# The prognostic potential of alternative transcript isoforms across human tumors

Juan L. Trincado[1], E. Sebestyén[2], H. Climente-González[1], A. Pagés[1], E. Eyras[1,3]

[1]Universitat Pompeu Fabra (UPF), Dr. Aiguader 88, E08003 Barcelona, Spain

[2]The Firc Institute of Molecular Oncology (IFOM), Via Adamello 16, 20139 Milan, Italy

[3]Catalan Institution for Research and Advanced Studies (ICREA), Passeig Lluís Companys 23, E08010 Barcelona, Spain

Correspondence to: eduardo.eyras@upf.edu

The authors declare no conflict of interests.

# Abstract

**Background**

Molecular signatures can improve tumor stage identification, which is essential for therapy selection and patient prognosis. These signatures are generally based on the expression changes that occur during cancer progression, which are related to the activation of specific aggressive phenotypes. However, it is not yet known whether specific transcript isoform expression patterns are informative for clinical stage and survival.

**Methods**

Using information theory and machine learning methods, we integrated RNA sequencing and clinical data from The Cancer Genome Atlas project to perform the first systematic analysis of the predictive potential of transcript relative abundances for stage and prognosis in 12 solid tumors. Additionally, we built predictive models for breast tumors with ER positive and negative status and for melanoma tumors with proliferative and invasive phenotypes.

**Results**

Tumor-specific models based of transcript isoforms accurately separate early from late stage and metastatic from non-metastatic tumors, and are predictive of survival in patients with undetermined stage. These models show comparable, sometimes better, accuracies compared with models based on gene expression or alternative splicing events, do not correlate with stromal or immune cell content of the samples, and indicate possible functional alterations in the involved genes. Additionally, we describe the transcriptome differences in breast tumors according to estrogen receptor status, and in melanoma tumors according to invasive or proliferative phenotypes, and derive accurate predictive models of stage and survival for each subtype.

**Conclusions**

Our analyses reveal new signatures that characterize tumor phenotypes and their progression independently of gene expression. Knowledge about the relative abundance of transcript isoforms in tumors can help predicting stage and clinical outcome, and thereby contribute towards current strategies of precision cancer medicine.

# Introduction

Tumors advance through stages that are generally characterized by their size and spread to lymph nodes and other parts of the body [1]. Establishing the stage of a tumor is critical to determine patient prognosis and to select the appropriate therapeutic strategy [2]. Even though stage is generally defined from a number of tests carried out on a patient, the information may sometimes be incomplete or inconclusive. Advances in the molecular characterization of tumors have lead to improvements in stage classification and clinical management of patients [3]. Although tumors originate primarily from genetic lesions, their progression involves other molecular alterations, some of which are related to the activation of specific aggressive phenotypes. These transformations, which may be triggered by environmental pressures, such as hypoxia, inflammation or metabolic stress, facilitate tumor spread and metastasis or lead to stem cell like properties like self-renewal [4,5], and are often reflected in gene expression changes. Accordingly, the development of gene expression signatures has been instrumental to complement and improve staging and prognosis determination [6-9]. Solid tumors present frequent alterations in the relative abundances of transcript isoforms independently of gene expression changes in comparison to normal tissues [10]. However, these, alterations, which could hold relevant novel mechanisms of tumor progression, remain largely unexplored as predictive signatures of tumor stage and survival. We investigated the predictive potential of the relative abundances of transcript isoforms for tumor staging and clinical outcome in 12 different tumor types, integrating RNA sequencing and clinical annotation data for 12 tumor types from The Cancer Genome Atlas (TCGA) project. We applied information theory and machine learning methodologies to build tumor-specific predictive models that accurately separate tumor stages and predict patient survival. Our analyses revealed new signatures that characterize tumor phenotypes and their progression independently of gene expression. Knowledge about the relative abundance of transcript isoforms in tumors can potentially help predicting stage and clinical outcome and contribute towards current molecular strategies in precision cancer medicine.

# Results

### Relative abundances of transcript isoforms are predictive of tumor stage

We considered the TNM and numbered stage (S) annotation for 4339 samples from 12 different tumor types from TCGA and compared samples according to each stage class independently (Additional file 1). For metastasis (M), we compared non-metastatic samples (M0) against metastatic ones (M1), whereas for the tumor size (T), lymph-node involvement (N) and stage (S) annotations, we compared early and late stages (Table 1) (Methods). To obtain transcript isoforms with the best discriminant power between clinical stage groups, we used information-based measures with a subsampling strategy to ensure balanced comparisons (Fig. 1a) (Fig. S1a in Additional file 2) (Additional file 3). Transcript isoforms that discriminate between early and late stages, or between M0 and M1, do not tend to coincide between the different tumor types, suggesting that the phenotypic properties of each tumor across stages is related to distinct molecular transformations. To characterize the functional relevance of the found discriminant isoforms, we performed an enrichment analysis of cancer hallmarks (Methods) (Additional file 4). Isoforms separating early and late stages for all clinical classes across all tumor types show enrichment in MYC targets and genes involved in oxidative phosphorylation (Fig. 1b). On the other hand, combining discriminant isoforms from different stage classes in the same tumor type, only 5 of the 12 tumor types tested have enriched hallmarks (Fig. 1c) and MYC targets appeared enriched in skin cutaneous melanoma (SKCM) and kidney papillary carcinoma (KIRP). Testing discriminant isoforms for each stage class and tumor type independently yielded frequent enrichment of MYC target, oxidative phosphorylation, mTORC signaling, DNA repair and Interferon response (Fig. S1b in Additional file 2). These results indicate that transcripts isoform changes potentially impact specific cancer-relevant pathways and may be frequently driven by MYC activity. We validated the discriminant isoforms for the metastatic state in SKCM by comparing their ΔPSI values with those measured between metastatic (SKMel147) [11] and non-metastatic (Mel505) [12] melanoma cells (Methods). Out of the 958 discriminant isoforms in SKCM, 817 had expression in the cell lines. From these, 504 (61.7%) show a change in the same direction and 253 of them have $|\Delta PSI| > 0.1$ in both comparisons (Fig. S1c in Additional file 2).

To build predictive models of tumor stage, we applied a multivariate feature selection method on the discriminant isoforms to obtain a non-redundant subset of predictive isoforms, which we used to build Logistic Model Trees (LMT) for each tumor type and stage class (Fig. 1a) (Additional file 5). Using cross-validation, the mean accuracy of the models in terms of the area under the ROC curve (AUC) is 0.783 (Fig. 1d), with similar average precision-recall values (Fig. S1d in Additional file 2). T-models show the best accuracies (mean AUC = 0.824), with the models for KIRP, kidney chromophobe (KICH), colon adenocarcinoma (COAD) and neck squamous cell carcinoma (HNSC) being the most accurate (mean AUC > 0.87). KIRP T-model includes an isoform for *PAX6*. Increased inclusion of exon 5 of this gene has been related to neuronal differentiation [13], which we see associated to late T stage (Fig. S2a in Additional file 5). The best N-models correspond to KIRP and prostate adenocarcinoma (PRAD) (mean AUC > 0.89). KIRP N-model includes an isoform in the MAP kinase MKNK1 (Fig. S2a in Additional file 5), suggesting that its alternative splicing may have a similar role to that for *MKNK2* in cancer [14]. PRAD N-model (mean AUC = 0.986) includes an isoform of *IDO1* (Fig. 1e), a gene related to anti tumor defense [15]. The best M-model corresponds to SKCM (mean AUC = 0.93) and includes an isoform change in the transmembrane gene *TM6SF1* (Fig. S2a in Additional file 5) and the tyrosine kinase *SYK* (Fig. 1e). In metastatic melanoma samples, *SYK* shows an increase in the abundance of the long form and a decrease of the short form similarly to what was observed in breast tumors [16]. Finally, the best S-models correspond to COAD, BRCA, KICH and Ovarian serous cystadenocarcinoma (OV) (AUC > 0.9). The S-model for OV includes an isoform in the cancer driver *GAS7* (Fig. 1e), whereas the BRCA S-model includes isoform changes in the transcription factor *PRDM16* and the tyrosine kinase *PTK2B*, which also appears in the BRCA M-model for BRCA. In general, there is no overlap between the different stage models. A notable exception is an isoform of *NSUN7* that appears in all models for KIRC with high PSI values at late-stage and an isoform of *SKA3* that appears in the N, T and S models for KIRP, with low PSI values at late stages (Additional file 6). The low general overlap is consistent with pathological transformations being associated with multiple molecular alterations.

**Transcript isoform changes are predictive of survival**

We hypothesized that if the transcript isoforms captured by our models provide clinically relevant information, we should find worse clinical outcomes for patients predicted to be at

late stage. We thus performed a blind test on those samples that lacked stage annotation, and therefore were not used for building the models, to predict the stage using the model for the corresponding tumor type (Fig. 2a) (Additional file 1). In the labeled patients stage does not always correlate with prognosis (Table 2). Accordingly, we performed the blind test only in tumor types for which we observed survival differences in annotated patients. There were 40 samples from COAD, 116 from lung adenocarcinoma (LUAD) and 80 from BRCA that lacked M annotation. After predicting with the M-model from each tumor type, we obtained a total of 227 patients predicted as M0 and 10 patients predicted as M1. Aggregating patients according to the predicted metastatic class yielded a significant difference in survival between the two groups (p-value = 0.0079) (Fig. 2c). On the other hand, there were 1 sample from COAD, 10 from LUAD, 82 from KIRP, 247 from KIRC and 74 from HNSC without N annotation. After predicting with the N-models from the corresponding tumor types, 356 and 58 patients were predicted as early and late N, respectively. Survival analysis with the aggregated patients yielded a significant difference between the two predicted groups (p-value = 0.013) (Fig. 2d). Finally, for the S stage, we predicted on a set of 91 samples without S annotation (8 from COAD, 18 from BRCA, 47 from HNSC, 11 from KIRP, 4 from LUSC, 2 from THCA, and 1 from LUAD). This resulted in 47 and 44 samples predicted as early and late, respectively, which showed no difference in survival (p-value = 0.479). These results represent an independent validation of our models and provide evidence that the relative abundances of transcripts can be predictive of tumor staging and prognosis.

**Potential functional impact of the isoform changes**

The majority of the isoform changes described do not affect known functional protein domains (Additional file 6) (Methods). Nonetheless, we found a number of potentially interesting alterations. In advanced stages, the zinc finger genes *ZNF772*, *ZNF256* and *ZNF805* potentially lose the Krüppel associated box (KRAB) domain, which is a repression domain generally involved in protein-protein interactions. The exon encoding this domain is skipped at late stages in KIRC (T and S models), HNSC (N and S models) and COAD (T model), respectively (Fig. S2b in Additional file 5). Additionally, *ZNF772* has an interaction with *PEX5*, which plays an essential role in peroxisomal protein import. Interestingly, *PEX5* expression decreases for late stages, further supporting a loss of the interaction. We also identified isoform changes that would affect the interaction domains SH2, commonly found in

proteins involved signal transduction of receptor tyrosine kinases. In particular, an isoform change in the Tyrosine-protein phosphatase non-receptor type 6 (*PTPN6*) in the M-model for SKCM is associated to a gain of an SH2 domain in metastatic samples. Interestingly, described interactors of *PTPN6* have a significant increase in gene expression in metastatic samples, suggesting a gain of function (Fig. S2c in Additional file 5). These and other cases (Additional file 6) indicate that some of the isoform changes may have a functional impact in cells and merit further investigation.

