Manuscript Title:

TCGEx: A visual interface for multifaceted analyses of The Cancer Genome Atlas gene expression data

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, *]

[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. TCGA data repository consists of genomic, transcriptomic, epigenetic data and clinical metadata from 11000+ patients across 33+ cancer types. Analysis of such large data sets require coding skills that are associated with a steep learning curve for most non-specialists. To enable a wider utilization of TCGA gene expression 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. TCGEx offers customizable options to tailor the analysis to different study contexts and helps generate publication-ready plots. Developed using R/Shiny framework, this open-source tool enables researchers with no programming expertise to analyze TCGA RNA and miRNA sequencing data from multiple angles. Pre-processed data in TCGEx contains cancer subtype information as well as previously reported intratumoral immune cell signatures making it possible to investigate the possible tumor-immune interactions. TCGEx, available at https://tcgex.iyte.edu.tr, provides an interactive solution to extract meaningful insights from TCGA data and guide scientific research.

Keywords: TCGA, transcriptomics, 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 sequencing, whole exome sequencing, and genotyping and methylation arrays amounting to 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 and proteomic analyses helped differentiate disease subsets with different clinical characteristics leading to the development of novel therapeutics. Gene expression profiling has also helped identify immune signatures within the tumor microenvironment (TME). For instance, a recent report characterized at least 6 different immunological states across all tumor types in humans and provided a framework for investigating cancer-immune cell interactions (6). Additionally, in skin cutaneous melanoma, immune-infiltrated and immune-devoid subsets have been defined that were characterized by differential survival outcomes (7). Follow-up studies have revealed that the genomic and transcriptomic features of tumors determine whether they will respond to targeted therapies and immunotherapeutics (8–10), suggesting that a better understanding of the heterogeneity within the TME is essential for improved therapeutics and clinical outcomes.

An efficient analysis of the large-scale genomic data in the TCGA repository requires programming skills and it is often challenging for the non-specialists. Multiple web-based software applications have been developed to facilitate TCGA data analysis including cBioPortal, Web-TCGA, UCSC Xena, and GEPIA2 (11–14). Although these visual analysis interfaces reduced the barriers to utilizing omics data in cancer research and have been used by numerous investigators in the field, there is a need for a centralized platform that integrates multiple methods to perform customized analyses. In particular, no interactive analysis tool to date is equipped with sophisticated functional genomic analysis capabilities that allow examining TCGA transcriptomics data. In this report, we introduce The Cancer Genome Explorer (TCGEx), an open-source web-based analysis platform written in R/Shiny, to address this need and increase the accessibility and reusability of the TCGA data.

The TCGEx program offers powerful features including the ability to perform multivariable Kaplan-Meier and Cox proportional hazards survival modeling, gene set enrichment analysis (GSEA), principal component analysis (PCA), annotated heatmap generation and hierarchical clustering, feature-to-feature correlation analysis, exploratory graphing of gene expression and clinical metadata, receiver operating curve (ROC) analysis, linear regression-based machine learning for investigating predictor variables against an arbitrary

response variable. Through interactive interfaces and customizable plotting options, TCGEx allows users to focus on specific data subsets and tailor the analysis to various study contexts. Furthermore, data from multiple TCGA projects can be aggregated to perform integrative analyses and investigate commonalities among cancers. TCGEx's efficient and comprehensive computational pipeline utilizes pre-processed data that was prepared by combining normalized RNA/miRNA sequencing data with clinical metadata and intratumoral immune signatures described by the leading studies in the field (6). 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 iterative analyses of all TCGA projects. Data files for this 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 (6). TCGEx analysis starts by selecting one or more TCGA projects in the data selection module. After data selection, relevant data objects are loaded into the session and passed to individual analysis modules **(Figure-1)**. TCGEx offers 10 analytical pipelines accessible through user-friendly 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) top correlating feature analysis; vi) hierarchical clustering and heatmap generation; vii) gene set enrichment analysis (GSEA); viii) receiving operating characteristics (ROC) analysis; ix) principal component analysis (PCA); and x) machine learning analysis with lasso, ridge, or elastic net regression. 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 input 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 that generate both the user interface and the server code required for computation. This design facilitates troubleshooting and will help incorporate our 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 the TCGA 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. TCGEx data objects were prepared by sample-level merging of counts-per-million-normalized log-transformed RNAseq and miRNA data and clinical metadata described by the TCGA studies. To facilitate downstream analyses, we removed genes that are not expressed in 25% or more of the samples in each TCGA project. Unlike some of the previously available tools for TCGA data analysis (13, 14), TCGEx allows users to select single or multiple TCGA projects 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. In these modules, users can apply further filtering to focus the analysis on specific tissue types such as primary or metastatic samples.

Survival Analysis Modules

Studying the association between gene expression and survival outcome is important for understanding key mechanisms of cancer progression (15). TCGEx features two distinct survival analysis methods: Kaplan-Meier (KM) and Cox proportional hazards (CoxPH) modeling (16, 17). In both modules, categorical and continuous variables can be examined. Categorical meta-variables can differ between TCGA projects but generally include tumor grade, mutation types, and patient demographic information, among other features. Continuous variables such as gene expression and intratumoral immune cell signatures (6) 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 (7, 18), genome-wide methylation state (7), oncogenic signaling (OncoSign) cluster type (19), genetic aberrations (2), genomic instability levels (5), intratumoral lymphocyte scores (7), and numerous other metadata features **(Supplementary Table-1)**. This module allows selecting specific categories to generate gene expression plots to emphasize interesting gene expression 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 (8) and miRNA-mRNA interactions (20). 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 (6). 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 and generate publication-ready figures. In addition to these functionalities, we anticipated that it can 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 across the data set 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 (21). We created an interactive heatmap module equipped with flexible options to facilitate visualization and analysis of TCGA RNAseq and miRNAseq data. Here, users can manually enter genes of interest, or select curated gene sets described in the Molecular Signatures Database (MSigDB) (22). 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 can 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 (23). 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 (24). 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 (24). 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 (22). 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 (24), 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 classification performance of variables in a binary classification system (25). 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) can be examined to assess the power and accuracy of the classifier. ROC analysis is commonly used in the field for investigating diagnostic and prognostic biomarkers (26–28). 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 identify novel predictors in cancer data.

