cell*, **173**, 400-416.e11.

744 30. Rich,J.T., Neely,J.G., Paniello,R.C., Voelker,C.C.J., Nussenbaum,B. and Wang,E.W. (2010) A PRACTICAL GUIDE
745 TO UNDERSTANDING KAPLAN-MEIER CURVES. *Otolaryngol Head Neck Surg*, **143**, 331.

746 31. Breslow,N.E. (1975) Analysis of Survival Data under the Proportional Hazards Model. *Int Stat Rev*, **43**, 45.

747 32. Berger,A.C., Korkut,A., Kanchi,R.S., Hegde,A.M., Lenoir,W., Liu,W., Liu,Y., Fan,H., Shen,H., Ravikumar,V., *et al.*
748 (2018) A Comprehensive Pan-Cancer Molecular Study of Gynecologic and Breast Cancers. *Cancer Cell*, **33**,
749 690-705.e9.

750 33. Akbani,R., Akdemir,K.C., Aksoy,B.A., Albert,M., Ally,A., Amin,S.B., Arachchi,H., Arora,A., Auman,J.T., Ayala,B.,
751 *et al.* (2015) Genomic Classification of Cutaneous Melanoma. *Cell*, **161**, 1681–1696.

752 34. Ciriello,G., Miller,M.L., Aksoy,B.A., Senbabaoglu,Y., Schultz,N. and Sander,C. (2013) Emerging landscape of
753 oncogenic signatures across human cancers. *Nat Genet*, **45**, 1127–1133.

754 35. Van Der Auwera,I., Limame,R., Van Dam,P., Vermeulen,P.B., Dirix,L.Y. and Van Laere,S.J. (2010) Integrated
755 miRNA and mRNA expression profiling of the inflammatory breast cancer subtype. *Br J Cancer*, **103**, 532–
756 541.

757 36. Sørlie,T., Perou,C.M., Tibshirani,R., Aas,T., Geisler,S., Johnsen,H., Hastie,T., Eisen,M.B., Van De Rijn,M., Jef-
758 frey,S.S., *et al.* (2001) Gene expression patterns of breast carcinomas distinguish tumor subclasses with clin-
759 ical implications. *Proc Natl Acad Sci U S A*, **98**, 10869–10874.

760 37. Liberzon,A., Subramanian,A., Pinchback,R., Thorvaldsdóttir,H., Tamayo,P. and Mesirov,J.P. (2011) Molecular
761 signatures database (MSigDB) 3.0. *Bioinformatics*, **27**, 1739.

762 38. Carter,S.L., Cibulskis,K., Helman,E., McKenna,A., Shen,H., Zack,T., Laird,P.W., Onofrio,R.C., Winckler,W.,
763 Weir,B.A., *et al.* (2012) Absolute quantification of somatic DNA alterations in human cancer. *Nature Bio-*
764 *technology 2012 30:5*, **30**, 413–421.

765 39. Subramanian,A., Tamayo,P., Mootha,V.K., Mukherjee,S., Ebert,B.L., Gillette,M.A., Paulovich,A., Pomeroy,S.L.,
766 Golub,T.R., Lander,E.S., *et al.* (2005) Gene set enrichment analysis: A knowledge-based approach for inter-
767 preting genome-wide expression profiles. *Proc Natl Acad Sci U S A*, **102**, 15545–15550.

768 40. Coleman,S., Xie,M., Tarhini,A.A. and Tan,A.C. (2023) Systematic evaluation of the predictive gene expression
769 signatures of immune checkpoint inhibitors in metastatic melanoma. *Mol Carcinog*, **62**, 77–89.

770 41. Metz,C.E. (1978) Basic principles of ROC analysis. *Semin Nucl Med*, **8**, 283–298.

771 42. Kreft,B.P., Müller-Miny,H., Sommer,T., Steudel,A., Vahlensieck,M., Novak,D., Müller,B.C. and Schild,H.H.
772 (1997) Diagnostic value of MR imaging in comparison to CT in the detection and differential diagnosis of
773 renal masses: ROC analysis. *Eur Radiol*, **7**, 542–547.

774 43. Guo,W., Zhu,L., Zhu,R., Chen,Q., Wang,Q. and Chen,J.Q. (2019) A four-DNA methylation biomarker is a supe-
775 rior predictor of survival of patients with cutaneous melanoma. *Elife*, **8**.

776 44. Lim,S.Y., Lee,J.H., Gide,T.N., Menzies,A.M., Guminski,A., Carlino,M.S., Breen,E.J., Yang,J.Y.H., Ghazanfar,S.,
777 Kefford,R.F., *et al.* (2019) Circulating cytokines predict immune-related toxicity in melanoma patients re-
778 ceiving anti-PD-1–based immunotherapy. *Clinical Cancer Research*, **25**, 1557–1563.

779 45. Yeung,K.Y. and Ruzzo,W.L. (2001) Principal component analysis for clustering gene expression data. *Bioinfor-*
780 *matics*, **17**, 763–774.

781 46. Park,S., Supek,F. and Lehner,B. (2018) Systematic discovery of germline cancer predisposition genes through
782 the identification of somatic second hits. *Nature Communications 2018 9:1*, **9**, 1–13.

783 47. Shastry,K.A. and Sanjay,H.A. (2020) Machine Learning for Bioinformatics. 10.1007/978-981-15-2445-5_3.

784 48. Xu,C. and Jackson,S.A. (2019) Machine learning and complex biological data. *Genome Biol*, **20**, 1–4.

785 49. Friedman,J., Hastie,T. and Tibshirani,R. (2010) Regularization Paths for Generalized Linear Models via Coordi-
786 nate Descent. *J Stat Softw*, **33**, 1–22.

787 50. Tibshirani,R. (1996) Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society*,
788 **58**, 267–288.

