# Supplementary Information

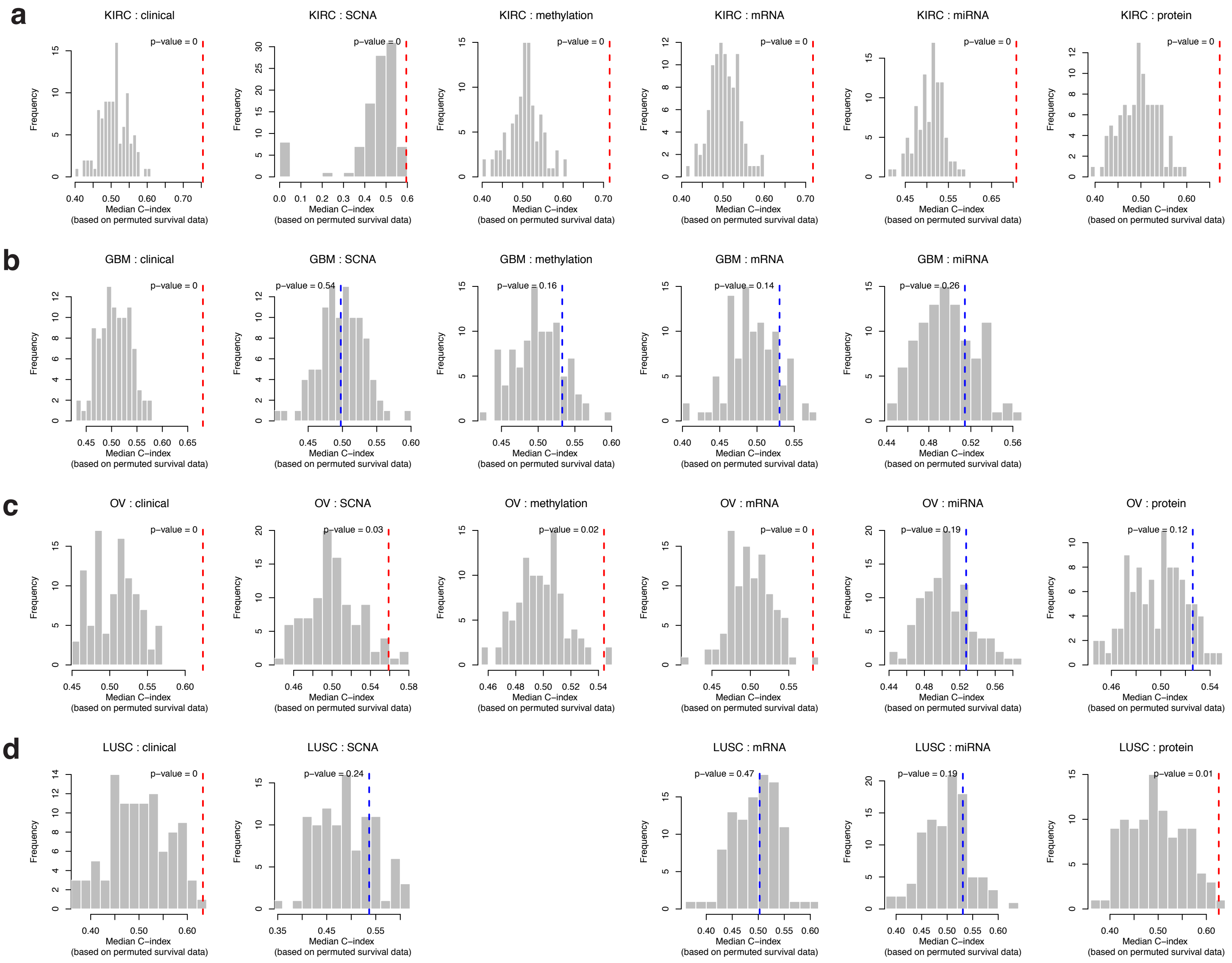
## Assessing the clinical utility of cancer genomic and proteomic data across tumor types

Table of contents:

