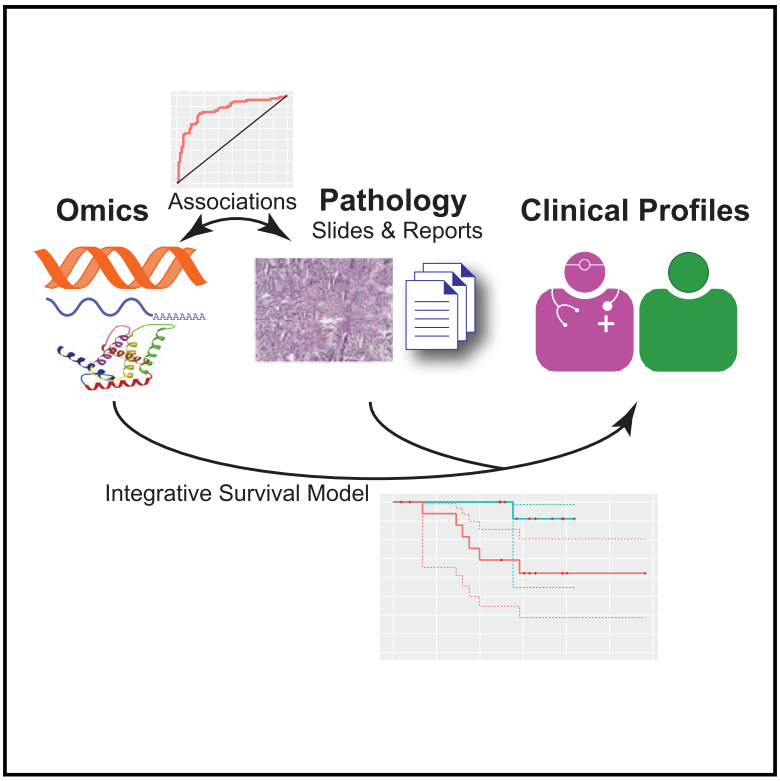
Report

Cell Systems

Association of Omics Features with Histopathology Patterns in Lung Adenocarcinoma

Graphical Abstract



Authors

Kun-Hsing Yu, Gerald J. Berry, Daniel L. Rubin, Christopher Ré, Russ B. Altman, Michael Snyder

Correspondence

mpsnyder@stanford.edu

In Brief

Integrative omics-histopathology analyses identified the gene and protein expression patterns associated with lung adenocarcinoma differentiation. Regularized machine-learning models using both transcriptomics and histopathology information better predicted the survival outcomes of stage I lung adenocarcinoma patients, with the results replicated in an independent cohort.

Highlights

- Gene and protein expression levels predicted lung adenocarcinoma grade
- Quantitative histopathology features correlated with omics classifications
- An integrative omics-pathology model better predicted stage I patients' prognoses
- The improved survival prediction results were replicated in an independent cohort

Yu et al., 2017, Cell Systems *5*, 620–627
December 27, 2017
https://doi.org/10.1016/j.cels.2017.10.014

CellPress

Cell Systems
Report

Association of Omics Features with Histopathology Patterns in Lung Adenocarcinoma

Kun-Hsing Yu,[1,2,3] Gerald J. Berry,[4] Daniel L. Rubin,[1,5,6,7] Christopher Ré,[8] Russ B. Altman,[1,2,8,9] and Michael Snyder[2,10,*]
[1]Biomedical Informatics Program, Stanford University, Stanford, CA 94305-5479, USA
[2]Department of Genetics, Stanford University, Stanford, CA 94305-5120, USA
[3]Department of Biomedical Informatics, Harvard Medical School, Boston, MA 02115, USA
[4]Department of Pathology, Stanford University, Stanford, CA 94305, USA
[5]Department of Biomedical Data Science, Stanford University, Stanford, CA 94305, USA
[6]Department of Radiology, Stanford University, Stanford, CA 94305-5105, USA
[7]Department of Medicine (Biomedical Informatics Research), Stanford University, Stanford, CA 94305-5479, USA
[8]Department of Computer Science, Stanford University, Stanford, CA 94305-9025, USA
[9]Department of Bioengineering, Stanford University, Stanford, CA 94305-4125, USA
[10]Lead Contact
*Correspondence: mpsnyder@stanford.edu
https://doi.org/10.1016/j.cels.2017.10.014

SUMMARY

Adenocarcinoma accounts for more than 40% of lung malignancy, and microscopic pathology evaluation is indispensable for its diagnosis. However, how histopathology findings relate to molecular abnormalities remains largely unknown. Here, we obtained H&E-stained whole-slide histopathology images, pathology reports, RNA sequencing, and proteomics data of 538 lung adenocarcinoma patients from The Cancer Genome Atlas and used these to identify molecular pathways associated with histopathology patterns. We report cell-cycle regulation and nucleotide binding pathways underpinning tumor cell dedifferentiation, and we predicted histology grade using transcriptomics and proteomics signatures (area under curve >0.80). We built an integrative histopathology-transcriptomics model to generate better prognostic predictions for stage I patients ($p = 0.0182 \pm 0.0021$) compared with gene expression or histopathology studies alone, and the results were replicated in an independent cohort ($p = 0.0220 \pm 0.0070$). These results motivate the integration of histopathology and omics data to investigate molecular mechanisms of pathology findings and enhance clinical prognostic prediction.

INTRODUCTION

Lung cancer causes more than 1.4 million deaths per year worldwide, and adenocarcinoma is the most common subtype (Jemal et al., 2011; Siegel et al., 2014). For decades, histopathology evaluation has been the definitive diagnostic method for lung cancer (Collins et al., 2007). However, the underlying molecular mechanisms for histological patterns are not fully understood (Gardiner et al., 2014; Zugazagoitia et al., 2014). In addition, whole-slide histopathology image scanning and high-throughput omics technologies generate terabytes of personal tumor profile per patient, but how to integrate these data to advance precision cancer medicine remain to be explored (Yu and Snyder, 2016).

Histopathology morphology has guided the diagnosis of lung cancer and defined subtypes of lung malignancy (Travis et al., 2011). To diagnose lung cancer, pathologists prepare microscopic slides from tissue samples, stain them with H&E, which non-specifically binds to nuclear acids and proteins, respectively (Fischer et al., 2008). These slides are observed under light microscopy, and the cyto-architectural features define the specific types and subtypes of lung tumors. Studies have shown that certain pathology annotations, such as the level of tumor cell dedifferentiation, are associated with survival outcomes (Harpole et al., 1995). However, this manual evaluation process involves some level of subjectivity (Raab et al., 2005), and it is difficult to integrate these visual findings with terabytes of omics information. Thus, how these visual patterns associated with their underlying biological processes remain largely unknown (Zugazagoitia et al., 2014).

Computer vision algorithms have attained exceptionally good performance for image classification (Danuser, 2011; Lawrence et al., 1997). Previously, investigators have defined many types of quantitative image features, including the size, perimeter, shape, eccentricity, and texture patterns of the cell nuclei and cytoplasm, to analyze pathology images objectively (Beck et al., 2011; Yu et al., 2016b). A number of image features are not easily identified by human evaluators, but they are significantly associated with cancer patients' diagnoses and prognoses (Beck et al., 2011). These results support the clinical utility of quantifying the morphological changes of tumor cells with an automated and objective algorithm.