789 51. Lu,Y., Zhou,Y., Qu,W., Deng,M. and Zhang,C. (2011) A Lasso regression model for the construction of mi-
790 croRNA-target regulatory networks. *Bioinformatics*, **27**, 2406–2413.

791 52. Sung,H., Ferlay,J., Siegel,R.L., Laversanne,M., Soerjomataram,I., Jemal,A. and Bray,F. (2021) Global Cancer
792 Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Coun-
793 tries. *CA Cancer J Clin*, **71**, 209–249.

794 53. Hong,X., Roh,W., Sullivan,R.J., Wong,K.H.K., Wittner,B.S., Guo,H., Dubash,T.D., Sade-Feldman,M., Wesley,B.,
795 Horwitz,E., *et al.* (2021) The Lipogenic Regulator SREBP2 Induces Transferrin in Circulating Melanoma Cells
796 and Suppresses Ferroptosis. *Cancer Discov*, **11**, 678–695.

797 54. Zhao,H., Li,Y., Wang,S., Yang,Y., Wang,J., Ruan,X., Yang,Y., Cai,K., Zhang,B., Cui,P., *et al.* (2014) Whole tran-
798 scriptome RNA-seq analysis: tumorigenesis and metastasis of melanoma. *Gene*, **548**, 234–243.

799 55. He,W., Kang,Y., Zhu,W., Zhou,B., Jiang,X., Ren,C. and Guo,W. (2020) FOXF2 acts as a crucial molecule in tu-
800 mours and embryonic development. *Cell Death & Disease 2020 11:6*, **11**, 1–11.

801 56. Morgese,F., Sampaolesi,C., Torniai,M., Conti,A., Ranallo,N., Giacchetti,A., Serresi,S., Onofri,A., Burattini,M.,
802 Ricotti,G., *et al.* (2020) Gender Differences and Outcomes in Melanoma Patients. *Oncol Ther*, **8**, 103.

803 57. Parker,B.S., Rautela,J. and Hertzog,P.J. (2016) Antitumour actions of interferons: implications for cancer ther-
804 apy. *Nature Reviews Cancer 2016 16:3*, **16**, 131–144.

805 58. Liberzon,A., Birger,C., Thorvaldsdóttir,H., Ghandi,M., Mesirov,J.P. and Tamayo,P. (2015) The Molecular Signa-
806 tures Database (MSigDB) hallmark gene set collection. *Cell Syst*, **1**, 417–425.

807 59. Tirosh,I., Izar,B., Prakadan,S.M., Wadsworth,M.H., Treacy,D., Trombetta,J.J., Rotem,A., Rodman,C., Lian,C.,
808 Murphy,G., *et al.* (2016) Dissecting the multicellular ecosystem of metastatic melanoma by single-cell RNA-
809 seq. *Science (1979)*, **352**, 189–196.

810 60. Li,H., van der Leun,A.M., Yofe,I., Lubling,Y., Gelbard-Solodkin,D., van Akkooi,A.C.J., van den Braber,M., Ro-
811 zeman,E.A., Haanen,J.B.A.G., Blank,C.U., *et al.* (2019) Dysfunctional CD8 T Cells Form a Proliferative, Dy-
812 namically Regulated Compartment within Human Melanoma. *Cell*, **176**, 775-789.e18.

813 61. van der Leun,A.M., Thommen,D.S. and Schumacher,T.N. (2020) CD8+ T cell states in human cancer: insights
814 from single-cell analysis. *Nature Reviews Cancer 2020 20:4*, **20**, 218–232.

815 62. Xing,Y., Ruan,G., Ni,H., Qin,H., Chen,S., Gu,X., Shang,J., Zhou,Y., Tao,X. and Zheng,L. (2021) Tumor Immune
816 Microenvironment and Its Related miRNAs in Tumor Progression. *Front Immunol*, **12**, 1–16.

817 63. Ekiz,H.A., Huffaker,T.B., Grossmann,A.H., Zac Stephens,W., Williams,M.A., Round,J.L. and O'Connell,R.M.
818 (2019) MicroRNA-155 coordinates the immunological landscape within murine melanoma and correlates
819 with immunity in human cancers. *JCI Insight*, **4**.

820 64. Hsin,J.P., Lu,Y., Loeb,G.B., Leslie,C.S. and Rudensky,A.Y. (2018) The effect of cellular context on miR-155-medi-
821 ated gene regulation in four major immune cell types. *Nat Immunol*, **19**, 1137–1145.

822 65. Gracias,D.T., Stelekati,E., Hope,J.L., Boesteanu,A.C., Doering,T.A., Norton,J., Mueller,Y.M., Fraietta,J.A.,
823 Wherry,E.J., Turner,M., *et al.* (2013) The microRNA miR-155 controls CD8+ T cell responses by regulating
824 interferon signaling. *Nat Immunol*, **14**, 593–602.

825 66. Yun,J., Claudia,W., Zhiya,Y., Jinhui,H., Sanjivan,G., V.,H.N., G.,T.W., C.,P.D., Zulmarie,F., Madhusudhanan,S., *et*
826 *al.* (2015) miR-155 augments CD8+ T-cell antitumor activity in lymphoreplete hosts by enhancing respon-
827 siveness to homeostatic γc cytokines. *Proceedings of the National Academy of Sciences*, **112**, 476–481.

828 67. Liu,D.D., Schilling,B., Liu,D.D., Sucker,A., Livingstone,E., Jerby-Amon,L., Zimmer,L., Gutzmer,R., Satzger,I., Lo-
829 quai,C., *et al.* (2019) Integrative molecular and clinical modeling of clinical outcomes to PD1 blockade in pa-
830 tients with metastatic melanoma. *Nat Med*, **25**, 1916–1927.

