Lung Cancer Computational Biology and Resources

Ling Cai,^{1,2,3,4} Guanghua Xiao,^{1,3,4,5} David Gerber,^{3,4,6} John D. Minna,^{3,6,7,8} and Yang Xie^{1,3,4,5}

¹Quantitative Biomedical Research Center, University of Texas Southwestern Medical Center, Dallas, Texas 75390, USA

²Children's Medical Center Research Institute, University of Texas Southwestern Medical Center, Dallas, Texas 75390, USA

³Harold C. Simmons Comprehensive Cancer Center, UT Southwestern Medical Center, Dallas, Texas 75390, USA

⁴Department of Population and Data Sciences, UT Southwestern Medical Center, Dallas, Texas 75390, USA

⁵Department of Bioinformatics, UT Southwestern Medical Center, Dallas, Texas 75390, USA

⁶Hamon Center for Therapeutic Oncology Research, UT Southwestern Medical Center, Dallas, Texas 75390, USA

⁷Department of Internal Medicine, University of Texas Southwestern Medical Center, Dallas, Texas 75390, USA

⁸Department of Pharmacology, University of Texas Southwestern Medical Center, Dallas, Texas 75390, USA

Correspondence: Yang.Xie@UTSouthwestern.edu

Comprehensive clinical, pathological, and molecular data, when appropriately integrated with advanced computational approaches, are transforming the way we characterize and study lung cancer. Clinically, cancer registry and publicly available historical clinical trial data enable retrospective analyses to examine how socioeconomic factors, patient demographics, and cancer characteristics affect treatment and outcome. Pathologically, digital pathology and artificial intelligence are revolutionizing histopathological image analyses, not only with improved efficiency and accuracy, but also by extracting additional information for prognostication and tumor microenvironment characterization. Genetically and molecularly, individual patient tumors and preclinical models of lung cancer are profiled by various high-throughput platforms to characterize the molecular properties and functional liabilities. The resulting multi-omics data sets and their interrogation facilitate both basic research mechanistic studies and translation of the findings into the clinic. In this review, we provide a list of resources and tools potentially valuable for lung cancer basic and translational research. Importantly, we point out pitfalls and caveats when performing computational analyses of these data sets and provide a vision of future computational biology developments that will aid lung cancer translational research.

Editors: Christine M. Lovly, David P. Carbone, and John D. Minna

Additional Perspectives on Lung Cancer: Disease Biology and Its Potential for Clinical Translation available at www.perspectivesinmedicine.org

Copyright © 2022 Cold Spring Harbor Laboratory Press; all rights reserved; doi: 10.1101/cshperspect.a038273

Cite this article as *Cold Spring Harb Perspect Med* 2022;12:a038273

L. Cai et al.

Lung cancer is the most lethal malignancy, claiming 1.8 million lives yearly worldwide (Thomas-White et al. 2018), with a 5-year survival rate under 20% (Bender 2014). Non-small-cell lung cancer (NSCLC), accounting for 85% of lung cancer cases, is comprised histologically of adenocarcinoma (ADC), squamous cell carcinoma (SQCC), and large-cell carcinoma (LCC). Small-cell lung cancer (SCLC) accounts for the remaining 15% of cases. These histologic subtypes of lung cancer are clinically, pathologically, molecularly, and biologically distinct. In the past couple of decades, we are seeing an explosion of data generated from these different research domains as efforts to understand lung cancer comprehensively for better treatment regimens. Here we review data resources, describe examples of analyses performed with such resources, and introduce a list of user-friendly informatics tools for lung cancer research in these different areas. Table 1 provides a summary of major data sources and tools, along with their URL links for lung cancer research described in this article.

CLINICAL DATA

Data Sources for Lung Cancer Clinical Data

Cancer registries are formal information systems that collect, store, and manage cancer patient data, including patient demographics, cancer characteristics, stage of disease, as well as treatment and outcome. The National Cancer Database (NCDB) and the Surveillance, Epidemiology, and End Results (SEER) are the two largest cancer registries in the United States. The NCDB, sponsored by the American College of Surgeons and the American Cancer Society, incorporates hospital registry data from more than 1500 Commission on Cancer (CoC)-accredited facilities. Currently, the NCDB contains more than 34 million historical records and covers more than 70% of newly diagnosed cancer cases in the United States (Bilimoria et al. 2008; Banerjee et al. 2014). SEER, supported by the Surveillance Research Program (SRP) in the National Cancer Institute's (NCI) Division of Cancer Control and Population Sciences (DCCPS), is a population-based registry. The

latest release, SEER 21, includes data from 21 region-specific registries. It contains more than 10 million historical records and covers about 35% of the U.S. population (Liang 2010).

Beyond cancer registries, there is a trend toward sharing de-identified historical clinical trial data for research purposes. For example, Project Data Sphere (PDS), supported by private funds, provides a platform for analysis of patient-level data from comparative arms of federally funded or industry-sponsored phase III cancer clinical trials (Ha et al. 2018). Thus far, it has assembled data from more than 160,000 cancer patients, from more than 17 data providers. Among these, 26 out of the 204 studies include lung cancer patients. Whereas access to the two cancer registries and PDS are all restricted and contingent upon specific data use agreements, SEER provides a few public-facing interactive tools for basic data exploration (Anders and Huber 2010). PDS allows users to view the metadata for each study, whereas NCDB requires additional credentialing for data access eligibility.

In the private sector, with a focus on real-world data (RWD), Flatiron Health has collected almost 3 million patient records from more than 280 community cancer centers and seven major academic centers in the United States. Data models across more than 20 tumor types for clinical and clinic-genomic data sets include mortality, progression, and response outcome variables. Recent lung cancer studies using Flatiron data have focused on real-world adoption of immunotherapy for advanced NSCLC (Khozin et al. 2019), generating real-world tumor burden end points (Griffith et al. 2019), and association between *KRAS* mutational status and outcomes with first-line immune checkpoint inhibitor in advanced NSCLC (Sun et al. 2021).

Examples of Research Studies Using Public Clinical Data

Substantial knowledge about how clinical, socioeconomic, and epidemiological factors affect lung cancer clinical outcomes may be gleaned from retrospective analyses of these cancer registry data and historical clinical trial data. Here we present a few such examples.

Table 1. Summary of major data sources and tools for lung cancer research described in this article

Type	Domain	Name	Description	URL
Data resource	Clinical data for patients	National Cancer Database (NCDB)	A clinical oncology database sourced from hospital registry data represent more than 70% of newly diagnosed cancer cases nationwide and more than 34 million historical records	www.facs.org/quality-programs/cancer/ncdb
Data resource with tools	Clinical data for patients	Surveillance, Epidemiology and End Results program (SEER)	A clinical oncology database that collects cancer incidence data from population-based cancer registries covering approximately 34.6% of the U.S. population	seer.cancer.gov
Data resource	Clinical data for patients	Project Data Sphere (PDS)	A platform to share, integrate, and analyze phase III comparator arms of historical cancer data, currently with more than 100,000 patients	projectdatasphere.org/projectdatasphere
Tools	Clinical data for patients	Small-cell lung cancer (SCLC) nomogram	Prognostic models for predicting survival in SCLC patients	lce.biohpc.swmed.edu/lungcancer/sclc_nomogram
Tools	Clinical data for patients	Non-small-cell lung cancer (NSCLC) nomogram	Prognostic models for predicting survival in NSCLC patients	lce.biohpc.swmed.edu/lungcancer/nomogram
Data resource	Image data for patient tumors	The Cancer Imaging Archive (TCIA)	An open-access database of medical images for cancer research	www.cancerimagingarchive.net
Data resource	Omics and clinical data for patient tumors	The Cancer Genome Atlas (TCGA)	Data from a landmark National Cancer Institute (NCI) program that performed multidimensional genomic and molecular characterization of more than 10,000 pancancer samples, including more than 1000 lung tumors	portal.gdc.cancer.gov
Tools	Omics and clinical data for patient tumors	TCGA tools	A list of tools developed by TCGA network researchers	www.cancer.gov/about-nci/organization/ccg/research/structuralgenomics/tcga/using-tcga/tools

Continued

Table 1. Continued

Type	Domain	Name	Description	URL
Data resource	Genomics and clinical data from patient tumors	The American Association for Cancer Research (AACR) Project, Genomics Evidence Neoplasia Information Exchange (GENIE)	An aggregation of existing and ongoing next-generation clinical sequencing data and associated pathology reports from multiple cancer centers in the United States, Canada, and Europe	www.synapse.org/#!Synapse:syn7222066/wiki/405659
Tools	Genomics and clinical data from patient tumors	cBioPortal (GENIE)	cBioPortal for exploring AACR GENIE data	genie.cbioportal.org
Data resource with tools	Omics and clinical data for patient tumors	cBioPortal	An open-access, open-source resource for interactive exploration of multidimensional cancer genomics data sets	www.cbioportal.org
Data resource with tools	Transcriptomics and clinical data for patient tumors	The Lung Cancer Explorer (LCE)	A lung cancer-specific data commons that collected transcriptomic and clinical data from more than 56 studies, including more than 6700 patients	lce.biohpc.swmed.edu/lungcancer/dataset.php
Data resource with tools	Omics data for NCI-60 cell lines	CellMiner	A web application for retrieval and integration of the molecular and pharmacological data sets for the NCI-60 cell lines	discover.nci.nih.gov/cellminer
Data resource with tools	Omics data for SCLC cell lines	SCLC CellMiner	A web application for retrieval and integration of the molecular and pharmacological data sets for a large number of SCLC cell lines	discover.nci.nih.gov/SclcCellMinerCDB
Data resource with tools	Omics and clinical data for cell lines	Cancer Cell Line Encyclopedia (CCLE)	A platform for public access to genomic data, analysis, and visualization for more than 1100 cell lines	portals.broadinstitute.org/ccle/data