Moreover, with the advent of the omics (including genomics, transcriptomics, and proteomics) revolution, there is the potential for understanding the molecular biology of histological phenotypes by integrating omics and morphological features of the tumor cells (Haspel et al., 2010; Wall and Tonellato, 2012; Wilkerson et al., 2012; Yuan et al., 2012). Omics studies

have provided insights into the molecular mechanisms of many cancer types (Dong et al., 2016; Snyder, 2016; Yu et al., 2016a; Yu and Snyder, 2016; Zhang et al., 2016), and have characterized the inter-individual differences in disease phenotypes (Clinical Lung Cancer Genome Project [CLCGP] and Network Genomic Medicine [NGM], 2013; Henry et al., 2016; Yu et al., 2017). The systematic integration of histomorphological studies and omics profiles is expected to provide further understandings of tumor cell morphology and potentially more accurate stratification of patients' prognoses (Beck et al., 2011; Liu et al., 2006; Yu and Snyder, 2016; Yuan et al., 2012).

Here we analyze lung adenocarcinoma samples and correlate cell morphology features from histopathology images with genomic, transcriptomic, or proteomic profiles to generate hypotheses about the biological processes associated with morphological changes and the molecular basis of cancer development. In addition, the integration of histopathology features and omics profiles improved the prediction accuracy of patient prognosis, which contributes to personalizing cancer treatment plans (Chin et al., 2011; Revannasiddaiah et al., 2014; Tang et al., 2014).

RESULTS

Patient Characteristics

We analyzed data from a total of 538 lung adenocarcinoma patients previously collected by The Cancer Genome Atlas (TCGA) project (Cancer Genome Atlas Research Network, 2014). These data included genetic variants identified by whole-exome sequencing, tumor transcriptomics profiles characterized by RNA sequencing, tumor proteomics information quantified by reverse-phase protein array, and clinical variables such as tumor stage and survival information. We also obtained digital whole-slide histopathology images of the primary tumors along with the accompanying pathology reports from the same TCGA dataset. We divided the TCGA dataset into distinct training and test sets for machine-learning approaches. To validate our survival prediction methods, we acquired RNA sequencing, histopathology annotations, and survival information of an independent lung adenocarcinoma patient cohort (n = 27) from the Mayo Clinic (Sun et al., 2014). Table S1 shows the patient characteristics of all participants in the TCGA cohorts under study. Table S2 shows the clinical profiles of stage I patients in both TCGA and Mayo Clinic cohorts for survival analysis. The tumor grade, stage I sub-classifications (stage IA and IB), and survival outcomes of stage I adenocarcinoma patients in the two cohorts were not significantly different (p values of 0.1833, 0.4362, and 0.3556, respectively).

We first processed the pathology images by applying an automated algorithm to convert the whole-slide histopathology scans into overlapping tiles, selected the regions of interest and discarded blank background, segmented the cells, and extracted quantitative features from the images, such as the size, shape, intensity distribution, and texture features from the identified tumor cells and tumor nuclei. Since there are tens to hundreds of cells per image tile, we calculated summary statistics including mean, median, percentiles, and standard deviations to capture the distribution of each basic quantitative feature. We next identified pathology grade from pathology reports and collected gene expression and protein expression data generated by RNA sequencing and reverse-phase protein array, respectively. The resulting histopathology and omics profiles served as the input to our machine-learning tasks (Figure 1A).

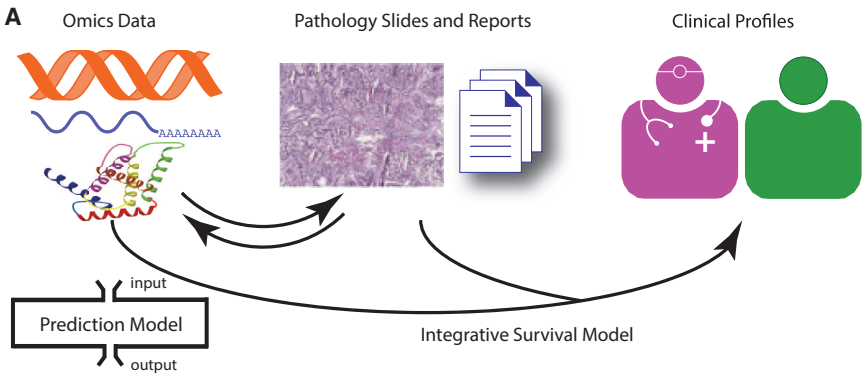
Genes Involved in Cell-Cycle Regulation and Nucleotide Binding Are Predictive of Histological Grade

With an aim of revealing the biological processes underlying tumor differentiation, we first used machine-learning methods to identify the correlations between pathology grade and global gene/protein expression profiles (Figure 1B). To reduce the impact from inter-rater variability on tumor grade, we divided the patient cohort into a higher-grade group (with poorly differentiated or moderately-to-poorly differentiated tumor) and a lower-grade group (with well-differentiated or moderately differentiated tumor) (Barletta et al., 2010), built transcriptomics and proteomics signatures for pathology grade in the training set (n = 300 for transcriptomics; n = 109 for proteomics), and evaluated the prediction models with the held-out test set (n = 128 for transcriptomics; n = 47 for proteomics).

We found that the gene expression profiles of 15 genes predicted the histopathology grade in the held-out test set, with an area under the receiver operating characteristic curve (AUC) of 0.80 ± 0.0067 (Figure 2A). This prediction performance was significantly better than a random classifier ($p < 0.001$), and each of the 15 features was significantly associated with histopathology grade (adjusted $p < 0.01$). The expression levels of the top genes associated with tumor grade are summarized in Figure S1A. All genes highly associated with tumor grade possessed significantly more gene-gene interactions compared with a null model consisting of random genes ($p < 0.0001$, Figures 2C and S2A). (Please see the STAR Methods section for the statistical methods.) KEGG pathway analysis showed that the differentially expressed genes between the two grade groups are enriched in cell-cycle, DNA replication, and p53 signaling pathways. Gene ontology (GO) enrichment analysis also revealed that these genes were highly enriched in mitosis, cell-cycle regulation, and nucleotide binding. Similarly, we identified a proteomics signature that correlated with pathology grade. Our classifiers, using a total of 15 proteins, attained AUCs approximately 0.81 ± 0.0071 in the test set, demonstrating that these protein expression profiles were indicative of pathology grade (Figure 2B). The abundance levels of the proteins indicative of tumor differentiation levels are outlined in Figure S1B. The prediction performance was significantly better than expected by chance ($p < 0.001$). These proteins have significant interactions among one another ($p < 0.0001$; Figures 2C and S2B). GO and KEGG analysis revealed that proteins predictive of tumor grade are enriched in cancer signaling pathways and regulation of cell development, pointing to the regulatory mechanisms related to tumor cell differentiation at the protein level. Taken together, our analyses suggest that genes participating in the cell-cycle and cancer signaling pathways contribute to the levels of tumor cell dedifferentiation.