831 68. Curti,A., Isidori,A., Russo,G. Lo, Garassino,M.C., Kalkusova,K., Taborska,P., Stakheev,D. and Smrz,D. (2022)
832 The Role of miR-155 in Antitumor Immunity. *Cancers 2022, Vol. 14, Page 5414*, **14**, 5414.

833 69. Hsin,J.-P., Lu,Y., Loeb,G.B., Leslie,C.S. and Rudensky,A.Y. (2018) The effect of cellular context on miR-155-me-
834 diated gene regulation in four major immune cell types. *Nat Immunol*, **19**, 1137–1145.

835 70. Newman,A.M., Liu,C.L., Green,M.R., Gentles,A.J., Feng,W., Xu,Y., Hoang,C.D., Diehn,M. and Alizadeh,A.A.
836 (2015) Robust enumeration of cell subsets from tissue expression profiles. *Nat Methods*, **12**, 453–457.

837

838

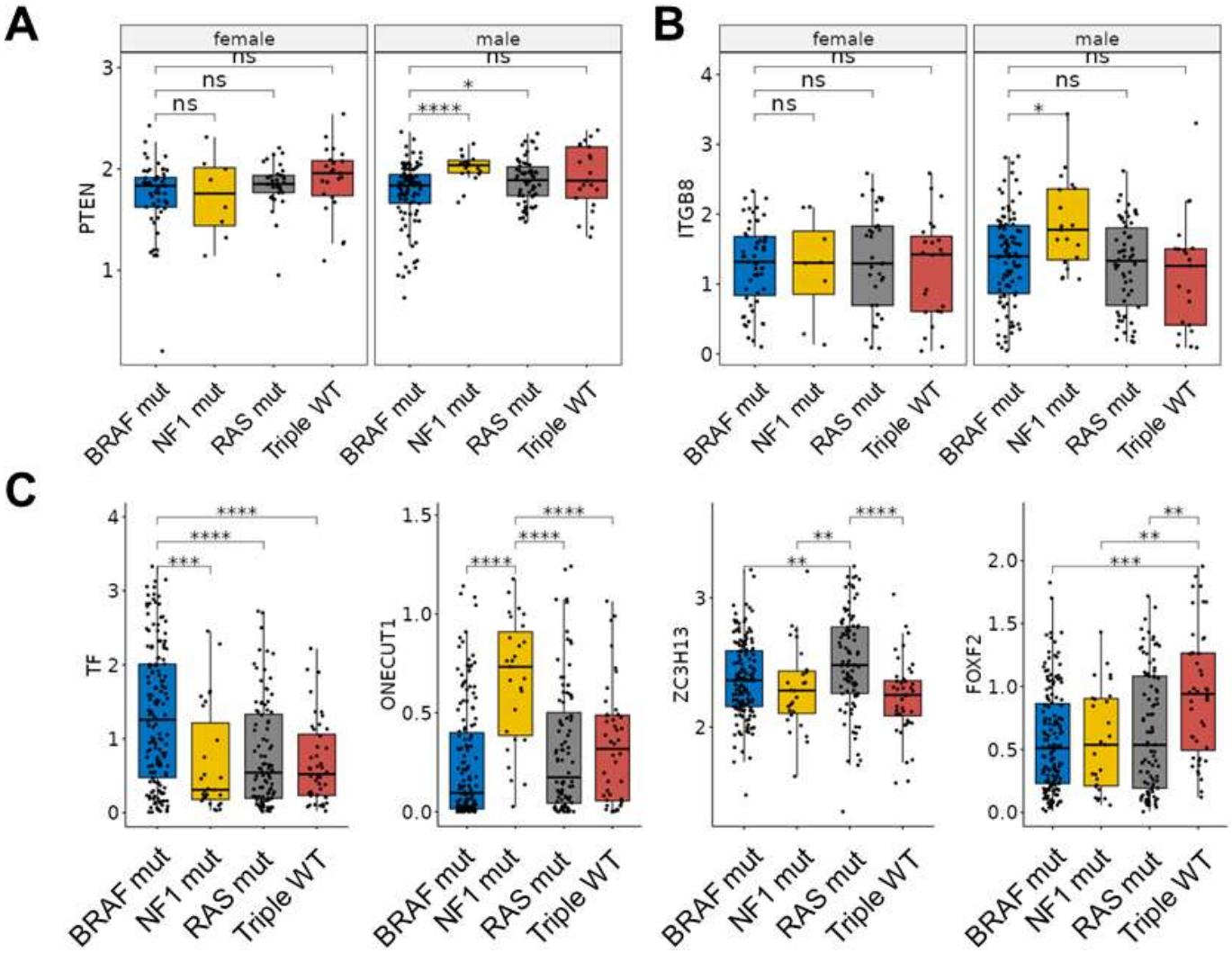


Figure S1. Expression profiles of subtype-selective genes in melanoma (related to Fig.2). A-B. Expression of previously reported genes distinguishing melanoma mutational subtypes may show sex-specific expression patterns. **C.** Other transcripts selectively expressed in mutational subtypes of melanoma are shown. T-test p-values are adjusted for multiple comparisons (p<0.0001, "****"; p<0.001, "***"; p<0.01, "**"; p<0.05, "*"; p>0.05, "ns")