Continued

Table 1. *Continued*

Type	Domain	Name	Description	URL
Data resource with tools	Transcriptomics and clinical data for patient tumors	Kaplan–Meier (KM) plotter for NSCLC	A lung cancer–specific data commons that collected microarray transcriptomic and clinical and survival data from 1927 NSCLC patients from 10 independent data sets	kmplot.com/analysis/index.php?p=service&cancer=lung
Data resource with tools	Omics and clinical data for cell lines	The Dependency Map (DepMap) portal	A continuation of the CCLE project, currently include CCLE data as well as gene dependency data and compound sensitivity data for CCLE cell lines	depmap.org/portal/download
Tools	Transcriptomics data for patient tumors	Genomic Regression Analysis of Coexpression (GRACE)	TCGA-based gene coexpression tool with confounding copy number impact removed	grace.biohpc.swmed.edu
Tools	Knowledge base for cell lines	Cellosaurus	Cell line knowledge base	web.expasy.org/cellosaurus
Tools	Omics data for cell lines	Functional Data Consistency Explorer (FDCE)	Consistency measure of cell line–derived functional screening data	lccl.shinyapps.io/FDCE

L. Cai et al.

Example 1. What is the prevalence and prognostic significance of a prior cancer diagnosis in individuals with lung cancer?

Historically, lung cancer clinical trials excluded patients with a prior cancer diagnosis, presumably due to a concern that these individuals would have worse clinical outcomes. A series of SEER-based studies found that up to 24% of patients with lung cancer have a prior cancer diagnosis, with the highest rates seen in earlier-stage lung cancer (Gerber et al. 2014; Laccetti et al. 2015, 2016; Pruitt et al. 2017). In propensity-score adjusted analysis, patients with prior cancer had slightly better all-cause (HR 0.93; 95% CI, 0.91–0.94) and lung cancer-specific (HR 0.81; 95% CI, 0.79–0.82) survival. In a simulated clinical trial-eligible population (age <75 yr, no comorbidity, receipt of systemic anticancer therapy for advanced disease), similar findings were noted. Regardless of the stage, timing, and type of prior cancer, a prior cancer diagnosis did not adversely affect all-cause or lung cancer-specific survival. As a result of these findings, prior cancer-related exclusion criteria have been removed from most lung cancer clinical trial protocols.

Example 2. How do type and case volume of health care facilities influence treatment selection and survival for patients with early-stage NSCLC?

In 2019, from analysis of NCDB data, Wang et al. showed that both the type and case volume of the facility where patients receive treatment influence treatment selection and clinical outcome substantially. Early-stage NSCLC cases from academic/research program facilities have the longest survival (median = 16.4 mo) and the highest rate of surgery (75%), while patients from community cancer program facilities had the shortest survival (median = 9.7 mo) and lowest rate of surgery (61%) (Wang et al. 2019a).

Example 3. Can we identify new biological and clinical factors from analysis of cancer registry data to improve lung cancer prognostic models?

In 2018, Yang et al. investigated whether differences in the primary locations of individual lung

ADCs are associated with patient prognosis. With 397,189 ADCs registered in NCDB, it was found that the main bronchus location for ADC is associated with worse prognosis, higher rate of lymph node involvement, and a higher rate of distant metastasis (Yang et al. 2018).

Example 4. How does a certain treatment affect patient outcome?

Does neoadjuvant (preoperative) chemotherapy administration benefit patient survival in stage II and III NSCLC? In 2018, MacLean et al. investigated this question using the NCDB. A survival advantage for use of neoadjuvant chemotherapy (administered prior to surgery) was found when compared to surgery alone, but there was no advantage for neoadjuvant chemotherapy when compared to adjuvant (administered after surgery) chemotherapy (MacLean et al. 2018).

Example 5. How do findings from observational studies challenge current guidelines?

SCLC is a highly aggressive disease, and the current National Comprehensive Cancer Network (NCCN) guidelines recommend surgery be considered only for truly localized disease, which represent only about 5% of total cases. However, several retrospective analyses from the NCDB have suggested that surgery may be beneficial even for late-stage SCLC cases (Combs et al. 2015; Wakeam et al. 2017; Hoda et al. 2018). These findings await prospective trials for confirmation.

Computational Tools

Nomograms are graphical calculating devices used for generating individual probabilities of clinical events by integrating diverse prognostic variables (Balachandran et al. 2015). They are widely applied in prognosis prediction in oncology and other medical fields for risk stratification, treatment planning, and clinical trial enrollment. Recently, two sets of user-friendly online nomogram tools for NSCLC as well as

SCLC were developed for the lung cancer research community (Wang et al. 2018, 2019d).

For SCLC, a nomogram prognostic model was developed using NCDB clinical data from 24,680 patients diagnosed from 2004 to 2011. This model was validated using an independent cohort of 9,700 SCLC patients diagnosed from 2012 to 2013, with an integrated area under the curve of 0.79 from the receiver operating characteristic (ROC) curve, superior to several previous models (Wang et al. 2018).

Whereas several prognostic models have been developed for advanced NSCLC, these models did not undergo external validation against multiple independent data sets. In 2019, Wang et al. developed a nomogram for advanced NSCLC using data from the control arms of four phase III randomized clinical trials, totaling 1,817 patients from the PDS lung cancer data sets (Wang et al. 2019d). The nomogram model was developed from 524 patients in one

study and extensively validated with patient data from three other trials. The online tool for implementing this nomogram in parallel to several other published prognostic models for NSCLC (Fig. 1; Wang et al. 2018, 2019d).

PATHOLOGICAL IMAGING DATA

The diagnosis, histological subtyping, and grading of lung cancer rely on careful histopathological evaluation of biopsy tissue or the resected tumor specimens by hematoxylin and eosin (H&E) staining as well as immunohistochemistry (IHC). The WHO 2015 lung tumor classification system (Travis et al. 2015), amended from the 2004 WHO system (Beasley et al. 2005), provides the most up-to-date hierarchical classification scheme. Such classification serves as the basis for molecular testing and treatment plans.

The morphological characteristics and spatial distribution of cells in the tumor and its mi-

A http://lce.biohpc.swmed.edu/lungcancer/sclc_nomogram/

Prognostic Model for Predicting Survival in Small Cell Lung Cancer Patients

Prognostic Models

Wang Model

Pan Model

Xiao Model

Xie Model for Extensive Stage

Xie Model for Limited Stage

About

Disclaimer

Citation

Input parameters

Age (0 <= value <= 120)

50

Gender

Male

Race

White

Spanish Origin?

FALSE

Charlson/Deyo Score

0

AJCC V8 TNM Stage

IA

Surgery?

No Surgery

Chemotherapy?

No Chemo

Radiation Therapy?

No Radiation

Laterality

Not a paired site

Submit

Please input parameters and press "Submit" button

B <http://lce.biohpc.swmed.edu/lungcancer/nomogram/>

Prognostic Model for Predicting Survival in Non Small Cell Lung Cancer Patients

Prognostic Models

Wang Model

Hoang Model

Paesmans Model

Albain Model

Finkelstein Model

About

Disclaimer

For Researchers

Input parameters

Gender

Male

Histology

Large cell carcinoma

ECOG score

0

☐ Skin metastasis

☐ Liver metastasis

☐ Peritoneum metastasis

Hemoglobin test (g/dL)

0

White blood cell (million/mL)

1

Lymphocyte percentage (%)

0

Submit

Please input parameters and press "Submit" button

Figure 1. Web interface for lung cancer nomogram tools. Different prognostic models for small-cell lung cancer (SCLC) (A) and non-small-cell lung cancer (NSCLC) (B) are provided for users. Upon parameter specification and request submission, a Kaplan–Meier plot will be generated along with a table of survival probability over different time periods.

L. Cai et al.

croenvironment are manifestations of the underlying disease biology and can be captured by histopathological images. However, microscopic examination of tissue slides requires experienced pathologists with sophisticated training to recognize the subtle histopathological patterns in the highly complex tissue images. Even then, the process is time consuming and may be subject to considerable inter- and intra-observer variations (van den Bent 2010; Wang et al. 2019c). Moreover, while slide imaging is becoming a routine clinical procedure, the limited throughput of human pathologists creates a bottleneck in digital pathology for fully extracting and using the information from the pathological images. In this section, we review public pathology image data sources and the informatics methods that are revolutionizing digital pathology analyses. In addition to the methods described here, readers may refer to a review paper from Wang et al. for a more comprehensive summary of other recent applications of deep learning methods in lung cancer pathology image analysis (Wang et al. 2019c).

Public Data Sources for Lung Cancer Pathology Images

Although lung cancer tissue slides and/or digitized pathology images are available in many hospitals and research institutes, there are few pathology image data sets for relatively large, clinically annotated, patient cohorts available in the public domain. The Cancer Imaging Archive (TCIA), funded by the NCI, is one of the largest such collections (Clark et al. 2013). It contains de-identified pathology image data sets for different cancer types and allows users to download entire slide images as .SVS files for research purposes. Users can search TCIA data “collections” by either cancer type (for example, using the key word “lung cancer”) or by image modality (for example, using the key word “pathology”). Within the TCIA, there are several relatively large data sets for lung cancer pathology:

- The Cancer Genome Atlas (TCGA) lung adenocarcinoma (LUAD) data set. TCGA-LUAD data set contains 1,337 tumor tissue

images from 523 LUAD patients. All the images were captured at 20× or 40× magnification and include both frozen and formalin-fixed, paraffin-embedded (FFPE) slides.