**No relation of isoform signatures with stromal and immune cell content**

To assess the purity of the samples as a potential confounding factor of the derived signatures, we tested the correlation between the PSI of the isoforms in our predictive models against signatures of stromal and immune cell content [17] (Methods). Overall we found low correlation values for the models. All models had mean $|R| < 0.4$ (Pearson) with stromal content, and all except the N-model in BRCA ($R=0.433$) had mean $|R|<0.4$ with immune cell content (Additional file 7) Additionally, most of the individual isoforms have low correlation with stromal or immune cell scores: From the 547 isoforms tested, 98% show a correlation $|R|<0.5$ (Pearson) (95% with $|R| < 0.4$) for both stromal and immune scores. Among the few cases with $|R|>0.5$ there is an isoform of *ENAH* (Fig. S2d in Additional file 5), which is present in the T-models in KIRP and COAD and that is identical to the one previously linked to an invasive phenotype [18]. Recent analyses show that clinical stages of the tumors do not correlated with tumor purity in the TCGA samples [19]. Our analysis further supports those results and indicates that isoform-based models of stage do not reflect stromal or immune cell content.

**No universal transcript isoform signature for tumor staging**

Our results prompt the question of whether there might be a universal signature of stage and survival based on transcript isoform changes. To test this, we grouped all annotated samples from the different tumor types according to the stage class and applied the same analyses as before. We could only build M and S models due to the lack of common isoforms with discriminant power for the other classes (Additional file 8). The average AUC values for M and S models were lower than before, with mean AUC of 0.5 and 0.685, respectively.

Aggregating samples from BRCA, COAD and LUAD, we observed a slight increase in accuracy (mean AUC = 0.702). Similarly, analyzing KIRC, KIRP and KICH samples together, the S-model achieves mean AUC = 0.809. In this case, approximately half of the isoforms were present in the previous models. Finally, analyzing the squamous tumors together (HNSC and LUSC), we derived N and S models with mean AUC = 0.72. For other combinations, we could not find accuracies greater than AUC = 0.5. This indicates that despite some overlapping features across tumor types, there is no common signature for all the tested tumor types.

**Comparison with gene and event based classifiers**

Molecular classifiers of stage proposed before have been based on gene expression [7, 8]. We thus compared the performance of our isoform-based models to models derived from gene expression (Methods). The overall accuracy is similar to the isoform-based models (average AUC values 0.783 and 0.781 for isoforms and genes, respectively) (Fig. 2d) (Additional file 9). Interestingly, isoforms show better mean accuracies for the M-model in LUAD (0.883 vs 0.535) (Fig. 2d upper left panel) and for the N-model in PRAD (0.986 vs 0.839) (Fig. 2d lower left panel), compared to gene models. In contrast, the gene-based S-model for THCA shows higher accuracy (0.529 vs 0.836) (Fig. 2d upper right panel). Gene and isoform based models generally involve different genes with only few exceptions, including the presence of CD72 in SKCM in both gene and isoform M-models, *PTGS2* and *VIPR1* in both models for N and T in THCA, *SLC14A1* in both S-models for COAD, and *DNASE1L3* in both S-models for KICH. As before, we tested whether the gene-based models can be predictive of survival for samples lacking stage annotation (Fig. 2a), and found a significant difference in survival between early and late stages predicted with the gene S-models (p-value = 0.0024) (Fig. 2e), whereas no significant difference was found for the M and N models (p-values = 0.983 and 0.161, respectively). These results show that isoform changes hold information about the tumor transformation that is independent and complementary to gene expression changes.

We also tested whether local alternative splicing events, as opposed to transcript isoform changes, could also be predictive of stage. We applied the same analysis pipeline using PSI values for all events in the same tumor samples (Methods). For most of the stage classes we observed similar or smaller accuracy values for events compared to isoform models (average

AUC 0.617 vs. 0.778, respectively) (Fig. 2d) (Additional file 9). Only 23.5% of the isoforms in models overlap with at least one alternative splicing event from the event-based models: 16.51% overlap with alternative 5'/3' splice-sites, mutually exclusive exons, retained introns or cassette exon events, and 6.54% overlap with alternative first or last exon events. Moreover, 82.39% of isoforms in models overlap with at least one of the pre-calculated alternative splicing event. This indicates that a considerable number of changes in exon-intron structures described by the isoform models that are predictive of tumor stage cannot be captured in terms of simple alternative splicing events.

**Isoform relative abundances as prognostic markers in ER-negative breast tumors**

Molecular subtypes in cancer have implications for prognosis and therapy that go beyond the staging system [6, 20, 21]. This is the case for breast tumors, for which the Estrogen receptor (*ESR1*) status is relevant for therapy and prognosis [3,7]. *ESR1* negative breast tumors have generally worse prognosis, and predictive signatures based on gene expression are generally less accurate than for *ESR1* positive tumors [7]. To test whether isoform-based signatures could be relevant for ER negative tumors, we considered the samples in the bottom (ER-) and top (ER+) 25% of the *ESR1* expression ranking (237 samples each) (Fig. 3a). First, we found 2591 discriminant transcript isoforms between ER+ and ER- subgroups (Fig. 3b) (Additional file 10). To validate these changes we used RNA-Seq data from the knockdown of *ESR1* and control in MCF7 cells [22]. From the 2337 isoforms with expression in the MCF7 experiments, 1123 (48%) show PSI changes in the same direction and 328 of them with $|\Delta PSI| > 0.1$ (Fig. S3a in Additional file 11). A predictive model with 81 of the 2591 discriminant isoforms to differentiate between ER+ and ER- samples yielded a mean AUC of 0.999 (Fig. 3c). Among the largest PSI changes we found an isoform of the MAP kinase *MAP3K7* (Fig. 3d), whose long isoform was linked before to apoptosis [23]. Interestingly, this long isoform is less abundant in ER- compared with ER+ samples.

The differences in their transcriptomes suggest that in the ER+ and ER- subtypes stage may be associated to distinct isoform signatures. To test this, we separated each subtype according to early and late stage (Table 3). ER- samples show significant differences in survival between early and late stages for N (p-value = 0.005) and S (p-value = 0.041) annotations (Fig. S3b and S3c in Additional file 11). Accordingly, a predictive model for stage in ER- samples may be relevant for prognosis. In contrast, ER+ samples do not show any significant

differences. Using our feature selection pipeline, we obtained 456 and 249 isoforms that best discriminate between early and late stages in the ER- and ER+ subsets, respectively (Additional file 10). The isoforms for ER- show enrichment in various cancer hallmarks, including DNA repair, Apoptosis and Epithelial-Mesenchymal transition (Fig. S3d in Additional file 11). In contrast, there were no enriched hallmarks associated to the isoforms in the ER+ subset. We built predictive models of stage for ER+ and ER- independently (Additional file 10). Cross-validation yielded average accuracies of AUC = 0.794 (ER-) and AUC = 0.756 (ER+) (Fig. 3c), with similar values for the precision-recall curves (Fig. S3e in Additional file 11). The S-model for ER- includes a member of the tumor necrosis factor receptor superfamily *TNFRS8* (Fig. 3e). Interestingly, expression of a gene from the same family, *TNFRSF17*, was related before to prognosis in ER- samples [3]. Unlike for the previous models, there were not enough unlabeled samples to perform a blind test. Taken together, these results suggest that RNA processing alterations can be informative of stage and prognosis in *ESR1* negative tumors.

**Isoform relative abundances characterize melanoma phenotypes and survival**

Clinical outcome of skin cutaneous melanoma (SKCM) remains poor due to its high degree of heterogeneity [24]. The microphthalmia-associated transcription factor (*MITF*) presents highly dynamic expression patterns in connection to proliferation and invasion in melanoma, with relevance for prognosis and therapy [25, 26]. Overexpression of *MITF* is connected with a proliferative phenotype, whereas downregulation is associated to an invasive phenotype [27]. This raises the question of whether there are independent isoform signatures related to each subtype. To test this, we considered the top and bottom 25% of melanoma samples according to *MITF* expression and pooled them into the MITF+ and MITF- sets, respectively (96 samples per set) (Fig. 4a). Although the MITF+ and MITF- subsets do not show a significant difference in survival, samples in the top and bottom 10% of the *MITF* expression ranking (36 samples per set) show a significant difference, with *MITF* overexpressed samples showing worse prognosis (p=0.029) (Fig. 4b). Our feature selection strategy (Fig. 1a) yielded 2387 discriminant isoforms between MITF+ and MITF- (Fig. 5c) (Additional file 10). We validated these isoforms by comparing their ΔPSI values with those obtained from the knockdown of *MITF* in melanoma cells compared to controls (Fig. S4a in Additional file 12) [12]. From the total of 2279 discriminant isoforms for which we found expression in the cell

lines, 1050 (46%) show a ΔPSI change in the same direction, with 865 of them having |ΔPSI|>0.1 (Fig 4c). The found discriminant isoforms are enriched for multiple cancer hallmarks, including EMT and the mTOR pathway (Fig. S4b in Additional file 12). To further characterized their differences, we built a predictive model to separate MITF+ and MITF- samples with 72 isoforms, which showed a mean AUC of 0.996 (Methods). This model includes isoforms for the cancer drivers *FOXO3* and *TPM1*, as well as for *RAB27A* (Fig. S4c in Additional file 12), a component of the melanosome that is transcriptionally regulated by *MITF* [28] and that is lowly included in MITF+ samples (Fig. S4d in Additional file 12). On the other hand, the *TPM1* isoform that is highly included in MITF+ is the same one linked before to tumor growth [29].

To test whether the melanoma phenotypes are associated to different molecular transformations during tumor progression, we studied the MITF+ and MITF- sets independently. As there was no staging annotation for melanoma samples, we built predictive models of survival. We selected samples in the top and bottom 40% according to days of survival (36 samples per group) and obtained the isoforms that best separate each survival group within each phenotype. The discriminant isoforms in the invasive phenotype (MITF-) are enriched for multiple cancer hallmarks, whereas the proliferative phenotype (MITF+) shows enrichment only for activation of *KRAS* signaling, which does not appear in the invasive phenotype (Fig. S4b in Additional file 12). To further characterize the melanoma phenotypes, we built predictive models of survival for each subset independently (Additional file 10). Cross-validation yielded for MITF+ (34 isoforms) and MITF- (46 isoforms) accuracies of AUC = 0.854 and 0.896, respectively (Fig. 4d) (Fig. S4e). The MITF+ model includes isoforms of the cancer drivers *CSF3R*, *ARHGEF6*, *ITGB7* and *IKZF1* (Additional file 10). On the other hand, the MITF- model includes isoforms in the cancer drivers *ARHGEF11*, *NCOA2* as well as in the MAP Kinase-Activating Death Domain gene *MADD* (Fig. 4e), for which it was shown that expression of the isoforms that skip exon 16 has anti-apoptotic effects [30]. Consistent with this, the PSI of the *MADD* isoform that skips exon 16 is higher in the group with worse prognosis, suggesting that the anti-apoptotic function of *MADD* is related to worse prognosis in invasive melanoma. Taken together, our results provide evidence of distinct transcript isoform abundance patterns linked to melanoma phenotypes and to phenotype-specific signatures of survival.