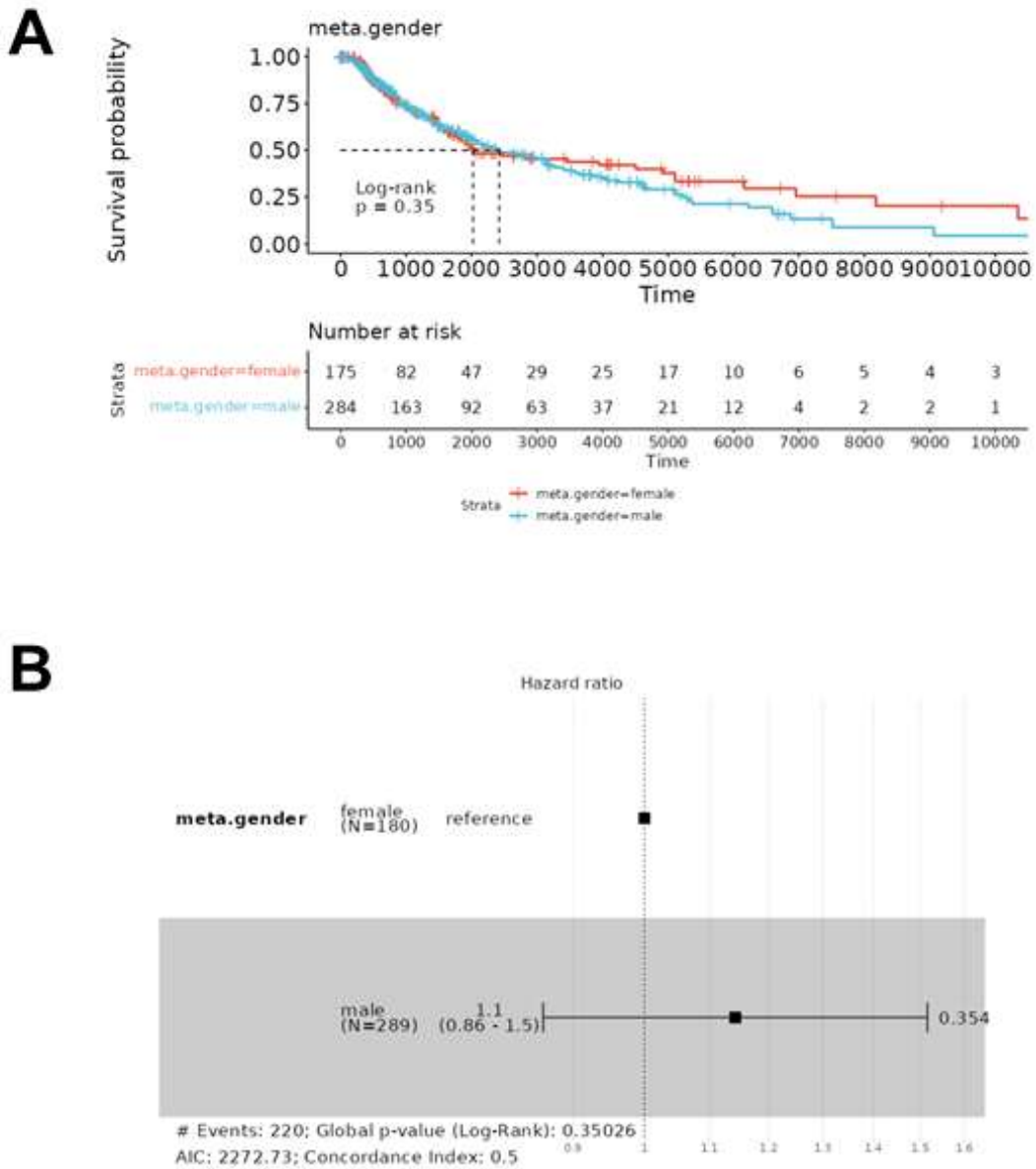
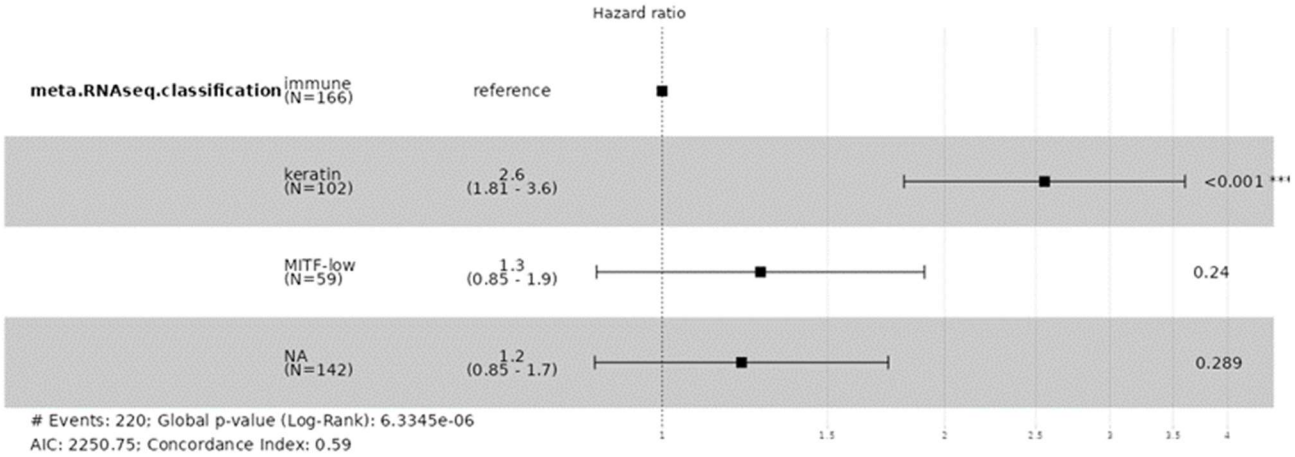


Figure S2. Modeling the effects of patient gender on melanoma survival (related to Fig.2). A. Kaplan-Meier survival analysis of gender groups in SKCM shows non-significant differences, although male patients tended to have decreased survival in longer follow-up. **B.** Cox proportional hazards survival analysis of gender effects in SKCM reveals no significant differences between the two groups. However, the male group has a slightly higher risk of death as seen in the forest plot.