**Supplementary Figures**

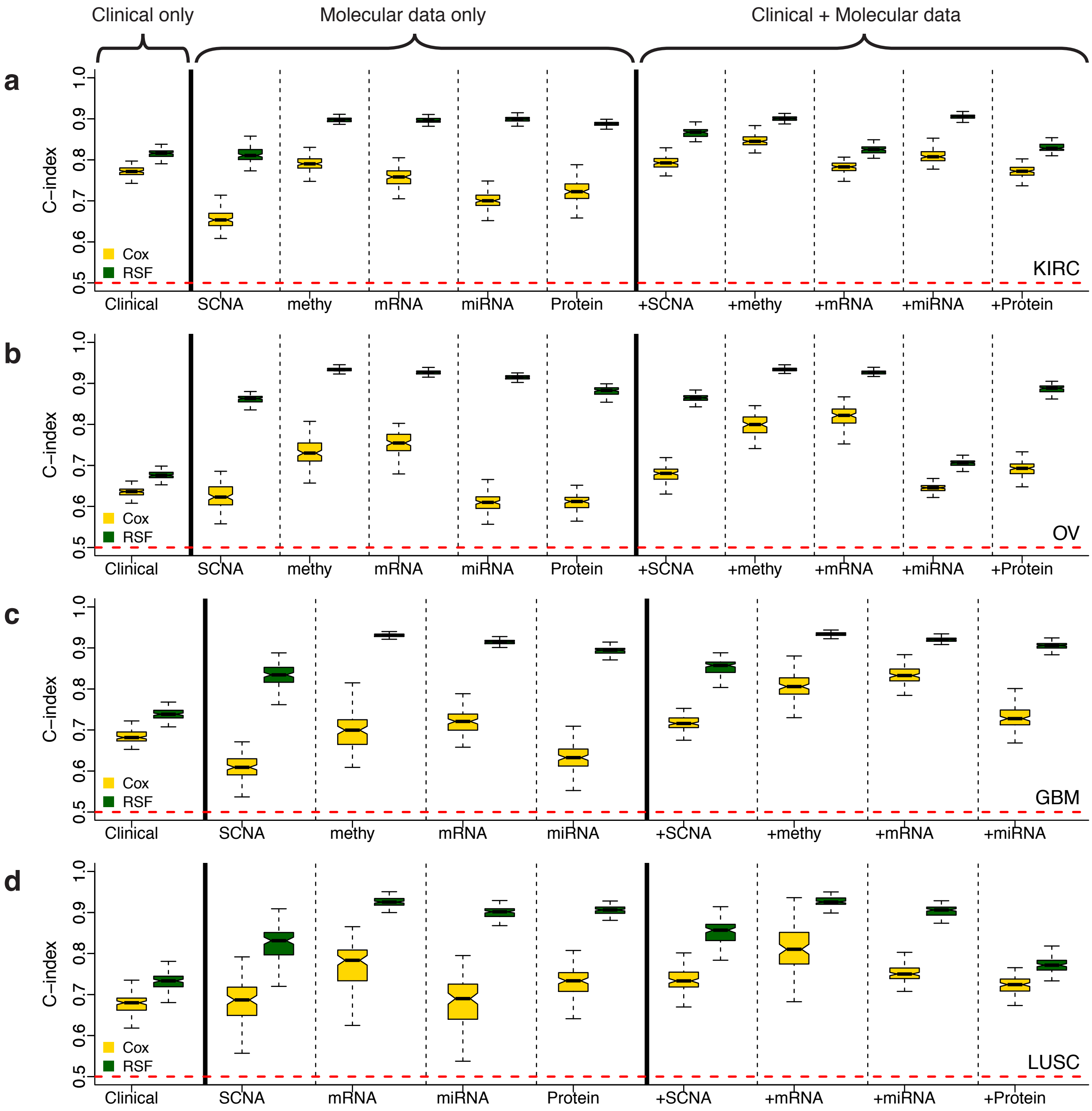
**Supplementary Tables**

# Supplementary Figure 1



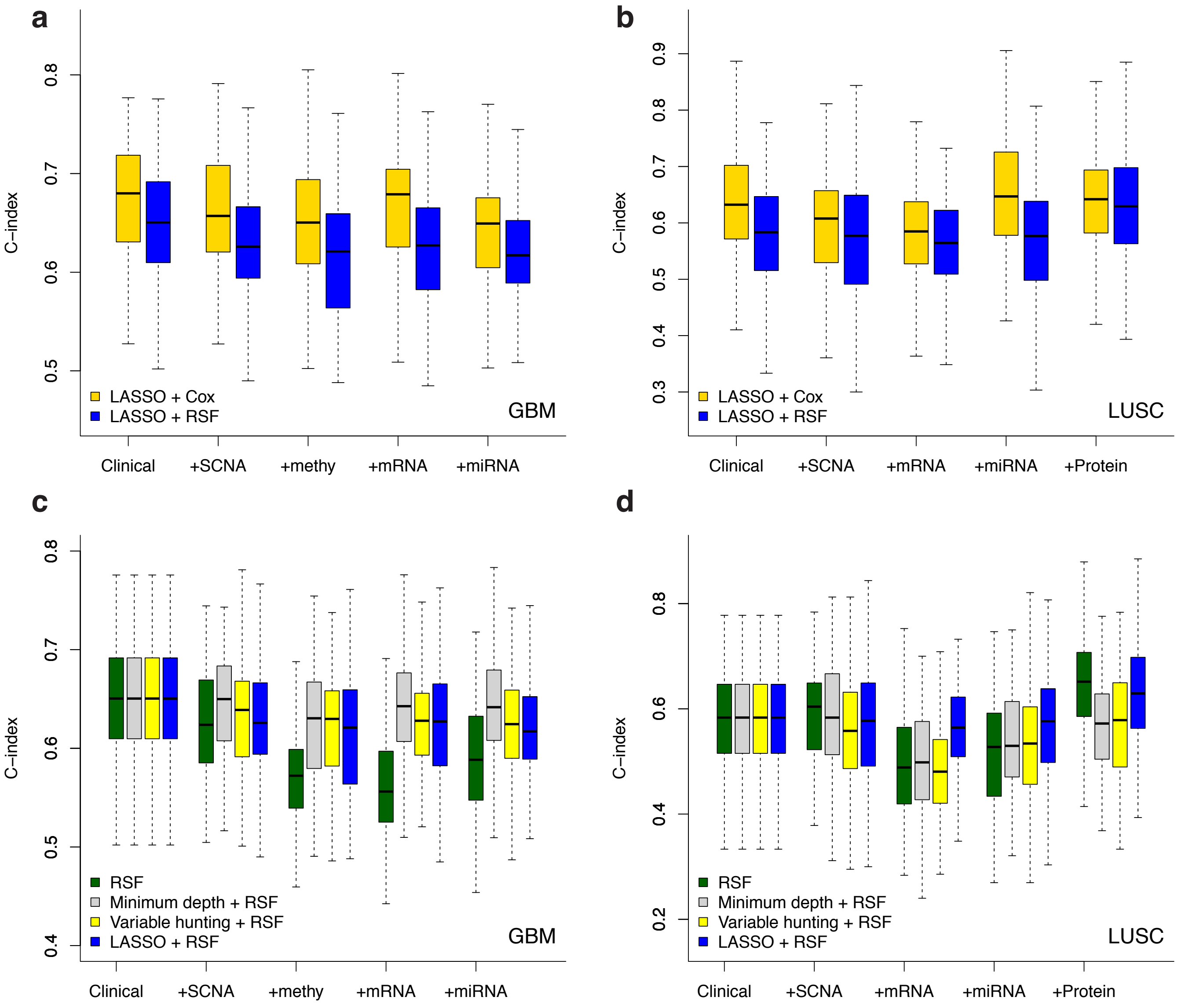
**Model performance assessment of clinical and individual molecular data types using permutation tests.** The histograms show the distributions of the median C-indexes of the 100 survival-permuted data for (a) KIRC, (b) GBM, (c) OV, and (d) LUSC. The median C-index values of the original survival data were marked with the vertical dashed lines. The *P*-values were calculated based on the permutation C-index distributions.