# Discussion

We described the first systematic analysis of the predictive potential of transcript relative abundances for stage and clinical outcome in multiple solid tumors. We derived novel molecular signatures of clinical stage for 12 different tumor types that can separate tumors according to stage or metastatic status. Importantly, a blind test on patients with unknown stage can separate patients according to survival. Although a multi-cancer signature of clinical outcome based on gene expression has been proposed [31], our results argue against a generic isoform-based signature for all tumors types. Rather, isoform changes appear linked to tumor-type specific processes, with several of them related to MYC activity, in concordance with recent findings [32]. Isoform-based models provide molecular signatures that are independent of gene expression and with comparable accuracies. Moreover, transcript isoforms provide better accuracies than local alternative splicing events and can describe more complex changes in exon-intron structures. We also extracted prognostic signatures for specific tumor subtypes in breast cancer and melanoma. We reported considerable transcript isoform changes between breast tumors according to estrogen receptor expression and between melanoma samples according to *MITF* expression. Moreover, the independent predictive models within each subtype contain isoforms from different genes, thereby highlighting the relevance of determining the transcriptome repertoire in tumor samples to derive accurate molecular signatures of tumor progression.

We observed partial reproducibility of the discriminant isoforms in experiments using cell lines. Transcriptional differences between cell lines and tumor tissues are thought to stem from the loss of the stromal and immune components by cells in culture [33]. We showed that there is low or no correlation between the isoform signatures and the stromal or immune cell content, hence we can discard that the models reflect the composition of stromal and immune cells in the tissue samples. It could be possible that part of the signatures reflect some of the interaction of cells with their environment in tumor tissues, which would be undetectable in cell lines. Our results thus support the notion that phenotypic states of tumor cells, like invasiveness, may be reflected on the relative abundance of transcript isoforms, some of which may be triggered by external cues, such as inflammation or metabolic stress [34]. On the other hand, the observed commonalities between tumor cells and tissues suggest that some of these alterations could be investigated further in cell lines.

It remains to be tested the clinical validity of our findings. Although we have shown that predicted late stages are associated with worse prognosis, it is not conclusive whether the proposed molecular signatures would actually improve current methodologies of stage determination. Our results indicate that isoform-based M and N models are generally accurate and often better than using gene expression. Those models may be especially useful, as they would indicate a metastasis or lymph node invasion before it is visible by other means. To test this, more validations on independent cohorts would be necessary. However, further studies are currently hampered by the scarcity of large enough datasets with clinical annotation comparable to TCGA [35]. We anticipate that transcript isoforms will be relevant to understand the progression of tumors beyond DNA and gene expression alterations and represents a potentially useful molecular tool to predict stage and clinical outcome, thereby complementing current molecular approaches in precision cancer medicine.

# Methods

### Datasets

Processed RNA sequencing data from The Cancer Genome Atlas (TCGA) (https://tcga-data.nci.nih.gov/tcga/) was compiled for 12 different tumor types: breast carcinoma (BRCA), colon adenocarcinoma (COAD), head and neck squamous cell carcinoma (HNSC), kidney chromophobe carcinoma (KICH), kidney renal clear cell carcinoma (KIRC), Kidney renal papillary carcinoma (KIRP), lung adenocarcinoma (LUAD), lung squamous cell carcinoma (LUSC), prostate adenocarcinoma (PRAD), skin cutaneous melanoma (SKCM), thyroid carcinoma (THCA) and ovarian carcinoma (OV). The abundance of every transcript per sample was calculated in transcripts per million (TPM) from the transcript-estimated reads counts and the isoform lengths. Genes were defined to be a set of transcripts that overlap in the same genomic locus and strand and share at least one splice-site (Additional file 1). A gene TPM was defined as the sum of TPMs for all transcripts in the gene. The relative abundance of each isoform (PSI), was calculated by normalizing the isoform TPM to the gene TPM. Only genes with a minimum TPM of 0.1 were considered. Additionally, we used RNA-Seq data from the knockdown of *ESR1* and controls in MCF7 cells (GSE53533) [22], from metastatic melanoma cells (SKMel147) and melanocytes (GSE68221) [11], and from the knockdown of *MITF* and controls using non-metastatic melanoma cells (Mel505)

(GSE61967) [12]. For each sample, transcript abundances were calculated with Sailfish [36]. Relative abundances (PSI) of transcripts were calculated as above and the ΔPSI values between conditions were calculated as the difference between conditions of the mean values from the replicates. Alternative splicing events and their PSI values were obtained from [37].

**Clinical data**

Clinical stage and survival information for patients was obtained from TGCA. We used the available annotation for the TNM staging system (www.cancerstaging.org/), where T followed by a number (1–4) describes the size of the tumor; N followed by a number (1–3) describes spread to lymph nodes according to number and distance; and M followed by 1 or 0 indicates whether the tumor has metastasized or not, respectively. We also considered the numbered stage annotation (S), which goes from 0 to 4, with each number corresponding approximately to a combination of the TNM numbers. When any of the stages were subdivided, only the label of the common class was included (e.g. T1a, T1b and T1c were considered as T1). Only patients with defined stage were used to build the predictive models.

**Selection of relevant features**

Discriminant isoforms, genes or events were calculated as follows. Only isoforms and events with a difference in mean relative abundance (PSI) of at least 0.1 in absolute value between two sample subgroups (e.g. stage groups or tumor subtypes) were kept. For genes, those with log-fold change of the mean gene TPM values between the two groups greater than 2 were kept. Next, a subsampling approach was used to compare the two clinical groups through 100 iterations, by extracting the same number of samples from each group randomly from the input dataset, using a minimum of 10 samples per group. For pooled tumor types, the same number of samples per tumor type was selected at each iteration step. Three different univariate discriminant measures (see below) were applied to each isoform. Additionally, at each iteration step a permutation of the group labels was performed and the univariate measures re-calculated. At the end of the 100 iterations, and for each univariate measure, two distributions of 100 points each are produced for each transcript, corresponding to the observed and expected values. Transcripts with a positive difference of the means of the two

distributions for all three measures were considered discriminant and were kept for further analysis. Three different information-based measures were applied in the subsampling: information gain (*IG*), gain ratio (*GR*) and symmetrical uncertainty (*SU*). *IG* is defined as the mutual information between the group labels of the training set *S* and the values of a feature (or attribute) A, e.g. an isoform: $IG(S,A) = MI(S,A) = H(S)-S(S|A)$, where $H(S)$ is Shannon's entropy according to the two sample classes, and $H(S|A)$ is the conditional entropy of *S* with respect to the attribute *A*. In our case, the group labels are the clinical stages (early, late) or survival groups (low, high), and the attribute values are the discretized isoform PSI values. *GR* is the mutual information of the group labels and the attribute, normalized by the entropy contribution from the proportions of the samples according to the partitioning by the attribute: $GR(S,A) = MI(S,A)/H(A)$. Finally, *SU* provides a symmetric measurement of feature correlation with the labels and it compensates possible biases from the other two measures: $SU(S,A) = 2 \cdot MI(S,A)/( H(S)+H(A) )$ [38]. For the gene expression classifiers, the gene TPM values were used for this analysis, whereas for events, we used the event PSIs as for the isoforms. In all cases, the continuous PSI or TPM values were discretized [39].

**Cancer hallmarks and drivers**

Enrichment analysis of the 50 cancer hallmarks from the Molecular Signatures Database v4.0 [40] was performed with the discriminant isoforms. For each hallmark, a Fisher exact test was performed with the genes with selected isoforms using as controls genes expressed (TPM>0.1) and with multiple transcripts A Benjamini-Hochberg correction was applied and only cases with FDR < 0.05 were kept. Known and predicted cancer drivers were obtained as described in [37].

**Predictive models**

Transcript isoforms that showed a positive difference between the means of the 100 observed and the 100 randomized values for all three univariate measures (*IG, GR, SU*) were analyzed with a Correlation Feature Selection (*CFS*) (Hall 2000). This subselects transcripts with similar discriminating power but lower redundancy among them (Hall 2000), thereby mitigating the problem of overfitting. This was repeated for each comparison between clinical

stages, survival groups, or tumor subtypes. Using the selected transcript isoforms, a Logistic Model Tree (*LMT*) was built with Rweka [41]. *LMTs* are classification trees with logistic regression functions at the leaves. The accuracy of the classifiers was evaluated using the area under the receiver-operating characteristic (ROC) curve or AUC. Additionally, we considered the area under the precision-recall curves (PRC). AUC and PRC take values between 0 (worst prediction) and 1 (best prediction). These values were estimated for each classifier through a 10-fold cross validation, repeated 100 times. The same approach was used for gene and event models.

**Blind tests**

For each samples without stage annotation, which was not used to build the models, we predicted the missing stage, early/late or metastatic/non-metastatic, using the corresponding model for the same tumor type. The survival of these newly predicted samples was then aggregated to test the differences between early and late (or metastatic and non-metastatic) predicted samples. The blind test was performed using only those tumor types that already showed significant differences in the survival between early and late stages for the annotated samples (Table 2). This analysis was not performed for T-models, as all samples had a T annotation.

**Survival analysis**

Survival curves were calculated with the Kaplan-Meier method and compared between patient subsets using a Cox proportional hazards regression model [42]. Survival was measured as date of death minus collection date for deceased patients and as last contact date minus collection date for the other patients.

**Stromal and Immune cell content analysis**

To estimate a stromal and immune signature for a set of samples from a tumor type, we collected a list of stromal and immune signature genes based on [17]. We transformed the RSEM read counts of these two gene lists into a gene set score using GSVA [43] for each sample. Using the resulting scores per sample, we then calculated the Pearson correlations of

the stromal and immune GSVA scores with the isoform PSIs using all tumor samples, including intermediate stages.

**Domain analysis**

To assess whether isoform changes could potentially affect any structural features of the encoded proteins we predicted Prosite patterns (version 20.111, Feb-2015) [44] and Pfam domains [45] using pfscan and InterProScan 5.8-49.0 [46], respectively. We then compared the predicted features on the transcript from our stage models with the predicted features on the transcript from the same gene that undergoes the largest PSI change on the opposite direction.

## List of abbreviations used

PSI: percent/proportion spliced in, IG: information gain, GR: gain ration, SU: symmetrical uncertainty, CFS: correlation feature selection, ER: estrogen receptor, LMT: logistic model tree, ROC: receiver operating characteristic, AUC: area under de ROC curve, PRC: area under the precision-recall curve.