Principal component analysis module

Dimensionality reduction techniques such as principal component analysis (PCA) are commonly employed in transcriptomics (29). 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 inputs to be utilized in various study contexts. Similar to other modules, one can specify sample types and select a custom list of genes to prepare the gene expression matrix for PCA. Here, the module conveniently allows selecting all genes in the RNAseq/miRNAseq data or specifying genes from MSigDB pathways. Although PCA accounts for both highly and lowly variable genes, we provided a pre-filtering option to speed up the analysis by removing lowly variable genes. In PCA plots, it is usually helpful to annotate individual points based on specific characteristics. For instance, one may want to see whether samples with certain genetic features form a separate cluster in the PCA space (30). To address this need, this module allows color-coding data points based on categorical

or continuous variables, the latter of which is categorized at the median value to form “high” and “low” groups. PCA module generates customizable and interactive graphics and allows exporting publication-ready plots.

Machine learning module

Machine learning (ML) refers to a range of applications and algorithms that aim to extract relevant and useful information from data. In the context of bioinformatics, ML is utilized to perform classification, prediction, and feature selection from biological data (31, 32). Regularized regression, a type of supervised machine learning technique, is a derivative of linear regression which allows one to simultaneously create a model and perform feature selection in high dimensional data. It is optimized to minimize the sum of squared residuals while penalizing the generated model coefficient estimates. The penalty term is applied to the model equation to reduce model complexity and make the prediction with the limited mean squared error. Ridge, lasso, and elastic net regression are all types of regularized regression methods with varying strengths depending on the underlying structure of the data. Ridge regression shrinks the estimated coefficients without making them zero and assigns correlating parameters with similar coefficients (33). Lasso regression on the other hand can shrink model coefficients to zero and therefore performs feature selection (34). Elastic net regression can refer to a middle ground in between ridge and lasso regression and it is ideal for data sets in which the number of predictor variables significantly exceeds the number of samples and/or there are a high number of correlating variables. The optimal degree of penalization for each approach can be selected using cross-validation across the data set. The ML module in TCGEx provides a user-friendly interface for utilizing these regularized linear regression methods on the TCGA transcriptomics data. Users choose response and predictor variables for the model and specify the ML algorithm by setting the penalization parameter alpha. The response variable (i.e., dependent variable) in these analyses can be constructed from a custom list of genes or MSigDB gene sets by taking the average of the expression values. While the predictor variables (i.e., independent variables) can be specified similarly, we also provided an option to select all miRNAs as predictors to make the pipeline convenient for miRNA-focused analyses (35). This way, one can easily examine, for instance, miRNAs that are most closely associated with a biological pathway of interest. The ML module can use the entire data to train a model or split it into training and testing subsets to evaluate the overall model accuracy (32). The findings of the analyses are displayed through interactive graphs which show penalized coefficients and the mean squared error of the cross-validated model across various levels of regularization. Taken together, the ML module provides a customizable platform for sophisticated analysis of the TCGA transcriptomics data.

Results and Discussion:

In this section, we provide a use-case scenario to demonstrate how TCGEx can aid hypothesis-driven research. While the analyses herein will demonstrate the basic capabilities of the pipeline on the TCGA Skin

Cutaneous Melanoma (SKCM) primary and metastatic tumor data, users can easily adapt this approach to various study contexts and examine other TCGA data sets. Melanoma is an aggressive form of cancer that develops in pigment-producing cells of the skin, although it can rarely be observed in other organs including the eye and gastrointestinal tract. More than 320,000 new cutaneous melanoma cases are reported globally each year and 55,000+ lives are lost because of this disease (36). UV exposure is a significant risk factor for the development of melanoma, and most cases are diagnosed at an advanced stage where the cancer cells have already metastasized to nearby lymph nodes or distant organs. Accordingly, the TCGA-SKCM study is composed of mainly metastatic tumor biopsies, as shown in the TCGEx data selection module **(Fig.2a)**. These samples were collected from comparable numbers of male and female patients in various age groups **(Fig.2b)**. The landmark TCGA paper describing the genetic features of SKCM classified tumors into subtypes based on mutations in BRAF, NF1, and RAS genes (7). These subtypes are characterized by distinct sets of accompanying genomic aberrations including PTEN loss in BRAF-mutant tumors, KIT amplification in triple wild-type (WT) tumors, Akt3 amplification in RAS-mutant and wild-type tumors which are readily observed in transcriptomics data **(Fig.2c)**. When these graphs are faceted by patient gender, some of these genes showed sex-dependent differences in their expression patterns **(Fig.S1a, b)**. Interestingly, BRAF-mutant melanomas expressed higher levels of transferrin (TF) which was recently shown to protect circulating tumor cells from oxidative stress leading to BRAF inhibitor resistance **(Fig.S1c)** (37). On the other hand, NF1-mutant melanoma was distinctly marked by the high expression of hepatocyte-specific transcription factor ONECUT1 which was shown to be associated with tumor progression **(Fig.S1c)** (38). Our analyses also revealed that ZC3H13, a component of the RNA methyltransferase complex, is enriched in RAS-mutant melanoma; and triple-WT tumors selectively expressed FOXF2, a transcription factor that plays a complex role in regulating oncogenic signaling pathways **(Fig.S1c)** (39). These findings suggest that distinct molecular mechanisms may be responsible for driving different tumor subtypes in melanoma and indicate possible intervention points.

BRAF-mutant melanoma constitutes about 50% of the cases in the TCGA cohort followed by RAS-mutant (~30%), NF1-mutant, and triple-WT tumors (~10% each). Of these 4 groups, BRAF-mutant melanoma has the most favorable prognosis **(Fig.2d)**, although this may be confounded by the younger age of the patients in this group (7). While melanoma incidence is comparable between both sexes, females were reported to have a reduced risk of mortality (40). Examination of the TCGA-SKCM cohort did not show a significant survival difference between the two groups; however, a slightly increased risk was observed for the males **(Fig.S2a, b)**. Notably, when we examined only BRAF and NRAS-mutant tumors, males were associated with poorer survival outcomes, especially in the NRAS-mutant subgroup **(Fig.2e)**. These findings suggest that sex-specific factors may differentially influence the prognosis in different mutation subtypes.