## Supplementary Figure 2



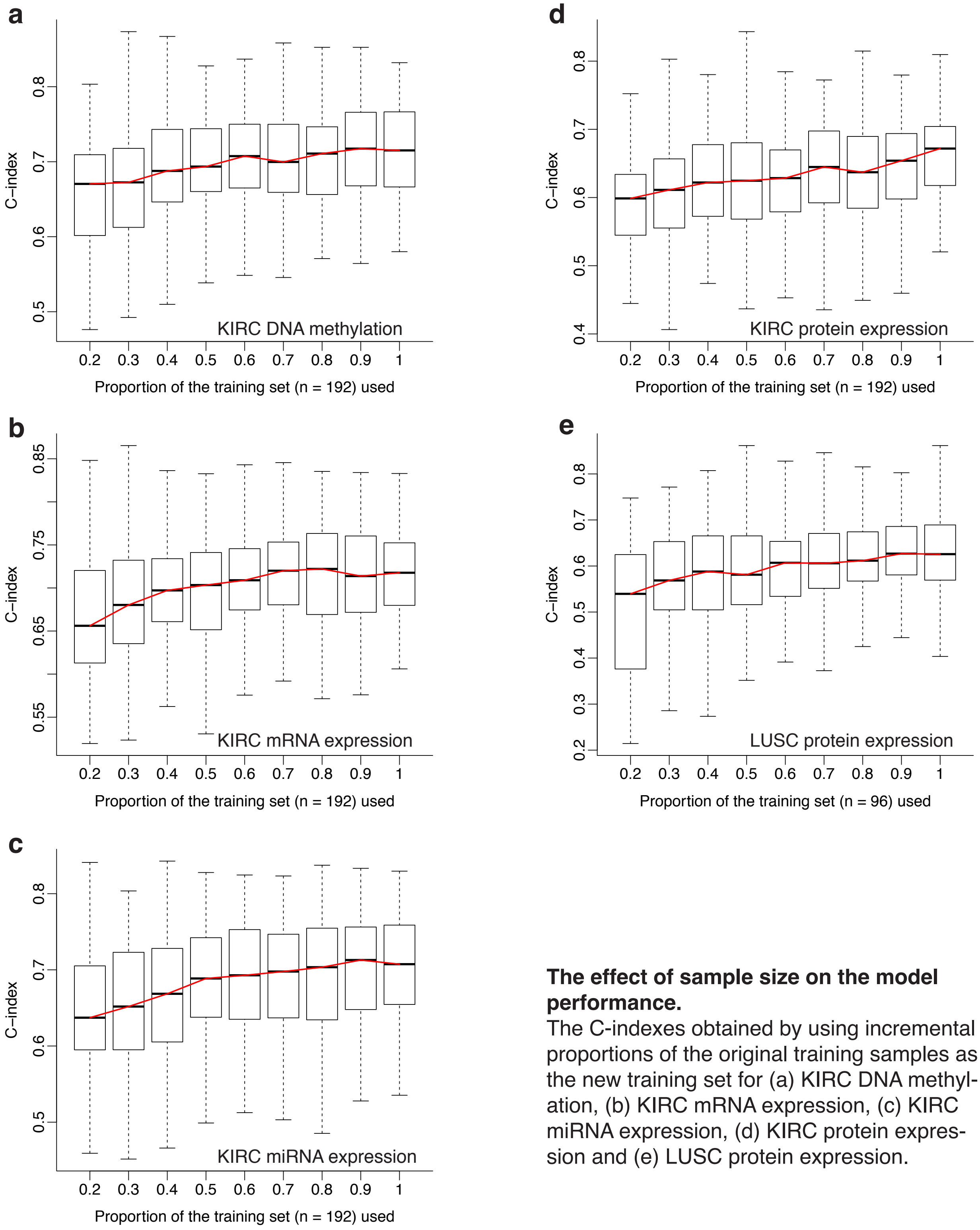
**Comparison of the training performance of clinical variables, molecular data and their combinations by Cox and RSF.**
The training C-indexes by models from clinical variables, individual molecular data alone or in combination with clinical variables in (a) KIRC, (b) GBM, (c) OV and (d) LUSC.

## Supplementary Figure 3



**The effect of learning algorithms on model performance.**
The C-indexes obtained by using the same LASSO approach before Cox and RSF for molecular+clinical data for (a) GBM and (b) LUSC. The C-indexes obtained by using different feature selection methods before RSF for molecular+clinical data for (c) GBM and (d) LUSC.