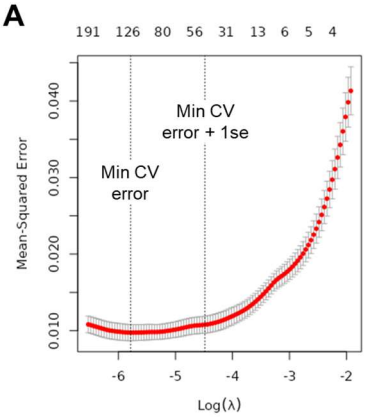


853

854 **Figure S3. Survival modeling of transcriptomic subsets of melanoma (related to Fig.3).** The
855 forest plot shows the hazard ratios of keratin, MITF-low, and unannotated groups in comparison to the
856 immune subset appointed as the reference group. The keratin subset is associated with a statistically
857 significant poor outcome in melanoma compared to the immune subset. MITF-low group tended to have
858 poorer outcomes compared to the immune subset, although this difference was not statistically signifi-
859 cant.

860

A



B

miRNA	Coefficient
hsa.miR.7702	0.0792
hsa.miR.3614.5p	0.0719
hsa.miR.155.5p	0.0512
hsa.miR.142.3p	0.0469
hsa.miR.5586.5p	0.0324
hsa.miR.181a.3p	0.0307
hsa.miR.100.5p	0.0266
hsa.miR.6125	0.0262
hsa.miR.503.5p	0.0260
hsa.miR.150.5p	0.0246
hsa.miR.29b.2.5p	0.0245
hsa.miR.150.3p	0.0203
hsa.miR.342.5p	0.0186
hsa.miR.4728.3p	0.0162
hsa.miR.3614.3p	0.0119
hsa.miR.181c.5p	0.0100
hsa.miR.29c.5p	0.0094
hsa.miR.342.3p	0.0080
hsa.let.7b.3p	0.0069
hsa.miR.1228.3p	0.0067
hsa.miR.542.5p	0.0057
hsa.miR.598.3p	0.0048
hsa.miR.424.3p	0.0039
hsa.miR.21.3p	0.0026
hsa.miR.30a.5p	0.0024
hsa.miR.29c.3p	0.0017

miRNA	Coefficient
hsa.miR.3170	-0.0360
hsa.miR.6509.5p	-0.0235
hsa.miR.4636	-0.0204
hsa.miR.3680.3p	-0.0191
hsa.miR.92a.3p	-0.0187
hsa.miR.340.3p	-0.0172
hsa.miR.187.3p	-0.0144
hsa.miR.19a.3p	-0.0140
hsa.miR.5699.5p	-0.0133
hsa.miR.1.3p	-0.0101
hsa.miR.6783.3p	-0.0048
hsa.miR.129.5p	-0.0047
hsa.miR.3150b.3p	-0.0045
hsa.miR.4677.3p	-0.0032
hsa.miR.219a.2.3p	-0.0029
hsa.miR.6854.5p	-0.0012
hsa.miR.144.5p	-0.0005
hsa.miR.17.3p	-0.0002

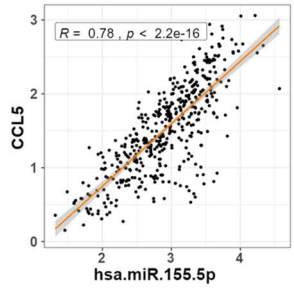
C

Positive correlation with miR-155-5p

Negative correlation with miR-155-5p

Genes	P-value	Pearson cor. coef.
CCL5	6.59E-92	0.778
NKG7	2.24E-86	0.763
CD2	1.68E-85	0.76
TRAC	2.89E-84	0.756
TRBC2	1.87E-83	0.754
CD8A	7.18E-83	0.752
MIR155HG	9.23E-83	0.752
CD3D	1.22E-82	0.752
CD8B	1.35E-82	0.752
TIGIT	3.33E-82	0.75
UCK2	1.66E-15	-0.364
CASKIN1	1.45E-15	-0.365
SNHG29	1.02E-15	-0.367
PTK7	5.52E-16	-0.37
KCTD15	5.20E-16	-0.37
KIAA1549	3.07E-17	-0.385
GTF2IRD1	1.47E-17	-0.388
SOCS7	6.11E-18	-0.392
ADGRA3	1.88E-23	-0.448
AL031778.1	1.09E-23	-0.45

