Unraveling the Molecular Basis of Lung Adenocarcinoma Dedifferentiation and Prognosis by Integrating Omics and Histopathology

Kun-Hsing Yu, MD, PhD[1,2,3], Gerald J. Berry, MD[4], Daniel L. Rubin, MD, MS[2], Christopher Ré, PhD[5], Russ B. Altman, MD, PhD[2,3,6], Michael Snyder, PhD[3]
[1]Department of Biomedical Informatics, Harvard Medical School, Boston, MA; [2]Biomedical Informatics Program, Stanford University, Stanford, CA; [3]Department of Genetics, Stanford University, Stanford, CA; [4]Department of Pathology, Stanford University, Stanford, CA; [5]Department of Computer Science, Stanford University, Stanford, CA; [6]Department of Bioengineering, Stanford University, Stanford, CA

Introduction

Adenocarcinoma accounts for more than 40% of lung malignancy, and microscopic pathology evaluation is indispensable for its diagnosis[1]. However, how histopathology findings relate to molecular abnormalities remains largely unknown[2]. With the advancement of transcriptomics and proteomics profiling technologies, there is the potential for understanding the molecular biology of histological phenotypes by integrating omics and morphological features of the tumor cells. In this study, we identified the molecular mechanisms underlying histopathology aberrations in lung adenocarcinoma and established integrative models for prognosis prediction, which will contribute to personalizing cancer treatment plans.

Methods

We obtained hematoxylin and eosin stained whole-slide histopathology images, pathology grade, stage, RNA-sequencing, and proteomics data of 538 lung adenocarcinoma patients from The Cancer Genome Atlas (TCGA). 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). Breiman's random forest, which can model non-linear relationship, was used to correlate transcriptomics and proteomics profiles with pathology grade. Information gain ratio was employed to select the top features using the data in the training set. KEGG pathway analysis was performed to identify the biological pathway associated with the selected features. The model was trained on 80% of the cases and tested on the untouched 20%. To build survival models for stage I patients, LASSO-Cox proportional hazards models were employed. Current prognostic methods, including tumor stage, grade, and a previously-reported gene expression signature[3] were used as the baseline for comparison. Integrative LASSO-Cox models were built using the previously-reported gene expression signature, pathology grades, and patient age. Leave-one-out cross-validation was used to evaluate the performance of our prediction models in the TCGA cohort. An independent cohort from Mayo Clinic (n=27) was used to further validate the survival model[4]. The same procedure described above was used to predict patients' survival outcomes in this validation set[5].

Results

We found that the 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 1A). Similarly, we identified a proteomic signature that attained AUCs approximately 0.81±0.0071 in predicting the tumor grade in the test set (Figure 1B). Enrichment analysis revealed that proteins predictive of tumor grade were enriched in cancer signaling pathways and regulation of cell development, pointing to the regulatory mechanisms related to tumor differentiation at the protein level.

We further integrated omics and histopathology data to build regularized Cox proportional hazards models to predict stage I patients' survival. Neither the distinction between stage IA and stage IB (P=0.878) nor grade (P=0.158) could accurately distinguish patients with different survival outcomes. A previously reported gene set could not reliably predict the survival outcomes of stage I patients in either the TCGA or the Mayo Clinic cohort (P=0.1097±0.0096 and P=0.0560±0.0108 respectively, adjusted for patient age). We built an integrative histopathology-transcriptomics model to generate better prognostic predictions (P=0.0182±0.0021; Figure 2A) compared with gene expression or histopathology studies alone, and the results were validated in the Mayo Clinic cohort (P=0.0220±0.0070; Figure 2B)[5].