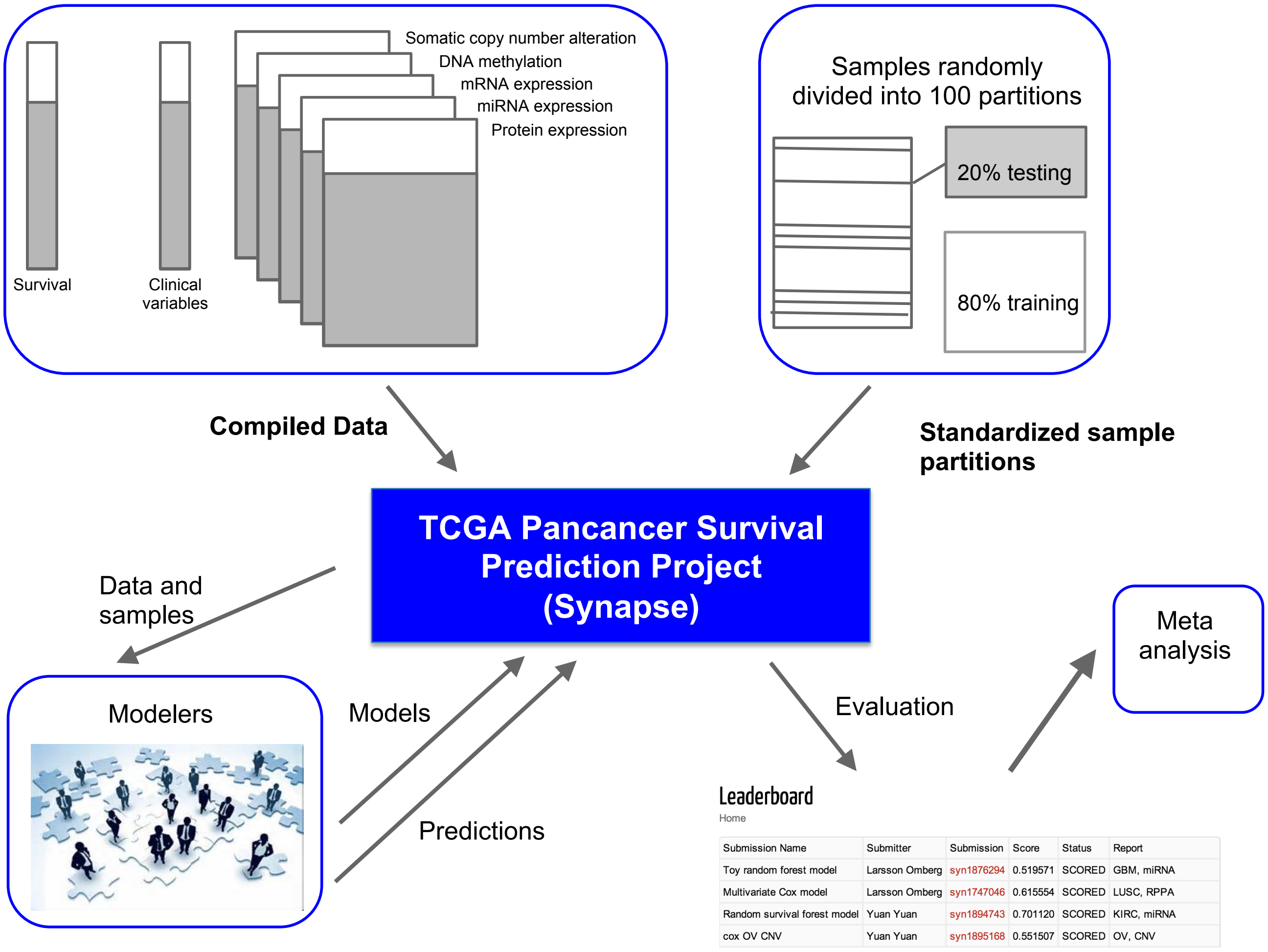
## Supplementary Figure 4



**The effect of sample size on the model performance.**

The C-indexes obtained by using incremental proportions of the original training samples as the new training set for (a) KIRC DNA methylation, (b) KIRC mRNA expression, (c) KIRC miRNA expression, (d) KIRC protein expression and (e) LUSC protein expression.