D



E

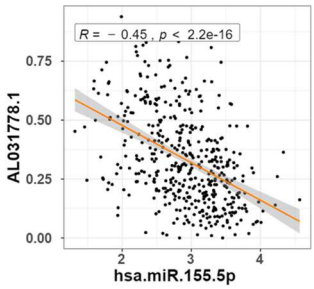
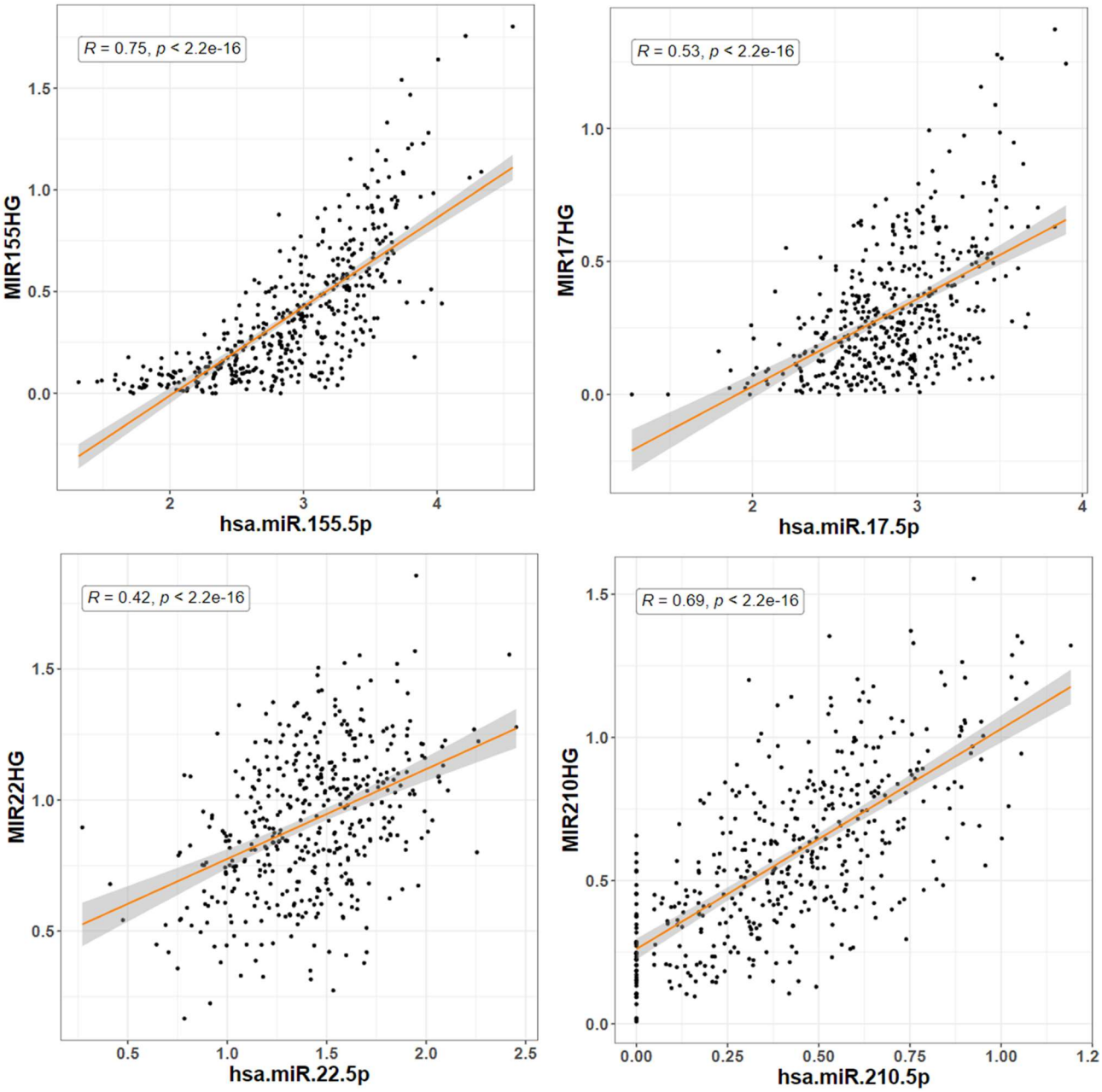


Figure S4. Lasso analysis of miRNAs against IFNγ signaling in melanoma (related to Fig.4). A. Mean-squared cross-validation (CV) error as a factor of increasing model penalty is shown. **B.** miRNA predictors with positive and negative coefficients at the lambda value corresponding to minimum CV error + 1 standard error are shown. **C.** Transcripts with the top positive and negative correlations with miR-155-5p are shown. **D-E.** TCGEx scatterplot module was used to visualize top correlators of miR-155-5p.