## Authors' contributions

EE proposed and supervised the study, JLT carried out work. AP and ES contributed with some of the software components and processed datasets. JLT and EE wrote the paper with inputs from ES. All authors read and approved the manuscript.

## Description of additional data files

The following additional data are available with the online version of this paper:

Additional file 1. Patient data, blind predictions and gene-isoform annotations.

Additional file 2. Figure S1.

Additional file 3. Discriminant transcript isoforms between stages and patient groups

Additional file 4. Enriched cancer hallmarks

Additional file 5. Figure S2.

Additional file 6. Predictive models of stage, affected domains and overlaps between models

Additional file 7. Correlation of predictive isoforms with stromal and immune scores.

Additional file 8. Predictive models of stage in pooled tumor types

Additional file 9. Predictive models of stage based on gene expression and splicing events

Additional file 10. Models in ER+/ER- breast and MITF+/MITF- melanoma tumors.

Additional file 11. Figure S3.

Additional file 12. Figure S4.

## Competing interests

The authors declare no competing interests.

## Acknowledgements

We would like to thank V. Moreno, J. Yokota, and R. Pio for useful discussions and A. Rubio and L. Montuenga for comments on the manuscript. This work was supported by grants BIO2014-52566-R and Consolider RNAREG (CSD2009-00080) from the MINECO (Spanish Government), by AGAUR (2014-SGR1121) and by the Sandra Ibarra Foundation for Cancer (FSI2013).

# References


1. Sobin, L. H (2003). TNM: evolution and relation to other prognostic factors. Semin. Surg. Oncol. 21, 3–7.
2. Dancey JE, Bedard PL, Onetto N, Hudson TJ. The genetic basis for cancer treatment decisions. Cell. 2012 Feb 3;148(3):409-20.
3. Teschendorff AE, Miremadi A, Pinder SE, Ellis IO, Caldas C. An immune response gene expression module identifies a good prognosis subtype in estrogen receptor negative breast cancer. Genome Biol. 2007;8(8):R157.
4. Ottewell PD, O'Donnell L, Holen I. Molecular alterations that drive breast cancer metastasis to bone. Bonekey Rep. 2015 Mar 18;4:643
5. Suvà ML, Riggi N, Bernstein BE. Epigenetic reprogramming in cancer. Science. 2013 Mar 29;339(6127):1567-70
6. Eschrich S, Yang I, Bloom G, Kwong KY, Boulware D, Cantor A, Coppola D, Kruhøffer M, Aaltonen L, Orntoft TF, Quackenbush J, Yeatman TJ. Molecular staging for survival prediction of colorectal cancer patients. J Clin Oncol. 2005 May 20;23(15):3526-35.
7. Taherian-Fard A, Srihari S, Ragan MA. (2015). Breast cancer classification: linking molecular mechanisms to disease prognosis. Brief Bioinform. 16(3):461-74.
8. Okayama H, Schetter AJ, Ishigame T, Robles AI, Kohno T, Yokota J, Takenoshita S, Harris CC. The expression of four genes as a prognostic classifier for stage I lung adenocarcinoma in 12 independent cohorts. Cancer Epidemiol Biomarkers Prev. 2014 Dec;23(12):2884-94.
9. Sadanandam A, Lyssiotis CA, Homicsko K, Collisson EA, Gibb WJ, Wullschleger S, Ostos LC, Lannon WA, Grotzinger C, Del Rio M, Lhermitte B, Olshen AB, Wiedenmann B, Cantley LC, Gray JW, Hanahan D (2013). A colorectal cancer classification system that associates cellular phenotype and responses to therapy. Nat Med. 19(5):619-25.
10. Sebestyén E, Zawisza M, Eyras E (2015). Detection of recurrent alternative splicing switches in tumor samples reveals novel signatures of cancer. Nucleic Acids Res. 43(3):1345-56.
11. Vardabasso C, Gaspar-Maia A, Hasson D, Pünzeler S et al. (2015). Histone Variant H2A.Z.2 Mediates Proliferation and Drug Sensitivity of Malignant Melanoma. *Mol Cell* 59(1):75-88.

12. Laurette P, Strub T, Koludrovic D, Keime C, Le Gras S, Seberg H, Van Otterloo E, Imrichova H, Siddaway R, Aerts S, Cornell RA, Mengus G, Davidson I (2015). Transcription factor MITF and remodeller BRG1 define chromatin organisation at regulatory elements in melanoma cells. Elife 2015 Mar 24;4.
13. Azuma N, Tadokoro K, Asaka A, Yamada M, Yamaguchi Y, Handa H, Matsushima S, Watanabe T, Kohsaka S, Kida Y, Shiraishi T, Ogura T, Shimamura K, Nakafuku M. (2005) The Pax6 isoform bearing an alternative spliced exon promotes the development of the neural retinal structure. Hum Mol Genet. 14(6):735-45.
14. Adesso L, Calabretta S, Barbagallo F, Capurso G, Pilozzi E, Geremia R, Delle Fave G, Sette C. (2013). Gemcitabine triggers a pro-survival response in pancreatic cancer cells through activation of the MNK2/eIF4E pathway. Oncogene 32(23):2848-57.
15. Metz, R, Smith C, DuHadaway JB, Chandler P, Baban B, Merlo LM, Pigott E, Keough MP, Rust S, Mellor AL, Mandik-Nayak L, Muller AJ, Prendergast GC (2014). IDO2 is critical for IDO1-mediated T-cell regulation and exerts a non-redundant function in inflammation. *International immunology* 26(7):357-67.
16. Prinos, P., Garneau, D., Lucier, J.-F., Gendron, D., Couture, S., Boivin, M., … Elela, S. A. (2011). Alternative splicing of SYK regulates mitosis and cell survival. *Nature Structural & Molecular Biology*, *18*(6), 673–9.
17. Yoshihara K, Shahmoradgoli M, Martínez E, Vegesna R, Kim H, Torres-Garcia W, Treviño V, Shen H, Laird PW, Levine DA, Carter SL, Getz G, Stemke-Hale K, Mills GB, Verhaak RG. Inferring tumour purity and stromal and immune cell admixture from expression data. Nat Commun. 2013;4:2612.
18. Pino MS, Balsamo M, Di Modugno F, Mottolese M, Alessio M, et al. (2008) Human Mena+11a isoform serves as a marker of epithelial phenotype and sensitivity to epidermal growth factor receptor inhibition in human pancreatic cancer cell lines. Clin Cancer Res 14: 4943–4950.
19. Aran D, Sirota M, Butte AJ. Systematic pan-cancer analysis of tumour purity. Nat Commun. 2015 Dec 4;6:8971. doi: 10.1038/ncomms9971.
20. De Sousa E Melo F, Wang X, Jansen M, Fessler E, Trinh A, de Rooij LP, de Jong JH, de Boer OJ, van Leersum R, Bijlsma MF, Rodermond H, van der Heijden M, van Noesel CJ, Tuynman JB, Dekker E, Markowetz F, Medema JP, Vermeulen L. Poor-prognosis colon cancer is defined by a molecularly distinct subtype and develops from serrated precursor lesions. Nat Med. 2013 May;19(5):614-8.

21. Choi W, Porten S, Kim S, Willis D, Plimack ER, Hoffman-Censits J, Roth B, Cheng T, Tran M, Lee IL, Melquist J, Bondaruk J, Majewski T, Zhang S, Pretzsch S, Baggerly K, Siefker-Radtke A, Czerniak B, Dinney CP, McConkey DJ (2014). Identification of distinct basal and luminal subtypes of muscle-invasive bladder cancer with different sensitivities to frontline chemotherapy. Cancer Cell 10;25(2):152-65.
22. Caizzi L, Ferrero G, Cutrupi S, Cordero F, Ballaré C, Miano V, Reineri S, Ricci L, Friard O, Testori A, Corà D, Caselle M, Di Croce L, De Bortoli M (2014). Genome-wide activity of unliganded estrogen receptor-α in breast cancer cells. PNAS 111(13):4892-7.
23. Venables JP, Klinck R, Koh C, Gervais-Bird J, Bramard A, Inkel L, Durand M, Couture S, Froehlich U, Lapointe E, Lucier JF, Thibault P, Rancourt C, Tremblay K, Prinos P, Chabot B, Elela SA (2009). Cancer-associated regulation of alternative splicing. Nat Struct Mol Biol. 16(6):670-6.
24. TCGA. Cancer Genome Atlas Network. Genomic Classification of Cutaneous Melanoma. Cell. 2015 Jun 18;161(7):1681-96.
25. Hoek KS, Goding CR. 2010. Cancer stem cells versus phenotype-switching in melanoma. Pigment Cell & Melanoma Research 23:746–759.
26. Li FZ, Dhillon AS, Anderson RL, McArthur G, Ferrao PT (2015). Phenotype switching in melanoma: implications for progression and therapy. Front Oncol. 5:31.
27. Hoek KS, Eichhoff OM, Schlegel NC, Döbbeling U, Kobert N, Schaerer L, Hemmi S, Dummer R. (2008). In vivo switching of human melanoma cells between proliferative and invasive states. Cancer Res. 2008 Feb 1;68(3):650-6.
28. Chiaverini C, Beuret L, Flori E, Busca R, Abbe P, Bille K, Bahadoran P, Ortonne JP, Bertolotto C, Ballotti R (2008). Microphthalmia-associated transcription factor regulates RAB27A gene expression and controls melanosome transport. J Biol Chem 283, 12635-12642.
29. Babaei-Jadidi R, Li N, Saadeddin A, Spencer-Dene B, Jandke A, Muhammad B, Ibrahim EE, Muraleedharan R, Abuzinadah M, Davis H, Lewis A, Watson S, Behrens A, Tomlinson I, Nateri AS. FBXW7 influences murine intestinal homeostasis and cancer, targeting Notch, Jun, and DEK for degradation. J Exp Med. 2011 Feb 14;208(2):295-312.
30. Lefave CV, Squatrito M, Vorlova S, Rocco GL, Brennan CW, Holland EC, et al. (2011). Splicing factor hnRNPH drives an oncogenic splicing switch in gliomas. EMBO J 2011;30:4084–97.