Correlation of Quantitative Histopathology Features with TP53 Mutation and Histological Sub-classifications

Next, we investigated the associations between quantitative histopathology measurements and omics data, as well as previously established histological sub-classifications. To quantify



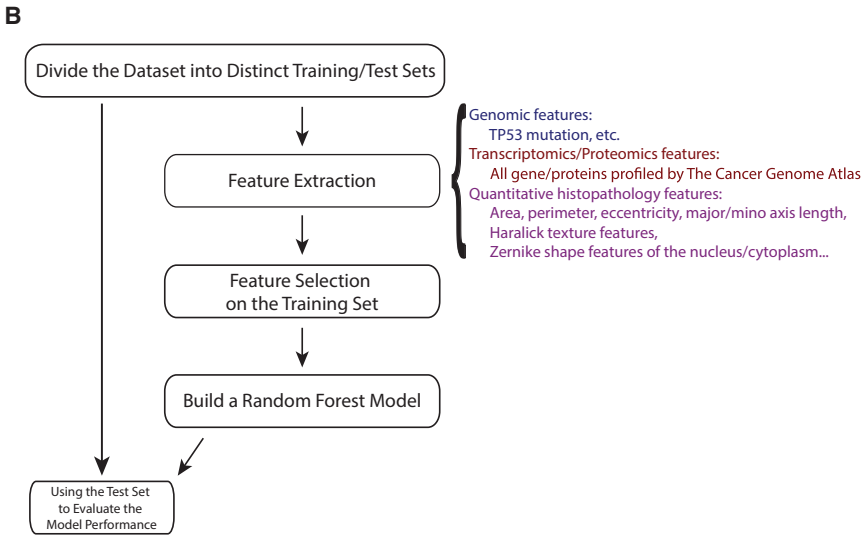


Figure 1. A Summary of Methods

(A) Model for data integration of this study. We processed the genomics, transcriptomics, and proteomics profiles of the primary tumor of lung adenocarcinoma patients and extracted quantitative histopathology features with a fully automated computational algorithm. The associations between functional omics and histopathology profiles were then analyzed to better understand the biology of this cancer. We further utilized both elements to generate an improved clinical prediction framework for lung adenocarcinoma patients. (B) A flow diagram of the machine-learning approach for classification. We divided the datasets into distinct training and test sets, extracted genomic, transcriptomic, proteomic, and histopathology features from the tumor samples, selected the top features, built random forest models, and used the untouched test set to evaluate the model performance.

the histopathology changes, we previously developed an automated method to identify the tumor nucleus and cytoplasm patterns. The extracted features were shown to associate with patient diagnosis and prognosis (Yu et al., 2016b).

TP53 mutation in lung adenocarcinoma has been associated with poorer prognosis (Ahrendt et al., 2003; Gu et al., 2016). We correlated the TP53 mutation status with the established quantitative morphological features. Our results showed that TP53 mutation was significantly associated with the pixel intensity distribution in the cytoplasm, as well as the texture features in the tumor nuclei (adjusted Wilcoxson rank-sum test $p < 0.05$; Table S3). Transcriptomic analysis showed that TP53 mutation was correlated with dysregulation of genes participating in the DNA replication, mismatch repair, and cell-cycle pathways (hypergeometric test Benjamini-Hochberg adjusted $p < 0.05$; Table S4).

We further associated quantitative histological features with sub-classifications of lung adenocarcinoma patients. Previously, researchers defined a few tumor sub-classifications associated with the genomic and transcriptomic patterns of lung adenocarcinoma, including acinar predominant, papillary predominant, and solid predominant tumors (Cancer Genome Atlas Research Network, 2014). When correlating these sub-classifications with quantitative histological features, a texture feature of the tumor nucleus was significantly different among the sub-classifications, after correcting for multiple tests (adjusted Wilcoxon rank-sum test $p = 0.0254$; Table S5). Five image features that quantified the radial distribution of pixels were marginally significant in their associations with these sub-classifications (adjusted Wilcoxon rank-sum test $p = 0.054$), and clustering analysis identified some heterogeneity in patients with the same sub-classification (Figure S3). In addition, 68 quantitative image features were associated with the purity of tumor (Table S6). Despite the wide range of purity score in the TCGA cohort, a least absolute shrinkage and selection operator (LASSO) regression model with the quantitative image features showed a moderate correlation between the histopathology-estimated purity scores and those measured by sequencing (Spearman's correlation coefficient = 0.323; $p < 0.0001$).

Integrative Model for Survival Prediction in Patients with Stage I Lung Adenocarcinoma

Next, we explored the use of omics and histopathology data to build regularized Cox proportional hazards models (Tibshirani, 1997) to predict patient survival. Patients with pathology stage I generally have better survival outcomes than patients with stage II or higher (log rank test $p < 0.001$; Figure 3A). However, the survival outcomes of stage I patients are very diverse and difficult to predict. After being diagnosed with stage I lung adenocarcinoma, more than half of this patient population died within 5 years, but there are approximately 15% of stage I patients who survived 10 years or more after the initial diagnosis. In addition, the clinical distinction between stage IA and stage IB did not reliably distinguish patients with different survival outcomes ($p = 0.878$; Figure 3B), and the differences in overall survival among lung adenocarcinoma patients with stage II or