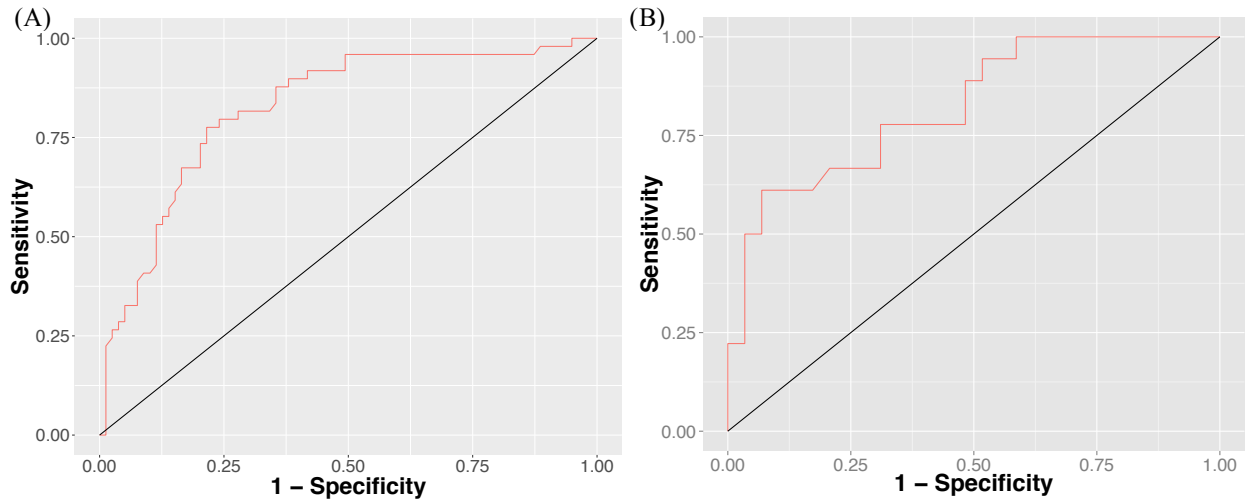

Figure 1. Functional omics profiles predicted the dedifferentiation levels of lung adenocarcinoma. (A) The expression levels of fifteen genes (CCNA2, CDC20, CDCA8, CENPW, CYB5R1, FAM72A, INCENP, KIF1A, KIF18B, MYBL2, RFC4, SPAG5, TACC3, TTK, UBE2C) selected by information gain ratio predicted pathology grade with an AUC of 0.80 ± 0.0067. (B) Fifteen proteomics features predicted pathology grade with an AUC of 0.81 ± 0.0071.

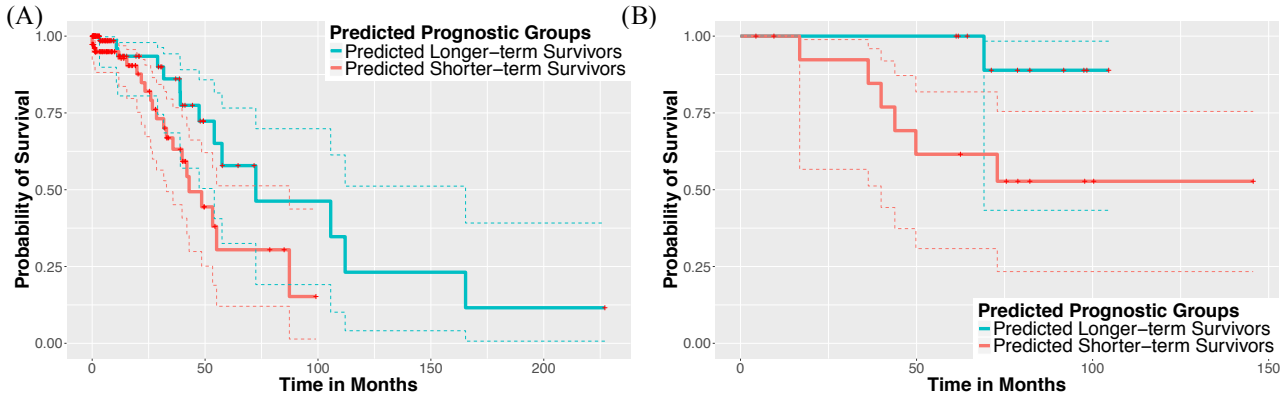

Figure 2. (A) Integrative models with gene expression profiles and pathology information predicted the survival outcomes of stage I lung adenocarcinoma patients in the TCGA cohort (P=0.0182±0.0021, n=222). (B) The results were validated in the Mayo Clinic stage I lung adenocarcinoma cohort (P=0.0220±0.0070, n=27). There is some overlap in the survival curves after 65 months. Red asterisks indicated censored data.

Discussion

Our results demonstrated promising biological applications and prognostic utilities of considering both omics and histopathology features. Pathway analyses on these transcriptomics and proteomics patterns suggested that the level of cancer cell differentiation was related to mitosis and cell division pathways, which were consistent with the observation that higher-grade tumors generally have more atypical mitosis[6]. In addition, an integrative model using gene expression, pathology, and clinical data performed better than each of the components individually, indicating the utility of multi-modality data integration in building survival models. Further studies are needed to compare the performance of different machine learning models and validate the results in large cohorts. Our developed algorithms are likely extensible to other tumor types or complex diseases.

References

1. Collins LG, Haines C, Perkel R, Enck RE. Lung cancer: diagnosis and management. Am Fam Physician. 2007 Jan 1;75(1):56-63.
2. Gardiner N, Jogai S, Wallis A. The revised lung adenocarcinoma classification-an imaging guide. J Thorac Dis. 2014 Oct;6(Suppl 5):S537-46.
3. Bianchi F, Nuciforo P, Vecchi M, Bernard L, Tizzoni L, Marchetti A, Buttitta F, Felicioni L, Nicassio F, Di Fiore PP. Survival prediction of stage I lung adenocarcinomas by expression of 10 genes. J Clin Invest. 2007 Nov;117(11):3436-44.
4. Sun Z, Wang L, Eckloff BW, Deng B, Wang Y, Wampfler JA, Jang J, Wieben ED, Jen J, You M, Yang P. Conserved recurrent gene mutations correlate with pathway deregulation and clinical outcomes of lung adenocarcinoma in never-smokers. BMC Med Genomics. 2014 Jun 4;7:32.
5. Yu KH, Berry GJ, Rubin DL, Ré C, Altman RB, Snyder M. Association of Omics Features with Histopathology Patterns in Lung Adenocarcinoma. Cell Syst. 2017 Nov 13. pii: S2405-4712(17)30484-2. doi: 10.1016/j.cels.2017.10.014.
6. Kadota K, Suzuki K, Kachala SS, Zabor EC, Sima CS, Moreira AL, Yoshizawa A, Riely GJ, Rusch VW, Adusumilli PS, Travis WD. A grading system combining architectural features and mitotic count predicts recurrence in stage I lung adenocarcinoma. Mod Pathol. 2012 Aug;25(8):1117-27