31. Martinez-Ledesma E, Verhaak RG, Treviño V (2015). Identification of a multi-cancer gene expression biomarker for cancer clinical outcomes using a network-based algorithm. Sci Rep. 5:11966.
32. Koh CM, Bezzi M, Low DH, Ang WX, Teo SX, Gay FP, Al-Haddawi M, Tan SY, Osato M, Sabò A, Amati B, Wee KB, Guccione E. MYC regulates the core pre-mRNA splicing machinery as an essential step in lymphomagenesis. Nature. 2015 Jul 2;523(7558):96-100.
33. Vincent KM, Findlay SD, Postovit LM (2015). Assessing breast cancer cell lines as tumour models by comparison of mRNA expression profiles. Breast Cancer Res. 17:114.
34. Gillies RJ, Verduzco D, Gatenby RA. Evolutionary dynamics of carcinogenesis and why targeted therapy does not work. Nat Rev Cancer. 2012 Jun 14;12(7):487-93.
35. Rubin MA (2015). Health: Make precision medicine work for cancer care. Nature. 520(7547):290-1.
36. Patro R, Mount SM, Kingsford C (2014). Sailfish enables alignment-free isoform quantification from RNA-seq reads using lightweight algorithms. Nat Biotechnol. 32(5):462-4.
37. Sebestyén E, Singh B, Miñana B, Pagès A, Mateo M, Pujana MA, Valcarcel J, Eyras E. (2015b) Large-scale analysis of genome and transcriptome alterations in multiple tumors unveils novel cancer-relevant splicing networks. *bioRxiv* (2015): 023010. http://dx.doi.org/10.1101/023010 (Under revision).
38. Hall M. Correlation-based Feature Selection for Discrete and Numeric Class Machine Learning. 2000 ICML'00 Proceedings of the Seventeenth International Conference of Machine Learning, pages 359-366.
39. Fayyad, U, and Keki I. (1993) "Multi-interval discretization of continuous-valued attributes for classification learning." (1993).Proceedings of the thirteen joint conference of Artificial Intelligence, pages 1022-1027. Morgan Kaufmann
40. Liberzon, A., Birger, C., Thorvaldsdóttir, H., Ghandi, M., Mesirov, J. P., & Tamayo, P. (2015). The Molecular Signatures Database Hallmark Gene Set Collection. *Cell Systems*, *1*(6), 417-425
41. Landwehr, N., Hall, M., & Frank, E. (2005). Logistic Model Trees. *Machine Learning*, *59*(1-2), 161–205.
42. Cox, DR (1972). "Regression Models and Life-Tables". Journal of the Royal Statistical Society, Series B 34 (2): 187–220.
43. Hänzelmann S, Castelo R, Guinney J (2013). GSVA: gene set variation analysis for microarray and RNA-seq data. BMC Bioinformatics 14:7.

44. Sigrist CJ, de Castro E, Cerutti L, Cuche BA, Hulo N, Bridge A, Bougueleret L, Xenarios I. New and continuing developments at PROSITE. Nucleic Acids Res. 2013 Jan;41(Database issue):D344-7.
45. Finn RD, Bateman A, Clements J, Coggill P, Eberhardt RY, Eddy SR, Heger A, Hetherington K, Holm L, Mistry J, Sonnhammer EL, Tate J, Punta M (2014). Pfam: the protein families database. Nucleic Acids Res. 42(Database issue):D222-30.
46. Jones P, Binns D, Chang HY, Fraser M, Li W, McAnulla C, McWilliam H, Maslen J, Mitchell A, Nuka G, Pesseat S, Quinn AF, Sangrador-Vegas A, Scheremetjew M, Yong SY, Lopez R, Hunter S. InterProScan 5: genome-scale protein function classification. Bioinformatics. 2014 May 1;30(9):1236-40.

# Tables

**Table 1. Number of samples analyzed for each tumor type and stage.**

| | | T | | N | | M | | S | |
|---|---|---|---|---|---|---|---|---|---|
| **Tumor type** | **Acronym** | **Early** | **Late** | **Early** | **Late** | **Early** | **Late** | **Early** | **Late** |
| Breast invasive carcinoma | BRCA | 256 (T1) | 147 (T3,T4) | 455 (N0) | 171 (N2,N3) | 836 (M0) | 15 (M1) | 164 (S1) | 15 (S4) |
| Colon adenocarcinoma | COAD | 45 (T1,T2) | 31 (T4) | 149 (N0) | 39 (N2) | 179 (M0) | 33 (M1) | 40 (S1) | 34 (S4) |
| Head and neck squamous cell carcinoma | HNSC | 35 (T1) | 110 (T4) | 166 (N0) | 166 (N2,N3) | | | 77 (S1,S2) | 169 (S4) |
| Kidney chromophobe | KICH | 20 (T1) | 19 (T3,T4) | | | | | 20 (S1) | 19 (S3,S4) |
| Kidney renal clear cell carcinoma | KIRC | 245 (T1) | 186 (T3,T4) | 233 (N0) | 16 (N1) | 419 (M0) | 77 (M1) | 240 (S1) | 78 (S4) |
| Kidney renal papillary carcinoma | KIRP | 71 (T1) | 38 (T3,T4) | 23 (N0) | 16 (N1,N2) | | | 66 (S1) | 38 (S3,S4) |
| Lung squamous cell carcinoma | LUSC | 93 (T1) | 59 (T3,T4) | 242 (N0) | 37 (N2,N3) | | | 195 (S1) | 76 (S3,S4) |
| Lung adenocarcinoma | LUAD | 137 (T1) | 57 (T3,T4) | 281 (N0) | 70 (N2,N3) | 307 (M0) | 22 (M1) | 99 (S1) | 242 (S3,S4) |
| Ovarian serous cystadenocarcinoma | OV | | | | | | | 18 (S2) | 243 (S4) |
| Prostate adenocarcinoma | PRAD | 69 (T2) | 93 (T3,T4) | 129 (N0) | 14 (N1) | | | | |
| Skin cutaneous melanoma | SKCM | | | | | 68 (M0) | 17 (M1) | | |
| Thyroid Carcinoma | THCA | 137 (T1) | 179 (T3,T4) | 220 (N0) | 211 (N1) | | | 270 (S1) | 48 (S4) |

The number of samples used for the comparison early vs late are indicated for each annotation T, N, M, S. Stages I, II, III and IV are indicated as S1, S2, S3 and S4. Comparisons were performed between the earliest and latest available stage groups, with some exceptions for which adjacent stages were added to have enough samples for comparison. Empty cells correspond to cases not tested due to lack of sufficient samples or complete lack of annotation in the samples.

**Table 2. Survival analysis between early and late stage patient groups.**

| Tumor type | T | N | M | S |
|---|---|---|---|---|
| BRCA | P=0.375 | P=0.00012 | P=0.008 | P=0.0007 |
| COAD | P=0.0011 | P=0.011 | P=1.48e-05 | P=0.012 |
| HNSC | P=0.051 | P=0.0137 | | P=2.49e-07 |
| KICH | P=0.00896 | | | P=0.00896 |
| KIRC | P=2e-15 | P=0.0125 | P=0 | P=0 |
| KIRP | P=0.0043 | P=0.005 | | P=8.86e-007 |
| LUSC | P=0.029 | P=0.071 | | P=0.025 |
| LUAD | P=7.02e-09 | P=3.26e-06 | P=0.165 | P=7.02e-09 |
| OV | | | | P=0.0537 |
| PRAD | P=0.456 | P=1 | | |
| SKCM | | | P=0.418 | |
| THCA | P=0.324 | P=0.597 | | P=2.49e-07 |

P-values from the survival test comparing the patient subsets from Table 1. The p-values were obtained using a Cox proportional hazards regression model. Light gray cells indicate comparisons with no significant difference in patient survival. Empty cells correspond to cases not tested due to lack of sufficient samples (see Table 1).

**Table 3. ER-negative (ER-) and ER-positive (ER+) breast tumor subgroups**.

| | T | | N | | S | |
|---|---|---|---|---|---|---|
| **BRCA subtype** | Early | Late | Early | Late | Early | Late |
| **ER-** | 72 (T1) | 48 (T3, T4) | 122 (N0) | 37 (N2, N3) | 48 (S1) | 55 (S3, S4) |
| **ER+** | 54 (T1) | 29 (T3, T4) | 130 (N0) | 36 (N2, N3) | 31 (S1) | 43 (S3, S4) |

The number of samples used for the comparison early vs late are indicated for each annotation T, N ans S. Stages I, II, III and IV are indicated as S1, S2, S3 and S4. In some cases, more than one clinical stage is included in a patient group to have sufficient samples. Due to the insufficient number of annotated samples, it was not possible to build M-models.

# Figure Legends

**Figure 1. (a)** Workflow to obtain discriminant transcript isoforms and predictive models. Given two patient groups, we subsampled two equal sized subsets, one from each group (e.g. metastatic and non-metastatic), which were compared using information-based measures, denoted as $I_{iso}$. At each iteration step, the group labels were randomized to obtain an expected measure, denoted as $I_{rand}$. After 100 iterations, two distributions were produced for each isoform corresponding to observed ($I_{iso}$ ) and expected ($I_{rand}$ ) values. Transcript isoforms with a difference of mean PSI values >0.1 in absolute value between the two patient groups and with a positive difference of the means of the observed and expected distributions for all information-based measures used were then considered as discriminant, which were then used to evaluate enriched cancer hallmarks. Discriminant isoforms were further filtered for redundancy with a Correlation Feature Selection strategy to build a predictive model, which was evaluated using cross-fold validation (Methods). **(b)** Enriched hallmarks in the set of discriminant isoforms for each stage class, metastasis (M), tumor size (T), lymph-node involvement (N) and overall staging (S), using all isoforms selected across all tumor types. **(c)** Enriched hallmarks for each tumor type using all discriminant isoforms selected across all stage classes in each tumor type independently. **(d)** Accuracies of the classifiers for each tumor type for the T, N, M and S annotation, given as the distributions of the areas under the receiving operating characteristic (ROC) curves (AUC). The variation on each bar indicates the minimum and maximum AUC values. Some models are absent due to lack of sufficient samples (Table 1). **(e)** PSI distributions for the transcript isoforms of *IDO1* in PRAD, *SYK* in SKCM and *GAS7* in OV, for the N, M and S models, respectively (Wilcoxon test p-values < 0.001).

**Figure 2. (a)** Illustration of the blind test on unlabeled patients. Patients without annotated stage were predicted using the model of the corresponding tumor type, for each of the stage classes independently. Patients predicted as early or late were collected into two separate groups and tested for differences in survival. This test was performed for each stage class independently and only using tumor types that showed an association between stage and survival in the labeled patients (Table 2). Figures **(b)** and **(c)** show survival (Kaplan-Meyer) plots associated to the test for M- and N- models, respectively. They indicate the survival percentage (y axis) versus survival in months (x axis) based on the predicted stage on the

unannotated samples using the classifier for each corresponding tumor type. The p-value in each plot corresponds to the Cox regression between the two groups and HR indicates the hazards ratio. **(d)** Accuracies of the transcript isoform models (I) compared to the gene (G) and event (E) models. Accuracies are given as boxplots for the distribution of AUC values (y axis) from a 10-fold cross-validation for each tumor type (x axis) for the M, S, N and T models. Tumors for which stage data was missing are not shown (Table 1). **(e)** Survival (Kaplan-Meyer) plot of the early and late stage predictions performed with the gene-based S-models on unannotated samples. The p-value corresponds to the Cox regression between the two groups and HR indicates the hazards ratio.