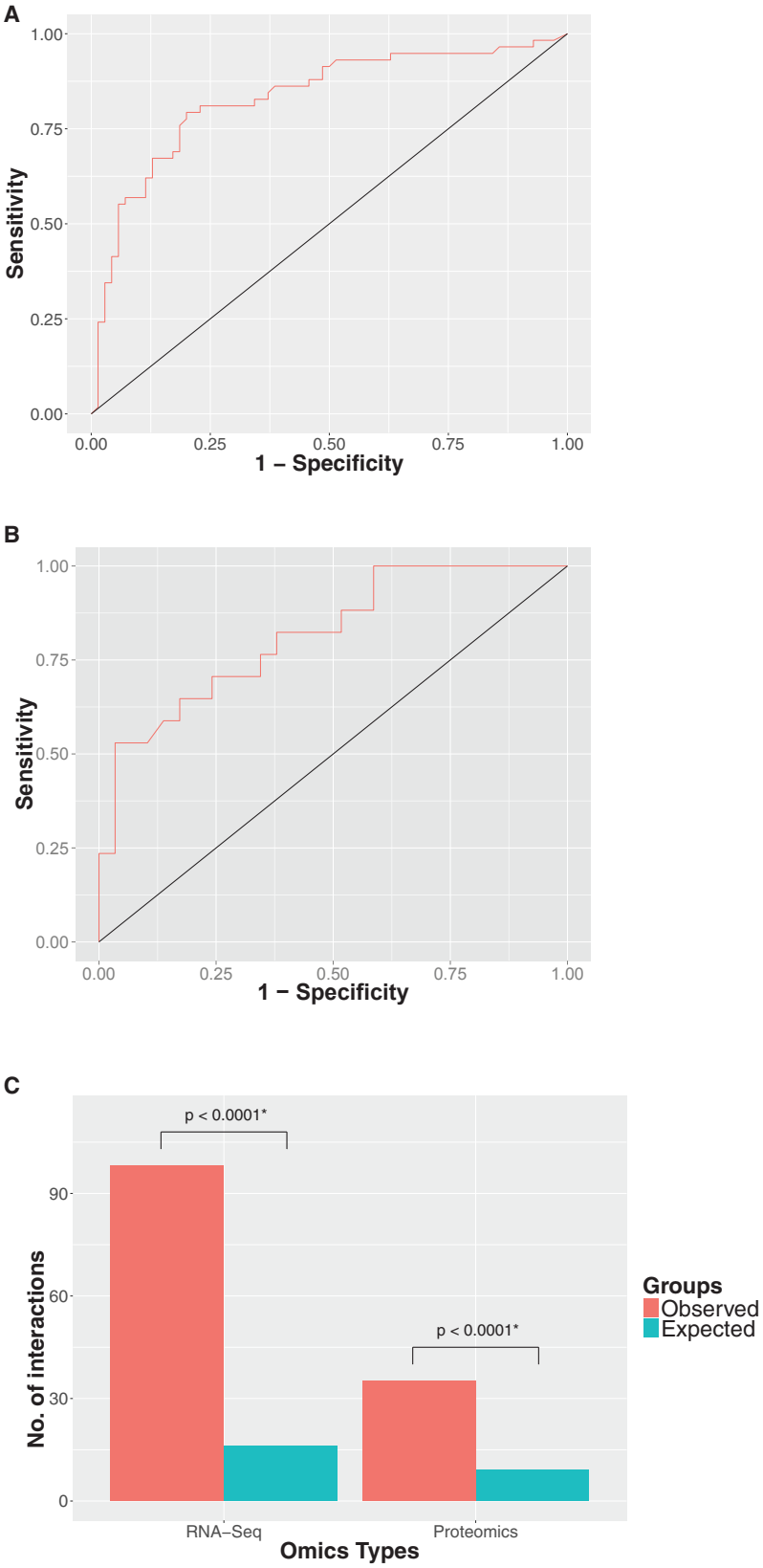
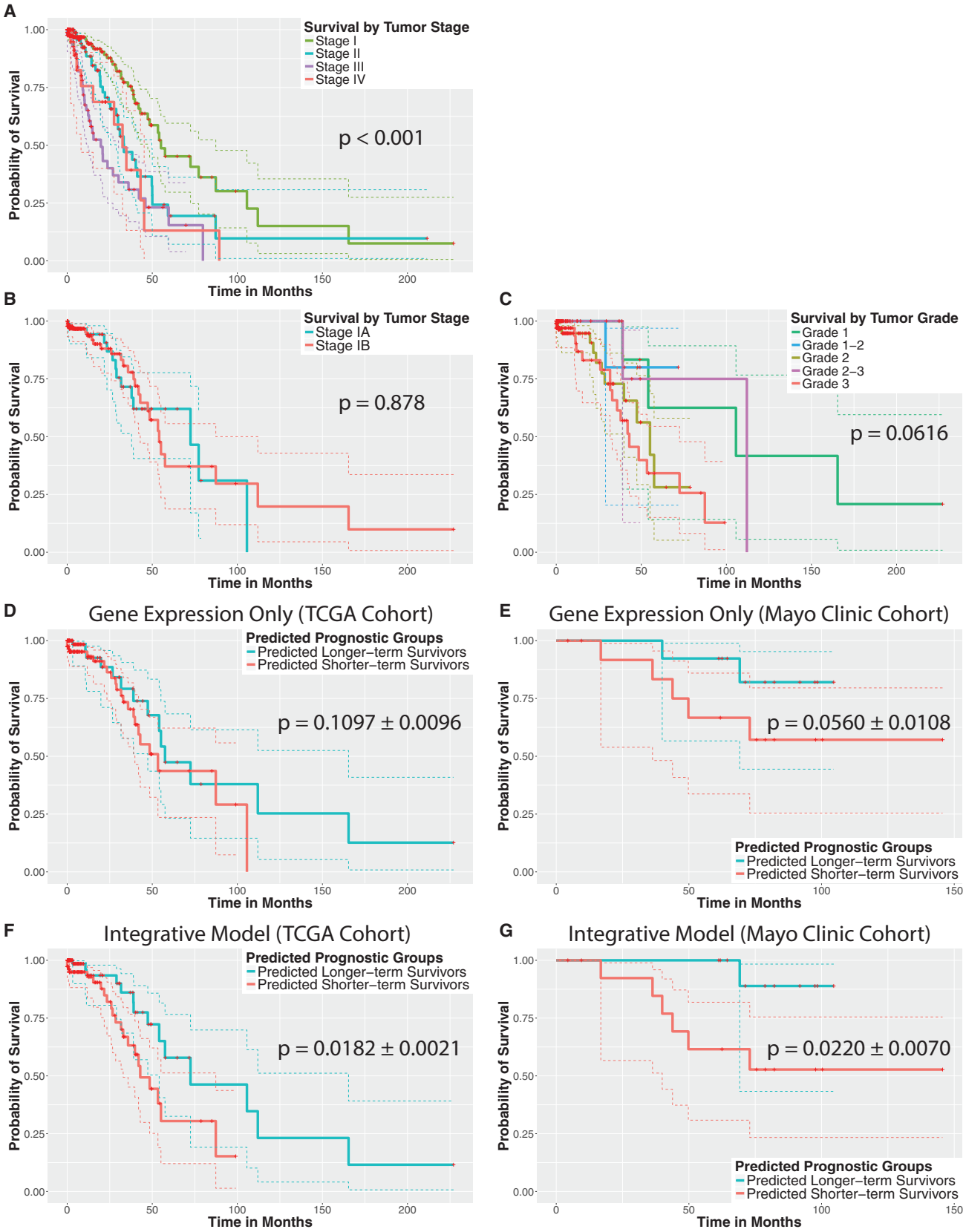


Figure 2. Functional Omics Profiles Predicted the Dedifferentiation Levels of Lung Adenocarcinoma

(A) The expression levels of 15 genes selected by information gain ratio accurately predicted pathology grade, with an area under the ROC curve (AUC) approximately 0.80 ± 0.0067.

(B) Fifteen proteomics features predicted histology grade with good accuracy. A panel of protein markers predicted pathology grade with an AUC 0.81 ± 0.0071.

(C) Dysregulated genes and proteins associated with tumor grade were enriched in gene-gene/protein-protein interactions. The observed numbers of gene-gene/protein-protein interactions and the expected numbers were shown for the transcriptomic and proteomic analyses.



(legend on next page)

higher tumor was not statistically significant in our cohort ($p = 0.139$ among stage IIa, IIb, IIIa, IIIb, and IV; Figure S4A) either. Furthermore, tumor grade alone did not significantly correlate with stage I patient survival ($p > 0.06$; Figures 3C and S4B).