In addition to defining potential genetic drivers in melanoma, the TCGA-SKCM manuscript also identified three distinct tumor subsets based on gene expression profiles as immune, keratin, and MITF-low groups (7). The immune subset was characterized by the higher expression of immune cell-specific genes **(Fig.3a)**, while the keratin and MITF-low subsets were characterized by the high and low expression of epithelial genes, respectively, in the absence of a clear immune infiltration signature. Of these three groups, the immune subset exhibits the most favorable survival outcome **(Fig.3b, S3)**, which is consistent with the positive role of inflammation in SKCM reported previously (6). Interferon-gamma (IFNγ) is a key inflammatory cytokine released from T and NK cells to eliminate tumor cells. IFNγ exerts its multifaceted functions within the TME by signaling through IFNγ-receptors (IFNGR) and regulating dozens of downstream genes in tumor cells and the neighboring immune cells (41). We examined IFNγ response gene expression signature in SKCM subsets and observed higher levels of most of these transcripts in the immune subset **(Fig.3c)** (42), and as expected, these trends were parallel by the elevated IFNγ levels and CD8+ T cell signatures within the TME. We next wanted to investigate the relationship between the intratumoral CD8+ T cell score and the T cell receptor (TCR) diversity (6). This is of interest because multiple studies have suggested that the clonally expanded T cells within melanoma are associated with a spectrum of dysfunctional phenotypes (43, 44). Increasing CD8+ T cell signature within TCGA-SKCM tumors positively correlated with the overall TCR diversity, suggesting a polyclonal T cell infiltrate and the presence of T cell clones with different antigenic specificities **(Fig.3d)** (45). Notably, increasing TCR diversity correlated with higher levels of cytotoxic effector molecule perforin (PRF1) and decreasing genetic signatures of M2 macrophages that are known to exert pro-tumorigenic functions **(Fig.3e)** (6). 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 (46). 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 as the response variable. 26 positive and 18 negative coefficients remained in the optimally regularized model (log lambda + 1 standard error) **(Fig.4a, S4)**. 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.S5)**. Since a single miRNA can control dozens of targets and miR-155-5p correlates with immunity in SKCM (47), 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 **(Fig.4c)**. 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 KEGG gene sets from the MSigDB database (42) and noted that the expression of genes associated with cytokine-receptor interaction 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 (47–50).

Taken together, TCGEx provides a powerful and flexible interface for investigating the TCGA gene expression data through sophisticated analyses. The use case scenario provided for SKCM here can be adapted to study many other aspects of cancers including genetic drivers in microsatellite-stable or unstable subsets of colon cancer and their immune associations (5), prognostic markers in different mutational and histological subtypes of breast cancers (4), and molecular characteristics of previously defined miRNA subtypes in ovarian cancer (3), among others. Furthermore, TCGEx allows integrative analyses of multiple cancer types to study the molecular underpinnings of human cancer. Visual analysis interfaces like TCGEx increase the accessibility and reusability of high-throughput cancer data and facilitate scientific research by serving as a bridge between bench scientists and bioinformaticians. TCGEx, publicly and freely accessible on the web, not only reduces barriers to data analysis for researchers with no programming experience but also can be implemented into larger analysis pipelines to facilitate hypothesis-driven research.

Acknowledgments:

We thank the 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 patients, and their families participating in the TCGA project, and the TCGA initiative as a whole for making this repository publicly available. 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, and 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 datasets 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.