**Figure 3. (a)** Ranking (x axis) of breast tumor (BRCA) samples according to *ESR1* expression (gene TPM) (y axis). ER+ and ER- subsets were defined as the top and bottom 25% of the ranking, respectively, leaving out samples in the inter-quartile range (IQR). **(b)** Heatmap of PSI values, from 0 (blue) to 1 (yellow), for the top 35 isoforms that separate ER+ and ER- subsets. Isoforms are labeled by gene name (y axis). Samples are clustered according to the PSI values using Euclidean distance and Ward's method. **(c)** Accuracies in terms of AUC values (y axis) from a 10-fold cross-validation for the isoform models for the comparison of ER+ and ER- samples, and for the comparison of early and late N, S and T stages within ER+ or ER- subsets. The variation on each bar indicates the minimum and maximum AUC values. **(d)** PSI distribution of the *MAP3K7* isoform that changes significantly between the ER+ and ER- BRCA sets (Wilcoxon test p-value < 2.2E-16). **(e)** PSI distribution of the isoform in *TNFRS8* that changes between early and late S stage in ER- samples (Wilcoxon test p-value = 0.1046).

**Figure 4. (a)** Ranking (x axis) of melanoma (SKCM) samples according to *MITF* expression (gene TPM) (y axis). We indicate the top and bottom 10% and 25% of the samples used for analyses. **(b)** Survival (Kaplan-Meyer) plot for the top and bottom 10% of the samples according to the ranking of *MITF* expression. The p-value corresponds to the Cox regression between the two groups and HR indicates the hazards ratio. **(c)** Heatmap of PSI values, from 0 (blue) to 1 (yellow), for the top 30 discriminant isoforms according to |ΔPSI| value between the MITF+ and MITF- subgroups. Isoforms are labeled by gene name (y axis). Samples are clustered according to the PSI values using Euclidean distance and Ward's method. **(d)** Accuracy given in terms of the distribution of area under the ROC curve (AUC) values (y

axis) from a 10-fold cross-validation for (from left to right in the x axis) the survival model for MITF+, MITF- as well as for the separation between MITF+ and MITF- subgroups using 25% (Q1 vs Q4) or 10% (D1 vs D10) of the top and bottom samples in the ranking of *MITF* expression. The bars show the minimum, mean and maximum AUC values. **(e)** Distribution of PSI values for the isoform in *MADD* that is predictive of prognosis in the MITF- subgroup (Wilcoxon test p-value = 7.781e-05).

**Figure S1**. **(a)** Information-based feature selection methods provide a robust and conservative measure of the discriminant power of features. The plots compare the information gain (IG) (upper panel), gain ratio (GR) (middle panel) and symmetrical uncertainty (SU) (lower panel) (y axes) with the Wilcoxon-test p-value after multiple-testing correction using Benjamini-Hochberg method (x axes) for the distribution of PSI values for two patient subgroups. In this case, the data corresponds to the multiple subsampling of 20 patients per group, ER-positive and ER-negative, from breast tumor samples. Each dot corresponds to one isoform in each of the subsamples. **(b)** Enriched hallmarks (y axis) for each stage class, metastasis (M), tumor size (T), lymph-node involvement (N) and overall staging (S), in each tumor type (x axis), using all discriminant isoforms found in each case. Only significant cases (corrected Fisher test p-value < 0.05) are shown. **(c)** Comparison of the ΔPSI values for the discriminant isoforms between metastatic and non-metastatic SKMC samples (x axis) with the ΔPSI values from the comparison of the metastatic (SKMel147) and non-metastatic (Mel505) melanoma cells (y axis). We indicate in blue or red those isoforms with the same or opposite change direction, respectively. Dark and light colors indicate |ΔPSI|>0.1 and |ΔPSI|<0.1, respectively. The correlation (Pearson R) is given for isoforms in dark blue. **(d)** Accuracy of the models in terms of the areas under the precision-recall curves (PRC) (y axis) for the late-stage classes (i.e. precision is measured as the proportion of predicted late stage samples that are correctly predicted). The bars show the minimum, mean and maximum values of the area of the precision-recall curves. Some models are absent due to lack of sufficient samples (Table 1).

**Figure S2. (a)** PSI distributions of some of the isoforms in the predictive models. From left to right, *PAX6* isoform in the KIRP T-model for KIRP (Wilcoxon test p-value = 2.695e-06), *MKNK1* isoform in the KIRP N-model (Wilcoxon test p-value = 0.0004), *TM6SF1* isoform in the SKCM M-model (Wilcoxon test p-value = 1.813e-05), *PRDM16* isoform (Wilcoxon test p-value = 0.0001) and *PTKB* isoform (Wilcoxon test p-value = 0.005) in BRCA S-model. The

y-axis indicates the PSI value in each sample separated according to early and late stages (x-axis). **(b)** The isoforms in *ZNF772*, *ZNF256* and *ZNF805* that lacks the KRAB domain increase PSI abundance at late stages in KIRC (T and S models), HNSC (N and S models) and COAD (T model), respectively. The left panel shows the *ZNF256* locus with the annotated domains and local alternative splicing events, as well as the PSI distributions of its two isoforms at early and late stages in HNSC. The right panel shows the expression levels of *PEX5*, interactor of *ZNF772*, in early and late stages in KIRC (Wilcoxon test p-value < 0.01). **(c)** Left panel: The *PTPN6* isoform uc010sfr lacks the exon that encodes the first SH2 domain, whereas uc001qsb includes this exon. In the boxplot we show that the abundance of the isoform uc001qsb decreases at M1. Right panel: Expression levels for 10 interactors of *PTPN6* in M0 and M1 SKCM samples (Wilcoxon p-values < 0.05). **(d)** Left panel: XY-plot of the PSI values (y axis) of the *ENAH* isoform that appears in the T-models of KIRP and COAD, and the stromal score (x axis), across all COAD tumor samples. Pearson correlation with stromal score R=-0.59 and with immune score R=-0.41. Right panel: PSI distribution of the same *ENAH* isoform in early and late T-stages in KIRP and COAD (Wilcoxon test p-value < 0.001).

**Figure S3. (a)** Comparison of the ΔPSI values for the discriminant isoforms between ER+ and ER- samples (x axis) with the ΔPSI values from the comparison of the control and knockdown of *ESR1* in MCF7 cells (y axis). We indicate in blue or red those isoforms with the same or opposite change direction, respectively. Dark and light colors indicate |ΔPSI|>0.1 and |ΔPSI|<0.1, respectively. The correlation (Pearson R) is given for isoforms in dark blue. Plots **(b)** and **(c)** show the survival (Kaplan-Meyer) curves for the ER- samples according to early and late N and S stages, respectively. The p-value in each plot corresponds to the Cox regression between the two groups and HR indicates the hazards ratio. **(d)** Enriched cancer hallmarks for the set of discriminant isoforms between ER+ and ER- subsets (ER+_ER-) and for the set of isoforms separating early and late stages in ER- (ER-). For this latter comparison isoforms associated to N, T and S stages were combined into early and late subgroups. **(e)** Accuracy of the of the models in terms of the areas of the precision-recall curves (PRC) (y axis) for the comparison between ER+ and ER- subgroups and for the comparison of early vs late stage classes in each subtype, ER+ and ER-, for N, S and T annotation. The precision is measured as the proportion of predicted late stage samples or ER+ samples that are correctly predicted. The bars show the minimum, mean and maximum values of the area of the

precision-recall curves.

**Figure S4. (a)** Comparison of the ΔPSI values for the discriminant isoforms between MITF+ and MITF- melanoma tissue samples (x axis) with the ΔPSIs obtained from the comparison of the control and knockdown of *MITF* in Mel505 cells (y axis). We indicate in blue or red those isoforms with the same or opposite change direction, respectively. Dark and light colors indicate |ΔPSI|>0.1 and |ΔPSI|<0.1, respectively. The correlation (Pearson R) is given for isoforms in dark blue. **(b)** Enriched cancer hallmarks (y axis) (corrected Fisher test p-value < 0.05) using the discriminant isoforms in the comparison MITF- vs MITF+ and comparing low and high survival subgroup of patients within each subtype MITF+ or MITF-. Enriched hallmarks were the same using the top and bottom 10% or 25% samples according to *MITF* expression to define the subtypes. **(c)** PSI distributions of the *TPM1* isoform (left panel) and *RAB27A* isoform (right panel) that separate the two melanoma subtypes, MITF+ and MITF- (Wilcoxon test p-values = 5.293e-9 and 4.86e-12, respectively). The plots indicate the PSI values (y-axis) for the isoforms in MITF+ and MITF- samples (x-axis). **(d)** Genomic locus for RAB27A indicating the annotated isoforms; uc002acr.2 decreases PSI in MITF+ (Fig. S4c in Additional file 12), whereas uc002acp.2 increases PSI in MITF+ (Additional file 10). **(e)** Accuracy given in terms of the areas under the precision-recall curves (PRC) (y axis) from a 10-fold cross-validation for (from left to right in the x axis) the survival model for MITF+, MITF- as well as for the separation between MITF+ and MITF- subgroups using 25% (Q1 vs Q4) or 10% (D1 vs D10) of the top and bottom samples in the ranking of *MITF* expression The bars show the minimum, mean and maximum values of the area of the precision-recall curves.