Previously, researchers have proposed gene expression profiles associated with survival outcomes in stage I lung adenocarcinoma patients (Bianchi et al., 2007). However, the reported gene set together with known clinical variables could not reliably distinguish the survival outcomes of stage I patients in either the TCGA or the Mayo Clinic cohort ($p = 0.1097 \pm 0.0096$ and $p = 0.0560 \pm 0.0108$, respectively, adjusted for patient age; Figures 3D and 3E).

We built integrative models by employing gene expression, histopathology grade, and patient age as input features of the regularized Cox proportional hazards model. The integrative model performed better than gene expression or histopathology alone in prognostic prediction ($p = 0.0182 \pm 0.0021$, adjusted for patient age; Figure 3F) on cross-validation in the TCGA cohort. We further replicated this integrative prediction method in the Mayo Clinic cohort ($p = 0.0220 \pm 0.0070$, adjusted for patient age; Figure 3G), which confirmed the improved performance of our integrative method. Since the Mayo Clinic cohort was not involved in building the survival prediction method, these results suggested the generalizability of our prognostic stratification framework. These results indicated the efficacy of combining the information from multiple sources and modalities in improving cancer prognosis prediction.

DISCUSSION

Our results demonstrate promising biological applications and prognostic uses of considering both omics and histopathology features. We investigate the correlation of functional omics profiles with pathology grade, revealing both genes and proteins associated with tumor grade. Pathway analyses on these transcriptomics and proteomics patterns suggested that the level of cancer cell differentiation was related to mitosis and cell division pathways. This finding is consistent with the observation that higher-grade tumors generally have higher mitotic figures, i.e., the number of cells undergoing mitosis observed by light microscopy, and more atypical mitosis (Kadota et al., 2012; Poleri et al., 2003). The slight difference between the enrichments from the gene- and the protein-level analyses might originate from the fact that gene expression levels can be altered by post-transcriptional modifications. Our methods can be used to identify the molecular mechanisms driving other clinically important pathology findings in other complex diseases.

There are several limitations of this work. One limitation is that all patients are from medical centers in the United States. Participants in our cohorts came from 11 participating medical centers across the country but are predominantly Caucasians. Results from other studies have shown different genetic alterations in lung adenocarcinoma in other ethnic groups (Koivunen et al., 2008; Shi et al., 2014). Thus, it would be interesting to systematically analyze the functional omics and histopathology in patients of other ethnicities. In addition, the Mayo Clinic dataset only contains 27 patients, and all of them were never-smokers. Although our integrative methods showed significant improvement in survival prediction in this cohort, the improvement over gene expression or histopathology only model was smaller than that in the TCGA test set. Further validation with a larger cohort is needed.

In summary, this work systematically correlated histopathology patterns with omics findings and developed models to predict survival outcomes of lung adenocarcinoma patients. The developed algorithms are likely extensible to other tumor types or complex diseases.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- CONTACT FOR REAGENT AND RESOURCE SHARING
- METHOD DETAILS
  - Extracting Genomic, Transcriptomic, Proteomic, Histopathology, and Clinical Features of Lung Adenocarcinoma Patients
- QUANTIFICATION AND STATISTICAL ANALYSIS
  - Correlating Omics Profiles with Histopathology Annotations by Machine Learning Methods
  - Genetic Aberrations, Tumor Purity and Their Correlations with Quantitative Histopathology
  - Prognostic Prediction
  - Evaluation of Prognostic Prediction Models
- DATA AND SOFTWARE AVAILABILITY

Figure 3. Integrative Models with Gene Expression Profiles and Pathology Information Predicted the Survival Outcomes of Stage I Lung Adenocarcinoma Patients

Red asterisks indicated censored data.

(A) Lung adenocarcinoma patient survival stratified by tumor stage. Stage I patients generally have better prognoses ($p < 0.001$), but there are significant inter-individual differences in their survival outcomes.

(B) Survival outcomes of stage IA and stage IB lung adenocarcinoma patients. This refinement in the staging system could not distinguish patients with different prognoses in this cohort ($p = 0.878$).

(C) Stage I lung adenocarcinoma patient survival stratified by tumor grade. Grade alone could not predict patient survival reliably ($p = 0.0616$).

(D) A previously reported gene set could not distinguish longer-term survivors ($n = 112$) from shorter-term survivors ($n = 110$) with statistical significance in the TCGA stage I lung adenocarcinoma cohort ($p = 0.1097 \pm 0.0096$).

(E) The same set of genes could not distinguish patient survival in the Mayo Clinic stage I lung adenocarcinoma cohort either ($p = 0.0560 \pm 0.0108$; 13 predicted longer-term survivors; 14 predicted shorter-term survivors).

(F) Integrating pathology with gene expression profiles better predicted patient survival in the TCGA stage I lung adenocarcinoma cohort ($p = 0.0182 \pm 0.0021$; 110 predicted longer-term survivors; 112 predicted-shorter-term survivors).

(G) The improved performance of the integrative survival prediction method is replicated in the Mayo Clinic stage I lung adenocarcinoma cohort ($p = 0.0220 \pm 0.0070$; 11 predicted longer-term survivors; 16 predicted shorter-term survivors).

SUPPLEMENTAL INFORMATION

Supplemental Information includes four figures, six tables, and one data file and can be found with this article online at https://doi.org/10.1016/j.cels.2017.10.014.

AUTHOR CONTRIBUTIONS

K.-H.Y. conceived, designed, performed the analyses, interpreted the results, and wrote the manuscript. G.J.B., D.L.R., C.R., R.B.A., and M.S. interpreted the results and edited the manuscript. C.R., R.B.A., and M.S. supervised the work.

ACKNOWLEDGMENTS

The authors would like to thank Dr. Matt van de Rijn for his valuable advice on the study design and interpretation of results, and Ce Zhang, Feiran Wang, Andrej Krevl, and Rok Sosic for their assistance on the computation framework. K.-H.Y. is a Harvard Data Science Postdoctoral Fellow, Howard Hughes Medical Institute International Student Research Fellow, and Winston Chen Stanford Graduate Fellow. This work was supported in part by grants from the NIH, grant numbers 5U24CA160036, 5P50HG007735, 1U01CA190214, and U01CA142555, and the Defense Advanced Research Projects Agency (DARPA) Simplifying Complexity in Scientific Discovery (SIMPLEX) grant number N66001-15-C-4043 and Data-Driven Discovery of Models contract number FA8750-17-2-0095. The authors thank the AWS Cloud Credits for Research, the Microsoft Azure Research Award, and the Nvidia GPU Grant Program for their support. M.S. is a cofounder and member of the scientific advisory board for Personalis, Sensomics, and Qbio. He is a member of the scientific advisory board of Epinomics and Genapsys.

Received: December 22, 2016
Revised: July 30, 2017
Accepted: October 19, 2017
Published: November 15, 2017

REFERENCES

Ahrendt, S.A., Hu, Y., Buta, M., McDermott, M.P., Benoit, N., Yang, S.C., Wu, L., and Sidransky, D. (2003). p53 mutations and survival in stage I non-small-cell lung cancer: results of a prospective study. J. Natl. Cancer Inst. *95*, 961–970.

