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

Yuan Yuan[1,2,14], Eliezer M Van Allen[3,4,14], Larsson Omberg[5,14], Nikhil Wagle[3,4], Ali Amin-Mansour[4], Artem Sokolov[6], Lauren A Byers[7], Yanxun Xu[8], Kenneth R Hess[9], Lixia Diao[2], Leng Han[2], Xuelin Huang[9], Michael S Lawrence[4], John N Weinstein[2,10], Josh M Stuart[6], Gordon B Mills[10], Levi A Garraway[3,4,11,15], Adam A Margolin[5,13,15], Gad Getz[4,11,12,15] & Han Liang[1,2,15]

© 2014 Nature America, Inc. All rights reserved.

Molecular profiling of tumors promises to advance the clinical management of cancer, but the benefits of integrating molecular data with traditional clinical variables have not been systematically studied. Here we retrospectively predict patient survival using diverse molecular data (somatic copy-number alteration, DNA methylation and mRNA, microRNA and protein expression) from 953 samples of four cancer types from The Cancer Genome Atlas project. We find that incorporating molecular data with clinical variables yields statistically significantly improved predictions (FDR < 0.05) for three cancers but those quantitative gains were limited (2.2–23.9%). Additional analyses revealed little predictive power across tumor types except for one case. In clinically relevant genes, we identified 10,281 somatic alterations across 12 cancer types in 2,928 of 3,277 patients (89.4%), many of which would not be revealed in single-tumor analyses. Our study provides a starting point and resources, including an open-access model evaluation platform, for building reliable prognostic and therapeutic strategies that incorporate molecular data.

The Cancer Genome Atlas (TCGA) project has yielded many biological insights through generating genomic, transcriptomic, epigenomic and proteomic data from a large number of patient samples in many cancer types[1–6]. However, the potential clinical utility of these data in aggregate remains largely unknown.

Large-scale molecular profiling data may be informative for multiple aspects of oncology practice. One key application for patients with primary disease is accurate prognosis, which helps stratify patients into different risk groups and choose both treatment and surveillance strategies. Traditionally, prognosis is based on clinical variables such as age and tumor stage. Recently, extensive efforts have been made to incorporate molecular information for better prognosis. For example, ER, PR, HER2 protein levels and HER2 genomic amplification are important biomarkers in breast cancer, which have demonstrated high value in clinical use[7]. However, owing to the high cost of molecular profiling on a large scale, previous studies have either focused on a small number of selected genes or have used only single-platform genomic data (e.g., microarrays). By convention, such studies have been limited to a single cancer lineage. Another important clinical application is to choose targeted therapies based on the alteration spectrum in an individual patient's tumor. Multiple efforts have been initiated to apply high-throughput sequencing data in clinical strategies[8,9], although alterations in clinically actionable genes have not been fully cataloged. Knowledge of this catalog may inform target selection for drug development as well as clinical trial design and identify patient populations that may benefit from emerging targeted therapeutics.

The overall goal of this study was to address how and to what extent TCGA molecular data could affect oncology practice. Thus, we evaluated two closely related but distinct aspects of clinical utility—prognostic utility (that is, predicting patient survival using various types of high-throughput molecular data across multiple tumor lineages) and therapeutic utility (that is, identifying the spectrum of somatic alterations in clinically actionable genes, which in the future may inform treatment selection). First, we examined the performance of molecular data (somatic copy-number alteration (SCNA), DNA methylation and mRNA, microRNA and protein expression) alone or in combination with clinical variables in predicting censored or dichotomized patient survival data for four TCGA cancer types with high-quality overall survival data. Furthermore, to facilitate a broader community effort, we developed an open-access platform that allows researchers to build and evaluate survival prediction models on these data sets. We did not intend to generate prognostic models ready for clinical

[1]Graduate Program in Structural and Computational Biology and Molecular Biophysics, Baylor College of Medicine, Houston, Texas, USA. [2]Department of Bioinformatics and Computational Biology, The University of Texas MD Anderson Cancer Center, Houston, Texas, USA. [3]Department of Medical Oncology, Dana-Farber Cancer Institute, Boston, Massachusetts, USA. [4]Broad Institute of MIT and Harvard, Cambridge, Massachusetts, USA. [5]Sage Bionetworks, Seattle, Washington, USA. [6]Department of Biomolecular Engineering, University of California, Santa Cruz, California, USA. [7]Department of Thoracic/Head & Neck Medical Oncology, The University of Texas MD Anderson Cancer Center, Houston, Texas, USA. [8]Division of Statistics and Scientific Computing, The University of Texas at Austin, Austin, Texas, USA. [9]Department of Biostatistics, The University of Texas MD Anderson Cancer Center, Houston, Texas, USA. [10]Department of Systems Biology, The University of Texas MD Anderson Cancer Center, Houston, Texas, USA. [11]Harvard Medical School, Boston, Massachusetts, USA. [12]Massachusetts General Hospital, Cancer Center and Department of Pathology, Boston, Massachusetts, USA. [13]Present address: Department of Biomedical Engineering, Computational Biology Program, Oregon Health and Science University, Portland, Oregon, USA. [14]These authors contributed equally to this work. [15]These authors jointly supervised this work. Correspondence should be addressed to H.L. (hliang1@mdanderson.org) or G.G. (gadgetz@broadinstitute.org) or A.A.M. (margolin@ohsu.edu).

Received 31 May 2013; accepted 28 May 2014; published online 22 June 2014; doi:10.1038/nbt.2940

use, but rather we sought to provide insights into how to improve such models by incorporating informative molecular data. Second, we investigated the current spectrum of potentially clinically actionable alterations (somatic point mutations and small insertions/deletions) across 12 TCGA tumor types. By analyzing molecular data from multiple cancer types, we were able to evaluate prognostic models and identify alterations that would not have been obtained with single-tumor data sets.

RESULTS

Assessment of the prognostic power of diverse molecular data

We focused on four TCGA cancer types: kidney renal clear cell carcinoma (KIRC)[6], glioblastoma multiforme (GBM)[1], ovarian serous cystadenocarcinoma (OV)[2] and lung squamous cell carcinoma (LUSC)[4]. These cancer types were chosen because their TCGA data sets included survival data with adequate follow-up time and sufficient samples characterized by multiple types of molecular data. The TCGA cohorts have overall survival patterns similar to those reported in previous publications[10–13]. For each cancer type, we compiled a core sample set in which each sample has information available for the overall survival time, clinical variables (e.g., gender, age, tumor stage and grade) and at least four out of the five types of molecular data related to gene expression ((i) SCNA: Affymetrix Human SNP Array 6.0, ~100 arm or focal alterations; (ii) DNA methylation: Illumina DNA Methylation microarray, ~20,000 genes; (iii) mRNA expression: Agilent 244 K microarray or Illumina mRNA-seq, ~20,000 genes; (iv) microRNA (miRNA) expression: Agilent Human miRNA-specific microarray or Illumina miRNA-seq, >500 microRNAs; (v) protein expression: reverse-phase protein array, ~170 proteins) (**Table 1**).

Table 1 Overview of TCGA samples and high-throughput characterization platform information by cancer type

Cancer	Overall survival	SCNA	Methy	mRNA	miRNA	Protein	Core set
GBM		SNP_6	27k	AgilentG4502A	H-miRNA_8x15K	RPPA[a]	
	565	563	287	492	491	214	210
KIRC		SNP_6	450k	HiseqV2	GA+Hiseq	RPPA	
	500	493	283	469	454	480	243
OV		SNP_6	27k	AgilentG4502A	H-miRNA_8x15K	RPPA	
	563	559	600	558	586	412	379
LUSC		SNP_6	450k[a]	HiseqV2	GA+Hiseq	RPPA	
	305	343	225	220	351	195	121

For each cancer type, the first row shows the platforms and the second row shows the sample counts. SNP_6: Affymetrix Genome-Wide Human SNP Array 6.0; 27k: Illumina Infinium Human DNA Methylation 27K, 450k: Illumina Infinium Human DNA Methylation 450K; AgilentG4502A: Agilent 244K Custom Gene Expression G4502A; HiseqV2: Illumina HiSeq 2000 RNA Sequencing V2; H-miRNA_8x15K: Agilent 8 × 15K Human miRNA-specific microarray platform; GA+Hiseq: Illumina Genome Analyzer/HiSeq 2000 miRNA sequencing platform; RPPA: MD Anderson reverse phase protein array.
[a]The data type was not included in that cancer type.

For each core sample set, we applied Monte Carlo cross-validation and assessed the predictive power of individual molecular data types or clinical variables using the concordance index (C-index)[14]. The C-index is a nonparametric measure to quantify the discriminatory power of a predictive model: a C-index of 1 indicates perfect prediction accuracy and a C-index of 0.5 is as good as a random guess (Online Methods). We compiled candidate features from molecular data or clinical data for each cancer type and randomly split the core set into training and test sets 100 times (**Fig. 1a**). We built the predictive models from the training set using two well-established but highly complementary methods: (i) Cox, the multivariate Cox proportional hazards model with L1 penalized log partial likelihood (LASSO)[15] for feature selection; and (ii) random survival forest (RSF)[16].