- TCGA-lung squamous cell carcinoma (LUSC) data set. TCGA-LUSC data set contains 1,272 tumor tissue images from 511 LUSC patients. As with the TCGA-LUAD data set, all the images in TCGA-LUSC data set were captured at 20× or 40× magnification and include both frozen and FFPE slides.
- The National Lung Screening Trial (NLST) data set. NLST (National Lung Screening Trial Research Team et al. 2011) contains tumor tissue images from 441 lung cancer patients, including both LUAD and LUSC. All the images were captured at 40× magnification from FFPE tissue slides.

It is worth noting that the precise number of images and patients in these data sets may vary depending on the time of data download and quality control criteria.

Examples of Method Development for Lung Cancer Digital Pathology Analysis

Example 1. Extraction of morphological features

In 2016–2017, two milestone studies (Yu et al. 2016b; Luo et al. 2017) demonstrated the feasibility of using morphological features from tumor tissue slides for lung cancer prognosis. These studies opened a promising new direction for applying digital analysis to clinical images. In one study (Luo et al. 2017), objective and quantitative computational approaches were developed to analyze morphological features. Tissue pathology images from TCGA LUAD and LUSC cohorts were analyzed. Features extracted from pathology images were used to develop statistical models to survival. In a follow-up study, these methods and results were further validated in two additional independent data sets (Luo et al. 2019).

Example 2. Tumor region recognition

Following these initial studies, more advanced computational approaches have been developed

to analyze different aspects of histopathological images data. For example, an automated tumor region recognition method was developed and applied to analyze pathology imaging data from LUAD patients from the NLST data set (National Lung Screening Trial Research Team et al. 2011). This method also quantifies tumor topological features using an AI-based model. The resulting tumor region heat maps can help pathologists locate tumor regions in pathology tissue images swiftly and accurately. Furthermore, it was observed that certain tumor topological features are associated with patient prognosis (Wang et al. 2020). This study sheds light on how AI-based digital pathology analysis may help quantify and pinpoint disease-relevant phenotypes.

Example 3. Visualization of cellular spatial organization and interactions

The spatial distributions and organization of different cell types could reveal a cancer cell's growth pattern, its relationships with the tumor micro-environment (TME), and the host immune response, all of which are important characteristics of cancer. To facilitate the study of cellular spatial organization and interactions, ConvPath, a pathology image analysis pipeline was developed (Wang et al. 2019b). ConvPath segments cell nuclei, classifies cell type, and extracts morphologic features to predict prognosis in lung cancer. It may be used to visualize spatial distributions of tumor, stromal, and immune cells in standard H&E-stained lung cancer pathology images. Another deep learning-based computation model, Histology-based Digital (HD)-Staining (Wang et al. 2020), was developed to identify and distinguish nuclei of tumor, stroma, lymphocyte, macrophage, and red blood cells, as well as karyorrhexis from pathology images of lung cancer. Compared to ConvPath, HD-Staining (Fig. 2) is faster and more suitable for whole-slide-image (WSI) analysis. This tool was used to classify cell nuclei and extract 48 cell spatial organization-related features that characterize the TME and predict patient outcomes. The model was developed from the NLST data set, and independently validated in the TCGA-LUAD data set.

Example 4. Quantification of spatial patterns from tissue images

The deep-learning methods discussed above enabled the large-scale identification and classification of individual cells from standard H&E pathology images, which provides new opportunities to study cell spatial organization and interactions. However, due to the complexity and heterogeneity of lung cancer pathology images, it is important to use rigorous statistical methods to characterize the complex distribution and interactions among different cell types. Advanced statistical methods have been developed to quantify such complex spatial distributions (Li et al. 2019c) and cell-cell interactions (Li et al. 2019b) in heterogeneous lung cancer pathology images.

MOLECULAR AND GENOMIC PROFILING DATA FROM PRIMARY TUMORS

Public Data Sources for Lung Cancer Molecular Data

For human NSCLC cases, the most comprehensive, high-dimensional, and high-throughput molecular characterization to date has been performed by the TCGA for LUAD (Cancer Genome Atlas Research Network 2014) and LUSC (Cancer Genome Atlas Research Network 2012). Over the past decade, TCGA has collected molecular profiling data including mutation, copy number, DNA methylation, microRNA, mRNA, reverse phase protein array (RPPA), as well as clinical and pathological slide image data for more than 11,000 tumors from 33 of the most common types of cancer. More than 500 tumor samples were characterized for LUAD and LUSC, respectively.

For SCLC, such data sets are inherently limited by the small proportion of cases for which adequate tissue is available, as only a small minority of cases ever undergo surgical resection, limiting both the feasibility and generalizability of these efforts. The most comprehensive characterization to date was conducted by George et al. (2015) with mutation analysis and gene expression profiling for more than 100 tumors.

L. Cai et al.

Figure 2. An example of computational analysis of digital pathology data of lung cancer. (A) Flow chart of pathology image analysis HD-Staining pipeline. Cell classification result overlaid on the H&E image. (RCNN) Mask Regional-Convolutional Neural Network. Refer to color legends in B for cell type annotation. (B) Nuclei segmentation and classification across whole slide image. (Upper panel) Original pathology image. (Lower panel) Detected and classified nuclei overlay on top of the original pathology image.

Other groups have contributed SCLC gene expression data sets (Rudin et al. 2012; Jiang et al. 2016), but the overall number of studies and the depth of molecular characterization remains low compared to NSCLC.

Prior to the advent of the TCGA era, numerous studies were conducted to examine the gene expression profiles of lung tumors. For example, to generate prognostic models for early-stage LUAD, Shedden and colleagues profiled tumors from more than 440 patients using microarray technology (Director's Challenge Consortium for the Molecular Classification of Lung Adenocarcinoma et al. 2008). With a special interest in examining how germline and placental genes predict lung cancer prognosis, Rousseaux et al. performed expression profiling for more than 300 patient tumors (Rousseaux et al. 2013). Although many such data sets have been deposited to the Gene Expression Omnibus (GEO), substantial harmonization is needed for cross data set comparison of both transcriptomic and clinical data. An important example of this is "KM-Plotter" for NSCLC, which features an online, user-friendly accessible version (Györfy et al. 2013). KM-Plotter contains microarray mRNA expression and clinical data, allowing the user to determine survival (mainly after surgical resection) of NSCLC patients according to mRNA expression profiles of individual or combinations of genes, with additional annotation for histology, stage, gender, and smoking status. Although KM-Plotter is easy to use, its data-pooling strategy has substantial caveats; it obscures study-to-study reproducibility and can inflate statistical significance, leading to false-positive hits (Bravata and Olkin 2001). To address this problem, the lung cancer data commons Lung Cancer Explorer (LCE), featuring a web application populated by a centralized lung cancer database, has been developed and used widely (Cai et al. 2019a). LCE houses the largest collection of lung tumor expression data, culled from 56 studies generated by 23 genome-wide expression profiling platforms encompassing more than 6,700 patients (Cai et al. 2019a), including histological subtyping following the WHO 2015 lung tumor classification system (Travis et al. 2015), as well as patient demographics, diagno-

sis, adjuvant therapy status, smoking status, recurrence-free and overall survival time and status, mutation status of selected key cancer genes, etc. Users are free to download the processed and curated data from LCE's data set web page (Fig. 3).

In the post-TCGA era, NCI has funded a new initiative, the Human Tumor Atlas Network (HTAN), to construct three-dimensional atlases of the cellular, morphological, and molecular features of human cancers over time, including characterizations performed at single-cell resolution (Rozenblatt-Rosen et al. 2020). For lung cancer, HTAN efforts have been carried out for lesions of LUAD (Krysan et al. 2019), subsolid and solid lesions in the LUAD spectrum (Yanagawa et al. 2020), lesions of LUSC (Teixeira et al. 2019), early-stage LUAD (Sinjab et al. 2021), LUAD transition to metastasis (Laughney et al. 2020), and SCLC biopsies (Chan et al. 2020). Outside HTAN, a few other studies have also performed scRNA-seq profiling for NSCLC tumors (Kim et al. 2020; Wu et al. 2021), studied the innate immune landscape in LUAD (Lavin et al. 2017), and investigated therapy-induced evolution of NSCLC (Maynard et al. 2020). These data provide unprecedented resolution for examination of intratumor heterogeneity and the tumor microenvironment. Another new NCI-funded pan-cancer program, the Clinical Proteomic Tumor Analysis Consortium (CPTAC), focuses on proteogenomic characterization. CPTAC has now published data sets for two LUAD cohorts (Chen et al. 2020; Gillette et al. 2020) and one LUSC cohort (Satpathy et al. 2021).

Beyond the data generated from research studies, in the era of precision medicine, routine oncology practice also produces massive quantities of genomic data. The American Association for Cancer Research (AACR)-sponsored project Genomics Evidence Neoplasia Information Exchange (GENIE), a multiphase, multiyear, national and international project, is developing a registry aggregating and linking clinical-grade cancer genomic data with clinical outcomes from tens of thousands of cancer patients treated at participating institutions (AACR Project GENIE Consortium 2017).