Figure 1

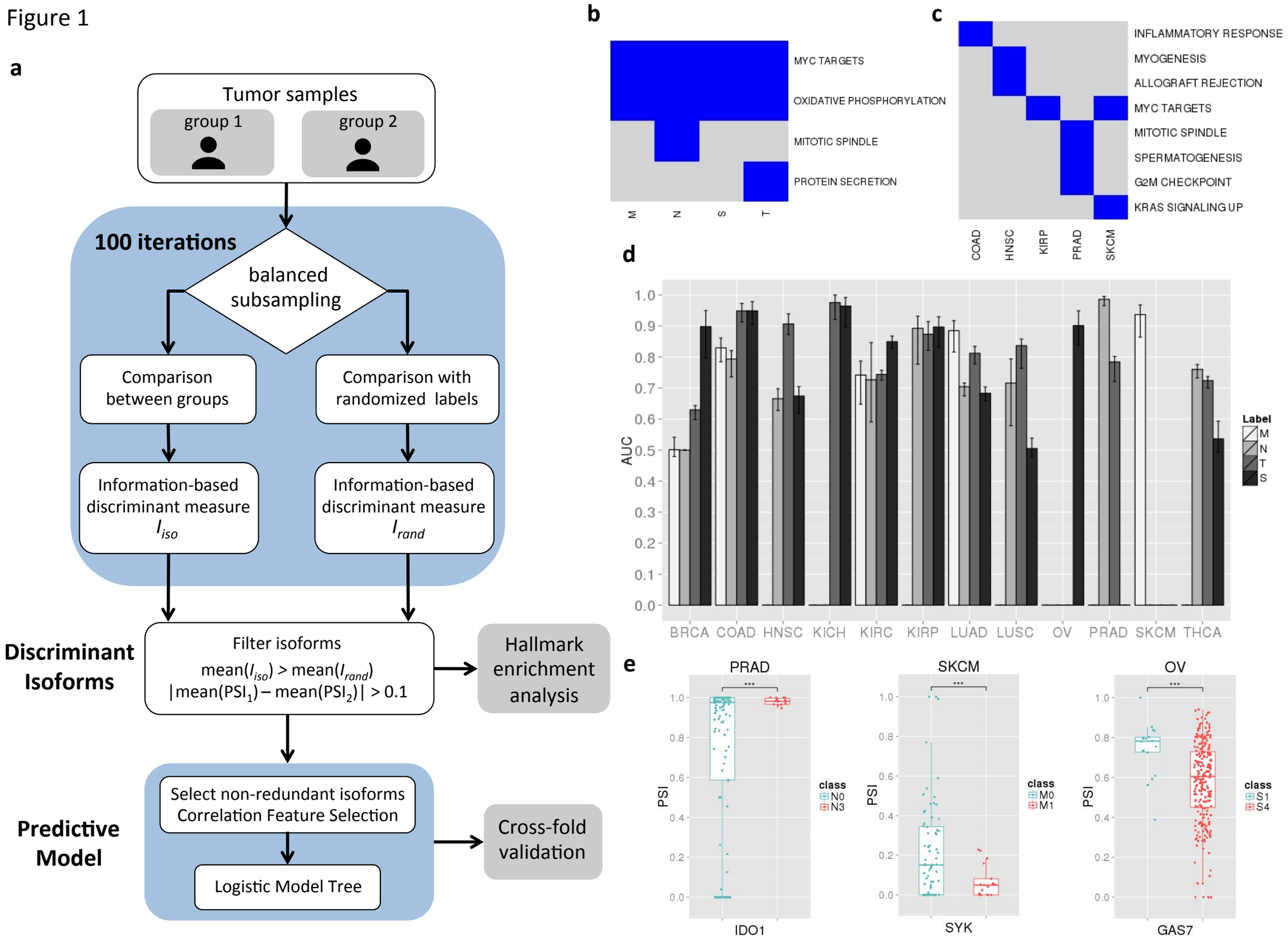

a
Tumor samples
group 1
group 2
100 iterations
balanced subsampling
Comparison between groups
Comparison with randomized labels
Information-based discriminant measure $I_{iso}$
Information-based discriminant measure $I_{rand}$
Discriminant Isoforms
Filter isoforms
$mean(I_{iso}) > mean(I_{rand})$
$|mean(PSI_1) - mean(PSI_2)| > 0.1$
Hallmark enrichment analysis
Predictive Model
Select non-redundant isoforms
Correlation Feature Selection
Logistic Model Tree
Cross-fold validation
b
MYC TARGETS
OXIDATIVE PHOSPHORYLATION
MITOTIC SPINDLE
PROTEIN SECRETION
M
N
S
T
c
INFLAMMATORY RESPONSE
MYOGENESIS
ALLOGRAFT REJECTION
MYC TARGETS
MITOTIC SPINDLE
SPERMATOGENESIS
G2M CHECKPOINT
KRAS SIGNALING UP
COAD
HNSC
KIRP
PRAD
SKCM
d
AUC
Label
M
N
T
S
BRCA
COAD
HNSC
KICH
KIRC
KIRP
LUAD
LUSC
OV
PRAD
SKCM
THCA
e
PRAD
SKCM
OV
PSI
class
N0
N3
M0
M1
S1
S4
IDO1
SYK
GAS7

Figure 2

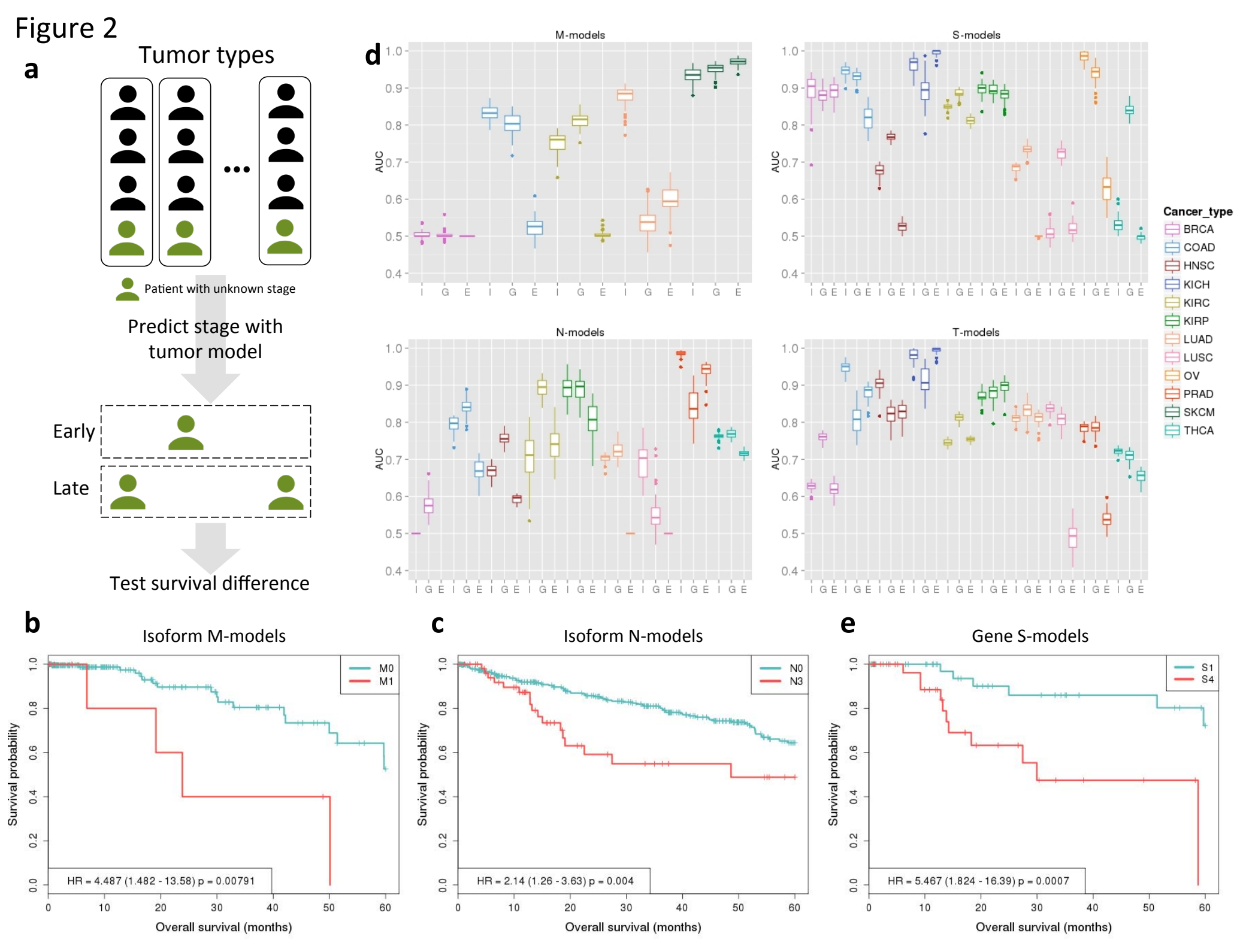

a
Tumor types
Patient with unknown stage
Predict stage with tumor model
Early
Late
Test survival difference
d
M-models
S-models
N-models
T-models
AUC
I G E
Cancer_type
BRCA
COAD
HNSC
KICH
KIRC
KIRP
LUAD
LUSC
OV
PRAD
SKCM
THCA
b
Isoform M-models
M0
M1
HR = 4.487 (1.482 - 13.58) p = 0.00791
c
Isoform N-models
N0
N3
HR = 2.14 (1.26 - 3.63) p = 0.004
e
Gene S-models
S1
S4
HR = 5.467 (1.824 - 16.39) p = 0.0007
Survival probability
Overall survival (months)