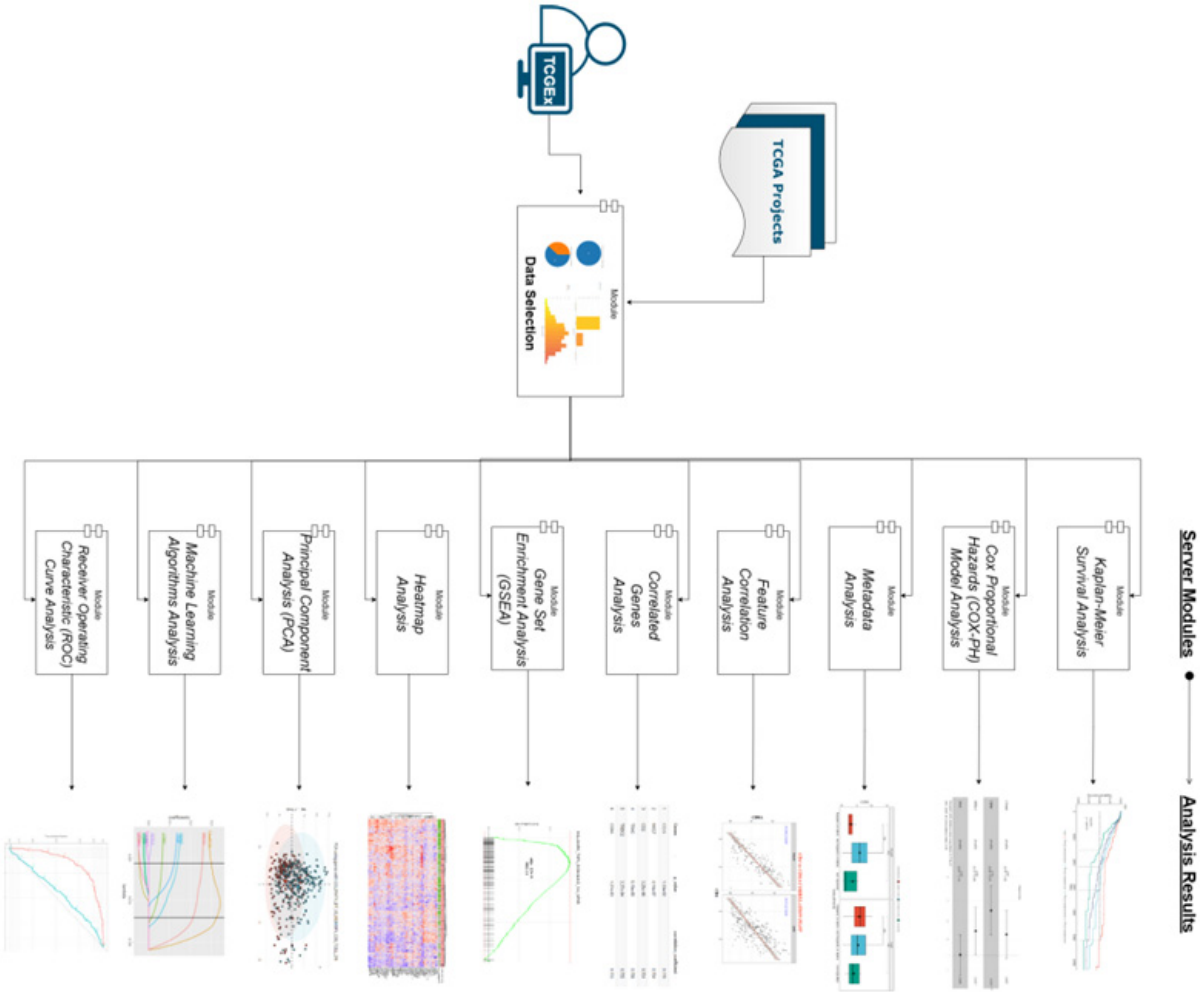


Figure 1. TCGEx workflow overview. TCGEx pipeline begins with the user selection of pre-processed transcriptomics data from a single TCGA project or multiple projects for integrated analyses. The data selection module shows descriptive statistics including the sample types, gender, and age distribution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

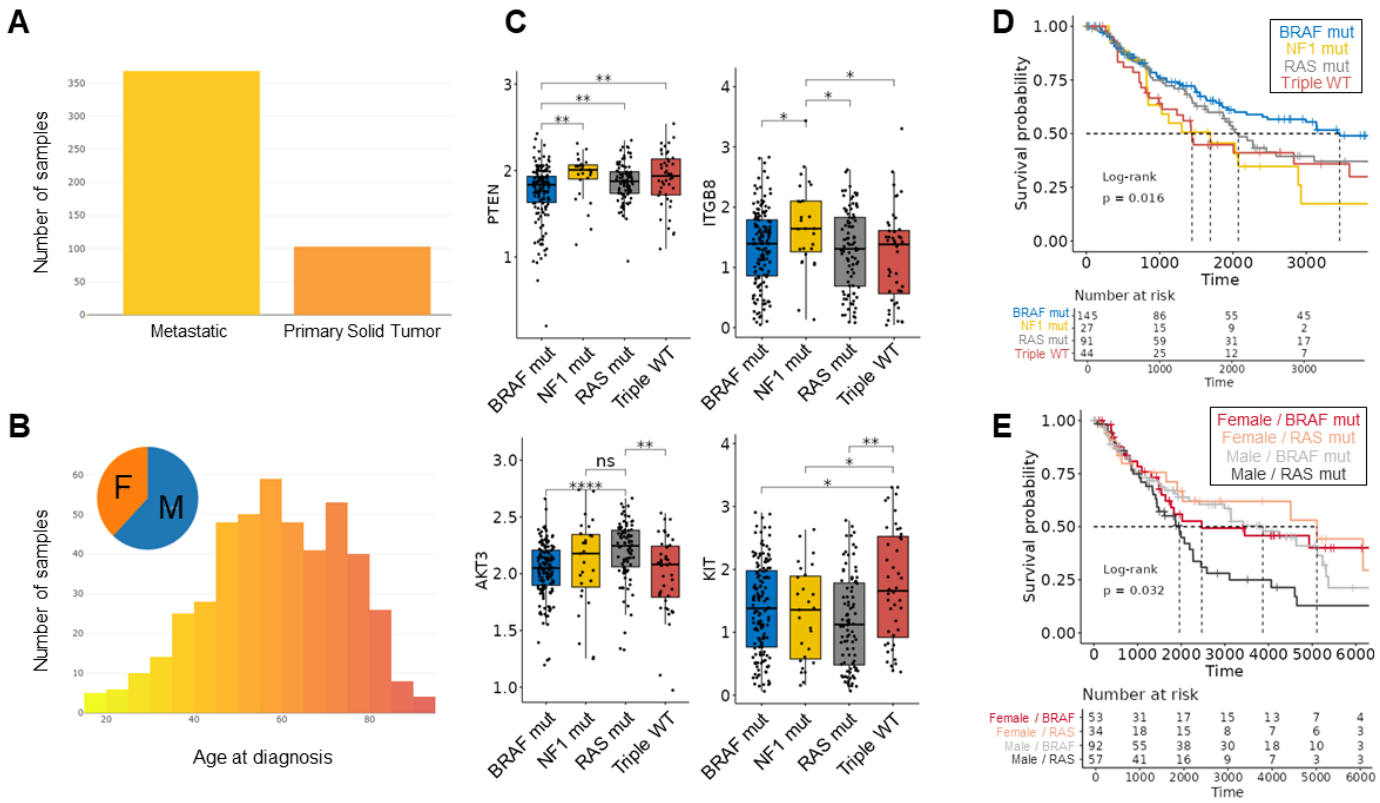


Figure 2. Analysis of mutational subsets in skin cutaneous melanoma. A. The TCGA-SKCM project consists of mostly metastatic samples. Analyses described herein are performed on both metastatic and primary tumors. **B.** Comparable numbers of male and female individuals from a range of age groups are included in the SKCM patient cohort. **C.** Expression profiles of PTEN, ITGB8, AKT3, and KIT can distinguish mutational subtypes of SKCM, suggesting different molecular signaling landscapes exist in disease subsets (7). Statistical comparisons are performed using the t-test although users have the option to select non-parametric Wilcoxon test as well. The p-values are adjusted for multiple comparisons (p<0.0001, "****"; p<0.001, "***"; p<0.01, "**"; p<0.05, "*"; p>0.05, "ns") **D.** Kaplan-Meier survival analysis revealed that BRAF-mutant SKCM has the most favorable clinical outcome. 10-year survival curves are shown where dashed lines indicate the median survival time points in each group. The risk table provided below the survival curve denotes the number of surviving patients (i.e. at-risk patients) at different follow-up times. The log-rank p-value shown on the graph tests the null hypothesis that none of the groups are different from one another. **E.** Survival modeling using only BRAF- and RAS-mutant melanoma patients suggests that males in the RAS-mutant group have poorer survival outcomes compared to BRAF-mutant males and females in general. TCGEx provides users with flexible options to change the color scheme and the range of time axis in the graphs.