For each cancer type, the clinical-variable-only models showed substantial predictive power, with C-indexes significantly higher than 0.5 (range: 0.624–0.754; $P = 0$) (**Fig. 1b–e** and **Supplementary Fig. 1**). In 9 out of 18 cases, the models built from individual molecular data sets alone showed statistically significant predictive power (**Supplementary Fig. 1**), but in only one case, the model built from LUSC protein expression data had predictive power similar to that of the corresponding clinical-variable-only model (**Fig. 1e**, C-index 0.632 versus 0.626, $P = 0.40$, Wilcoxon signed rank test). The relative predictive power of individual molecular data sets strongly depended on the cancer type; for example, the prognostic power was much higher for KIRC than for the other three cancer types. In general, the trends observed with the Cox models were similar to those observed using the RSF models.

To examine whether genomic and proteomic data can provide additional prognostic power when used with clinical variables, we built predictive models by integrating clinical variables with each type of molecular data (both gene-level features and molecular subtype features) (Online Methods). Notably, the integrated models resulted in statistically significantly improved predictive power compared to those clinical-variable-only models in three cancer types, including mRNA, microRNA and protein expression in KIRC, miRNA expression in OV and protein expression in LUSC (one-sided Wilcoxon signed rank test, KIRC clinical + mRNA: $P < 3.3 \times 10^{-3}$, false-discovery rate (FDR) < 0.035; clinical + miRNA: $P < 1.2 \times 10^{-4}$, FDR < 2.1×10^{-3}; clinical + protein: $P < 8.4 \times 10^{-5}$, FDR < 2.1×10^{-3}; OV clinical + miRNA: $P < 7.0 \times 10^{-5}$, FDR < 2.1×10^{-3}; LUSC clinical + protein: $P < 7.9 \times 10^{-4}$, FDR < 0.011) (**Fig. 1b–e**). However, in terms of quantitative gain (i.e., the median value of Somers' D[14] across the 100 splits, a measurement for C-index change), the increase was limited (KIRC clinical + mRNA: 4.0%, clinical + miRNA: 7.4%, clinical + protein: 2.2%; OV clinical + miRNA: 13.7%; LUSC clinical + protein: 23.9%). In addition, we examined the effects of machine learning algorithms, feature selection and sample size of the training set on model performance (**Supplementary Results** and **Supplementary Figs. 2–4**).

To facilitate a broader community effort for such modeling, we developed an open-access platform that allows researchers to evaluate and submit survival prediction models in a "collaborative competition" research framework (**Supplementary Fig. 5**). The homepage of the TCGA Pan-Cancer Survival Prediction challenge can be accessed in Synapse (doi:10.7303/syn1710282). The site contains all models used in this study, including provenance records and transparent source code that allows each model to be inspected, rerun or improved upon. Each model is linked to a standard set of metadata annotations that provide online querying capability (e.g., corresponding to the cancer type or learning algorithm), allowing for the comparison of models based on user-defined criteria. The C-index scores for each model, as reported here, are displayed in the form of a real-time leaderboard.

Biological insights from top-performing prognostic models

For the top prognostic models highlighted in **Figure 1**, we further examined important molecular features included in each model to gain some mechanistic insights. The LUSC protein expression model is the only case where the molecular data alone showed a performance similar to that of the clinical-variable-only model. Features in the model with high predictive ability[16] were dominated by proteins involved in DNA repair and microsatellite instability (e.g., MSH2) and metabolism (e.g., ACC1) (**Supplementary Table 1**).

Molecular data in five integrative models conferred additional prognostic power given clinical variables (**Fig. 1**). Notably, in four

© 2014 Nature America, Inc. All rights reserved.

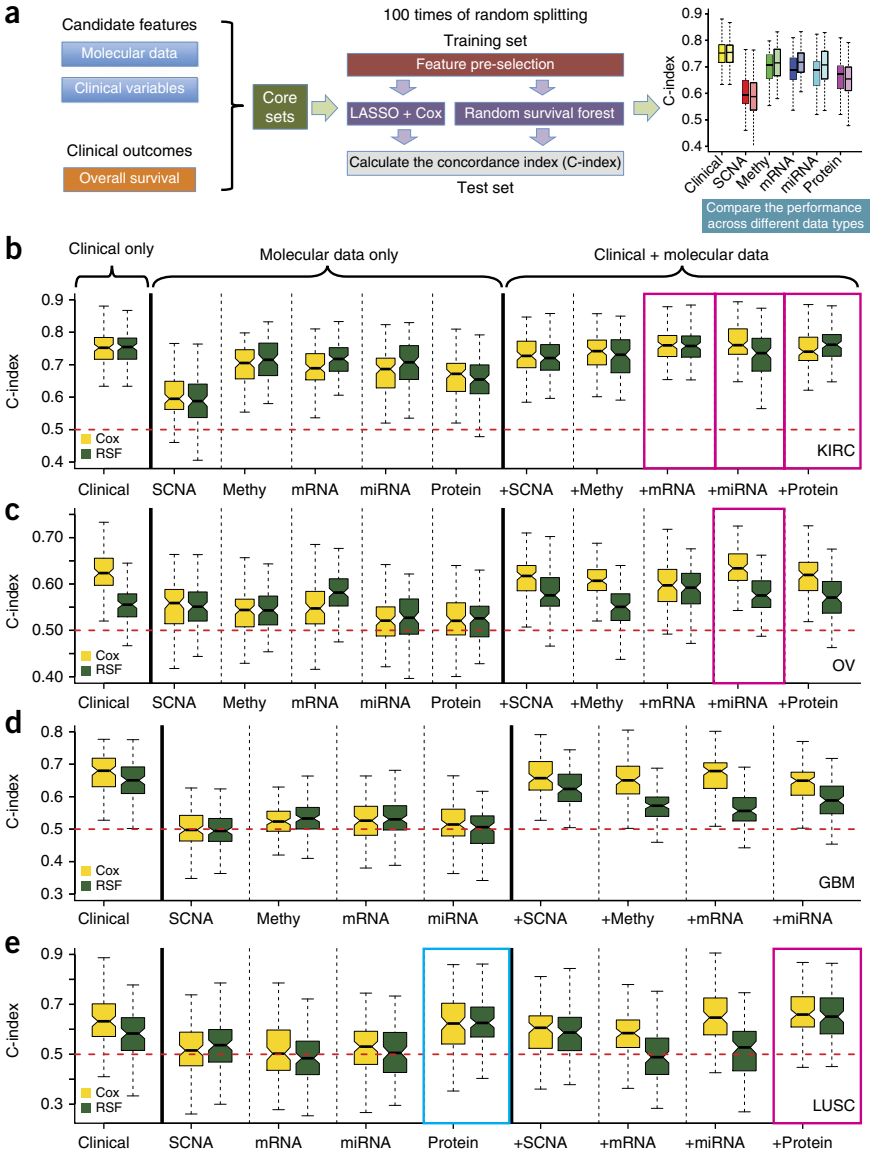


Figure 1 Comparison of the survival predictive power of clinical variables, molecular data and their combinations. (**a**) An overview of the computational approach. For each data type, box plots in different colors indicate results using different algorithms. (**b**–**e**) C-indexes by models trained from clinical variables, individual molecular data alone or in combination with clinical variables in KIRC (N_{total} = 243) (**b**), OV (N_{total} = 379) (**c**), GBM (N_{total} = 210) (**d**) and LUSC (N_{total} = 121) (**e**). For each cancer type, during each of the 100 times of random splitting, 80% of the total samples were used to train the model and the remaining 20% as the test set for C-index calculations. The blue box highlights the model built from individual molecular data that shows comparable performance to that based on clinical variables (two-sided Wilcoxon signed rank test, $P > 0.05$); and the magenta boxes highlight the models integrating molecular data and clinical variables that show better performance than that based on only clinical variables (one-sided Wilcoxon signed rank test, FDR < 0.05). ((The boundaries of the box mark the first and third quartile, with the median in the center, and whiskers extending to 1.5 interquartile range from the boundaries.) The red dashed lines marked the C-index equivalent to random guess (C-index = 0.5).)

of these models, the only contributing molecular feature was the molecular subtype derived from the corresponding expression data (through consensus non-negative matrix factorization (NMF)[17]). Molecular subtypes can be regarded as higher-level assemblies of individual gene features and therefore may act as a more robust predictor than an individual marker or small marker sets. Indeed, the NMF subtypes (derived from OV miRNA expression, LUSC protein expression and KIRC mRNA, and protein expression data, respectively) showed distinct survival patterns in the respective cancer types (log-rank test, **Fig. 2c**, $P < 0.043$; **Fig. 2e**, $P < 8.2 \times 10^{-3}$; **Supplementary Fig. 6a**, $P < 9.8 \times 10^{-5}$; **Supplementary Fig. 7a**, $P < 1.1 \times 10^{-4}$).

Given the limited availability of suitable independent data in the public domain, we evaluated the performance of the OV clinical + miRNA model. Using the multiclass classifier built from TCGA OV miRNA expression data (**Fig. 2a**, Online Methods, the area under the receiver operating characteristic curve (AUC) = 0.98, **Fig. 2b**), we recovered the survival pattern of the NMF subtypes observed for the TCGA core set (**Fig. 2c**, log-rank test, $P < 0.043$) in an independent cohort[18] (**Fig. 2d**, log-rank test, $P < 6.3 \times 10^{-3}$): the patients in cluster 3 have better survival than those in clusters 1 and 2 (prognostic miRNAs in each cluster are shown in **Supplementary Table 2**). Further, applying this trained model to the independent cohort yielded the expected improvement for including the miRNA NMF subtypes.