Figure 3

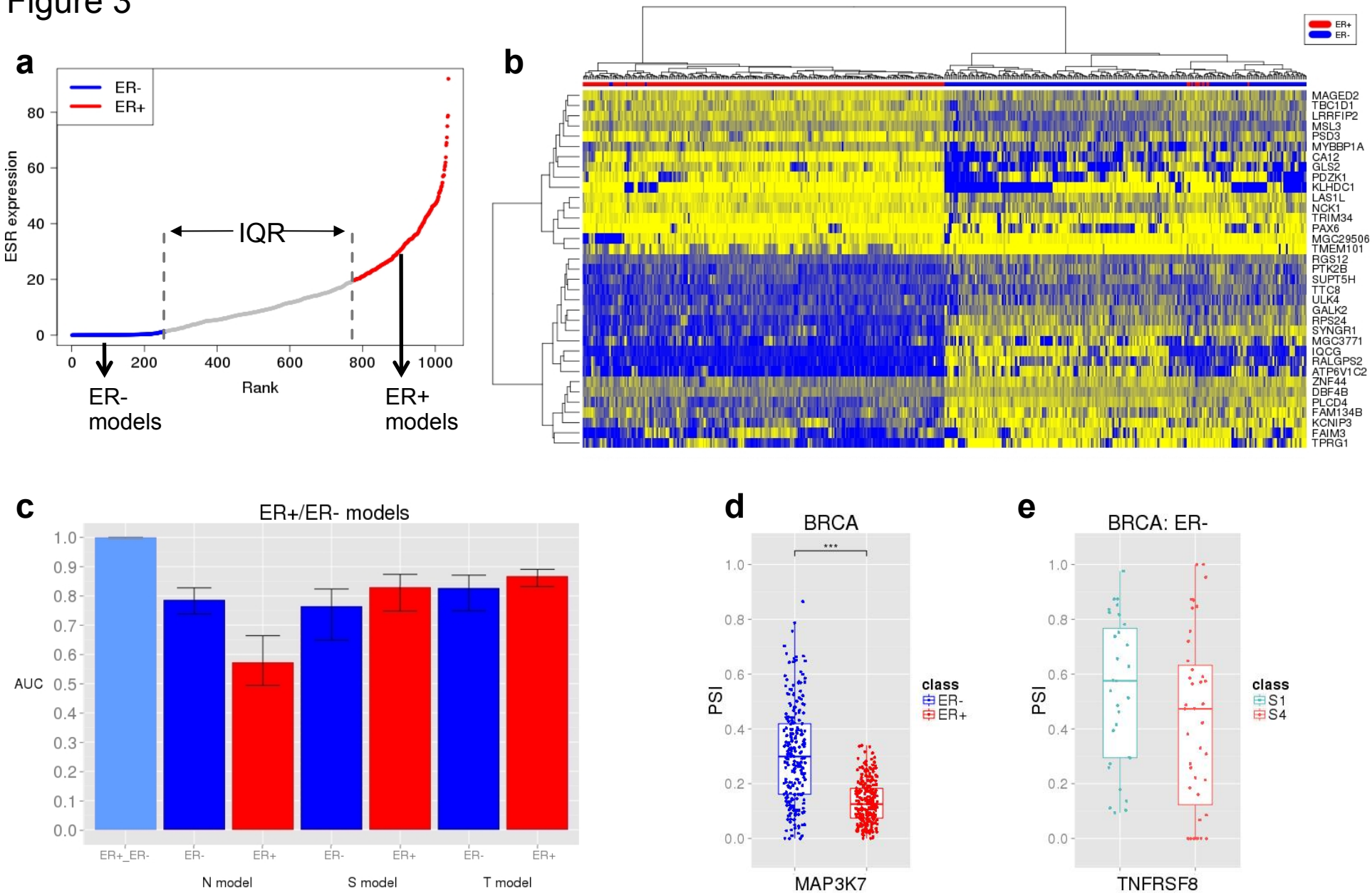
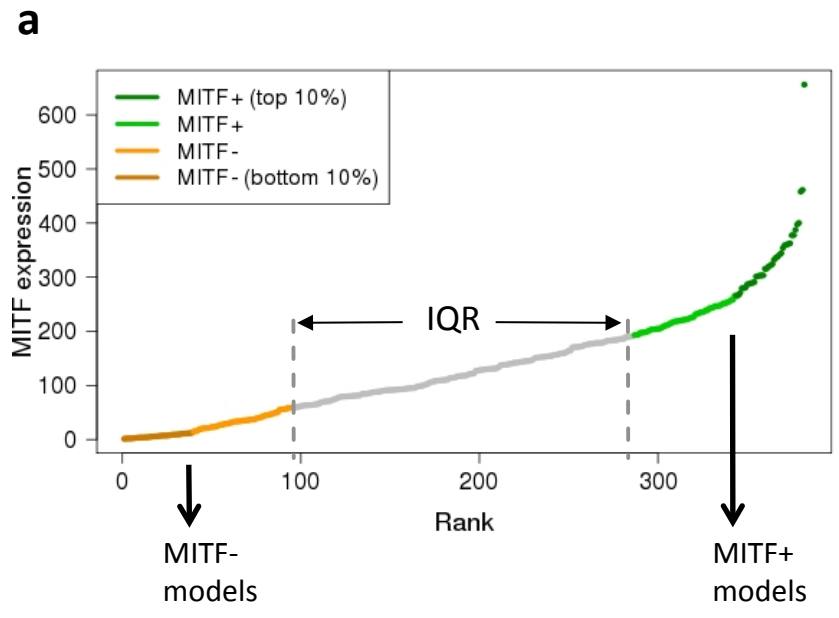
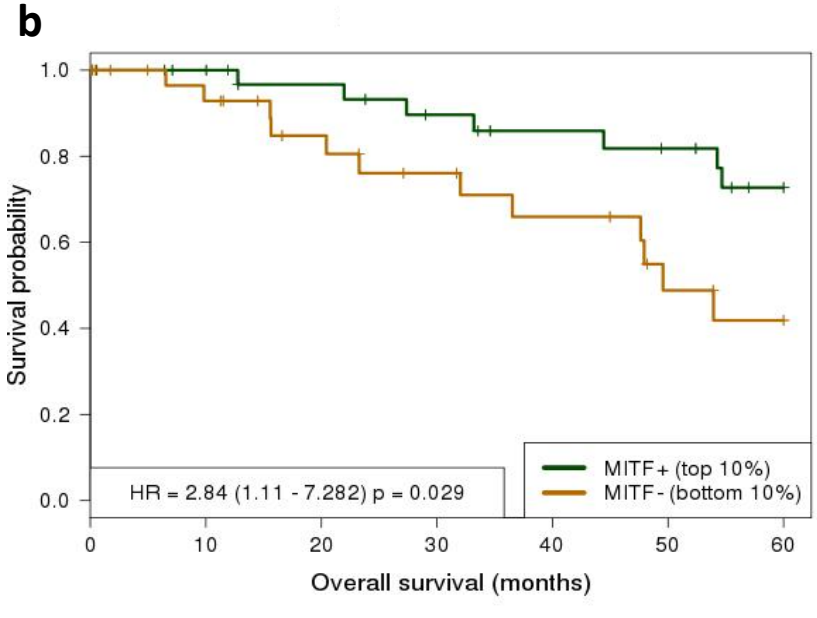
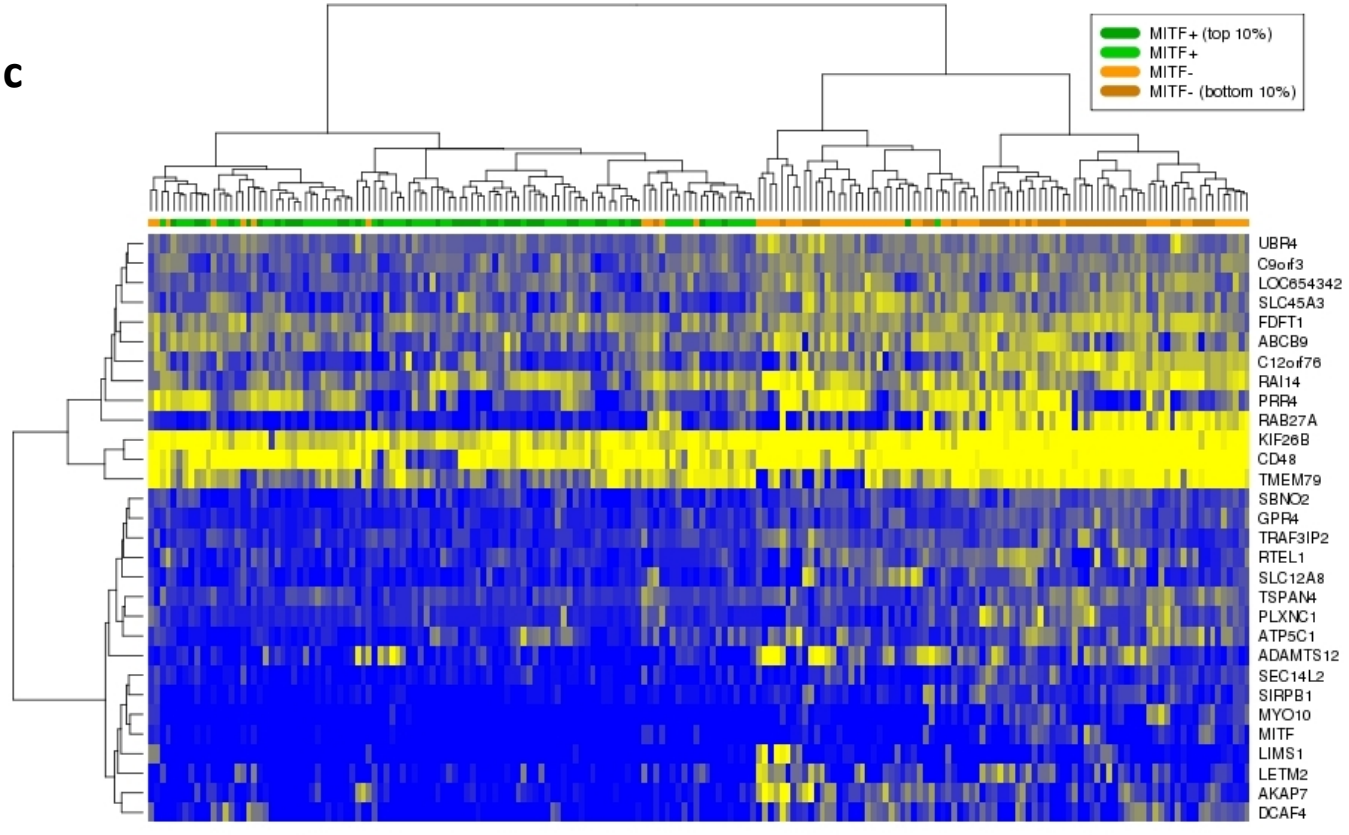
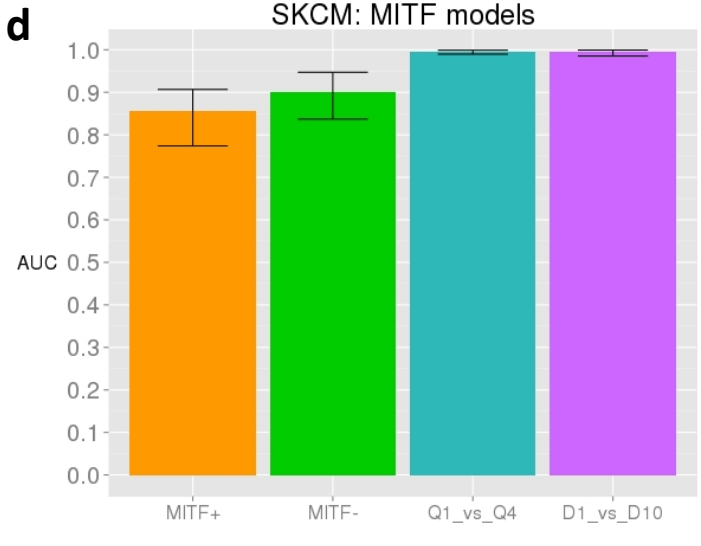

a
ER-
ER+
ESR expression
IQR
Rank
ER-
models
ER+
models
b
ER+
ER-
MAGED2
TBC1D1
LRRFIP2
MSL3
PSD3
MYBBP1A
CA12
GLS2
PDZK1
KLHDC1
LAS1L
NCK1
TRIM34
PAX6
MGC29506
TMEM101
RGS12
PTK2B
SUPT5H
TTC8
ULK4
GALK2
RPS24
SYNGR1
MGC3771
IQCG
RALGPS2
ATP6V1C2
ZNF44
DBF4B
PLCD4
FAM134B
KCNIP3
FAIM3
TPRG1
c
ER+/ER- models
AUC
ER+_ER-
ER-
ER+
ER-
ER+
ER-
ER+
N model
S model
T model
d
BRCA
***
PSI
class
ER-
ER+
MAP3K7
e
BRCA: ER-
PSI
class
S1
S4
TNFRSF8

Figure 4


a
MITF+ (top 10%)
MITF+
MITF-
MITF- (bottom 10%)
MITF expression
IQR
Rank
MITF-
models
MITF+
models


b
Survival probability
HR = 2.84 (1.11 - 7.282) p = 0.029
MITF+ (top 10%)
MITF- (bottom 10%)
Overall survival (months)


c
MITF+ (top 10%)
MITF+
MITF-
MITF- (bottom 10%)
UBR4
C9orf3
LOC654342
SLC45A3
FDFT1
ABCB9
C12orf76
RAI14
PRR4
RAB27A
KIF26B
CD48
TMEM79
SBNO2
GPR4
TRAF3IP2
RTEL1
SLC12A8
TSPAN4
PLXNC1
ATP5C1
ADAMTS12
SEC14L2
SIRPB1
MYO10
MITF
LIMS1
LETM2
AKAP7
DCAF4


d
SKCM: MITF models
AUC
MITF+
MITF-
Q1_vs_Q4
D1_vs_D10

e
SKCM
***
PSI
class
Better_progn
Worse_progn
MADD