Barletta, J.A., Yeap, B.Y., and Chirieac, L.R. (2010). Prognostic significance of grading in lung adenocarcinoma. Cancer *116*, 659–669.

Beck, A.H., Sangoi, A.R., Leung, S., Marinelli, R.J., Nielsen, T.O., van de Vijver, M.J., West, R.B., van de Rijn, M., and Koller, D. (2011). Systematic analysis of breast cancer morphology uncovers stromal features associated with survival. Sci. Transl. Med. *3*, 108ra113.

Bianchi, F., Nuciforo, P., Vecchi, M., Bernard, L., Tizzoni, L., Marchetti, A., Buttitta, F., Felicioni, L., Nicassio, F., and Di Fiore, P.P. (2007). Survival prediction of stage I lung adenocarcinomas by expression of 10 genes. J. Clin. Invest. *117*, 3436–3444.

Breiman, L. (2001). Random forests. Mach. Learn. *45*, 5–32.

Cancer Genome Atlas Research Network. (2014). Comprehensive molecular profiling of lung adenocarcinoma. Nature *511*, 543–550.

Carpenter, A.E., Jones, T.R., Lamprecht, M.R., Clarke, C., Kang, I.H., Friman, O., Guertin, D.A., Chang, J.H., Lindquist, R.A., Moffat, J., et al. (2006). CellProfiler: image analysis software for identifying and quantifying cell phenotypes. Genome Biol. *7*, R100.

Chin, L., Andersen, J.N., and Futreal, P.A. (2011). Cancer genomics: from discovery science to personalized medicine. Nat. Med. *17*, 297–303.

Clinical Lung Cancer Genome Project (CLCGP); Network Genomic Medicine (NGM) (2013). A genomics-based classification of human lung tumors. Sci. Transl. Med. *5*, 209ra153.

Collins, L.G., Haines, C., Perkel, R., and Enck, R.E. (2007). Lung cancer: diagnosis and management. Am. Fam. Physician *75*, 56–63.

Danuser, G. (2011). Computer vision in cell biology. Cell *147*, 973–978.

Dong, F., Davineni, P.K., Howitt, B.E., and Beck, A.H. (2016). A BRCA1/2 mutational signature and survival in ovarian high-grade serous carcinoma. Cancer Epidemiol. Biomarkers Prev. *25*, 1511–1516.

Fischer, A.H., Jacobson, K.A., Rose, J., and Zeller, R. (2008). Hematoxylin and eosin staining of tissue and cell sections. CSH Protoc. *2008*, pdb.prot4986.

Franceschini, A., Szklarczyk, D., Frankild, S., Kuhn, M., Simonovic, M., Roth, A., Lin, J., Minguez, P., Bork, P., von Mering, C., et al. (2013). STRING v9.1: protein-protein interaction networks, with increased coverage and integration. Nucleic Acids Res. *41*, D808–D815.

Friedman, J., Hastie, T., and Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. J. Stat. Softw. *33*, 1–22.

Gardiner, N., Jogai, S., and Wallis, A. (2014). The revised lung adenocarcinoma classification-an imaging guide. J. Thorac. Dis. *6*, S537–S546.

Gu, J., Zhou, Y., Huang, L., Ou, W., Wu, J., Li, S., Xu, J., Feng, J., and Liu, B. (2016). TP53 mutation is associated with a poor clinical outcome for non-small cell lung cancer: evidence from a meta-analysis. Mol. Clin. Oncol. *5*, 705–713.

Harpole, D.H., Jr., Herndon, J.E., 2nd, Wolfe, W.G., Iglehart, J.D., and Marks, J.R. (1995). A prognostic model of recurrence and death in stage I non-small cell lung cancer utilizing presentation, histopathology, and oncoprotein expression. Cancer Res. *55*, 51–56.

Haspel, R.L., Arnaout, R., Briere, L., Kantarci, S., Marchand, K., Tonellato, P., Connolly, J., Boguski, M.S., and Saffitz, J.E. (2010). A call to action: training pathology residents in genomics and personalized medicine. Am. J. Clin. Pathol. *133*, 832–834.

Henry, W.S., Hendrickson, D.G., Beca, F., Glass, B., Lindahl-Allen, M., He, L., Ji, Z., Struhl, K., Beck, A.H., Rinn, J.L., et al. (2016). LINC00520 is induced by Src, STAT3, and PI3K and plays a functional role in breast cancer. Oncotarget *7*, 81981–81994.

Jemal, A., Bray, F., Center, M.M., Ferlay, J., Ward, E., and Forman, D. (2011). Global cancer statistics. CA Cancer J. Clin. *61*, 69–90.

Kadota, K., Suzuki, K., Kachala, S.S., Zabor, E.C., Sima, C.S., Moreira, A.L., Yoshizawa, A., Riely, G.J., Rusch, V.W., Adusumilli, P.S., et al. (2012). A grading system combining architectural features and mitotic count predicts recurrence in stage I lung adenocarcinoma. Mod. Pathol. *25*, 1117–1127.

Koivunen, J.P., Kim, J., Lee, J., Rogers, A.M., Park, J.O., Zhao, X., Naoki, K., Okamoto, I., Nakagawa, K., Yeap, B.Y., et al. (2008). Mutations in the LKB1 tumour suppressor are frequently detected in tumours from Caucasian but not Asian lung cancer patients. Br. J. Cancer *99*, 245–252.

Lawrence, S., Giles, C.L., Tsoi, A.C., and Back, A.D. (1997). Face recognition: a convolutional neural-network approach. IEEE Trans. Neural. Netw. *8*, 98–113.

Li, B., and Dewey, C.N. (2011). RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. BMC Bioinformatics *12*, 323.

Liaw, A., and Wiener, M. (2002). Classification and regression by randomForest. R News *2*, 18–22.

Liu, H., Kho, A.T., Kohane, I.S., and Sun, Y. (2006). Predicting survival within the lung cancer histopathological hierarchy using a multi-scale genomic model of development. PLoS Med. *3*, e232.

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler, D., Gabriel, S., Daly, M., et al. (2010). The genome analysis toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. *20*, 1297–1303.

Poleri, C., Morero, J.L., Nieva, B., Vazquez, M.F., Rodriguez, C., de Titto, E., and Rosenberg, M. (2003). Risk of recurrence in patients with surgically resected stage I non-small cell lung carcinoma: histopathologic and immunohistochemical analysis. Chest *123*, 1858–1867.

Raab, S.S., Grzybicki, D.M., Janosky, J.E., Zarbo, R.J., Meier, F.A., Jensen, C., and Geyer, S.J. (2005). Clinical impact and frequency of anatomic pathology errors in cancer diagnoses. Cancer *104*, 2205–2213.

Revannasiddaiah, S., Thakur, P., Bhardwaj, B., Susheela, S.P., and Madabhavi, I. (2014). Pulmonary adenocarcinoma: implications of the recent advances in molecular biology, treatment and the IASLC/ATS/ERS classification. J. Thorac. Dis. *6*, S502–S525.