For the molecular subtypes defined by LUSC protein expression, pMEK1 and pMAPK and the downstream target pS6 were among the top markers expressed at higher levels in patients with shorter survival times (clusters 2 and 3, **Fig. 2e**,**f**). Clinical and preclinical data suggest that MEK inhibitors are active in specific subsets of non–small cell lung cancer, such as KRAS-mutated lung adenocarcinomas[19,20]. Our results suggest that patients with high-risk forms of LUSC have relatively greater activation of the RAS/MEK/MAPK pathway and that MEK targeting warrants further exploration in this population as well. In addition, the mTOR and Src pathways may also be more active in cluster 3, whereas DNA-repair protein levels were low in both clusters 2 and 3 (**Fig. 2f**). Gene signatures associated with KIRC mRNA expression subtypes were aligned with their reported biological roles and survival patterns. Many proteins involved in acute-phase response signaling and several pro-metastatic matrix metalloproteases were highly expressed in the groups with worse survival outcomes (clusters 2 and 3)[21,22], whereas death receptor signaling proteins were downregulated in these groups[23] (**Supplementary Fig. 6a**,**b**). The survival pattern of the NMF subtypes by KIRC protein expression also matches the survival correlations of individual protein markers (**Supplementary Fig. 7a**,**b**).

Finally, the KIRC clinical + miRNA model was the only integrative model for which individual gene features, instead of a molecular subtype derived from the complete expression data set, provided additional prognostic power. Each of the six miRNAs comprising the signature was significantly correlated with survival, and their hazard ratio matched with previously reported roles in cancer progression. Although upregulation of miR-21, which has growth-promoting activity, is associated with a worse prognosis[24,25], the remaining miRNAs (miR-192, miR-101,

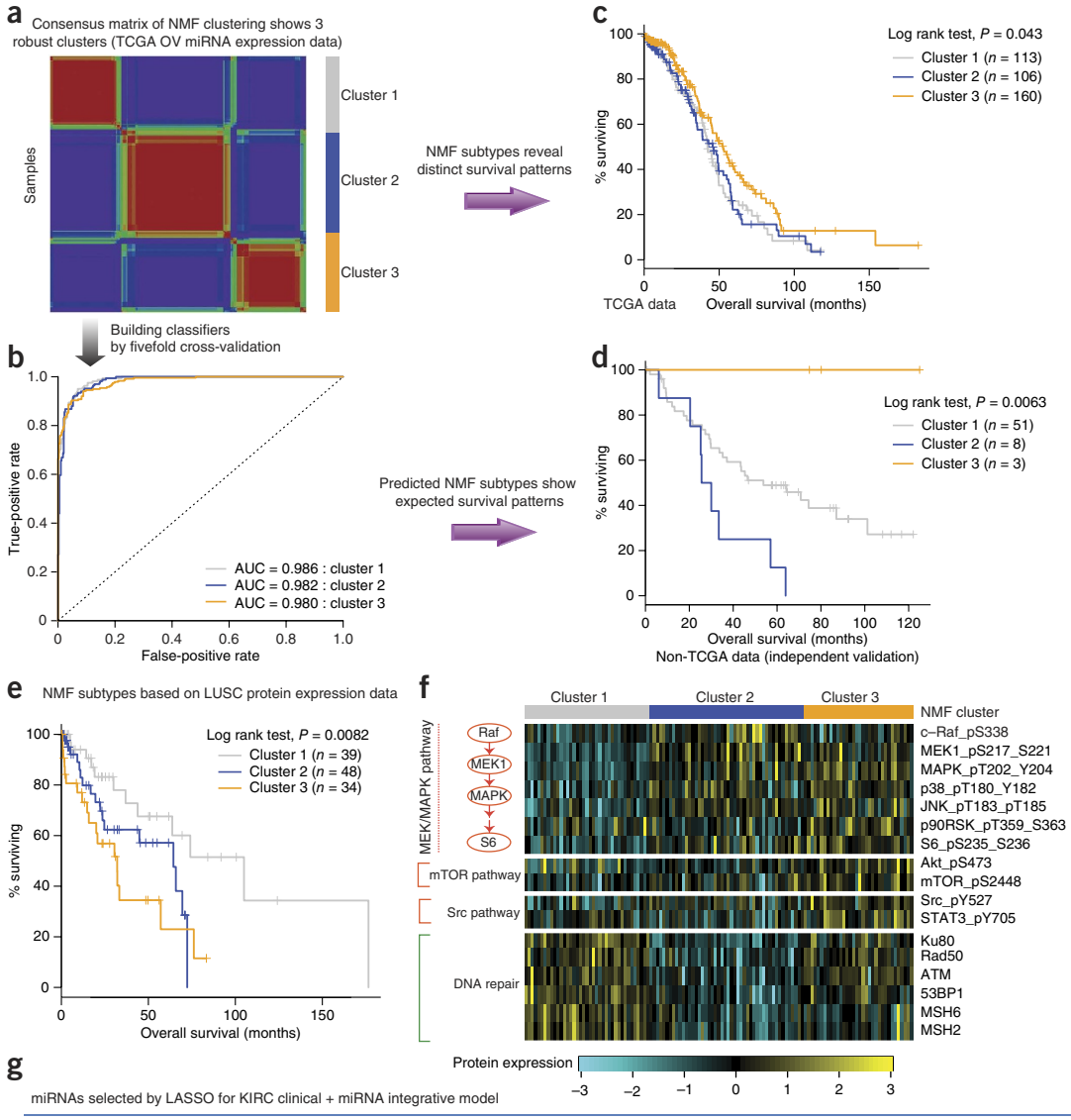
© 2014 Nature America, Inc. All rights reserved.

let-7a, let-7f and miR-143) suppress tumor growth, with higher expression being associated with a better prognosis[26–31] (**Fig. 2g**).

Patient survival prediction using cross-tumor models

To test whether molecular data could identify commonalities across different tumor types, we assessed whether a model trained using molecular data in one cancer type could predict survival in other cancer types that share the same type of molecular data generated by the same platform (Online Methods). In the vast majority of cases, the C-index was around 0.5, suggesting little predictive power across tumor types (**Supplementary Fig. 8**).

However, a model trained from OV SCNA data was predictive of survival for patients with KIRC, with a median C-index of 0.67 (**Fig. 3a** and **Supplementary Fig. 9**). Furthermore, given the



g

miRNAs selected by LASSO for KIRC clinical + miRNA integrative model

Feature	Hazard ratio (95% CI)	Wald's test *P* value	Functional remarks
miR-192	0.81 (0.75 – 0.90)	2.4×10^{-5}	Suppresses tumor progression in renal cell carcinoma[26].
miR-21	2.16 (1.57 – 2.98)	2.7×10^{-6}	Higher miR-21 levels were associated with higher stage and grade, shorter disease-free and overall survival rates in renal cell carcinoma[24,25].
miR-101	0.52 (0.40 – 0.69)	4.8×10^{-6}	Functions as a negative regulator of EZH2 expression and renal cancer cell proliferation[31].
let-7a	0.60 (0.47 – 0.77)	6.5×10^{-5}	Functions as a tumor suppressor in renal cell carcinoma by targeting c-myc[28].
let-7f	0.74 (0.63 – 0.87)	2.4×10^{-4}	Inhibits tumor invasion and metastasis by targeting MYH9 in human gastric cancer[27].
miR-143	0.77 (0.65 – 0.90)	1.4×10^{-3}	Functions as a tumor suppressor in human esophageal squamous cell carcinoma[29] and human bladder cancer T24 cells[30].

Figure 2 Biological insights from the top prognostic models. (**a**) Consensus non-negative matrix factorization (NMF) clustering of the TCGA OV miRNA expression data reveals three molecular subtypes (clusters). (**b**) The ROC curves of the multiclass classifier against NMF subtypes trained from the TCGA OV miRNA expression data through fivefold cross-validation. (**c**) The Kaplan-Meier plot of the patients from the TCGA OV core set stratified by OV miRNA-expression NMF subtypes. (**d**) The Kaplan-Meier plot of the patients from the independent OV cohort stratified by predicted miRNA NMF subtypes using the classifier in **b**. (**e**) The Kaplan-Meier plot of the patients from the LUSC core set stratified by LUSC protein-expression NMF subtypes. (**f**) The top differentially expressed protein markers among LUSC protein-expression NMF subtypes grouped by pathways or functions. (**g**) The miRNAs selected by LASSO for the KIRC clinical + miRNA integrative model.

© 2014 Nature America, Inc. All rights reserved.