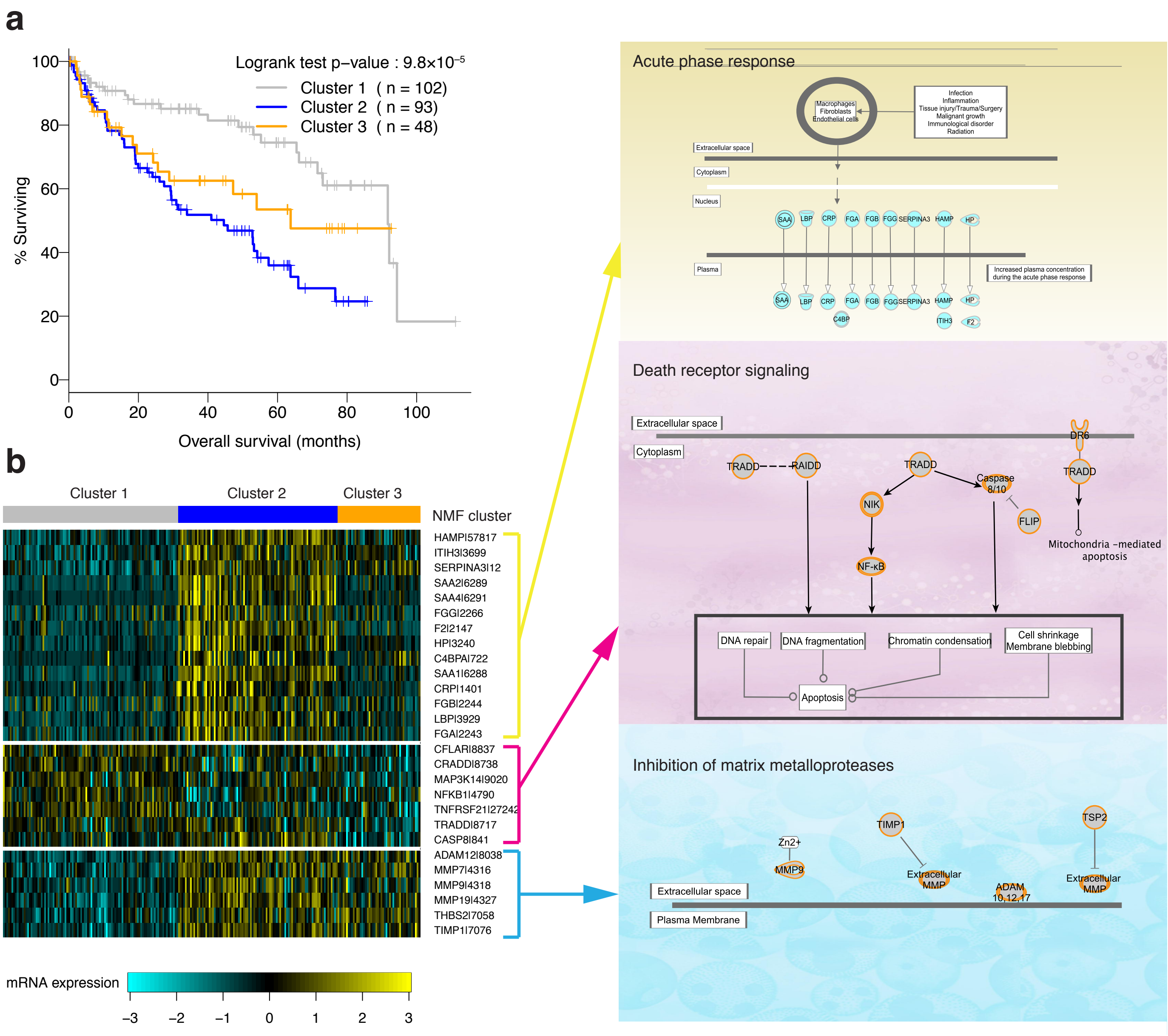
## Supplementary Figure 5



| Submission Name | Submitter | Submission | Score | Status | Report |
|---|---|---|---|---|---|
| Toy random forest model | Larsson Omberg | syn1876294 | 0.519571 | SCORED | GBM, miRNA |
| Multivariate Cox model | Larsson Omberg | syn1747046 | 0.615554 | SCORED | LUSC, RPPA |
| Random survival forest model | Yuan Yuan | syn1894743 | 0.701120 | SCORED | KIRC, miRNA |
| cox OV CNV | Yuan Yuan | syn1895168 | 0.551507 | SCORED | OV, CNV |

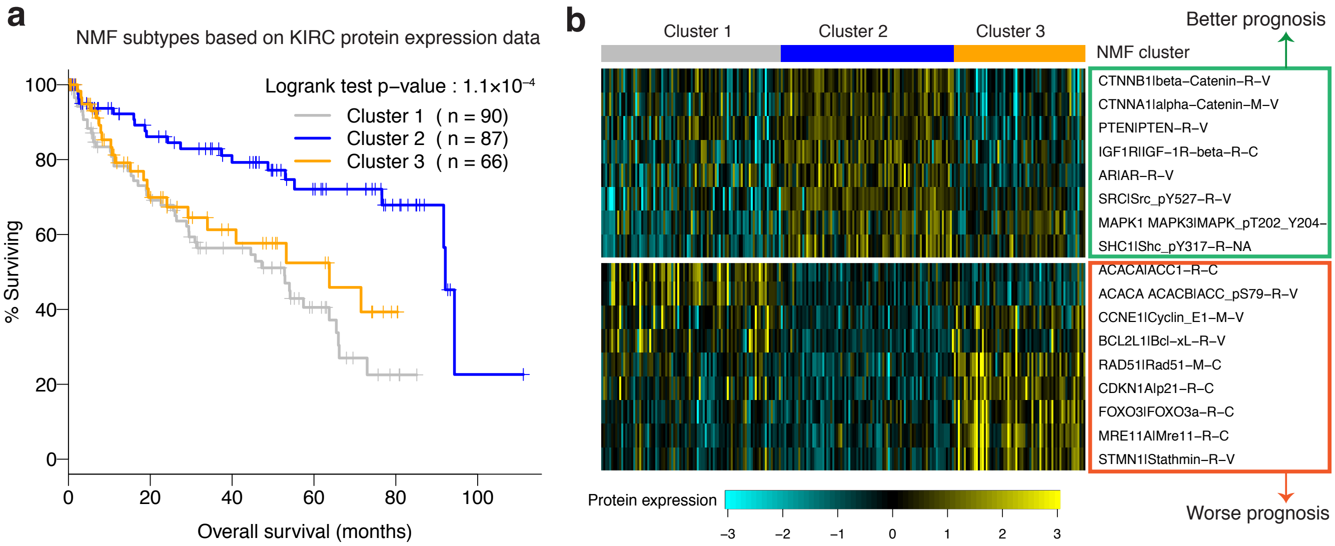
**Scheme of TCGA Pan-Cancer Survival Prediction Project.**