L. Cai et al.

Figure 3. Data summary for studies collected into the lung cancer explorer database. This summary describes the data sets and features of the lung cancer database that feeds into the Lung Cancer Explorer. Gene expression data and clinical data were collected from 56 studies that include more than 6,700 patients. For each study and each variable, a pie chart is used to summarize the data. The color scheme for the pie chart sectors is provided below the gridded pie charts. (Figure reprinted from Cai et al. 2019a courtesy of a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution, and reproduction in any medium or format, as long as credit is given to the original source.)

Examples of Research Using These Data for Prognosis and Treatment Response Analyses

These large public data sets have enabled numerous discoveries and research progress in lung cancer. Here, we present two examples of using these molecular data with clinical annotations for lung cancer prognosis and treatment response predictions.

Example 1. Prognostic gene signature for resected NSCLC

Prognostic gene signatures have been developed in many previous studies. However, these published gene expression signatures seldom overlap, and there is a lack of an unbiased evaluation. A systematic review and meta-analysis (Tang et al. 2017) of published mRNA prognostic signatures for resected NSCLC evaluated 1,927 early-stage NSCLC patients collected from 15 studies using three evaluation metrics (hazard ratios, concordance scores, and time-dependent receiver-operating characteristic curves). This study identified numerous promising prognostic gene signatures in lung cancer, independent of clinical risk factors.

Example 2. Predictive gene signature for adjuvant chemotherapy benefits

Conventional adjuvant chemotherapies (ACTs) remain widely used after surgical resection of NSCLC. The ability to identify a subgroup of patients who would benefit from ACT using a certifiable clinical assay (CLIA) might provide an effective means of improving clinical outcomes. A systems biology and integrative analysis approach identified a 12-gene predictive signature for ACT benefits in NSCLC (Xie et al. 2019), which was successfully validated in two independent data sets. The predicted benefit group showed significant improvement in survival after ACT, whereas the predicted nonbenefit group showed no change in survival. Analyses integrating genetic aberration, genome-wide RNAi data, and mRNA expression data were carried out to refine a functional gene set that predicts which resectable patients with NSCLC

will benefit from ACT. To apply this mRNA gene signature in clinical practice, a clinical-grade assay was developed based on analysis of FFPE samples using the NanoString nCounter platform to measure mRNA expression of this 12-gene set (Xie et al. 2019).

Analysis Tools for Lung Cancer Patient Tumor Data

A number of databases—Oncomine (Rhodes et al. 2007), cBioPortal (Gao et al. 2013), KmPlot (Szász et al. 2016), Prognoscan (Mizuno et al. 2009), PROGgene (Goswami and Nakshatri 2014), etc., have been developed to re-annotate and aggregate pancancer data sets and facilitate data exploration and analysis. Similarly, TCGA network researchers and others have also developed a plethora of tools for exploring the TCGA data (Bettencourt and Ribeiro 2008; Zhang et al. 2019). For lung cancer specifically, we have developed a data commons—the LCE (Cai et al. 2019a). Here we provide a few example utilities from LCE to demonstrate how to analyze historical lung cancer molecular data and the pitfalls to avoid.

Example 1. Meta-analysis and inter-study heterogeneity

LCE houses data from 56 studies. Meta-analysis is used to summarize results from multiple data sets. This approach avoids the inflated statistical significance often found for pooled-sample analysis. Importantly, LCE data sets have been specifically harmonized to facilitate meta-analysis. Results from meta-analysis not only combine statistical power from multiple data sets, but also provide heterogeneity statistics to reveal the degree of reproducibility across studies on a per-gene basis. The LCE includes tools for meta-analysis of tumor versus normal expression differences, as well as survival association with gene expression.

Example 2. Data set-specific customized analysis

LCE also allows users to choose a specific data set and perform highly customizable group comparison analysis and survival analysis. For

L. Cai et al.

survival analysis, we recognize that although traditional median-based dichotomization cutoff often provides the best statistical power, in some cases gene expression follows bimodal distribution as a result of oncogenotypes or lineage predisposition. To address this issue, LCE provides model-based clustering for data-driven group specification (Scrucca et al. 2016).

Example 3. Histology-specific analysis

From systematic analysis of LUAD and LUSC, we observed certain histology-specific differences (Cai et al. 2019b). For instance, while tumor versus normal expression differences are highly similar between LUAD and LUSC, survival association with gene expression are much noisier in LUSC than in LUAD. We therefore designed LCE to allow users to include all tumors, or select specific histological subtypes, for meta-analysis and gene coexpression analysis.

Example 4. Whole transcriptome association results

While many existing tools are designed for investigation of single genes, with LCE, results from genome-wide meta-analysis are provided for tumor versus normal expression differences, as well as gene expression survival association analysis (Cai et al. 2019b). We found that for both LUAD and LUSC, major histocompatibility complex II genes, which are usually expressed on “professional” antigen-presenting immune cells, are down-regulated in tumor compared to normal tissues, and lower expression of these genes is associated with worse survival (Fig. 4A; Cai et al. 2019b). This observation supports the concept of “immune evasion” employed by lung cancer and other malignancies. Importantly, these initial findings can be easily validated in independent data sets with the LCE meta-analysis tool (Fig. 4B).

PRECLINICAL MODEL DATA

While patient tumor characterization directly informs clinically relevant cancer biology, preclinical models, including cell lines, patient-derived xenografts (PDXs), and genetically engineered

mouse models (GEMMs) are amenable to experimentation and hence serve as essential tools for lung cancer researchers to gain mechanistic insights into cancer biology, as well as to screen and test for therapeutic agents. Consideration of GEMMs of lung cancer is a large topic unto itself and is covered in the literature by Estape et al. (2021). In this section, we focus on cell lines.

Cell Lines

Cancer cell lines are by far the most widely used preclinical models for lung cancer research. The popularity stems from their properties of being relatively inexpensive, technically simple, cryo-preserveable, easy to expand and distribute, and amenable to large-scale global profiling and in vitro experimentation such as drug screening, while preserving the oncogenotypes and differentiated cell properties of the original tumor (Gazdar et al. 2016). With defined protocols for tumor procurement and culture, the development of methods and media to cultivate tumor-derived cell lines (Gazdar et al. 2010), and new technologies allowing the development of organoid cultures and PDXs in immune-deprived mice directly from patient biopsy specimens or from circulating tumor cells (CTCs), approximately 500 preclinical models of NSCLC and SCLC have been generated over the past several decades (Gazdar et al. 2010). In addition, approximately 100 immortalized human bronchial epithelial cell (HBEC) and small airway epithelial cell (HSAEC) lines have been developed and used in a variety of assays to study lung cancer pathogenesis (Sato et al. 2013). The wide distribution of these preclinical models has had a tremendous impact on lung cancer clinical translational research (Mulshine et al. 2020). Importantly, the establishment of SCLC cell lines and PDXs provide critical tools for the field of SCLC research that would have otherwise been impeded by the short supply of resected tumors (Gazdar et al. 2010; Drapkin et al. 2018).

Resources and Tools for Cell Lines

In large-scale cell-line-based pancancer profiling projects such as NCI-60, GSK, CCLE, and

Figure 4. Systematic meta-analysis using Lung Cancer Explorer (LCE) and investigation of a single gene. (A) Relationship between tumor-normal gene expression difference and survival association in adenocarcinoma (ADC) and squamous cell carcinoma (SQCC). For all genes, summary standardized tumor-normal gene expression differences from the meta-analysis were used as the x-axis values, and summary z-scores from the survival association meta-analysis based on the Cox-PH model were used as the y-axis values. Results from ADC and SQCC were plotted separately. Moderately positive correlation with a Pearson correlation $r = 0.48$ was observed for ADC, while that of SQCC was close to 0. Selected gene sets were highlighted, including the Myc targets from HALLMARK_MYC_TARGETS_V2 in the hallmark gene set collection and major histocompatibility complex (MHC) class II protein complex components in the c5.cc GO cellular component gene set collection. Ellipsoidal boundary wraps were around the 75% highest density/minimum volume for all genes (gray) or selected genes (non-gray colors). *HLA-DOB* that encodes a component of the MHC class II complex was selected for further investigation. (B) Forest plot generated by LCE shows that based on meta-analysis, higher expression of this gene associates with better survival in most of the studies. (TE) Treatment estimate, (seTE) standard error of treatment estimate. (Figure is reprinted from Cai et al. (2019b) under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.)

L. Cai et al.

DepMap, lung cancer cell lines always account for the largest share within the panels of different lineages (Kim et al. 2014). Similar to TCGA, multidimensional characterization has been extensively performed for these lung cancer cell lines. Mutation and copy number alterations have been determined; molecular data, including DNA methylation, microRNA, RNA expression, and RPPA, have been generated (Gu et al. 2017; McMillan et al. 2018; Ghandi et al. 2019); metabolic characterization (metabolomics) (Li et al. 2019a) and parallel isotope-tracing experiments (Chen et al. 2019) were performed; several large-scale compound screens (Barretina et al. 2012; Basu et al. 2013; Iorio et al. 2016; Polley et al. 2016; Yu et al. 2016a; McMillan et al. 2018) and gene dependency screens (McDonald et al. 2017; Tsherniak et al. 2017) were conducted; scRNA-seq was also performed to capture cellular heterogeneity (Kinker et al. 2020). Some of these data can be accessed from integrative cell line databases such as CellMiner (Reinhold et al. 2012), SCLC-CellMiner (Tlemsani et al. 2020), CCLE, and DepMap (Table 1). However, the cell line names used in different studies can sometimes be different. This issue can be resolved by mapping to the Research Resource Identifiers (RRIDs) in Cellosaurus, a cell line knowledge database. Cellosaurus catalogs cell lines from different species and provides extensive information on cell line and donor characteristics, such as synonyms for the cell lines, STR profile for fingerprinting, hierarchical relationships with other cell lines, associated publications, etc. (Bairoch 2018). With redundant characterization from different sources, cross-data consistency is another important issue, especially for functional screening data sets where poor reproducibility was often seen (Safikhani et al. 2016). We have found that compound or gene dependency screening data are often less consistent for those without functionally important targets. To address this issue, we developed a functional data consistency explorer to allow query of consistency measures for individual compound or gene (Cai et al. 2021b).