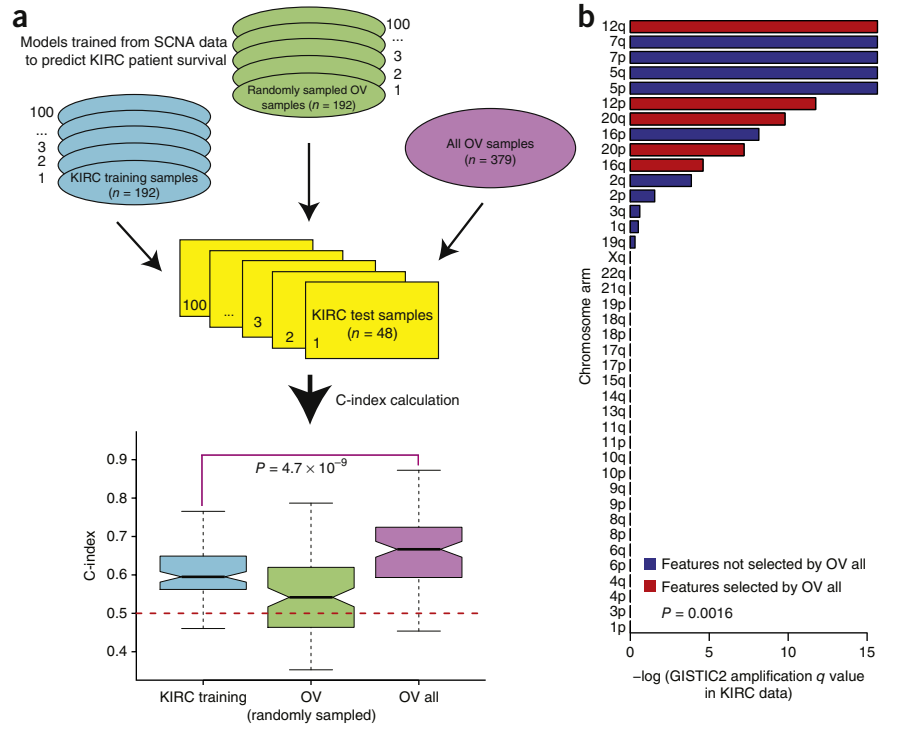


Figure 3 Models trained from OV SCNA data can predict survival of individuals with KIRC. (**a**) From left to right: C-index for the models trained from KIRC SCNA data ($N_{training}$ = 192), C-index for the models trained from SCNA of OV sample sets with the same size as the KIRC training sets ($N_{training}$ = 192), and C-index for the model trained from SCNA of the whole OV core set ($N_{training}$ = 379). The OV model trained from the whole OV core set showed higher predictive power than the model trained from SCNA data of independent KIRC training samples (two-sided Wilcoxon signed rank test, $P = 4.7 \times 10^{-9}$). (The boundaries of the box mark the first and third quartile, with the median in the center, and whiskers extending to 1.5 interquartile range from the boundaries.) (**b**) The bar plot of amplification q value of arm-level SCNA features from GISTIC2. The features included in the model trained from OV SCNA data are shown in red. The KIRC q values of the features selected from OV SCNA data were lower than those not selected (two-sided Wilcoxon rank sum test, $P = 1.6 \times 10^{-3}$). The red dashed lines marked the C-index equivalent to random guess (C-index = 0.5).

same KIRC test sets, the OV model showed higher predictive power than a model trained from SCNA data of the KIRC training samples (**Fig. 3a**, median C-index, 0.67 versus 0.59, $P = 4.7 \times 10^{-9}$, Wilcoxon signed rank test). When we randomly sampled the same number of OV samples as the KIRC training sets to build the predictive model, the C-index dropped from 0.67 to 0.54, suggesting that the higher predictive power was largely due to the larger sample size of the OV core set. We further confirmed this pattern using an independent approach and an independent sample partition (**Supplementary Fig. 10**). Closer examination revealed one common feature ("12q") among the features selected from SCNA of the whole OV core set and KIRC core set, which may be crucial for the cross-tumor predictive power. In addition to 12q, there were other four arm-level SNCA features included in the model trained from OV SCNA data, including 12p, 16q, 20p and 20q, all of which showed significant amplification of q values in KIRC according to GISTIC2 (ref. 32) (**Fig. 3b**). Indeed, the KIRC q values of the features selected from OV SCNA data were lower than those not selected (**Fig. 3b**, $P = 1.6 \times 10^{-3}$, Wilcoxon rank sum test). The shared biological features identified above provide key insights into mechanistic connections between the two cancer types.

Factors affecting prediction of dichotomized survival data

In addition to analysis on censored survival data, we examined the power of molecular data in predicting dichotomized overall survival data. Unlike censored survival data, there are many machine-learning algorithms for classifying binary clinical outcomes. Although the process of dichotomization will lose some information, this practice enables us to systematically survey many modeling scenarios and assess the effect of different factors on predicting survival data. For each cancer type, we dichotomized the censored continuous survival data through a designated cutoff time (survival milestone), and then constructed a series of models from individual molecular data alone or with clinical variables by (i) using eight common classification algorithms, (ii) applying two feature pre-selection strategies and (iii) including different numbers of final features in the model. In total, we assessed the performance of >5,000 models through tenfold cross-validation based on the threshold-independent AUC score.

Figure 4a–d and **Supplementary Table 3** show the AUC score for each algorithm, with the optimal setting for each data set and each cancer. Overall, as observed for continuous survival data, the predictive power of molecular data strongly depended on the cancer type. Clinical variables showed better performance than individual molecular data except for LUSC, where the protein expression data showed better performance than the clinical variables given the best-performing algorithms (**Fig. 4d**). Moreover, the integration of molecular data with clinical variables improved the predictive power, especially for the following data sets: DNA methylation and protein expression in KIRC using most algorithms (**Fig. 4a**); mRNA expression in GBM using K-nearest neighbor (KNN), nearest centroid (NC) and support vector machine (SVM) (**Fig. 4b**); and protein expression in OV and LUSC using most algorithms (**Fig. 4c**,**d**). To quantify the effects of specific factors on survival prediction, we performed an analysis of variance (ANOVA) on the AUC scores, and found that cancer type, data type and their interactions were the three dominant sources of variability, respectively explaining 35.7%, 17.4% and 11.8% of the variability of the prediction performance (**Fig. 4e**). In contrast, the effect of machine-learning algorithms was moderate (5.2%). We obtained similar results when the sample size was kept consistent across all cancer types (**Fig. 4e**). These results were consistent with the recent microarray quality control (MAQC)-II study[33].

Somatic alterations in clinically relevant genes

Finally, we assessed the therapeutic utility of TCGA data by analyzing somatic mutations and small insertion/deletions (indels) in 3,277 patients across 12 tumor types. We applied a heuristic algorithm[34,35] to score the clinical importance of each alteration in 121 clinically relevant genes. Clinically relevant genes were defined as those that, when somatically altered, may predict resistance or response to a therapy and/or have diagnostic or prognostic relevance for a particular tumor type[36]. It is important to emphasize that not all aberrations in

© 2014 Nature America, Inc. All rights reserved.