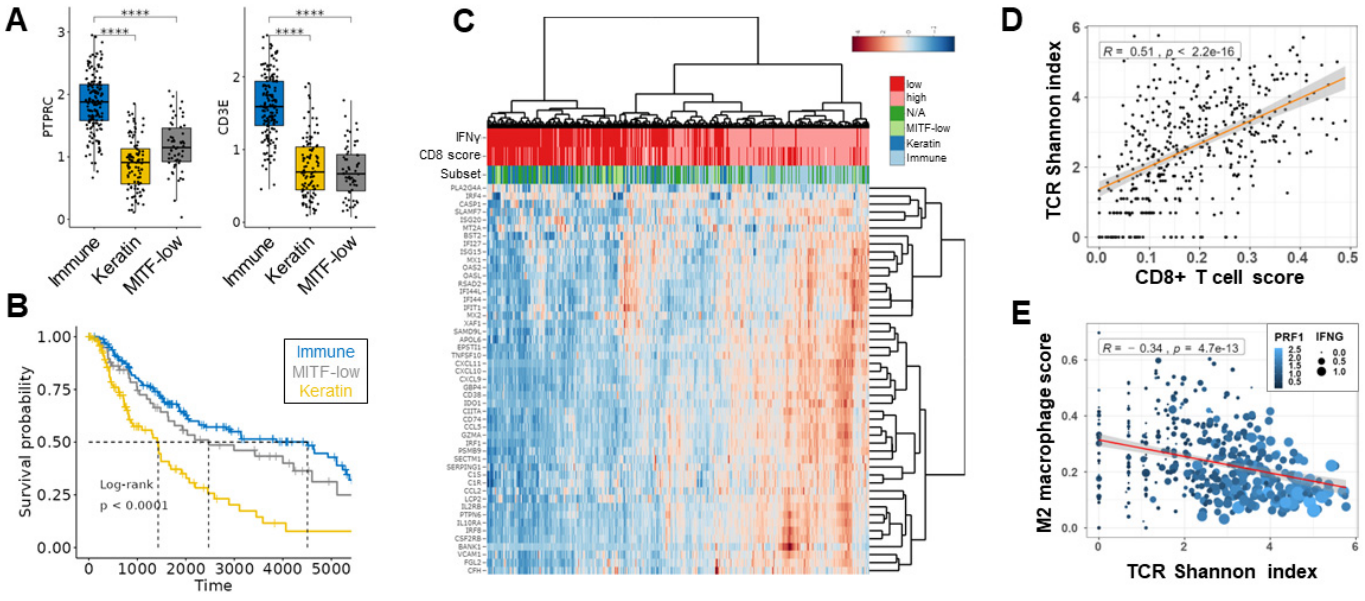


Figure 3. Analysis of the melanoma intratumoral immune landscape. A. The immune-enriched melanoma subset previously defined based on multi-gene expression profiles (7) is characterized by higher levels of PTPRC (CD45, a pan-leukocyte marker) and CD3E (a T cell marker) transcripts. T-test p-values are adjusted for multiple comparisons (p<0.0001, "****"; p<0.001, "***"; p<0.01, "**"; p<0.05, "*"; p>0.05, "ns"). **B.** Immune subset of the SKCM is associated with an improved survival outcome compared to keratin and MITF-low subsets. **C.** Heatmap visualization of the top 25% variably expressed genes from the MSigDB Hallmark IFNγ response gene set (42) in primary and metastatic melanoma is shown (50 out of 200 genes are plotted). The annotation rows on top of the heatmap were prepared by categorizing IFNγ transcript level and CIBERSORT CD8+ T cell score (51) at their median values. Melanoma subsets defined previously based on genome-wide expression patterns are also shown. CPM-normalized and log-transformed transcript counts are scaled by row to better show expression differences in data subsets. The default TCGEx parameters were used for hierarchical clustering (Euclidean distance and Ward.D2 clustering algorithm for both genes and samples). **D.** Increasing intratumoral CD8+ T cell signature is associated with a higher TCR diversity suggesting a polyclonal T cell infiltrate. **E.** The higher TCR diversity in melanoma tumors correlates with increased levels of T cell effector genes including perforin (PRF1) and IFNγ and decreased pro-tumorigenic M2 macrophage signatures (6).