Considerations for Analyses and Important Knowledge Gaps to Be Addressed Going Forward

Besides patient-derived lung cancer cell lines, large-scale profiling of PDXs is also underway (Woo et al. 2019). scRNA-seq of CTC-derived xenografts (CDXs) has been performed to understand therapy-induced evolution for SCLC (Stewart et al. 2020). A number of lung cancer GEMMs have been cataloged in the mouse tumor biology database (Krupke et al. 2017), some with links to gene expression profiling data. scRNA-seq data sets from healthy lung of human (Travaglini et al. 2020) and mice (Montoro et al. 2018; Plasschaert et al. 2018) have also become available, enabling discovery of lineage-specific properties of tumor that can be tied to its cell-of-origin (Cai et al. 2021a). Although some of these data can be accessed through existing public portals, they will need to be adapted into a user-friendly format to allow widespread use in the lung cancer translational research community.

Whereas these high-dimensional and high-content data from preclinical models provide unprecedented opportunities to accelerate lung cancer research, they also feature limitations. Copy number changes found in cancer can greatly confound gene expression and coexpression analyses (Fig. 5); it would be important to use copy number adjusted gene expression data for association analyses involving cancer transcriptomic data (Cai et al. 2017). Given the availability of multiple data sets that characterize the same sets of cell lines for the same molecular features, consistency must be examined (Fig. 6A,B). Although we have developed a tool for functional screening data, tools for other molecular profiling data remains to be developed. The nascent field of proteomic and epigenomic data integration into multi-omics analyses (Teixeira et al. 2019; Wang and Wang 2019; LaFave et al. 2020; Mullen et al. 2020; Meng and Jin 2021; Zhao et al. 2021) will require new computational tools and approaches. Given the growing importance of immunotherapy and understanding the mechanisms behind lung cancer immune evasion, new clinical and molecular data sets and

A

GRACE

Genomic Regression Analysis of Coordinated Expression

B

Figure 5. *EIF2D* coexpressing genes with standard method or Genomic Regression Analysis of Co-Expression (GRACE). (A) In lung squamous cell carcinoma (LUSC), with standard method, top coexpressing genes for *EIF2D* are dominated by chromosome neighbors. (B) When we used GRACE, a method to control for copy number-associated expression changes in coexpression analysis, the top genes are replaced by ribosomal genes that are non-neighbors but with similar function in protein translation (see grace.biohpc.swmed.edu/Analysis.php for additional details). (Panels A and B reprinted from Cai et al. 2017 under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution, and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source.)

preclinical models are needed to study this crucial area. Similarly, new preclinical models are needed to study tumor cell–tumor microenvironmental interactions and to develop new immune checkpoint blockade approaches. Whereas syngeneic mouse models have been

widely used in these studies, it will clearly be important to study human models and tools for analyses of deposited molecular data sets from “humanized mice.” Finally, although patient data and preclinical model data are readily available, we lack tools for assessing the cross-

Figure 6. Considerations for using lung cancer preclinical model data. (A,B) Consistency between two data sources for different features. Reverse phase protein array (RPPA) measurement for Rb protein is more consistent than that for phospho-Rb (A); for compound sensitivity where inconsistency is often seen, EGFR inhibitor Erlotinib looks more consistent than chemotherapy drug Paclitaxel, and EGFR mutants reported by two sources more reliably predict sensitivity to Erlotinib than mutants reported by a single source (B). (C) Examples of different consistency between preclinical models and patient samples in cancer versus noncancer gene expression difference analyses. Expression of genes encoding aurora kinase A shows the best consistency, where up-regulation in cancer samples and genetically engineered mouse model (GEMM) encoding gene has higher expression in cancer in patient samples and noncancer samples for fatty acid-binding protein 4 samples but not in the cell line settings; the difference between cancer and noncancer samples for fatty acid-binding protein 4 encoding gene expression is observed in patient samples but not in GEMM or cell lines.

model consistency and fidelity of preclinical models in representing human cancer through side-by-side comparison. In Figure 6C, we show examples of cancer versus the normal differential gene expression exhibit different degrees of agreement in cell lines, GEMMs, and patients. For tools built upon preclinical model data sets, these factors should be considered. Doing so will allow researchers to mine rich omics data generated from preclinical models, while discerning the caveats and selecting the best preclinical models for their research studies.

CONCLUDING REMARKS

We have reviewed several large-scale data sets that characterize lung cancer clinically, pathologically, molecularly, and biologically. We have also provided examples of analyses performed with such data sets and tools built around these resources to demonstrate how researchers may explore and exploit existing lung cancer data. Many of the described data sets result from collaborations across multiple publicly funded institutions. The availability of such data to the computational biologists forged the development of informatics tools to help basic and translational researchers fully exploit the data. With the private sector now sequencing tumors and overseeing clinical trials of precision medicine, a vast amount of collected data has yet to be released to the public. We hope these tremendously useful data sets will also become widely accessible and will stimulate algorithm/tool development, and inspire creative secondary analyses from the lung cancer research community for us to better understand and treat lung cancer.

ACKNOWLEDGMENTS

This work is supported by the National Institutes of Health (5R01CA152301, P30CA142543, P50CA70907, R35GM136375, 1R01GM15473, and 1R01CA172211) and the Cancer Prevention and Research Institute of Texas (RP190107 and RP180805). We thank Ms. Jessie Norris for proofreading this manuscript.

REFERENCES

*Reference is also in this collection.

- AACR Project GENIE Consortium. 2017. AACR project GENIE: powering precision medicine through an international consortium. *Cancer Discov* 7: 818–831. doi:10.1158/2159-8290.CD-17-0151
- Anders S, Huber W. 2010. Differential expression analysis for sequence count data. *Gen Biol* 11: R106. doi:10.1186/gb-2010-11-10-r106
- Bairoch A. 2018. The cellosaurus, a cell-line knowledge resource. *J Biomol Tech* 29: 25–38. doi:10.7171/jbt.18-2902-002
- Balachandran VP, Gonen M, Smith JJ, DeMatteo RP. 2015. Nomograms in oncology: more than meets the eye. *Lancet Oncol* 16: e173–e180. doi:10.1016/S1470-2045(14)71116-7
- Banerjee S, Carlin BP, Gelfand AE. 2014. *Hierarchical modeling and analysis for spatial data*. Chapman & Hall, London.
- Barretina J, Caponigro G, Stransky N, Venkatesan K, Margolin AA, Kim S, Wilson CJ, Lehar J, Kryukov GV, Sonkin D, et al. 2012. The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. *Nature* 483: 603–607. doi:10.1038/nature11003
- Basu A, Bodycombe NE, Cheah JH, Price EV, Liu K, Schaefer GI, Ebright RY, Stewart ML, Ito D, Wang S, et al. 2013. An interactive resource to identify cancer genetic and lineage dependencies targeted by small molecules. *Cell* 154: 1151–1161. doi:10.1016/j.cell.2013.08.003
- Beasley MB, Brambilla E, Travis WD. 2005. The 2004 World Health Organization classification of lung tumors. *Semin Roentgenol* 40: 90–97. doi:10.1053/j.ro.2005.01.001
- Bender E. 2014. Epidemiology: the dominant malignancy. *Nature* 513: S2–S3. doi:10.1038/513S2a
- Bettencourt LMA, Ribeiro RM. 2008. Real time Bayesian estimation of the epidemic potential of emerging infectious diseases. *PLoS ONE* 3: e2185. doi:10.1371/journal.pone.0002185
- Bilimoria KY, Stewart AK, Winchester DP, Ko CY. 2008. The National Cancer Data Base: a powerful initiative to improve cancer care in the United States. *Ann Surg Oncol* 15: 683–690. doi:10.1245/s10434-007-9747-3
- Bravata DM, Olkin I. 2001. Simple pooling versus combining in meta-analysis. *Eval Health Prof* 24: 218–230. doi:10.1177/01632780122034885
- Cai L, Li Q, Du Y, Yun J, Xie Y, DeBerardinis RJ, Xiao G. 2017. Genomic regression analysis of coordinated expression. *Nat Commun* 8: 2187. doi:10.1038/s41467-017-02181-0
- Cai L, Lin S, Girard L, Zhou Y, Yang L, Ci B, Zhou Q, Luo D, Yao B, Tang H, et al. 2019a. LCE: an open web portal to explore gene expression and clinical associations in lung cancer. *Oncogene* 38: 2551–2564. doi:10.1038/s41388-018-0588-2
- Cai L, Luo D, Yao B, Yang DM, Lin S, Girard L, DeBerardinis RJ, Minna JD, Xie Y, Xiao G. 2019b. Systematic analysis of gene expression in lung adenocarcinoma and squamous cell carcinoma with a case study of FAM83A and FAM83B. *Cancers (Basel)* 11: 886.