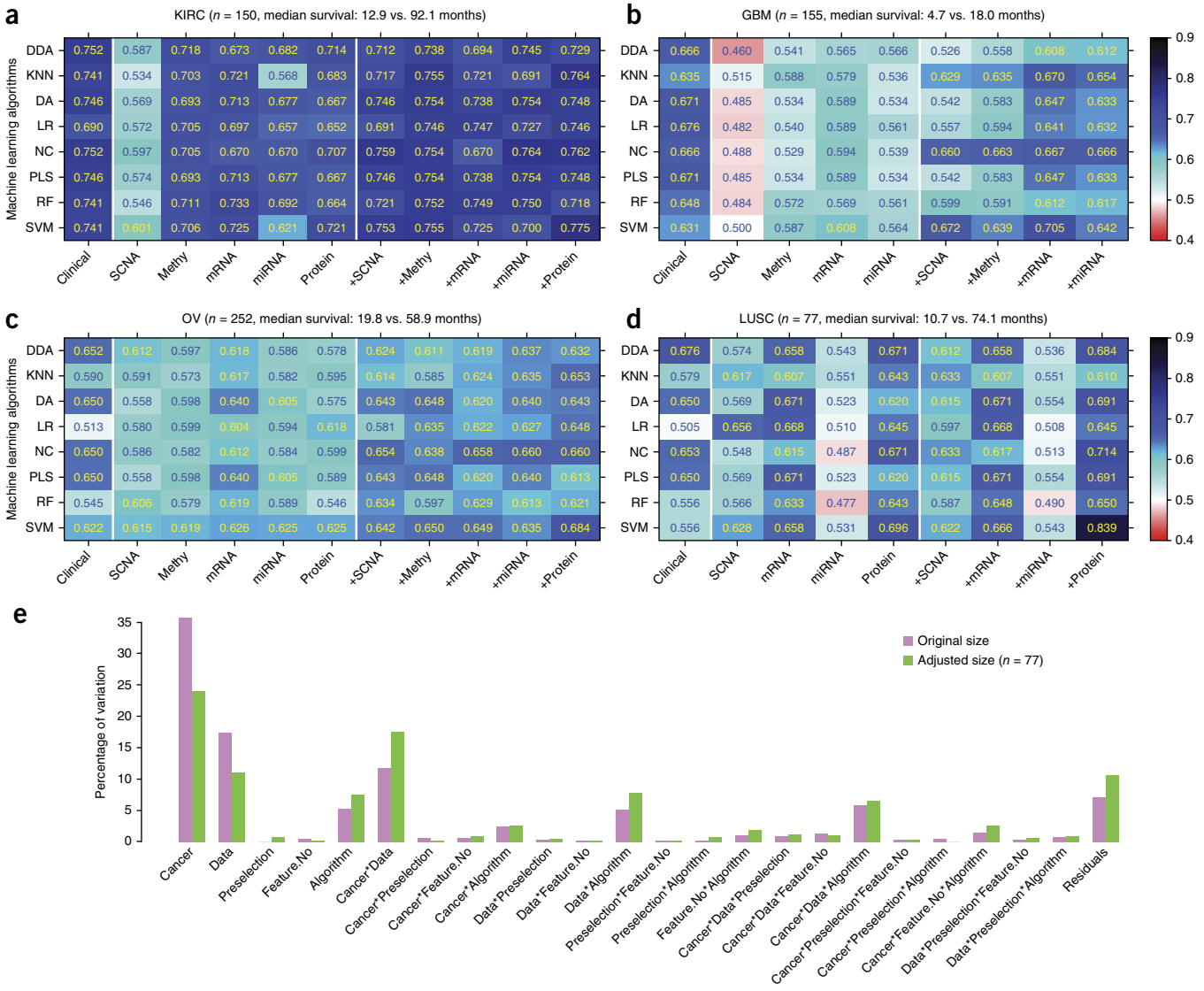


Figure 4 Predictive performance of clinical variables, molecular data and their combination on dichotomized survival data. (**a**–**e**) The best AUC achieved by each classification algorithm for each clinical/molecular/combination data set in KIRC (N_{total} = 150) (**a**), GBM (N_{total} = 155) (**b**), OV (N_{total} = 252) (**c**) and LUSC (N_{total} = 77) (**d**). DDA, diagonal discriminant analysis; KNN, K-nearest neighbor; DA, discriminant analysis; LR, logistic regression; NC, nearest centroid; PLS, partial least squares; RF, random forest; SVM, support vector machine. AUCs were calculated based on tenfold cross-validation. (**e**) Variation explained by modeling factors and their interactions.

clinically relevant genes will act as "drivers" and portend response to therapeutic targeting, and that a majority of the alterations in these clinically relevant genes remain variants of uncertain clinical significance and require further experimental and clinical evaluation.

In 89.4% (2,928/3,277) of the TCGA patient samples, 10,281 somatic nonsynonymous alterations (1.62% of all alterations, synonymous or nonsynonymous, in this combined cohort) in 121 clinically relevant cancer genes were observed (**Fig. 5a**,**b**). Of these, 1,287 alterations in 31.4% (1,028/3,277) of patients were observed in genomic hotspots that were tested for in representative prospective clinical settings using a panel that probes events in 41 genes[9]. As expected, by extending genomic profiling to cover all exons of the same gene set, we observed a large increase in the number of observed alterations in clinically relevant genes in all 12 tumor types (**Fig. 5c**,**d**). This result reflects the gap in the understanding of the clinical relevance of a majority of alterations in genes potentially linked to clinical actions. To exclude the effect of hypermutated tumors in this cohort (e.g., colorectal, endometrial)[3,5], we repeated the analysis for tumors with mutation rates of ≤10 mutations/Mb (n = 2,892), as suggested in previous TCGA studies. We observed a 3.5% (6,153 /177,977) rate of somatic alterations in the 121 clinically relevant genes (**Supplementary Fig. 11**).

These results highlight multiple themes pertaining to the use of genomics in clinical oncology. First, well-characterized clinically relevant alterations can rarely be observed in unexpected tumor types. For instance, one patient with cervical cancer harbored a somatic *BRAF*^{K601E} mutation (**Fig. 5e**). Although not previously reported in this tumor type, such a patient may warrant consideration for therapies that target these alterations[37]. This descriptive example highlights the potential benefit for comprehensive profiling in clinical settings, although prospective implementation of this approach is needed to determine the general applicability to clinical oncology.

Next, by combining mutation data from 12 tumor types, we observed a "tail" of low-frequency alterations in clinically relevant cancer genes that warrant clinical investigation but would not be

© 2014 Nature America, Inc. All rights reserved.

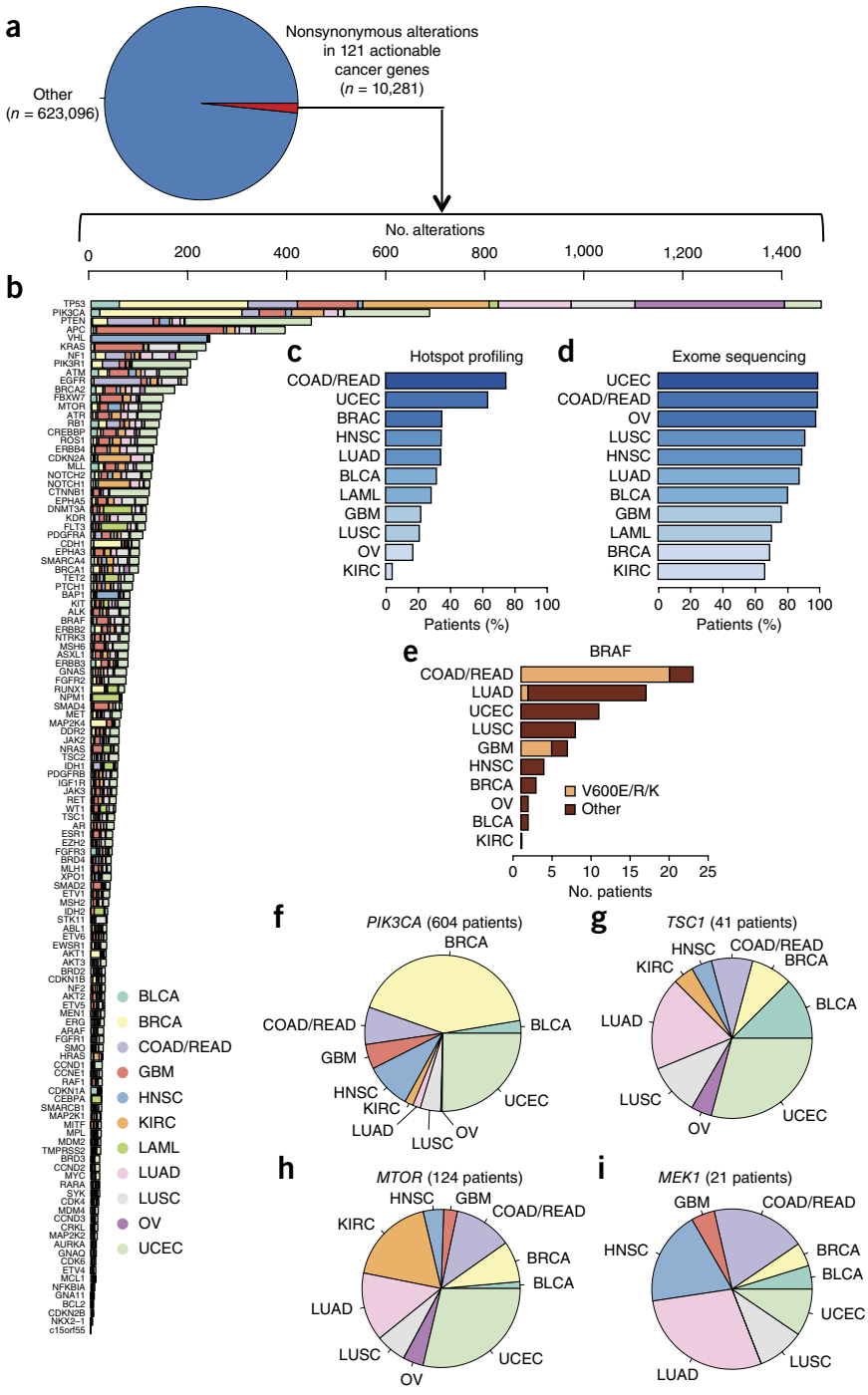


Figure 5 Alterations in clinically relevant genes across 12 tumor types. (**a**,**b**) Examination of mutations and indels in 3,277 patients representing 12 tumor types reveals a long tail of the frequency distribution of alterations in clinically relevant genes that warrant further exploration across 12 tumor types. (**c**,**d**) 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. (**e**) Hotspot alterations in known cancer genes occur at low frequencies in unexpected tumor types. (**f**–**i**) Alterations in emerging genes with potential clinical relevance are observed across tumor types: *PIK3CA* (**f**), *TSC1* (**g**), *MTOR* (**h**), *MEK1* (**i**). For a key to the tumor types, see **Supplementary Table 4**.

apparent with smaller, single-tumor cohorts. This is demonstrated by a series of observations involving genes in the phosphatidylinositide 3-kinase/mTOR signaling pathway. In *PIK3CA*, both hotspot alterations and those requiring preclinical and clinical evaluation were observed in 604 patients (**Fig. 5f**). Somatic mutations in *TSC1*, which have been implicated in everolimus (Afinitor) sensitivity in urothelial carcinomas[38,39], were noted in 41 patients with a diverse array of tumor types (**Fig. 5g**). Somatic alterations in *MTOR* itself, which may predict response and/or resistance to rapamycin analogs or mTOR catalytic domain inhibitors[40], were observed in 124 patients (**Fig. 5h**).