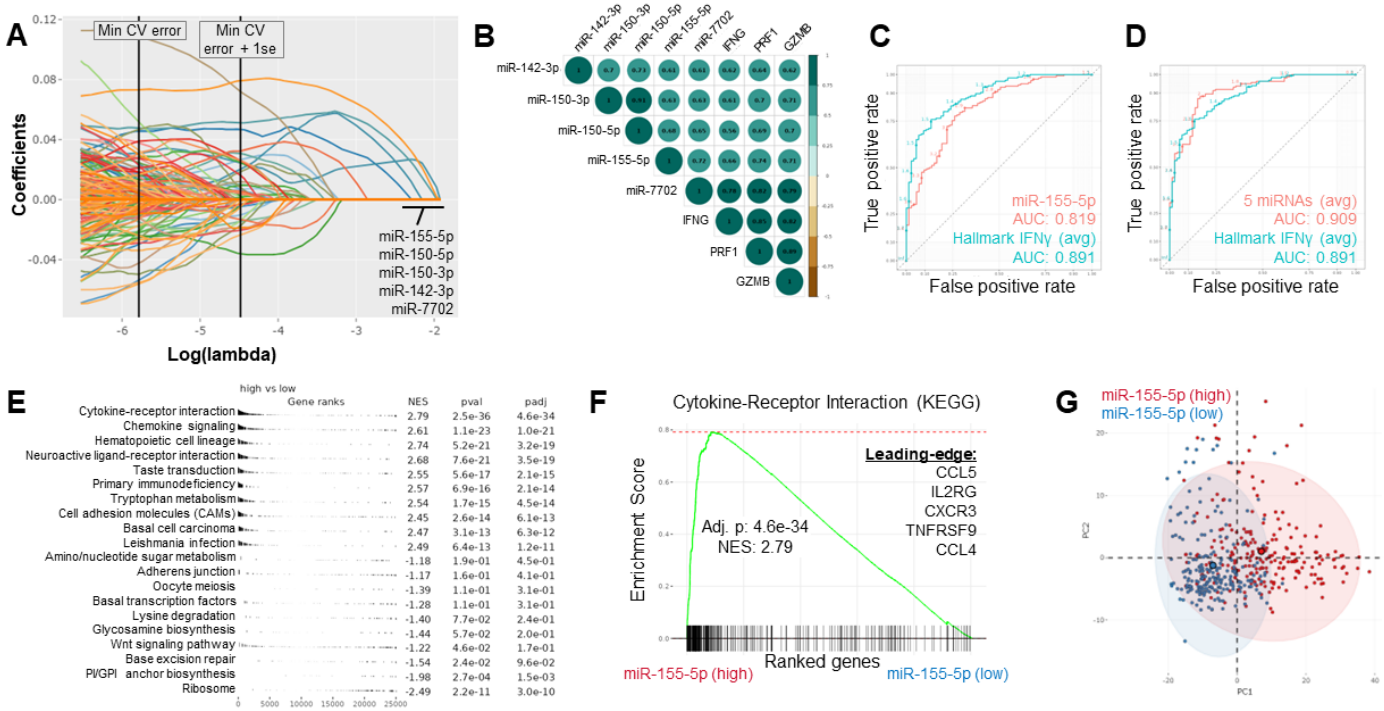


Figure 4. Machine-learning analysis of the associations between miRNAs and antitumor immunity. **A.** Lasso regression was performed by selecting IFNγ response score as the response variable and mature miRNA transcripts as predictor variables. 529 miRNA predictors in the pre-processed SKCM data entered the lasso model after splitting the data set into training (70%) and test (30%) subsets for cross-validation. The graph shows the coefficient regularization paths for individual predictors at increasing levels of penalty term lambda (λ). Vertical lines indicate the cutoffs where the cross-validation (CV) error was at its minimum value and within one standard error. 5 miRNAs that continued to have non-zero coefficients at arbitrarily high levels of regularization are shown. **B.** 5 miRNAs highlighted in panel (A) show positive correlations with each other and immune effector transcripts IFNγ, PRF1, and GZMB. **C.** ROC analysis with miR-155-5p and averaged Hallmark IFNγ response gene set was performed to test their power in predicting the immune subset of SKCM (binarized as 1 for correct grouping) over keratin and MITF-low groups (both binarized as 0 for incorrect grouping). AUC values indicate both predictors are strongly associated with the immune subset, although the IFNγ response gene set performed better compared to miR-155-5p. **D.** ROC analysis was performed as described in (C), with the averaged expression of 5 miRNAs (and IFNγ response) as predictors. In this analysis, the 5 miRNA-expression signature performed slightly better than the IFNγ response score for predicting the immune subset in melanoma. **E.** Gene set enrichment analysis was performed between miR-155-5p-high and -low subsets of melanoma (categorized at the median expression value) using genome-wide signal-to-noise ranked (24) transcripts and the KEGG gene sets (22). The top 10 positively and negatively enriched

gene sets are shown in the table where NES>0 indicates enrichment in the miR-155-5p-high subset, and "padj" shows the Benjamini-Hochberg adjusted p-value to account for multiple comparisons across the gene set. **F.** Enrichment plot showing the most significantly enriched KEGG gene set in the miR-155-5p-high melanoma samples where leading-edge genes contributing strongly to the score are also shown. **G.** Principal component analysis of melanoma samples was performed using 685 genes associated with the adaptive immune response GO annotation. Samples are categorized at the median expression of miR-155-5p and color-coded to demonstrate miR-155-5p-high and -low samples in the PCA space.

References:

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

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., Ouellette,B.F.F., Li,C.H., *et al.* (2020) Pan-cancer analysis of whole genomes. *Nature 2020 578:7793*, **578**, 82–93.

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) Integrated genomic analyses of ovarian carcinoma. *Nature 2011 474:7353*, **474**, 609–615.

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

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

6. 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., Eddy,J.A., *et al.* (2018) The Immune Landscape of Cancer. *Immunity*, **48**, 812-830.e14.

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

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

9. Ayers,M., Lunceford,J., Nebozhyn,M., Murphy,E., Loboda,A., Kaufman,D.R., Albright,A., Cheng,J.D., Kang,S.P., Shankaran,V., *et al.* (2017) IFN-γ-related mRNA profile predicts clinical response to PD-1 blockade. *Journal of Clinical Investigation*, **127**, 2930–2940.

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

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

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

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

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

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

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

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

18. 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.* (2018) A Comprehensive Pan-Cancer Molecular Study of Gynecologic and Breast Cancers. *Cancer Cell*, **33**, 690-705.e9.

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

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

21. Sørlie,T., Perou,C.M., Tibshirani,R., Aas,T., Geisler,S., Johnsen,H., Hastie,T., Eisen,M.B., Van De Rijn,M., Jeffrey,S.S., *et al.* (2001) Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. *Proc Natl Acad Sci U S A*, **98**, 10869–10874.

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

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

24. Subramanian,A., Tamayo,P., Mootha,V.K., Mukherjee,S., Ebert,B.L., Gillette,M.A., Paulovich,A., Pomeroy,S.L., Golub,T.R., Lander,E.S., *et al.* (2005) Gene set enrichment

analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci U S A*, **102**, 15545–15550.

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

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

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

28. 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., Kefford,R.F., *et al.* (2019) Circulating cytokines predict immune-related toxicity in melanoma patients receiving anti-PD-1–based immunotherapy. *Clinical Cancer Research*, **25**, 1557–1563.

29. Yeung,K.Y. and Ruzzo,W.L. (2001) Principal component analysis for clustering gene expression data. *Bioinformatics*, **17**, 763–774.

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

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

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

33. Friedman,J., Hastie,T. and Tibshirani,R. (2010) Regularization Paths for Generalized Linear Models via Coordinate Descent. *J Stat Softw*, **33**, 1–22.

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

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

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

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

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

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

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

41. Parker,B.S., Rautela,J. and Hertzog,P.J. (2016) Antitumour actions of interferons: implications for cancer therapy. *Nature Reviews Cancer 2016 16:3*, **16**, 131–144.

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

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

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

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

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

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

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

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

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

51. 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. (2015) Robust enumeration of cell subsets from tissue expression profiles. *Nat Methods*, **12**, 453–457.