L. Cai et al.

- Cai L, Liu H, Huang F, Fujimoto J, Girard L, Chen J, Li Y, Zhang YA, Deb D, Stastny V, et al. 2021a. Cell-autonomous immune gene expression is repressed in pulmonary neuroendocrine cells and small cell lung cancer. *Commun Biol* **4**: 314. doi:10.1038/s42003-021-01842-7
- Cai L, Liu H, Minna JD, DeBerardinis RJ, Xiao G, Xie Y. 2021b. Assessing consistency across functional screening datasets in cancer cells. *Bioinformatics* doi:10.1093/bioinformatics/btab423
- Cancer Genome Atlas Research Network. 2012. Comprehensive genomic characterization of squamous cell lung cancers. *Nature* **489**: 519–525. doi:10.1038/nature11404
- Cancer Genome Atlas Research Network. 2014. Comprehensive molecular profiling of lung adenocarcinoma. *Nature* **511**: 543–550. doi:10.1038/nature13385
- Chan JM, Quintanal-Villalonga Á, Gao V, Xie Y, Allaj V, Chaudhary O, Masilionis I, Egger J, Chow A, Walle T, et al. 2020. Single cell profiling reveals novel tumor and myeloid subpopulations in small cell lung cancer. *bioRxiv* doi:10.1101/2020.12.01.406363
- Chen PH, Cai L, Huffman K, Yang C, Kim J, Faubert B, Boroughs L, Ko B, Sudderth J, McMillan EA, et al. 2019. Metabolic diversity in human non-small cell lung cancer cells. *Mol Cell* **76**: 838–851.e5. doi:10.1016/j.molcel.2019.08.028
- Chen YJ, Roumeliotis TI, Chang YH, Chen CT, Han CL, Lin MH, Chen HW, Chang GC, Chang YL, Wu CT, et al. 2020. Proteogenomics of non-smoking lung cancer in east Asia delineates molecular signatures of pathogenesis and progression. *Cell* **182**: 226–244.e17. doi:10.1016/j.cell.2020.06.012
- Clark K, Vendt B, Smith K, Freymann J, Kirby J, Koppel P, Moore S, Phillips S, Maffitt D, Pringle M, et al. 2013. The cancer imaging archive (TCIA): maintaining and operating a public information repository. *J Digit Imaging* **26**: 1045–1057. doi:10.1007/s10278-013-9622-7
- Combs SE, Hancock JG, Boffa DJ, Decker RH, Detterbeck FC, Kim AW. 2015. Bolstering the case for lobectomy in stages I, II, and IIIA small-cell lung cancer using the National Cancer Data Base. *J Thorac Oncol* **10**: 316–323. doi:10.1097/JTO.0000000000000402
- Director's Challenge Consortium for the Molecular Classification of Lung Adenocarcinoma; Shedden K, Taylor JM, Enkemann SA, Tsao MS, Yeatman TJ, Gerald WL, Eschrich S, Jurisica I, Giordano TJ, et al. 2008. Gene expression-based survival prediction in lung adenocarcinoma: a multi-site, blinded validation study. *Nat Med* **14**: 822–827. doi:10.1038/nm.1790
- Drapkin BJ, George J, Christensen CL, Mino-Kenudson M, Dries R, Sundaresan T, Phat S, Myers DT, Zhong J, Igo P, et al. 2018. Genomic and functional fidelity of small cell lung cancer patient-derived xenografts. *Cancer Discov* **8**: 600–615. doi:10.1158/2159-8290.CD-17-0935
- * Estape AA, Foggetti G, Starrett J, Nguyen D, Politi K. 2021. Preclinical models for the study of lung cancer pathogenesis and therapy development. *Cold Spring Harbor Perspect Med* doi:10.1101/cshperspect.a037820
- Gao J, Aksoy BA, Dogrusoz U, Dresdner G, Gross B, Sumer SO, Sun Y, Jacobsen A, Sinha R, Larsson E, et al. 2013. Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal. *Sci Signal* **6**: pii. doi:10.1126/scisignal.2004088
- Gazdar AF, Gao B, Minna JD. 2010. Lung cancer cell lines: useless artifacts or invaluable tools for medical science? *Lung Cancer* **68**: 309–318. doi:10.1016/j.lungcan.2009.12.005
- Gazdar AF, Hirsch FR, Minna JD. 2016. From mice to men and back: an assessment of preclinical model systems for the study of lung cancers. *J Thorac Oncol* **11**: 287–299. doi:10.1016/j.jtho.2015.10.009
- George J, Lim JS, Jang SJ, Cun Y, Ozretić L, Kong G, Leenders F, Lu X, Fernández-Cuesta L, Bosco G, et al. 2015. Comprehensive genomic profiles of small cell lung cancer. *Nature* **524**: 47–53. doi:10.1038/nature14664
- Gerber DE, Laccetti AL, Xuan L, Halm EA, Pruitt SL. 2014. Impact of prior cancer on eligibility for lung cancer clinical trials. *J Natl Cancer Inst* **106**: dju302. doi:10.1093/jnci/dju302
- Ghandi M, Huang FW, Jané-Valbuena J, Kryukov GV, Lo CC, McDonald ER III, Barretina J, Gelfand ET, Bielski CM, Li H, et al. 2019. Next-generation characterization of the Cancer Cell Line Encyclopedia. *Nature* **569**: 503–508. doi:10.1038/s41586-019-1186-3
- Gillette MA, Satpathy S, Cao S, Dhanasekaran SM, Vasaikar SV, Krug K, Petralia F, Li Y, Liang WW, Reva B, et al. 2020. Proteogenomic characterization reveals therapeutic vulnerabilities in lung adenocarcinoma. *Cell* **182**: 200–225.e35. doi:10.1016/j.cell.2020.06.013
- Goswami CP, Nakshatri H. 2014. PROGgenev2: enhancements on the existing database. *BMC Cancer* **14**: 970. doi:10.1186/1471-2407-14-970
- Griffith SD, Tucker M, Bowser B, Calkins G, Chang CJ, Guardino E, Khozin S, Kraut J, You P, Schrag D, et al. 2019. Generating real-world tumor burden endpoints from electronic health record data: comparison of RE-CIST, radiology-anchored, and clinician-anchored approaches for abstracting real-world progression in non-small cell lung cancer. *Adv Ther* **36**: 2122–2136. doi:10.1007/s12325-019-00970-1
- Gu YF, Cohn S, Christie A, McKenzie T, Wolff N, Do QN, Madhuranthakam AJ, Pedrosa I, Wang T, Dey A, et al. 2017. Modeling renal cell carcinoma in mice: *Bap1* and *Pbrm1* inactivation drive tumor grade. *Cancer Discov* **7**: 900–917. doi:10.1158/2159-8290.CD-17-0292
- Györfy B, Surowiak P, Budczies J, Lánckzy A. 2013. Online survival analysis software to assess the prognostic value of biomarkers using transcriptomic data in non-small-cell lung cancer. *PLoS ONE* **8**: e82241. doi:10.1371/journal.pone.0082241
- Ha MJ, Banerjee S, Akbani R, Liang H, Mills GB, Do KA, Baladandayuthapani V. 2018. Personalized integrated network modeling of the Cancer Proteome Atlas. *Sci Rep* **8**: 14924. doi:10.1038/s41598-018-32682-x
- Hoda MA, Klikovits T, Klepetko W. 2018. Controversies in oncology: surgery for small cell lung cancer? It's time to rethink the case. *ESMO Open* **3**: e000366. doi:10.1136/esmoopen-2018-000366
- Iorio F, Knijnenburg TA, Vis DJ, Bignell GR, Menden MP, Schubert M, Aben N, Gonçalves E, Barthorpe S, Lightfoot H, et al. 2016. A landscape of pharmacogenomic interactions in cancer. *Cell* **166**: 740–754. doi:10.1016/j.cell.2016.06.017
- Jiang L, Huang J, Higgs BW, Hu Z, Xiao Z, Yao X, Conley S, Zhong H, Liu Z, Brohawn P, et al. 2016. Genomic land-