Finally, comprehensive profiling may also be useful for identifying patients who may be intrinsically resistant to certain therapies. For instance, 21 therapy-naive patients harbored *MEK1* somatic mutations (**Fig. 5i**); a subset of these mutations may predict resistance to RAF and/or MEK inhibitors in specific clinical contexts[41,42]. Critically, RAF and MEK inhibitors are currently being studied in numerous cancer types, so prospective knowledge of *MEK1* status may affect treatment selection for patients with *MEK*1 mutations. Broadly, these results demonstrate how global surveys of mutational patterns in clinically relevant genes may have an impact on clinical trial design and treatment selection.

DISCUSSION

In contrast to previous studies driven by a single cancer type or data type, we systematically evaluated patient survival prediction from different molecular data types and described the potential prognostic and/or therapeutic relevance revealed across multiple cancer types, raising several important issues related to the potential clinical utility of large-scale molecular data.

Currently, only a few gene expression–based molecular prognostic markers have been established in clinical practice. For the cancers surveyed here, none of the previously reported gene expression signatures are routinely used in current clinical practice in lung and kidney cancer. For GBM, although the status of a few molecular markers (e.g., MGMT promoter methylation) is frequently ordered for patients, that finding exerts limited influence on clinical decision making[43]. For OV, CA125 is the only marker accepted for clinical use[44]. Our systematic assessment helped address one key issue related to the lack of prognostic markers with clinical utility: statistical significance versus magnitude difference. Across the four TCGA patient cohorts, clinical variables appeared to be the most informative resources (C-index: 0.624–0.754); and molecular data alone often (9 out of 18 cases) had statistically significant predictive power above a random guess (C-index: 0.544–0.718). Given the clinical-variable-only models,

© 2014 Nature America, Inc. All rights reserved.

incorporating molecular data statistically boosts the model performance (5 out of 18 cases) in three cancer types, including mRNA, miRNA and protein expression, especially their molecular subtype information. However, the absolute magnitude gains were very limited (Somers' D, 2.2~22.9%; a 2.2% gain in Somers' D corresponds to a 2.2% increase of rank correlation coefficient between the predicted risk score and the actual survival of the patients), suggesting that the information content of clinical variables and molecular data are largely redundant in terms of patient survival stratification. This echoes the observation that the number of cancer prognostic molecular markers in clinical use is pitifully small, despite decades of protracted and tremendous efforts[45,46]. Currently, many investigators make conclusions about the utility of their markers of interest by heavily relying on *P* value rather than the size of the difference in patient outcomes[47]. Our study calls attention to the criteria of magnitude difference that should be emphasized in future publications of tumor prognostic markers.

Another important related issue is reliability and reproducibility. The literature of tumor biomarkers is plagued by publication bias and selective and/or incomplete reporting[46], which poses an acute challenge in the post-genomic era. In this regard, we have developed an open-access model-assessment platform for TCGA pan-cancer survival prediction, which will (i) reduce the barrier to analyzing TCGA data by providing access to well-curated, computable data sets used as inputs to all models in our study; (ii) increase the transparency and reproducibility of prognostic models by providing our models as re-runnable source code and providing this capability to other researchers; and (iii) improve the objectivity and rigor of future model assessments by providing a baseline set of model scores based on predefined criteria and evaluation scores posted on a real-time, publically available "leaderboard." Such an effort not only helps improve prognostic models though a community-based challenge[48,49] but also ensures transparency and reproducibility for tumor biomarker identification. We expect to seed such a community effort to release the whole data set, prognostic models and evaluation criteria for future studies of clinically usable prognostic models.

By exploring the spectrum of clinically actionable somatic alterations among 12 tumor types, we identified multiple instances where alterations in clinically relevant genes were observed in enough patients to rationalize clinical trial development across tumor types. This was true even if these alterations were rarely observed in any single tumor type, leading to so-called "bucket" trials. We also revealed how the potential applications of precision oncology can be expanded to numerous additional patients with more extensive forms of profiling. As the number of identified clinically actionable cancer genes and alterations continues to rise[8], prospective genomic profiling may inform individualized treatment plans for patients with metastatic or localized disease, as these patients are guided toward genomically driven clinical trials. Notably, many of the alterations observed in clinically actionable cancer genes have either no known functional effect that may be consistent with a clinical action or even have the converse effect. Through our cross-tumor analysis, we expect that many of these alterations will emerge in preclinical and clinical studies, thereby informing their relative impact in specific clinical settings from a predictive/prognostic standpoint when linked to relevant clinical outcomes.

Although our study provides important insights into the translation of biological data into clinical utility, it has some limitations. First, we employed purely data-mining approaches to prognostic modeling. Such a practice comes with a cost: we may miss some informative individual features that could be identified by a candidate gene approach driven by prior knowledge. Second, we did not analyze somatic mutations for prognostic utility because the mutation data are binary and sparse across the patient cohorts. New methods should be developed for assessing the prognostic power of large-scale mutation data. Third, effectively combining multiple types of molecular data remains a technical challenge owing to the overfitting issue and widespread co-linearity of large-scale biological data. Therefore, one important future direction is to build prognostic models that incorporate clinical variables and multiple types of molecular data, which may provide crucial complementary information. In that regard, more effective feature selection strategies should be developed. Finally, because TCGA patient samples were collected from multiple source sites for the purpose of comprehensive molecular profiling and were characterized at different centers, this practice may introduce both heterogeneity and bias. In addition, the resulting clinical annotation of patient samples may not be as rigorous and complete as those obtained from clinical trials. Therefore, further efforts, especially independent validations with clinical trial-grade follow-up, are crucial for assessing our findings from TCGA data.

METHODS

Methods and any associated references are available in the online version of the paper.

Accession codes. All core clinical and genomic/proteomic data used to construct survival models, as well as the training and test data set splits, are available at the Synapse homepage of our project (accession number syn1710282, doi:10.7303/syn1710282). The full Pan-Cancer data set is available at the Synapse Pan-Cancer home page (accession number syn300013, doi:10.7303/syn300013). Dichotomized survival data were deposited in Synapse (syn1748545).

Note: Any Supplementary Information and Source Data files are available in the online version of the paper.

ACKNOWLEDGMENTS

We gratefully acknowledge contributions from the TCGA Research Network and the TCGA Pan-Cancer Analysis Working Group (contributing consortium members are listed in the **Supplementary Note**). The TCGA Pan-Cancer Analysis Working Group is coordinated by J.M. Stuart, C. Sander and I. Shmulevich. This study was supported by the National Institutes of Health (CA143883 to G.B.M. and J.N.W., CA175486 to H.L., and CCSG grant CA016672); UTMDACC – G.S. Hogan Gastrointestinal Research Fund, NCI-MDACC Uterine SPORE career development award to H.L. and the Lorraine Dell Program in Bioinformatics for Personalization of Cancer Medicine to J.N.W.; Dana-Farber Leadership Council to E.M.V.A.; Conquer Cancer Foundation to E.M.V.A.

AUTHOR CONTRIBUTIONS

L.A.G., A.A.M., G.G. and H.L. conceived and designed the study; Y.Y., E.M.V.A., L.O., N.W., A. A.-M., A.S., L.A.B., Y.X., K.R.H., L.D., L.H., X.H., M.S.L., J.N.W., J.M.S., G.B.M., L.A.G., A.A.M., G.G. and H.L. performed data analysis; Y.Y., E.M.V.A., A.A.M. and H.L. wrote the manuscript with input from other authors; and H.L. supervised the whole project.

COMPETING FINANCIAL INTERESTS

The authors declare competing financial interests: details are available in the online version of the paper.

Reprints and permissions information is available online at http://www.nature.com/reprints/index.html.

1. Cancer Genome Atlas Research Network. Comprehensive genomic characterization defines human glioblastoma genes and core pathways. *Nature* **455**, 1061–1068 (2008).
2. Cancer Genome Atlas Research Network. Integrated genomic analyses of ovarian carcinoma. *Nature* **474**, 609–615 (2011).
3. Cancer Genome Atlas Research Network. Comprehensive molecular characterization of human colon and rectal cancer. *Nature* **487**, 330–337 (2012).

© 2014 Nature America, Inc. All rights reserved.