#

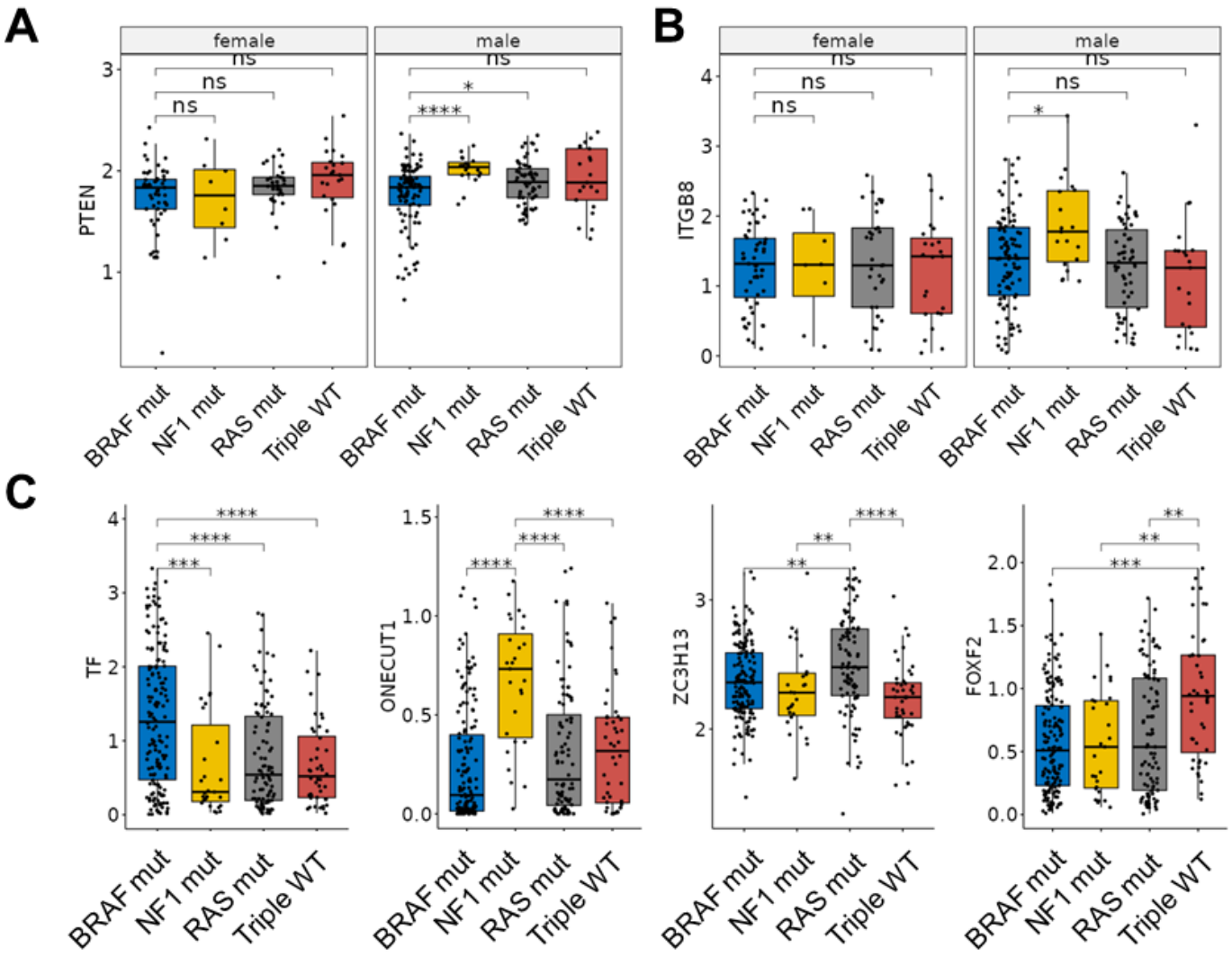


Figure S1. Expression profiles of subtype-selective genes in melanoma (related to Fig.2). A-B. Expression of previously reported genes distinguishing melanoma mutational subtypes may show sex-specific expression patterns. **C.** Other transcripts selectively expressed in mutational subtypes of melanoma 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")