- scape survey identifies SRSF1 as a key oncogene in small cell lung cancer. *PLoS Genet* **12**: e1005895. doi:10.1371/journal.pgen.1005895
- Khazin S, Miksad RA, Adami J, Boyd M, Brown NR, Gossai A, Kaganman I, Kuk D, Rockland JM, Pazdur R, et al. 2019. Real-world progression, treatment, and survival outcomes during rapid adoption of immunotherapy for advanced non-small cell lung cancer. *Cancer* **125**: 4019–4032. doi:10.1002/cncr.32383
- Kim N, He N, Yoon S. 2014. Cell line modeling for systems medicine in cancers (review). *Int J Oncol* **44**: 371–376. doi:10.3892/ijo.2013.2202
- Kim N, Kim HK, Lee K, Hong Y, Cho JH, Choi JW, Lee JJ, Suh YL, Ku BM, Eum HH, et al. 2020. Single-cell RNA sequencing demonstrates the molecular and cellular reprogramming of metastatic lung adenocarcinoma. *Nat Commun* **11**: 2285. doi:10.1038/s41467-020-16164-1
- Kinker GS, Greenwald AC, Tal R, Orlova Z, Cuoco MS, McFarland JM, Warren A, Rodman C, Roth JA, Bender SA, et al. 2020. Pan-cancer single-cell RNA-seq identifies recurring programs of cellular heterogeneity. *Nat Genet* **52**: 1208–1218. doi:10.1038/s41588-020-00726-6
- Krupke DM, Begley DA, Sundberg JP, Richardson JE, Neuhauer SB, Bult CJ. 2017. The Mouse Tumor Biology Database: a comprehensive resource for mouse models of human cancer. *Cancer Res* **77**: e67–e70. doi:10.1158/0008-5472.CAN-17-0584
- Krysan K, Tran LM, Grimes BS, Fishbein GA, Seki A, Gardner BK, Walser TC, Salehi-Rad R, Yanagawa J, Lee JM, et al. 2019. The immune contexture associates with the genomic landscape in lung adenomatous premalignancy. *Cancer Res* **79**: 5022–5033.
- Laccetti AL, Pruitt SL, Xuan L, Halm EA, Gerber DE. 2015. Effect of prior cancer on outcomes in advanced lung cancer: implications for clinical trial eligibility and accrual. *J Natl Cancer Inst* **107**: djv002. doi:10.1093/jnci/djv002
- Laccetti AL, Pruitt SL, Xuan L, Halm EA, Gerber DE. 2016. Prior cancer does not adversely affect survival in locally advanced lung cancer: a national SEER-medicare analysis. *Lung Cancer* **98**: 106–113. doi:10.1016/j.lungcan.2016.05.029
- LaFave LM, Kartha VK, Ma S, Meli K, Del Priore I, Lareau C, Naranjo S, Westcott PMK, Duarte FM, Sankar V, et al. 2020. Epigenomic state transitions characterize tumor progression in mouse lung adenocarcinoma. *Cancer Cell* **38**: 212–228 e213. doi:10.1016/j.ccell.2020.06.006
- Laughney AM, Hu J, Campbell NR, Bakhoum SF, Setty M, Lavallée VP, Xie Y, Masilionis I, Carr AJ, Kottapalli S, et al. 2020. Regenerative lineages and immune-mediated pruning in lung cancer metastasis. *Nat Med* **26**: 259–269. doi:10.1038/s41591-019-0750-6
- Lavin Y, Kobayashi S, Leader A, Amir ED, Elefant N, Bigenwald C, Remark R, Sweeney R, Becker CD, Levine JH, et al. 2017. Innate immune landscape in early lung adenocarcinoma by paired single-cell analyses. *Cell* **169**: 750–765.e17. doi:10.1016/j.cell.2017.04.014
- Li H, Ning S, Ghandi M, Kryukov GV, Gopal S, Deik A, Souza A, Pierce K, Keskula P, Hernandez D, et al. 2019a. The landscape of cancer cell line metabolism. *Nat Med* **25**: 850–860. doi:10.1038/s41591-019-0404-8
- Li Q, Wang X, Liang F, Xiao G. 2019b. A Bayesian mark interaction model for analysis of tumor pathology images. *Ann Appl Stat* **13**: 1708–1732.
- Li Q, Wang X, Liang F, Yi F, Xie Y, Gazdar A, Xiao G. 2019c. A Bayesian hidden Potts mixture model for analyzing lung cancer pathology images. *Biostatistics* **20**: 565–581. doi:10.1093/biostatistics/kxy019
- Liang F. 2010. A double Metropolis-Hastings sampler for spatial models with intractable normalizing constants. *J Stat Comput Simul* **80**: 1007–1022. doi:10.1080/00949650902882162
- Luo X, Zang X, Yang L, Huang J, Liang F, Rodriguez-Canales J, Wistuba II, Gazdar A, Xie Y, Xiao G. 2017. Comprehensive computational pathological image analysis predicts lung cancer prognosis. *J Thorac Oncol* **12**: 501–509. doi:10.1016/j.jtho.2016.10.017
- Luo X, Yin S, Yang L, Fujimoto J, Yang Y, Moran C, Kalhor N, Weissferdt A, Xie Y, Gazdar A, et al. 2019. Development and validation of a pathology image analysis-based predictive model for lung adenocarcinoma prognosis—a multi-cohort study. *Sci Rep* **9**: 6886. doi:10.1038/s41598-019-42845-z
- MacLean M, Luo X, Wang S, Kernstine K, Gerber DE, Xie Y. 2018. Outcomes of neoadjuvant and adjuvant chemotherapy in stage 2 and 3 non-small cell lung cancer: an analysis of the National Cancer Database. *Oncotarget* **9**: 24470–24479. doi:10.18632/oncotarget.25327
- Maynard A, McCoach CE, Rotow JK, Harris L, Haderk F, Kerr DL, Yu EA, Schenk EL, Tan W, Zee A, et al. 2020. Therapy-induced evolution of human lung cancer revealed by single-cell RNA sequencing. *Cell* **182**: 1232–1251.e22. doi:10.1016/j.cell.2020.07.017
- McDonald ER III, de Weck A, Schlabach MR, Billy E, Mavrikakis KJ, Hoffman GR, Belur D, Castelletti D, Frias E, Gampa K, et al. 2017. Project DRIVE: a compendium of cancer dependencies and synthetic lethal relationships uncovered by large-scale, deep RNAi screening. *Cell* **170**: 577–592.e10. doi:10.1016/j.cell.2017.07.005
- McMillan EA, Ryu MJ, Diep CH, Mendiratta S, Clemenceau JR, Vaden RM, Kim JH, Motoyaji T, Covington KR, Peyton M, et al. 2018. Chemistry-first approach for nomination of personalized treatment in lung cancer. *Cell* **173**: 864–878.e29. doi:10.1016/j.cell.2018.03.028
- Meng Y, Jin M. 2021. HFS-SLPPE: a novel hierarchical feature selection and second learning probability error ensemble model for precision cancer diagnosis. *Front Cell Dev Biol* **9**: 696359. doi:10.3389/fcell.2021.696359
- Mizuno H, Kitada K, Nakai K, Sarai A. 2009. Prognoscan: a new database for meta-analysis of the prognostic value of genes. *BMC Med Genomics* **2**: 18. doi:10.1186/1755-8794-2-18
- Montoro DT, Haber AL, Biton M, Vinarsky V, Lin B, Birket SE, Yuan F, Chen S, Leung HM, Villoria J, et al. 2018. A revised airway epithelial hierarchy includes CFTR-expressing ionocytes. *Nature* **560**: 319–324. doi:10.1038/s41586-018-0393-7
- Mullen DJ, Yan C, Kang DS, Zhou B, Borok Z, Marconett CN, Farnham PJ, Offringa IA, Rhie SK. 2020. TENET 2.0: identification of key transcriptional regulators and enhancers in lung adenocarcinoma. *PLoS Genet* **16**: e1009023. doi:10.1371/journal.pgen.1009023