4. Cancer Genome Atlas Research Network. Comprehensive genomic characterization of squamous cell lung cancers. *Nature* **489**, 519–525 (2012).
5. Cancer Genome Atlas Research Network. Integrated genomic characterization of endometrial carcinoma. *Nature* **497**, 67–73 (2013).
6. Cancer Genome Atlas Research Network. Comprehensive molecular characterization of clear cell renal cell carcinoma. *Nature* **499**, 43–49 (2013).
7. Weigel, M.T. & Dowsett, M. Current and emerging biomarkers in breast cancer: prognosis and prediction. *Endocr. Relat. Cancer* **17**, R245–R262 (2010).
8. Garraway, L. Genomics-driven oncology: framework for an emerging paradigm. *J. Clin. Oncol.* **31**, 1806–1814 (2013).
9. MacConaill, L. *et al.* Profiling critical cancer gene mutations in clinical tumor samples. *PLoS ONE* **4**, e7887 (2009).
10. Berchuck, A. *et al.* Patterns of gene expression that characterize long-term survival in advanced stage serous ovarian cancers. *Clin. Cancer Res.* **11**, 3686–3696 (2005).
11. Douillard, J.Y. *et al.* Adjuvant vinorelbine plus cisplatin versus observation in patients with completely resected stage IB-IIIA non-small-cell lung cancer (Adjuvant Navelbine International Trialist Association [ANITA]): a randomised controlled trial. *Lancet Oncol.* **7**, 719–727 (2006).
12. Heng, D.Y. *et al.* External validation and comparison with other models of the International Metastatic Renal-Cell Carcinoma Database Consortium prognostic model: a population-based study. *Lancet Oncol.* **14**, 141–148 (2013).
13. Johnson, D.R. & O'Neill, B.P. Glioblastoma survival in the United States before and during the temozolomide era. *J. Neurooncol.* **107**, 359–364 (2012).
14. Harrell, F., Lee, K. & Mark, D. Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Stat. Med.* **15**, 361–387 (1996).
15. Tibshirani, R. Regression shrinkage and selection via the lasso. *J. R. Stat. Soc., B* **58**, 267–288 (1996).
16. Ishwaran, H., Kogalur, U.B., Blackstone, E.H. & Lauer, M.S. Random survival forests. *Ann. Appl. Stat.* **2**, 841–860 (2008).
17. Brunet, J.-P., Tamayo, P., Golub, T. & Mesirov, J. Metagenes and molecular pattern discovery using matrix factorization. *Proc. Natl. Acad. Sci. USA* **101**, 4164–4169 (2004).
18. Shih, K. *et al.* A microRNA survival signature (MiSS) for advanced ovarian cancer. *Gynecol. Oncol.* **121**, 444–450 (2011).
19. Jänne, P.A. *et al.* Selumetinib plus docetaxel for KRAS-mutant advanced non-small-cell lung cancer: a randomised, multicentre, placebo-controlled, phase 2 study. *Lancet Oncol.* **14**, 38–47 (2013).
20. Ohashi, K. *et al.* Characteristics of lung cancers harboring NRAS mutations. *Clin. Cancer Res.* **19**, 2584–2591 (2013).
21. Falconer, J.S. *et al.* Acute-phase protein response and survival duration of patients with pancreatic cancer. *Cancer* **75**, 2077–2082 (1995).
22. Kallakury, B.V. *et al.* Increased expression of matrix metalloproteinases 2 and 9 and tissue inhibitors of metalloproteinases 1 and 2 correlate with poor prognostic variables in renal cell carcinoma. *Clin. Cancer Res.* **7**, 3113–3119 (2001).
23. Antoon, J.W. *et al.* Altered death receptor signaling promotes epithelial-to-mesenchymal transition and acquired chemoresistance. *Sci. Rep.* **2**, 539 (2012).
24. Faragalla, H. *et al.* The clinical utility of miR-21 as a diagnostic and prognostic marker for renal cell carcinoma. *J. Mol. Diagn.* **14**, 385–392 (2012).
25. Zaman, M.S. *et al.* Up-regulation of microRNA-21 correlates with lower kidney cancer survival. *PLoS ONE* **7**, e31060 (2012).
26. Khella, H. *et al.* miR-192, miR-194 and miR-215: a convergent microRNA network suppressing tumor progression in renal cell carcinoma. *Carcinogenesis* **34**, 2231–2239 (2013).
27. Liang, S. *et al.* MicroRNA let-7f inhibits tumor invasion and metastasis by targeting MYH9 in human gastric cancer. *PLoS ONE* **6**, e18409 (2011).
28. Liu, Y., Yin, B., Zhang, C., Zhou, L. & Fan, J. Hsa-let-7a functions as a tumor suppressor in renal cell carcinoma cell lines by targeting c-myc. *Biochem. Biophys. Res. Commun.* **417**, 371–375 (2012).
29. Ni, Y. *et al.* MicroRNA-143 functions as a tumor suppressor in human esophageal squamous cell carcinoma. *Gene* **517**, 197–204 (2012).
30. Noguchi, S. *et al.* MicroRNA-143 functions as a tumor suppressor in human bladder cancer T24 cells. *Cancer Lett.* **307**, 211–220 (2011).
31. Sakurai, T. *et al.* The enhancer of zeste homolog 2 (EZH2), a potential therapeutic target, is regulated by miR-101 in renal cancer cells. *Biochem. Biophys. Res. Commun.* **422**, 607–614 (2012).
32. Mermel, C. *et al.* GISTIC2.0 facilitates sensitive and confident localization of the targets of focal somatic copy-number alteration in human cancers. *Genome Biol.* **12**, R41 (2011).
33. Shi, L. *et al.* The MicroArray Quality Control (MAQC)-II study of common practices for the development and validation of microarray-based predictive models. *Nat. Biotechnol.* **28**, 827–838 (2010).
34. Van Allen, E.M., Wagle, N. & Levy, M.A. Clinical analysis and interpretation of cancer genome data. *J. Clin. Oncol.* **31**, 1825–1833 (2013).
35. Van Allen, E.M. *et al.* Whole-exome sequencing and clinical interpretation of FFPE tumor samples to guide precision cancer medicine. *Nat. Med.* doi:10.1038/nm.3559 (18 May 2014).
36. Wagle, N. *et al.* High-throughput detection of actionable genomic alterations in clinical tumor samples by targeted, massively parallel sequencing. *Cancer Discov.* **2**, 82–93 (2012).
37. Kim, K.B. *et al.* Phase II study of the MEK1/MEK2 inhibitor trametinib in patients with metastatic BRAF-mutant cutaneous melanoma previously treated with or without a BRAF inhibitor. *J. Clin. Oncol.* **31**, 482–489 (2013).
38. Iyer, G. *et al.* Genome sequencing identifies a basis for everolimus sensitivity. *Science* **338**, 221 (2012).
39. Krueger, D.A. *et al.* Everolimus for subependymal giant-cell astrocytomas in tuberous sclerosis. *N. Engl. J. Med.* **363**, 1801–1811 (2010).
40. Wagle, N. *et al.* Activating mTOR mutations in a patient with an extraordinary response on a phase I trial of everolimus and pazopanib. *Cancer Discov.* **4**, 546–553 (2014).
41. Van Allen, E.M. *et al.* The genetic landscape of clinical resistance to RAF inhibition in metastatic melanoma. *Cancer Discov.* **4**, 94–109 (2014).
42. Wagle, N. *et al.* MAP kinase pathway alterations in BRAF-mutant melanoma patients with acquired resistance to combined RAF/MEK inhibition. *Cancer Discov.* **4**, 61–68 (2014).
43. Holdhoff, M. *et al.* Use of personalized molecular biomarkers in the clinical care of adults with glioblastomas. *J. Neurooncol.* **110**, 279–285 (2012).
44. Sturgeon, C. *et al.* National Academy of Clinical Biochemistry laboratory medicine practice guidelines for use of tumor markers in testicular, prostate, colorectal, breast, and ovarian cancers. *Clin. Chem.* **54**, e11–e79 (2008).
45. McShane, L.M., Altman, D.G. & Sauerbrei, W. Identification of clinically useful cancer prognostic factors: what are we missing? *J. Natl. Cancer Inst.* **97**, 1023–1025 (2005).
46. McShane, L.M. & Hayes, D.F. Publication of tumor marker research results: the necessity for complete and transparent reporting. *J. Clin. Oncol.* **30**, 4223–4232 (2012).
47. Henry, N.L. & Hayes, D.F. Uses and abuses of tumor markers in the diagnosis, monitoring, and treatment of primary and metastatic breast cancer. *Oncologist* **11**, 541–552 (2006).
48. Bilal, E. *et al.* Improving breast cancer survival analysis through competition-based multidimensional modeling. *PLoS Comput. Biol.* **9**, e1003047 (2013).
49. Margolin, A. *et al.* Systematic analysis of challenge-driven improvements in molecular prognostic models for breast cancer. *Sci. Transl. Med.* **5**, 181re1 (2013).

© 2014 Nature America, Inc. All rights reserved.

ONLINE METHODS

Core data set compilation. We downloaded overall survival data and SCNA data from Firehose (https://confluence.broadinstitute.org/display/GDAC/Home). We obtained the clinical variables from TCGA Data Portal (https://tcga-data.nci.nih.gov/tcga/) and the molecular data (including DNA methylation, mRNA, miRNA and protein expression) from the Pan-cancer project (syn300013) on Synapse (http://www.synapse.org). Specifically, molecular data from the following platforms were used in our study. For SCNA, the platform is Affymetrix Genome-Wide Human SNP Array 6.0 (arm-level and focal-level copy number calls were derived from Firehose: https://confluence.broadinstitute.org/display/GDAC/Home). For DNA methylation, the platform is Illumina Infinium Human DNA Methylation 27K (for GBM and OV) or 450K (for KIRC and LUSC, we retained only the probes that most negatively correlated with gene expression according to Firehose). For mRNA expression, the platform was either Agilent 244K Custom Gene Expression G4502A (for GBM and OV) or Illumina HiSeq 2000 RNA Sequencing V2 (for KIRC and LUSC). For miRNA expression, the platform was either Agilent 8 × 15K Human miRNA-specific microarray platform (for GBM and OV) or Illumina Genome Analyzer/HiSeq 2000 miRNA sequencing platform (for KIRC and LUSC). For protein expression, the platform was the MD Anderson Reverse Phase Protein Array (RPPA) Core platform and both total protein and phosphorylated protein were included in this study as distinct molecular features. For each cancer type, we defined the sample intersection across all the platforms as the core sample set. We included neither DNA methylation data in the LUSC core nor protein expression data in the GBM core in order to preserve a statistically sufficient sample size. For each molecular data type, in additional to gene-level features, we included the NMF subtypes from Firehose analyses reported on January 16, 2013. All core clinical and genomic/proteomic data used to construct survival models, as well as the training and test data set splits, are available at the Synapse homepage of our project (accession number syn1710282, doi:10.7303/syn1710282). The full Pan-Cancer data set is available at the Synapse Pan-Cancer home page (accession number syn300013, doi:10.7303/syn300013).