## Supplementary Figure 6



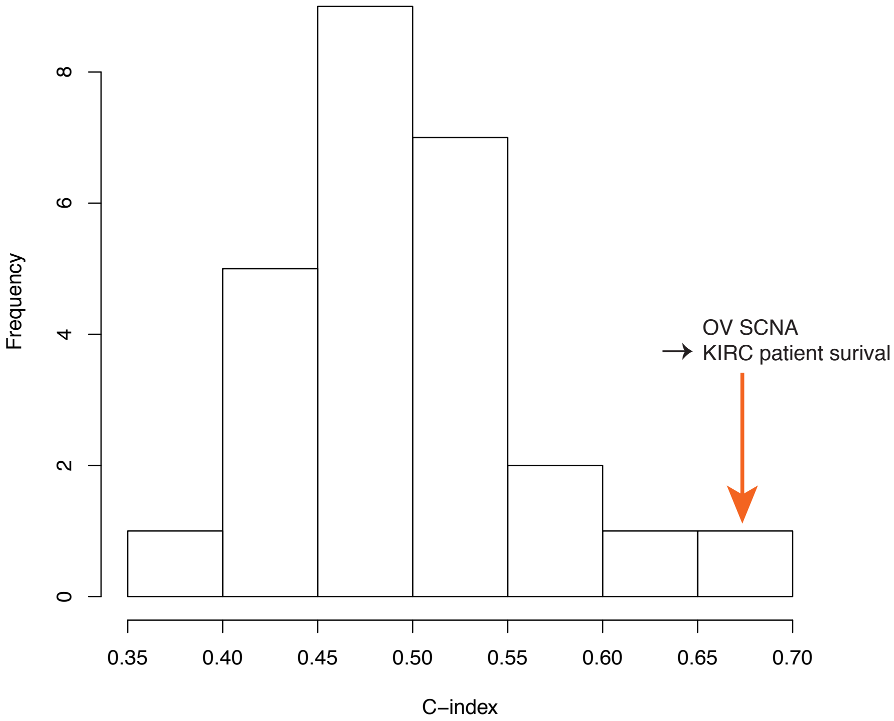
**Biological insights from the KIRC mRNA-expression subtype.**
(a) The Kaplan-Meier plot of the patients from the KIRC core set stratified by KIRC mRNA NMF subtypes. (b) The top differentially expressed genes among KIRC mRNA NMF subtypes grouped by the top enriched pathways.

# Supplementary Figure 7



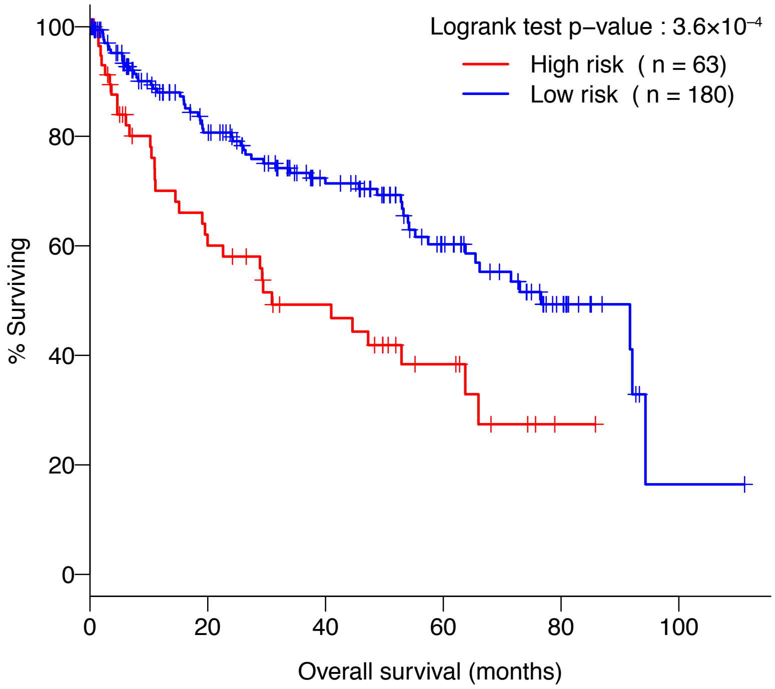
**Biological insights from the KIRC protein-expression subtype.**
(a) The Kaplan-Meier plot of the patients from the KIRC core set stratified by KIRC protein NMF subtypes. (b) The top differentially expressed protein markers among KIRC protein NMF subtypes.