Shi, Y., Au, J.S., Thongprasert, S., Srinivasan, S., Tsai, C.M., Khoa, M.T., Heeroma, K., Itoh, Y., Cornelio, G., and Yang, P.C. (2014). A prospective, molecular epidemiology study of EGFR mutations in Asian patients with advanced non-small-cell lung cancer of adenocarcinoma histology (PIONEER). J. Thorac. Oncol. *9*, 154–162.

Siegel, R., Ma, J., Zou, Z., and Jemal, A. (2014). Cancer statistics, 2014. CA Cancer J. Clin. *64*, 9–29.

Simon, N., Friedman, J., Hastie, T., and Tibshirani, R. (2011). Regularization paths for Cox's proportional hazards model via coordinate descent. J. Stat. Softw. *39*, 1–13.

Snyder, M. (2016). Genomics and personalized medicine: what everyone needs to know (Oxford University Press).

Sun, Z., Wang, L., Eckloff, B.W., Deng, B., Wang, Y., Wampfler, J.A., Jang, J., Wieben, E.D., Jen, J., You, M., et al. (2014). Conserved recurrent gene mutations correlate with pathway deregulation and clinical outcomes of lung adenocarcinoma in never-smokers. BMC Med. Genomics *7*, 32.

Szklarczyk, D., Franceschini, A., Wyder, S., Forslund, K., Heller, D., Huerta-Cepas, J., Simonovic, M., Roth, A., Santos, A., Tsafou, K.P., et al. (2015). STRING v10: protein-protein interaction networks, integrated over the tree of life. Nucleic Acids Res. *43*, D447–D452.

Tang, Y., He, Z., Zhu, Q., and Qiao, G. (2014). The 2011 IASLC/ATS/ERS pulmonary adenocarcinoma classification: a landmark in personalized medicine for lung cancer management. J. Thorac. Dis. *6*, S589–S596.

Tibshirani, R. (1997). The lasso method for variable selection in the Cox model. Stat. Med. *16*, 385–395.

Travis, W.D., Brambilla, E., Noguchi, M., Nicholson, A.G., Geisinger, K.R., Yatabe, Y., Beer, D.G., Powell, C.A., Riely, G.J., Van Schil, P.E., et al. (2011). International Association for the Study of Lung Cancer/American Thoracic Society/European Respiratory Society international multidisciplinary classification of lung adenocarcinoma. J. Thorac. Oncol. *6*, 244–285.

Wall, D.P., and Tonellato, P.J. (2012). The future of genomics in pathology. F1000 Med. Rep. *4*, 14.

Warth, A., Muley, T., Meister, M., Stenzinger, A., Thomas, M., Schirmacher, P., Schnabel, P.A., Budczies, J., Hoffmann, H., and Weichert, W. (2012). The novel histologic International Association for the Study of Lung Cancer/American Thoracic Society/European Respiratory Society classification system of lung adenocarcinoma is a stage-independent predictor of survival. J. Clin. Oncol. *30*, 1438–1446.

Wilkerson, M.D., Yin, X., Walter, V., Zhao, N., Cabanski, C.R., Hayward, M.C., Miller, C.R., Socinski, M.A., Parsons, A.M., Thorne, L.B., et al. (2012). Differential pathogenesis of lung adenocarcinoma subtypes involving sequence mutations, copy number, chromosomal instability, and methylation. PLoS One *7*, e36530.

Yu, K.H., Fitzpatrick, M.R., Pappas, L., Chan, W., Kung, J., and Snyder, M. (2017). Omics AnalySIs System for PRecision Oncology (OASISPRO): a web-based omics analysis tool for clinical phenotype prediction. Bioinformatics https://doi.org/10.1093/bioinformatics/btx572.

Yu, K.H., Levine, D.A., Zhang, H., Chan, D.W., Zhang, Z., and Snyder, M. (2016a). Predicting ovarian cancer patients' clinical response to platinum-based chemotherapy by their tumor proteomic signatures. J. Proteome Res. *15*, 2455–2465.

Yu, K.H., and Snyder, M. (2016). Omics profiling in precision oncology. Mol. Cell Proteomics *15*, 2525–2536.

Yu, K.H., Zhang, C., Berry, G.J., Altman, R.B., Re, C., Rubin, D.L., and Snyder, M. (2016b). Predicting non-small cell lung cancer prognosis by fully automated microscopic pathology image features. Nat. Commun. *7*, 12474.

Yuan, Y., Failmezger, H., Rueda, O.M., Ali, H.R., Graf, S., Chin, S.F., Schwarz, R.F., Curtis, C., Dunning, M.J., Bardwell, H., et al. (2012). Quantitative image analysis of cellular heterogeneity in breast tumors complements genomic profiling. Sci. Transl. Med. *4*, 157ra143.

Zhang, H., Liu, T., Zhang, Z., Payne, S.H., Zhang, B., McDermott, J.E., Zhou, J.Y., Petyuk, V.A., Chen, L., Ray, D., et al. (2016). Integrated proteogenomic characterization of human high-grade serous ovarian cancer. Cell *166*, 755–765.

Zugazagoitia, J., Enguita, A.B., Nunez, J.A., Iglesias, L., and Ponce, S. (2014). The new IASLC/ATS/ERS lung adenocarcinoma classification from a clinical perspective: current concepts and future prospects. J. Thorac. Dis. *6*, S526–S536.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Software and Algorithms		
CellProfiler	Carpenter et al., 2006	http://cellprofiler.org/
R randomForest package	Breiman, 2001	https://www.stat.berkeley.edu/~breiman/RandomForests/
R glmnet package	Friedman et al., 2010 Simon et al., 2011	https://web.stanford.edu/~hastie/glmnet/glmnet_alpha.html
Lung cancer feature extraction methods	Yu et al., 2016a, 2016b	https://www.nature.com/articles/ncomms12474
Other		
DNA-Seq, RNA-Seq, and proteomics data of the TCGA cohort	TCGA Data Portal	https://gdc.cancer.gov/
RNA-Seq data of the Mayo Clinic cohort	Sun et al., 2014	https://bmcmedgenomics.biomedcentral.com/articles/10.1186/1755-8794-7-32

CONTACT FOR REAGENT AND RESOURCE SHARING

Further information and requests for resources and reagents should be directed to and will be fulfilled by the Lead Contact, Michael Snyder (mpsnyder@stanford.edu).

METHOD DETAILS

Extracting Genomic, Transcriptomic, Proteomic, Histopathology, and Clinical Features of Lung Adenocarcinoma Patients

A high-quality data set for omics, histopathology, and clinical information of all 538 lung adenocarcinoma patients was obtained from The Cancer Genome Atlas (TCGA) data portal (Cancer Genome Atlas Research Network, 2014). The omics data were processed by standard bioinformatics pipelines (GATK(McKenna et al., 2010) for exome-sequencing, RSEM (Li and Dewey, 2011) for RNA-sequencing, and ArrayPro for reversed phase protein array) by the TCGA consortium. Whole-slide histopathology images, pathology reports, as well as clinical information were acquired for this patient cohort. To validate our clinical prediction method, an independent cohort of stage I lung adenocarcinoma patients (n=27) from Mayo Clinic was identified from the Gene Expression Omnibus (Sun et al., 2014). RNA-sequencing results and clinical variables were obtained and histopathology grade was manually extracted from the associated pathology reports. This study was retrospective and did not involve randomization or blinding. All samples with available data were included in the study.