Model training and performance comparison. For each core set, we randomly split the samples into two groups: 80% as the training set and 20% as the test set. On the training set, we first performed a pre-selection step to keep the top significant features correlated with overall survival (univariate Cox model, likelihood ratio test, $P < 0.05$). To obtain better convergence of the training model, we required that the retained feature number did not exceed the number of events (deaths) in the training set. We used two computational methods to train the models: (i) Cox: the Cox proportional hazards model with LASSO for feature selection, and (ii) RSF: random survival forest. The univariate and multivariate Cox models were built with the R package "survival"; the LASSO was performed using the R package "glmnet" and the penalty parameter λ was chosen based on the fivefold cross-validation within the training set; and the RSF models were built using the R package "RandomSurvivalForest" with the recommended default parameters. We then applied the models thereby obtained to the test set for prediction, and calculated the C-index using the R package "survcomp." For each core set, the above procedure was repeated 100 times to generate 100 C-indexes. To compare the performance across different data types, we first chose the better performing method (Cox or RSF) and then used its results based on the Wilcoxon signed rank test to calculate the *P* value (using 0.05 as the significance cutoff).

To assess the predictive power of integrating molecular data with clinical variables, we slightly modified the Cox method to include both clinical and molecular features. We used the clinical features (such as patient age and gender, tumor stage and grade, and Karnofsky performance score, upon availability) that were significantly correlated with survival (likelihood ratio test $P < 0.05$ in both the univariate Cox model and the full model with all clinical variables) as the baseline to build the clinical Cox model. We then combined the gene-level features that better fit the existing model (through performing a feature-selection step against the residuals) or the subtype features with the clinical variables to build a new multivariate Cox model. We performed the RSF method as before.

To evaluate the effect of feature selection, for the data exhibiting striking discrepancies (i.e., clinical + molecular data for GBM and LUSC), we applied different feature-selection methods before RSF: (i) the same LASSO approach as for Cox; (ii) minimal depth variable selection; and (iii) variable hunting. We calculated the C-indexes when applying these new models to the same test sets. To evaluate the effect of sample size, for the molecular models with substantial predictive power (median C-index > 0.6), we conducted a serial sampling of various portions (ranging from 0.2–1, with a increment of 0.1) of the original training samples as the new training set, from which we built the models using the same approach and calculated the C-index when applying these models to the same test set.

Building multiclass classifier from known NMF subtypes. Using the TCGA OV miRNA expression data as the explanatory variables and the three-class NMF subtypes derived from these expression data by Firehose as the response variable, we built the multiclass classifier (multinomial logistic regression model) using a scheme adapted from Yuan *et al.*[50]. We first cleaned out the expression data by retaining the common features between TCGA data and the independent data set and performed sample-wise centering. We then built the multiclass classifier from the TCGA data and evaluated its performance through fivefold cross-validation. During each of the five iterations, feature selection by LASSO (class = "multinomial") and tuning of the penalty parameter λ were based on 80% of the data, and the prediction (using the R package "glmnet") was made to the remaining 20% of the data, where class labels were assigned according to the class with the largest probability. The predictions from the five iterations were combined and the AUCs were calculated by the R package "ROCR." Finally, we trained the classifier using the whole TCGA data and applied it to the independent OV data set for the final prediction.

Analysis of important biological features in the top prognostic models. The Kaplan-Meier survival curves were drawn according to the samples' original NMF subtypes or predicted classes, and the log-rank test *P* values were calculated using the R package "survival." The association of individual features with survival (better or worse) was decided based on the hazard ratio (HR) from the univariate Cox model, where an HR greater than 1 represented a worse prognosis. Wald's test *P* values were used to assess the significance. For expression values obtained through sequencing (i.e., KIRC miRNA expression), the raw values were $\log_2$ transformed before the univariate Cox analysis. The enriched gene pathways were identified through IPA (Ingenuity Systems, http://www.ingenuity.com).

Cross–tumor-type survival prediction. For the cancers that shared the same platforms for the same type of genomic data (e.g., microarray for GBM and OV mRNA expression, RNA-seq for KIRC and LUSC mRNA expression), we first obtained the common features shared by any two cancers in a pairwise manner. Then we trained the Cox model from the shared molecular features of one cancer and applied the model trained to the same 100 test sample sets used in the global analysis of the other cancer. C-indexes were calculated accordingly. To test the effect of sample size, for the OV-KIRC case, we randomly sampled the same number of OV samples as the KIRC training size and repeated the whole analysis. We used the Wilcoxon signed rank test to compare the performance of the model trained with data from different cancers. We further confirmed this result using an independent approach and an independent sample partition (**Supplementary Fig. 11**). The sample partition was done using standard fivefold cross-validation on the KIRC data. For each fold, we trained nonparametric, unregularized Cox proportional hazards models using the glmnet package in R. We repeated the cross-validation process 30 times, randomly choosing a fivefold split each time. Finally, the KIRC patients were stratified using the SCNA signatures derived from OV, and the Kaplan-Meier curves were drawn based on the median risk score.

Dichotomization of survival data. For each cancer type, we dichotomized the censored continuous survival data by assigning a cutoff time (survival milestone) of 1 year for individuals with GBM, 2 years for LUSC, 3 years for OV and 4 years for KIRC. The individuals who lived beyond the cutoff time were labeled as 1; the deceased were labeled as 0. The individuals with survival times that were censored before the cutoff were excluded. The different cutoffs were chosen in order to reach a balance between the event ratio and the sample size. The dichotomized survival data were deposited in Synapse (syn1748545).

© 2014 Nature America, Inc. All rights reserved.

Classification algorithms for dichotomized survival data. For each dichotomized survival data set, we used two pre-selection strategies (ANOVA and shrinking centroids) and eight classification algorithms: diagonal discriminant analysis (DDA), K-nearest neighbor (KNN), discriminant analysis (DA), logistic regression (LR), nearest centroid (NC), partial least square (PLS), random forest (RF) and support vector machine (SVM). The performance was assessed in tenfold cross-validation, and AUCs were calculated as the measurement. For KNN, we varied the number of candidate neighbors from 1 to 9 (odd numbers only) and used the Euclidean distance as the measure of distance. For NC, we assigned equal prior probabilities. For DDA, DA and PLS, we chose the linear discriminant function with equal prior probabilities. For SVM, we chose the radial basis kernel. The parameter C (the cost) ranged from 1 to 1,001, in increments of 100, and gamma ranged from 10^{-10} to 10^{-2}, moving one decimal place per time. For RF, 1,000 trees were used. The other parameters were chosen by default. The number of features after pre-selection ranged from 10 to 50, in increments of 10.

Variability analysis for modeling factors. There are five factors that can potentially affect the performance of binary classification: cancer type, data type (clinical or individual molecular features), pre-selection strategies, the number of features after pre-selection and classification algorithms. We used ANOVA to assess the variability contributed by these factors and their interactions, and the Akaike information criterion in stepwise model selection for significant factors and interactions. The estimated variance components were then divided by their total in order to compare the proportion of variability explained by each modeling factor. To remove the effect of sample size, we performed a size-adjusted analysis, in which we kept the sample size consistent across all cancer types by randomly sampling 77 samples according to the smallest sample (which was LUSC) from KIRC, GBM and OV, and repeated the same procedure for the original dichotomized sets.

Identification of somatic alterations in clinically relevant genes. Somatic mutations and indels called from exome sequencing of matched tumor and normal genome pairs from 12 TCGA projects were aggregated using mutation annotation format (MAF) files from Synapse (syn1710680). Each alteration was ranked for clinical relevance using a heuristic algorithm[34,35]. Clinical actionability was defined at the gene level: any gene that, when somatically altered in cancer, predicted response or resistance to a specific therapy, had diagnostic potential or had prognostic significance, was considered a clinically actionable gene. These genes were derived by a review of the primary literature, consultation with experts and manual curation. A complete list is available at http://www.broadinstitute.org/cancer/cga/target. For the purpose of understanding the distribution of hotspot alterations in *BRAF* and *PIK3CA*, the following definitions were assigned: hotspot alterations in *BRAF* were defined as those leading to V600E, V600K and V600R protein changes. Similarly, hotspot alterations in *PIK3CA* were restricted to those that resulted in E545K and H1047R protein changes. All code for this effort was generated using the R statistical package.

50. Yuan, Y., Xu, Y., Xu, J., Ball, R. & Liang, H. Predicting the lethal phenotype of the knockout mouse by integrating comprehensive genomic data. *Bioinformatics* **28**, 1246–1252 (2012).

© 2014 Nature America, Inc. All rights reserved.

doi:10.1038/nbt.2940