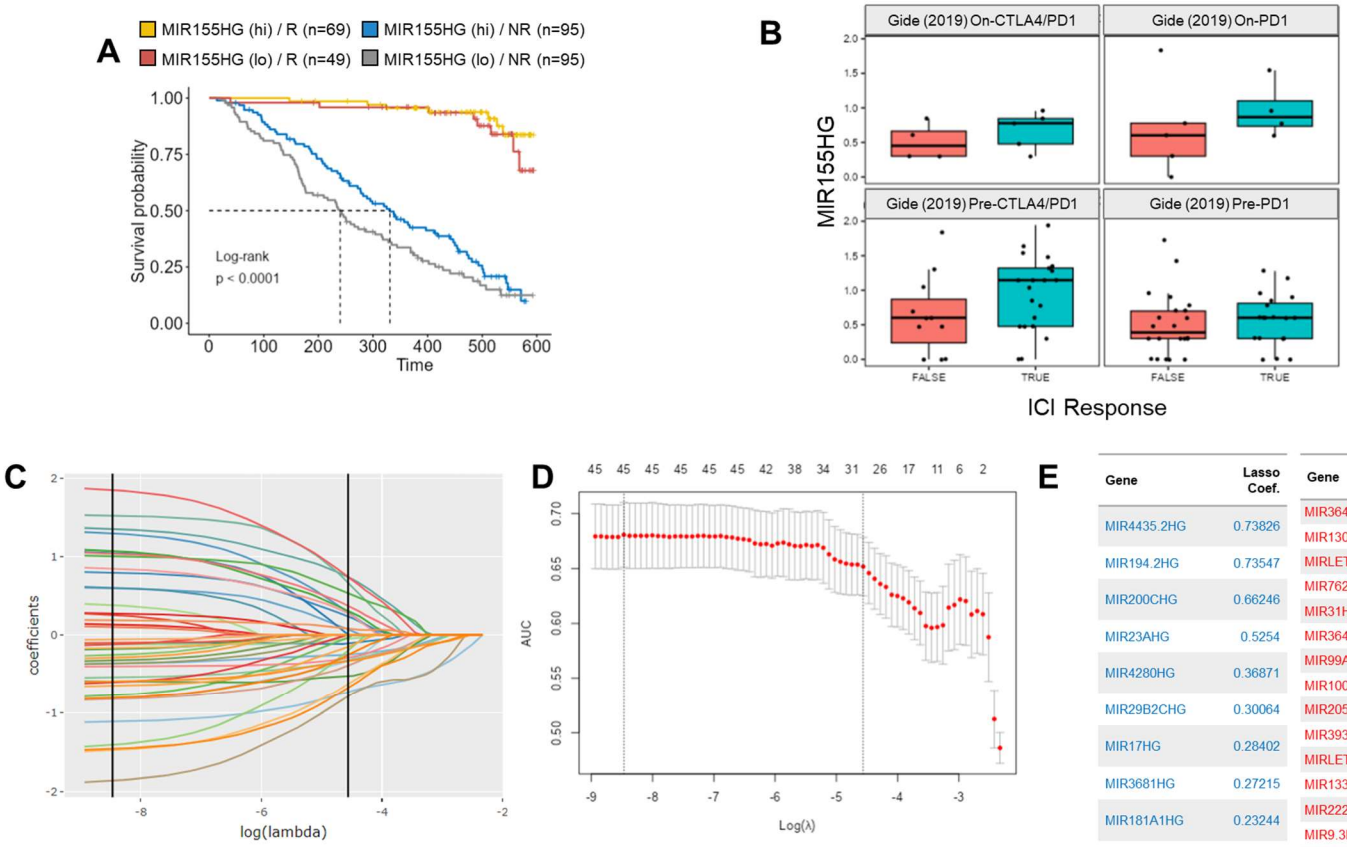


867

868

869 **Figure S5. Comparison of the expression levels of mature miRNAs and their host genes (re-**
870 **lated to Fig.5).** Scatterplots show the strong positive correlation between the host gene transcripts
871 (RNAseq data) and the mature miRNA levels (miRNAseq data) in tumor biopsies.

872



Gene	Lasso Coef.
MIR4435.2HG	0.73826
MIR194.2HG	0.73547
MIR200CHG	0.66246
MIR23AHG	0.5254
MIR4280HG	0.36871
MIR29B2CHG	0.30064
MIR17HG	0.28402
MIR3681HG	0.27215
MIR181A1HG	0.23244
MIR663AHG	0.07641
MIR155HG	0.06457

Gene	Lasso Coef.
MIR3648.1	-0.79195
MIR1302.2HG	-0.73349
MIRLET7A1HG	-0.68299
MIR762HG	-0.63343
MIR31HG	-0.53539
MIR3648.2	-0.40083
MIR99AHG	-0.34218
MIR100HG	-0.33294
MIR2052HG	-0.33087
MIR3936HG	-0.29763
MIRLET7BHG	-0.26012
MIR133A1HG	-0.25823
MIR222HG	-0.20291
MIR9.3HG	-0.15558
MIR1915HG	-0.11297
MIR205HG	-0.05933
MIR22HG	-0.03467

Figure S6. Investigating the relationship between immunotherapy response and MIR155HG and other miRNA host genes (related to Fig.6). A. Kaplan-Meier survival analysis demonstrates that higher MIR155HG levels (categorized at median) prior to treatment correspond to improved survival in melanoma patients; however, this trend was more emphasized in non-responders. **B.** Boxplots show the MIR155HG transcript levels in responders and non-responders during or prior to ICI treatment. **C-E.** Binomial lasso was performed using binarized response (CR/PR vs PD) as response variable and 45 miRNA host genes as predictors after low-expression filtering. Regularized coefficients (C), the cross-validated AUC of the penalized model (D), and non-zero coefficients at the minimum value of lambda within 1 standard error of the highest AUC (E) are shown.