QUANTIFICATION AND STATISTICAL ANALYSIS

Correlating Omics Profiles with Histopathology Annotations by Machine Learning Methods

Histopathology grade was manually extracted, due to their implications for patients' survival outcomes and their presence in most pathology reports (Barletta et al., 2010; Warth et al., 2012). To reduce the impact of inter-observer disagreement, pathology grades were binarized into a higher-grade group (poorly differentiated or moderately-to-poorly differentiated) or a lower-grade group (well differentiated or moderately differentiated) (Barletta et al., 2010). These group assignments led to relatively balanced groups, with at least 40% of cases in each group.

Breiman's random forest (Breiman, 2001; Liaw and Wiener, 2002) was used to correlate transcriptomics and proteomics profiles with pathology grade. To reduce the risk of overfitting, the information gain ratio of each feature is calculated and only the top features ranked by information gain ratio were selected and included in the model. Wilcoxon rank sum test, which does not rely on assumptions about the probability distributions of the variables, was performed to evaluate the expression difference of each of the selected feature, and the Benjamini-Hochberg procedure was performed to adjust for multiple tests. Unlike conventional machine learning methods that tend to select a minimal number of complementary features, this method ensured that the selected feature sets contain the individual transcriptomics or proteomics patterns correlated with the histopathology annotation of interest, which could be used for enrichment analysis.

To evaluate the performance of the resulting classifiers, the data set was divided into distinct training and test sets, with 80% of the cases in the training and 20% in the test set. There is no overlap between the training and test set. The top features were selected and the models were finalized using the training set. To ensure the robustness of the machine learning framework, the random partition process was repeated 20 times, generating distinct training and test sets each time with no overlaps between training and test data. The machine learning models were built using the training data and evaluated on the test set. The distribution of the area under the receiver operating characteristic curves (AUC) for the classifiers from repeated random partitions was reported.

To identify the biological pathways implicated in the selected lists of genes and proteins, we performed gene ontology (GO) enrichment analysis, KEGG pathway analysis, and network analysis using the String Database Tool (Szklarczyk et al., 2015). To estimate the enrichments in gene-gene interactions, the String Database Tool used a Poisson-Binomial variable to model the number of edges connecting the genes, and calculated the P-value of observing the number of gene-gene interactions under the null hypothesis that this gene set did not possess more gene-gene interactions than a random set (Franceschini et al., 2013). Significant gene-gene interactions often indicated that the selected genes participated in related molecular pathways. The gene expression and protein expression levels associated with tumor grade were visualized using heatmaps, and hierarchical clustering was employed to group genes/proteins with similar expression patterns.

Genetic Aberrations, Tumor Purity and Their Correlations with Quantitative Histopathology

The associations between quantitative histopathology image features and TP53 mutation status were investigated due to the clinical significance of TP53 mutation and the availability of patients with both TP53 mutation information and histopathology image data (Cancer Genome Atlas Research Network, 2014). To extract the quantitative features from the whole slide histopathology images, a fully automated computational framework was employed (Yu et al., 2016b). The framework employed the "IdentifyPrimaryObjects" and the "IdentifySecondaryObject" modules in CellProfiler to identify the lung tumor cells and tumor cell nuclei from the histopathology slides, and used the "Measure Image Area Occupied", "Measure Correlation", Measure Granularity", "Measure Image Intensity", "Measure Image Quality", "Measure Object Size Shape", "Measure Object Intensity", "Measure Object Radial Distribution", "Measure Object Neighbors", and "Measure Texture" modules to extract the size, shape, intensity distribution, and texture features from the identified tumor cells (Carpenter et al., 2006). A total of 694 basic quantitative image features for the tumor cells were extracted using this bioinformatics framework (Data S1). Wilcoxon rank sum test with Benjamini-Hochberg procedure was employed to identify the associations between the quantitative image features and TP53 mutation status.

Similar procedures were employed to characterize the correlations between quantitative histopathology image features and adenocarcinoma sub-classifications as well as tumor purity estimates in the TCGA cohort (Cancer Genome Atlas Research Network, 2014). When correlating with adenocarcinoma sub-classification, analysis of variance with Benjamini-Hochberg procedure was used to account for the multiple classes. Tumor purity estimates were binarized into two groups, where samples with absolute purity call less than 0.5 were categorized as the low purity group and those with absolute purity call greater than or equal to 0.5 were defined as the high purity group. Wilcoxon rank sum test with Benjamini-Hochberg procedure was employed to identify the associations between the quantitative image features and purity groups. A Least Absolute Shrinkage and Selection Operator (LASSO) regression model was built using the quantitative image features, and the Spearman's rank correlation coefficient, a non-parametric measure of rank correlation, was calculated to quantify the association between the predicted purity value and the absolute purity call.

Prognostic Prediction

Survival stratification by tumor stage and grade were evaluated with the log-rank test, which is non-parametric. A set of reported genes associated with stage I lung adenocarcinoma patient survival (Bianchi et al., 2007) were intersected with gene expression levels measured in the TCGA data set. LASSO-Cox proportional hazards models (Friedman et al., 2010; Simon et al., 2011; Tibshirani, 1997) were employed to handle right-censored survival information and avoid overfitting. The LASSO-Cox method used L1 regularization to push the coefficients of uninformative features to zero, which achieved the goal of feature selection while building the survival model.

Current clinical stratification methods using tumor stage and grade as well as a previously-reported gene expression signature (Bianchi et al., 2007) were used as the baseline for comparison. The survival stratifications among all stages as well as between stage IA and stage IB were investigated. In order to better predict the diverse clinical prognoses of stage I adenocarcinoma patients, integrative LASSO-Cox models were built using the previously reported gene expression signature (Bianchi et al., 2007), the pathology grades, and patient age as inputs. The regularization parameters in the LASSO-Cox models were optimized through cross-validation on the training set. After all parameters in the model were finalized, a survival index was calculated for each patient in the training set, and the median survival index in the training set was used as a threshold for distinguishing longer-term survivors from shorter-term survivors. Patients with missing pathology or omics data were discarded from the analysis. All models were adjusted for patient age.

Evaluation of Prognostic Prediction Models

To evaluate the performance of our prediction models in the TCGA cohort, leave-one-out cross-validation was employed. The log-rank test was used to determine the difference in survival outcomes between the predicted groups.

To further validate the survival model, an independent cohort from Mayo Clinic (Sun et al., 2014) was obtained and the gene expression and histopathology profiles of each patient in this cohort were analyzed. The same procedure described above was used to stratify patients in this replication set into two survival groups. Again, the log-rank test was used to determine the survival outcome difference between groups.

DATA AND SOFTWARE AVAILABILITY

The basic quantitative histopathology image features are provided in Data S1.