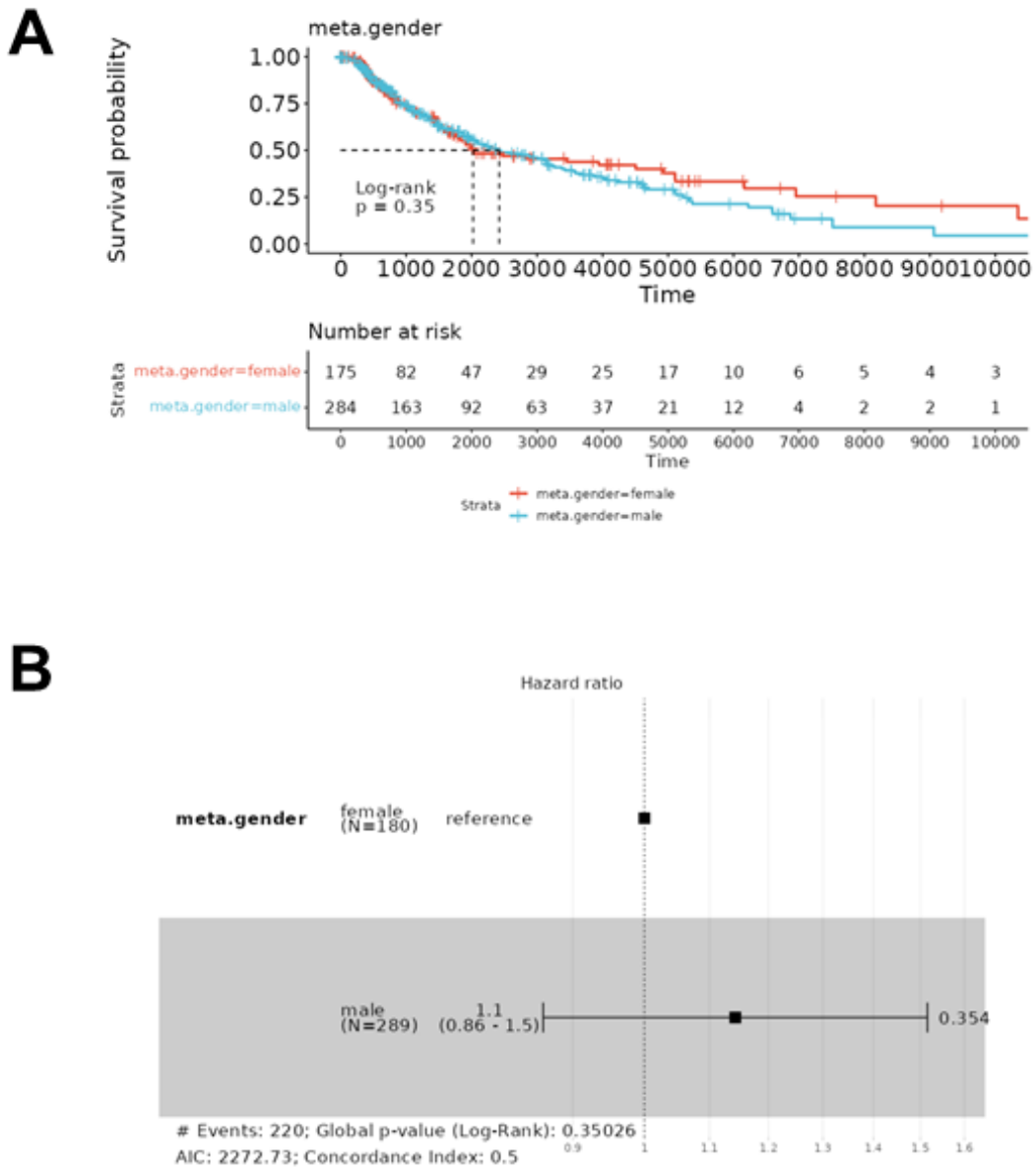


Figure S2. Modeling the effects of patient gender on melanoma survival (related to Fig.2). A. Kaplan-Meier survival analysis of gender groups in SKCM shows non-significant differences, although male patients tended to have decreased survival in longer follow-up. **B.** Cox proportional hazards survival 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.

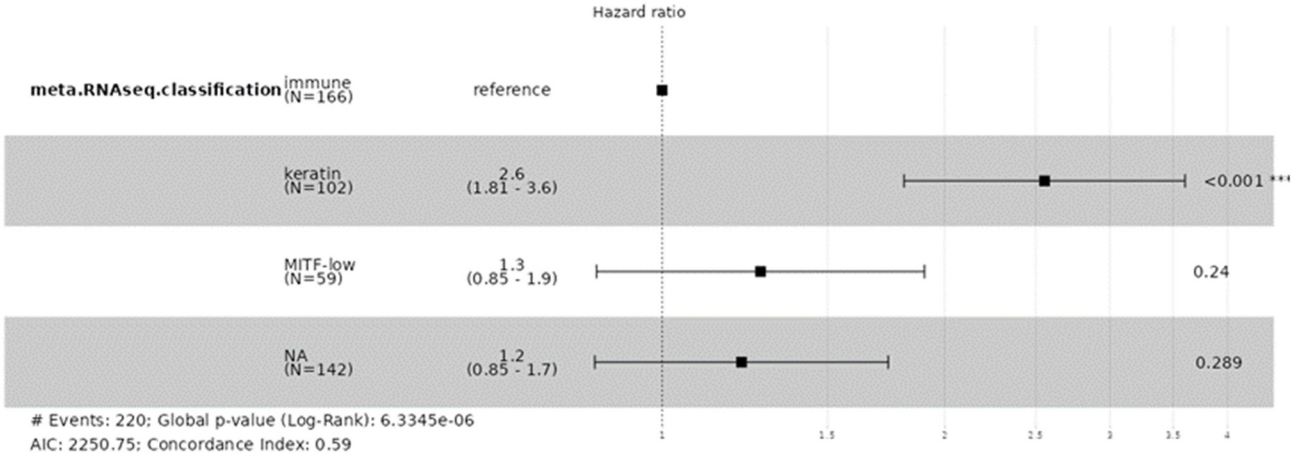
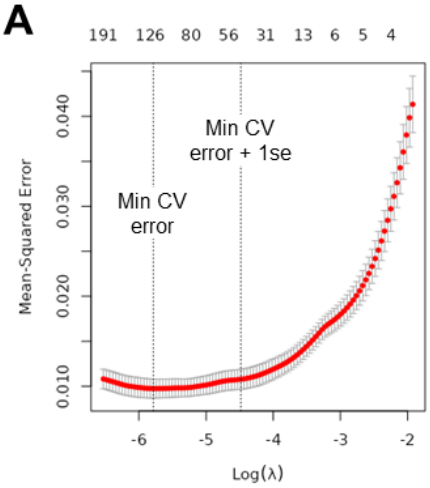


Figure S3. Survival modeling of transcriptomic subsets of melanoma (related to Fig.3). The forest plot shows the hazard ratios of keratin, MITF-low, and unannotated groups in comparison to the immune subset appointed as the reference group. The keratin subset is associated with a statistically significant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

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

C

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

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-C.** miRNA predictors with **(B)** positive and **(C)** negative coefficients at the lambda value corresponding to minimum CV error + 1 standard error (a more regularized model that is still within one standard error of the min CV error).

A

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

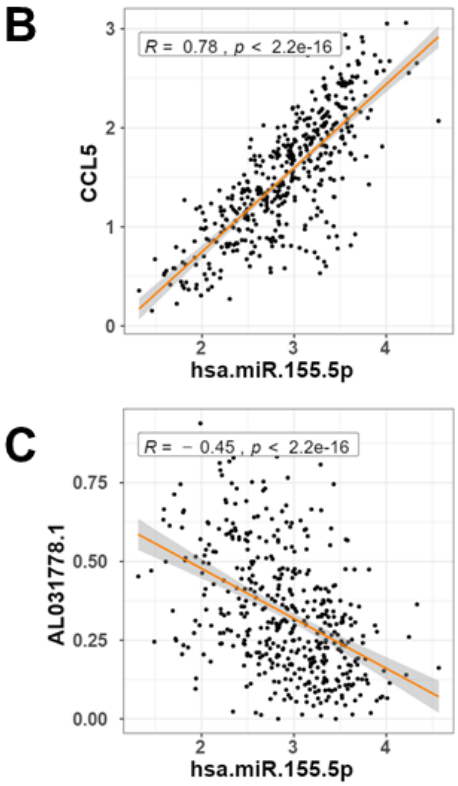


Figure S5. Positive and negative correlators of miR-155-5p in melanoma (related to Fig.4). A. TCGEx correlation module was used to find the top 10 positive and negative correlators of miR-155-5p and Pearson correlation coefficients and P-values are shown. **B-C.** Scatter plots are used to visualize the strong relationship between the expression levels of miR-155-5p and its top correlators.