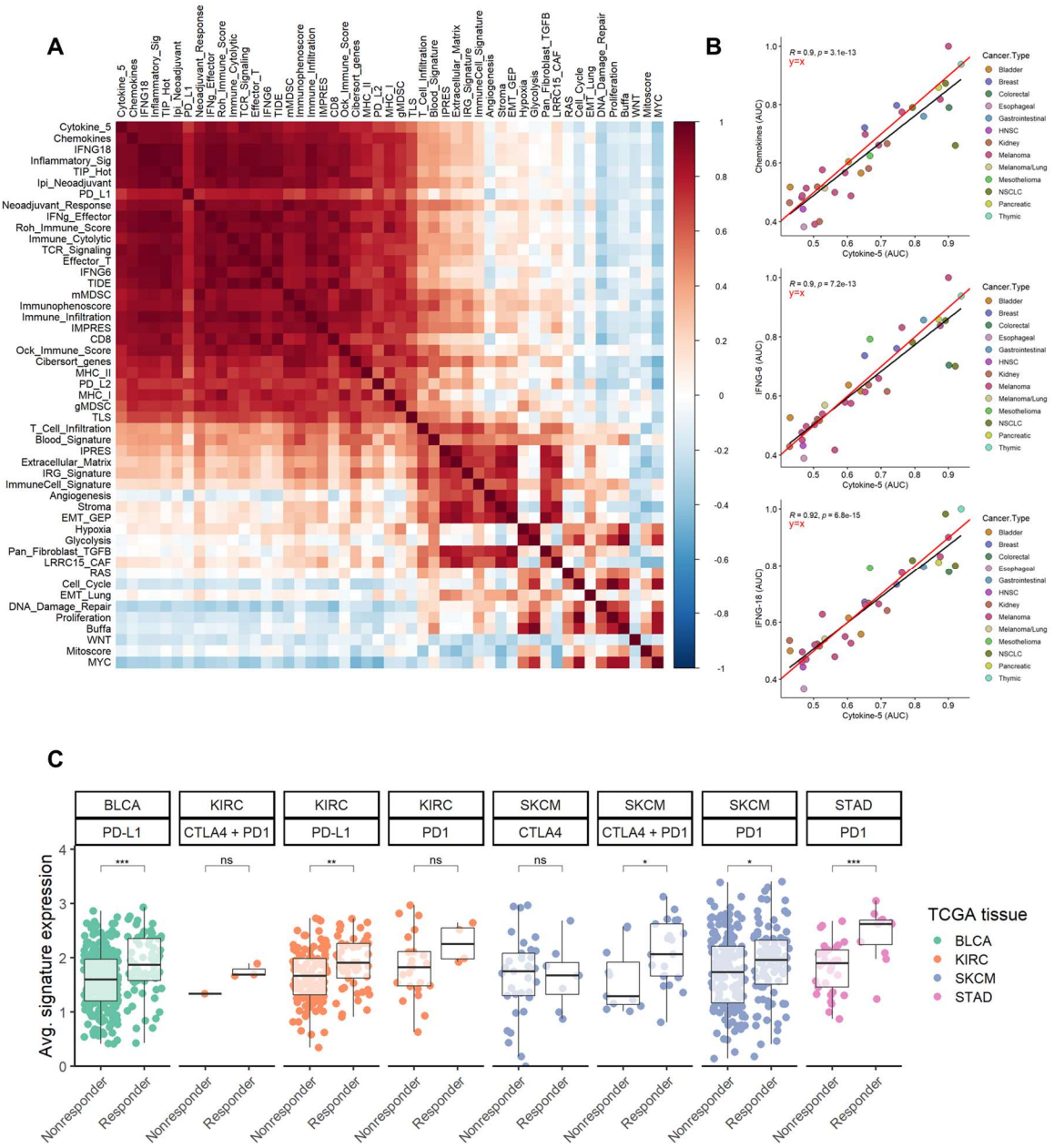


Figure S7. Examination of the Cytokine-5 signature as a new marker for immunotherapy response (related to Fig.7). A. The pairwise correlations between various signatures predicting immunotherapy response are depicted in a heatmap (source: IOSig platform). The Pearson correlation coefficients were calculated by using the AUROC values across the studies in the IOSig database. **B.** Pairwise correlations between the Cytokine-5 and IFNG6, IFN18 and Chemokine signatures are plotted. The best fitting line (black) and its equation is shown along with the y=x diagonal line (red) indicating equality. **C.** The boxplots illustrate the average Cytokine-5 signature expression in immunotherapy responders and non-responders across different cancer types.

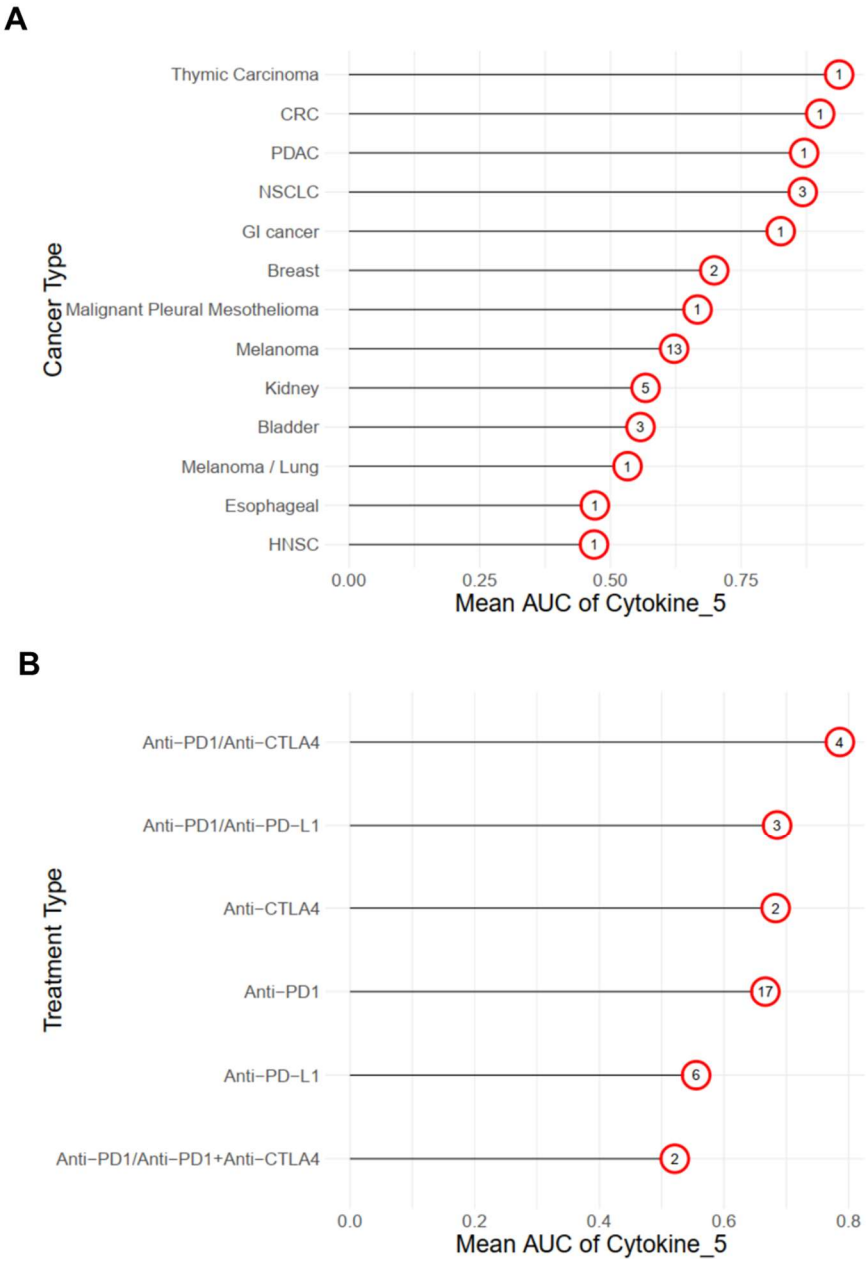


Figure S8. Examination of the Cytokine-5 signature across different cancers and immunotherapy modalities (related to Fig.7). A. The mean AUC values of Cytokine-5 signature was summarized for different cancer types (source: IOSig). The number of individual data sets for each cancer type is shown in circles. **B.** The mean AUC values of Cytokine-5 signature was summarized across different treatment modalities (source: IOSig). The numbers indicate the available data sets on the IOSig platform as in (A).