## Supplementary Figure 8



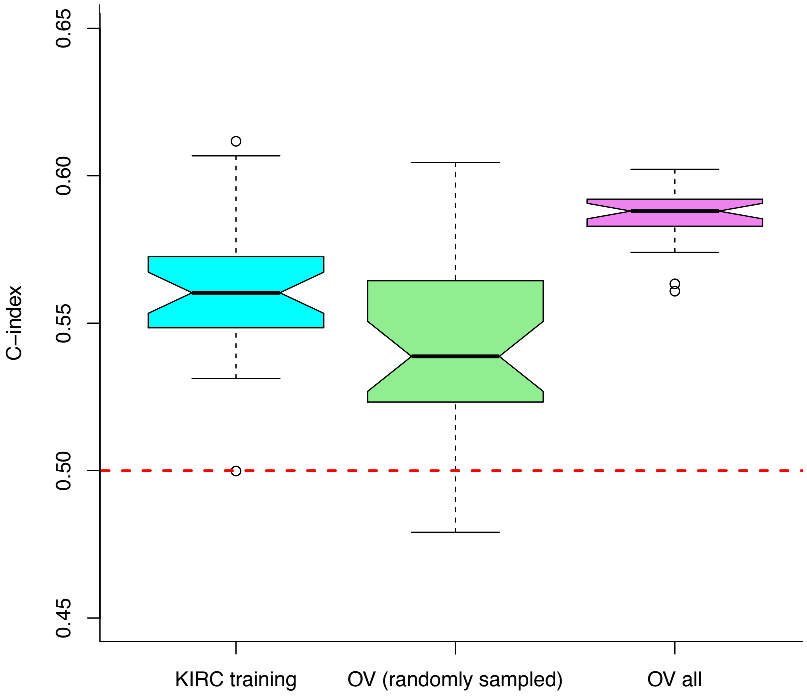
**Distribution of C-indexes from cross-tumor prediction.**
The two “NA” cases, where no feature passed the pre-selection for cross-tumor model construction, were not shown on the histogram.

## Supplemetary Figure 9



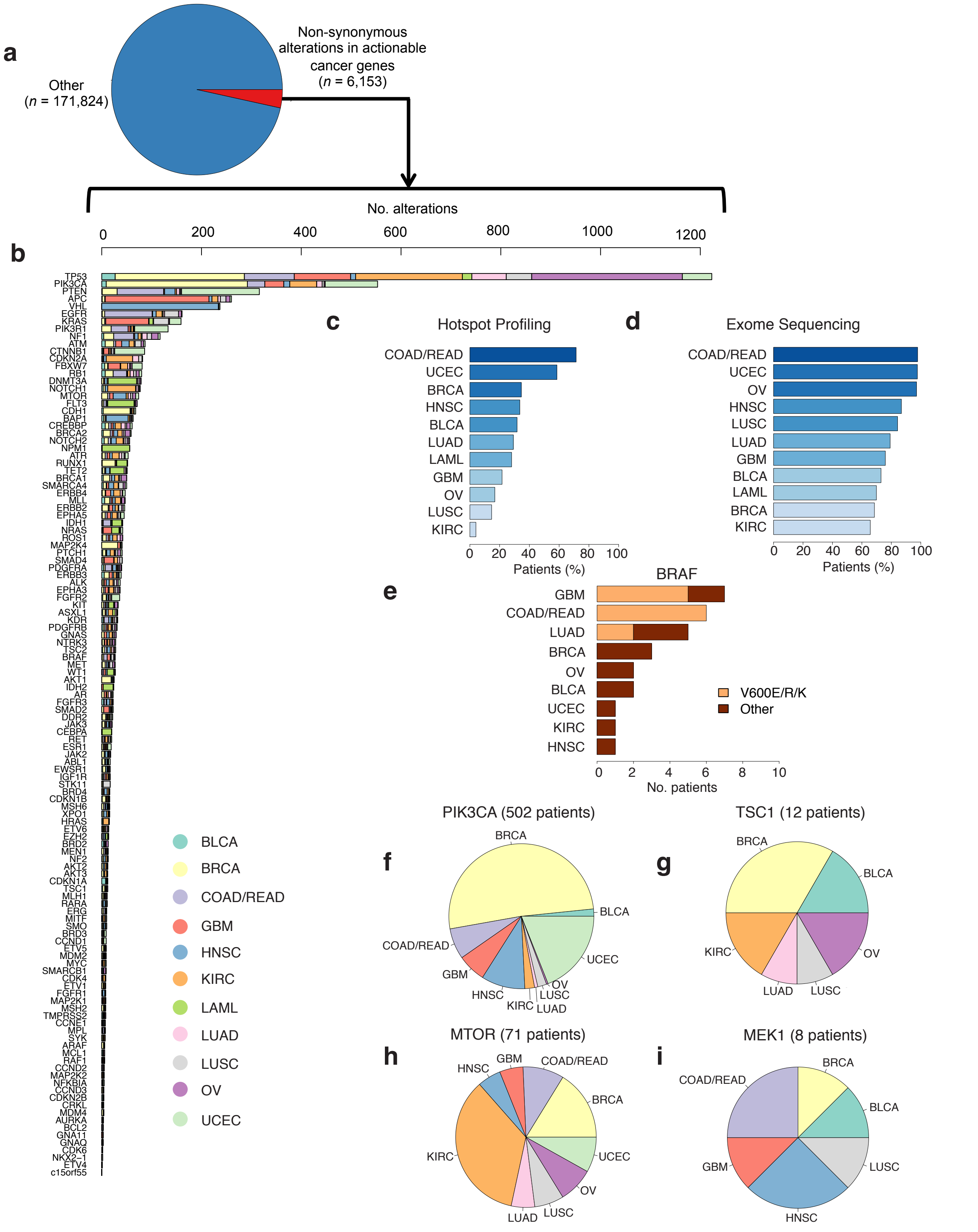
**The Kaplan-Meier plot of KIRC patients stratified by the risk scores predicted using the models trained from OV SCNA data.**

## Supplementary Figure 10



**Model trained from OV SCNA is predictive of the survival for KIRC patients by unregularized Cox model from 30 independent trials.** From left to right: C-indexes by unregularized Cox models trained from SCNA data of the KIRC training set; randomly sampled OV samples with the same size as the KIRC training set; and the whole OV core set.

# Supplementary Figure 11



**Alterations in clinically relevant genes across 12 tumor types in non-hypermutated tumors.** A subset analysis of mutations and indels in 2,892 patients representing 12 tumor types reveals a long tail of the frequency distribution of alterations in clinically relevant genes that warrant further exploration, focusing specifically on those with estimated mutation rates of ≤ 10 mutations/Mb (a-b). As with the larger set (**Fig. 5**), expanding tumor profiling beyond hotspot profiling technologies (c) to whole exome sequencing (d) increases the percentage of patients in all tumor types that may harbor clinically relevant alterations. While the distribution is not identical to the larger set, hotspot alterations in known cancer genes occur at low frequencies in unexpected tumor types (e); and alterations in emerging genes with potential clinical relevance are observed across tumor types (f-i). For a key to the tumor types, see **Supplementary Table 4**.

**Supplementary Table 1. The important features for LUSC protein-expression-based model by RSF.**

| Feature | HR | Wald_P | Pathway |
|---|---|---|---|
| MSH2.MSH2.M.C | 0.45 | 0.0046 | DNA repair/MSI |
| MSH6.MSH6.R.C | 0.48 | 0.0019 | DNA repair/MSI |
| MRE11A.Mre11.R.C | 8.74 | 0.0072 | DNA repair |
| CHEK2.Chk2_pT68.R.C | 0.14 | 0.0093 | DNA repair |
| XRCC5.Ku80.R.C | 0.46 | 0.0105 | DNA repair |
| GSK3A.GSK3B.GSK3.alpha.beta.M.V | 0.19 | 0.0022 | Metabolism |
| ACACA.ACC1.R.C | 0.59 | 0.0162 | Metabolism |
| ASNS.ASNS.R.C | 0.62 | 0.0147 | Metabolism |
| PRKAA1.AMPK_pT172.R.V | 0.42 | 0.0025 | Metabolism |
| COL6A1.Collagen_VI.R.V | 1.80 | 0.0130 | EMT/Stroma |
| CDH2.N.Cadherin.R.V | 5.37 | 0.0170 | EMT/Stroma |
| MAPK8.JNK_pT183_pT185.R.V | 4.44 | 0.0465 | Apoptosis signaling |
| CDC2.CDK1.R.V | 3.24 | 0.0096 | Apoptosis signaling |
| CASP3.Caspase.3_active.R.C | 0.17 | 0.0116 | Apoptosis signaling |
| FOXO3.FOXO3a_pS318_S321.R.C | 2.68 | 0.0038 | PI3K/AKT |
| PTCH1.PTCH.R.C | 1.80 | 0.0130 | Sonic Hedgehog signaling |

HR: hazard ratio; Wald_P: the *P*-value from Wald's test

**Supplementary Table 4. The short letter code for TCGA tumor types.**

| Short Letter Code | Tumor Type |
|---|---|
| BLCA | Bladder Urothelial Carcinoma |
| BRCA | Breast invasive carcinoma |
| COAD | Colon adenocarcinoma |
| GBM | Glioblastoma multiforme |
| HNSC | Head and Neck squamous cell carcinoma |
| KIRC | Kidney renal clear cell carcinoma |
| LAML | Acute Myeloid Leukemia |
| LUAD | Lung adenocarcinoma |
| LUSC | Lung squamous cell carcinoma |
| OV | Ovarian serous cystadenocarcinoma |
| READ | Rectum adenocarcinoma |
| UCEC | Uterine Corpus Endometrioid Carcinoma |

## Supplementary Text

### The effects of machine learning algorithm and feature selection on model performance

The overall predictive power of models built using the Cox and RSF methods are quite similar: for example, the predictive power of molecular data is generally high for KIRC but low for GBM. However, RSF models performed worse than Cox models in some scenarios (e.g., OV clinical variables only, GBM/LUSC clinical + molecular data), and we determined that the discrepancy arises from several factors. Compared with Cox, the models built by RSF consistently showed higher C-indexes on the training set (**Supplementary Fig. 2**), suggesting a higher likelihood of over-fitting. The feature selection scheme used by each method also contributes to the performance difference. Indeed, given the same feature set selected by LASSO, the performance difference between these two methods became smaller (**Supplementary Fig. 3a-b**). To further investigate the effect of feature selection, we examined two additional feature selection methods for RSF, minimal depth variable selection and variable hunting[25]. As shown in **Supplementary Figure 3c-d**, adding one feature selection step before RSF did not necessarily improve the model performance, and none of the feature selection methods appeared to be superior in all the cases.

### The effects of the training set sample size on model performance

In addition, we investigated the effect of the training set sample size on model performance. We conducted serial samplings and monitored the C-index as the training sample size varied for the cases where molecular data had substantial predictive power (median C-index > 0.6) (Online Methods). These cases include KIRC DNA methylation, KIRC mRNA expression, KIRC miRNA expression, and KIRC and LUSC protein expression. As expected, when the training sample size increased, there was a clear increase in the median C-index (**Supplementary Fig. 4**). For KIRC and LUSC protein expression, the C-index continued to improve up to the full sample set (**Supplementary Fig. 4d-e**), so a further increase in the number of training samples would likely boost the performance of these models. For KIRC DNA methylation, mRNA expression and miRNA expression, a high median C-index (C-index ≥ 0.7) could be achieved with fewer training samples (60%, 50% and 80% of the current training size, respectively) with the plateau being reached at a proportion of 0.9, 0.7, and 0.9, respectively (**Supplementary Fig. 4a-c**). For these cases, sample size might not represent a major factor limiting model performance.