L. Cai et al.

- Mulshine JL, Ujhazy P, Antman M, Burgess CM, Kuzmin I, Bunn PA Jr, Johnson BE, Roth JA, Pass HI, Ross SM, et al. 2020. From clinical specimens to human cancer preclinical models—a journey the NCI-cell line database—25 years later. *J Cell Biochem* **121**: 3986–3999. doi:10.1002/jcb.29564
- National Lung Screening Trial Research Team; Aberle DR, Adams AM, Berg CD, Black WC, Clapp JD, Fagerstrom RM, Gareen IF, Gatsonis C, Marcus PM, et al. 2011. Reduced lung-cancer mortality with low-dose computed tomographic screening. *N Engl J Med* **365**: 395–409. doi:10.1056/NEJMoa1102873
- Plasschaert LW, Žilionis R, Choo-Wing R, Savova V, Knehr J, Roma G, Klein AM, Jaffe AB. 2018. A single-cell atlas of the airway epithelium reveals the CFTR-rich pulmonary ionocyte. *Nature* **560**: 377–381. doi:10.1038/s41586-018-0394-6
- Polley E, Kunkel M, Evans D, Silvers T, Delosh R, Laudeman J, Ogle C, Reinhart R, Selby M, Connelly J, et al. 2016. Small cell lung cancer screen of oncology drugs, investigational agents, and gene and microRNA expression. *J Natl Cancer Inst* **108**: djw122. doi:10.1093/jnci/djw122
- Pruitt SL, Laccetti AL, Xuan L, Halm EA, Gerber DE. 2017. Revisiting a longstanding clinical trial exclusion criterion: impact of prior cancer in early-stage lung cancer. *Br J Cancer* **116**: 717–725. doi:10.1038/bjc.2017.27
- Reinhold WC, Sunshine M, Liu H, Varma S, Kohn KW, Morris J, Doroshow J, Pommier Y. 2012. CellMiner: a web-based suite of genomic and pharmacologic tools to explore transcript and drug patterns in the NCI-60 cell line set. *Cancer Res* **72**: 3499–3511. doi:10.1158/0008-5472.CAN-12-1370
- Rhodes DR, Kalyana-Sundaram S, Mahavisno V, Varambally R, Yu J, Briggs BB, Barrette TR, Anstet MJ, Kincaid-Beal C, Kulkarni P, et al. 2007. OncoPrint 3.0: genes, pathways, and networks in a collection of 18,000 cancer gene expression profiles. *Neoplasia* **9**: 166–180. doi:10.1593/neo.07112
- Rousseaux S, Debernardi A, Jacquiau B, Vitte AL, Vesin A, Nagy-Mignotte H, Moro-Sibilot D, Brichon PY, Lantuejoul S, Hainaut P, et al. 2013. Ectopic activation of germline and placental genes identifies aggressive metastasis-prone lung cancers. *Sci Transl Med* **5**: 186ra66. doi:10.1126/scitranslmed.3005723
- Rozenblatt-Rosen O, Regev A, Oberdoerffer P, Nawy T, Hupalowska A, Rood JE, Ashenberg O, Cerami E, Coffey RJ, Demir E, et al. 2020. The Human Tumor Atlas Network: charting tumor transitions across space and time at single-cell resolution. *Cell* **181**: 236–249. doi:10.1016/j.cell.2020.03.053
- Rudin CM, Durinck S, Stawiski EW, Poirier JT, Modrusan Z, Shames DS, Bergbower EA, Guan Y, Shin J, Guillory J, et al. 2012. Comprehensive genomic analysis identifies SOX2 as a frequently amplified gene in small-cell lung cancer. *Nat Genet* **44**: 1111–1116. doi:10.1038/ng.2405
- Safikhani Z, Smirnov P, Freeman M, El-Hachem N, She A, Rene Q, Goldenberg A, Birkbak NJ, Hatzis C, Shi L, et al. 2016. Revisiting inconsistency in large pharmacogenomic studies. *F1000Res* **5**: 2333. doi:10.12688/f1000research.9611.1
- Sato M, Larsen JE, Lee W, Sun H, Shames DS, Dalvi MP, Ramirez RD, Tang H, DiMaio JM, Gao B, et al. 2013. Human lung epithelial cells progressed to malignancy through specific oncogenic manipulations. *Mol Cancer Res* **11**: 638–650. doi:10.1158/1541-7786.MCR-12-0634-T
- Satpathy S, Krug K, Jean Beltran PM, Savage SR, Petralia F, Kumar-Sinha C, Dou Y, Reva B, Kane MH, Avanesian SC, et al. 2021. A proteogenomic portrait of lung squamous cell carcinoma. *Cell* **184**: 4348–4371.e40. doi:10.1016/j.cell.2021.07.016
- Scrucca L, Fop M, Murphy TB, Raftery AE. 2016. Mclust 5: clustering, classification and density estimation using Gaussian finite mixture models. *R J* **8**: 289–317. doi:10.32614/RJ-2016-021
- Sinjab A, Han G, Treekitkarnmongkol W, Hara K, Brennan PM, Dang M, Hao D, Wang R, Dai E, Dejima H, et al. 2021. Resolving the spatial and cellular architecture of lung adenocarcinoma by multiregion single-cell sequencing. *Cancer Discov* doi:10.1158/2159-8290.cd-20-1285
- Stewart CA, Gay CM, Xi Y, Sivajothi S, Sivakamasundari V, Fujimoto J, Bolisetty M, Hartsfield PM, Balasubramanian V, Chalishazar MD, et al. 2020. Single-cell analyses reveal increased intratumoral heterogeneity after the onset of therapy resistance in small-cell lung cancer. *Nat Cancer* **1**: 423–436. doi:10.1038/s43018-019-0020-z
- Sun L, Hsu M, Cohen RB, Langer CJ, Mamtani R, Aggarwal C. 2021. Association between KRAS variant status and outcomes with first-line immune checkpoint inhibitor-based therapy in patients with advanced non-small-cell lung cancer. *JAMA Oncol* **7**: 937–939. doi:10.1001/jamaoncol.2021.0546
- Szász AM, Lánczky A, Nagy A, Förster S, Hark K, Green JE, Boussiotas A, Busuttil R, Szabó A, Györfy B. 2016. Cross-validation of survival associated biomarkers in gastric cancer using transcriptomic data of 1,065 patients. *Oncotarget* **7**: 49322–49333. doi:10.18632/oncotarget.10337
- Tang H, Wang S, Xiao G, Schiller J, Papadimitrakopoulou V, Minna J, Wistuba II, Xie Y. 2017. Comprehensive evaluation of published gene expression prognostic signatures for biomarker-based lung cancer clinical studies. *Ann Oncol* **28**: 733–740. doi:10.1093/annonc/mdw683
- Teixeira VH, Pipinikas CP, Pennyquick A, Lee-Six H, Chandrasekharan D, Beane J, Morris TJ, Karpathakis A, Feber A, Breeze CE, et al. 2019. Deciphering the genomic, epigenomic, and transcriptomic landscapes of pre-invasive lung cancer lesions. *Nat Med* **25**: 517–525. doi:10.1038/s41591-018-0323-0
- Thomas-White K, Forster SC, Kumar N, Van Kuiken M, Putonti C, Stares MD, Hilt EE, Price TK, Wolfe AJ, Lawley TD. 2018. Culturing of female bladder bacteria reveals an interconnected urogenital microbiota. *Nat Commun* **9**: 1557. doi:10.1038/s41467-018-03968-5
- Tlemsani C, Pongor L, Girard L, Roper N, Elloumi F, Varma S, Luna A, Rajapakse VN, Sebastian R, Kohn KW, et al. 2020. SCLC_cellminer: integrated genomics and therapeutics predictors of small cell lung cancer cell lines based on their genomic signatures. *Cell Rep* **33**: 108296. doi:10.1016/j.celrep.2020.108296
- Travaglini KJ, Nabhan AN, Penland L, Sinha R, Gillich A, Sit RV, Chang S, Conley SD, Mori Y, Seita J, et al. 2020. A molecular cell atlas of the human lung from single-cell

- RNA sequencing. *Nature* **587**: 619–625. doi:10.1038/s41586-020-2922-4
- Travis WD, Brambilla E, Nicholson AG, Yatabe Y, Austin JHM, Beasley MB, Chirieac LR, Dacic S, Duhig E, Flieder DB, et al. 2015. The 2015 World Health Organization classification of lung tumors: impact of genetic, clinical and radiologic advances since the 2004 classification. *J Thorac Oncol* **10**: 1243–1260. doi:10.1097/JTO.0000000000000630
- Tsherniak A, Vazquez F, Montgomery PG, Weir BA, Kryukov G, Cowley GS, Gill S, Harrington WF, Pantel S, Krill-Burger JM, et al. 2017. Defining a cancer dependency map. *Cell* **170**: 564–576.e16. doi:10.1016/j.cell.2017.06.010
- van den Bent MJ. 2010. Interobserver variation of the histopathological diagnosis in clinical trials on glioma: a clinician's perspective. *Acta Neuropathol* **120**: 297–304. doi:10.1007/s00401-010-0725-7
- Wakeam E, Acuna SA, Leighl NB, Giuliani ME, Finlayson SRG, Varghese TK, Darling GE. 2017. Surgery versus chemotherapy and radiotherapy for early and locally advanced small cell lung cancer: a propensity-matched analysis of survival. *Lung Cancer* **109**: 78–88. doi:10.1016/j.lungcan.2017.04.021
- Wang Z, Wang Y. 2019. Extracting a biologically latent space of lung cancer epigenetics with variational autoencoders. *BMC Bioinformatics* **20**: 568. doi:10.1186/s12859-019-3130-9
- Wang S, Yang L, Ci B, Maclean M, Gerber DE, Xiao G, Xie Y. 2018. Development and validation of a nomogram prognostic model for SCLC patients. *J Thorac Oncol* **13**: 1338–1348. doi:10.1016/j.jtho.2018.05.037
- Wang S, Lai S, von Itzstein MS, Yang L, Yang DM, Zhan X, Xiao G, Halm EA, Gerber DE, Xie Y. 2019a. Type and case volume of health care facility influences survival and surgery selection in cases with early-stage non-small cell lung cancer. *Cancer* **125**: 4252–4259. doi:10.1002/cncr.32377
- Wang S, Wang T, Yang L, Yang DM, Fujimoto J, Yi F, Luo X, Yang Y, Yao B, Lin S, et al. 2019b. Convpath: a software tool for lung adenocarcinoma digital pathological image analysis aided by a convolutional neural network. *EBio-Medicine* **50**: 103–110. doi:10.1016/j.ebiom.2019.10.033
- Wang S, Yang DM, Rong R, Zhan X, Fujimoto J, Liu H, Minna J, Wistuba II, Xie Y, Xiao G. 2019c. Artificial intelligence in lung cancer pathology image analysis. *Cancers (Basel)* **11**: 1673.
- Wang T, Lu R, Lai S, Schiller JH, Zhou FL, Ci B, Wang S, Gao X, Yao B, Gerber DE, et al. 2019d. Development and validation of a nomogram prognostic model for patients with advanced non-small-cell lung cancer. *Cancer Inform* **18**: 117693511983754. doi:10.1177/1176935119837547
- Wang S, Rong R, Yang DM, Fujimoto J, Yan S, Cai L, Yang L, Luo D, Behrens C, Parra ER, et al. 2020. Computational staining of pathology images to study the tumor microenvironment in lung cancer. *Cancer Res* **80**: 2056–2066. doi:10.1158/0008-5472.CAN-19-1629
- Woo XY, Giordano J, Srivastava A, Zhao ZM, Lloyd MW, de Bruijn R, Suh YS, Patidar R, Chen L, Scherer S, et al. 2019. Conservation of copy number profiles during engraftment and passaging of patient-derived cancer xenografts. *Nat Genet* **53**: 86–99. doi:10.1038/s41588-020-00750-6
- Wu F, Fan J, He Y, Xiong A, Yu J, Li Y, Zhang Y, Zhao W, Zhou F, Li W, et al. 2021. Single-cell profiling of tumor heterogeneity and the microenvironment in advanced non-small cell lung cancer. *Nat Commun* **12**: 2540. doi:10.1038/s41467-021-22801-0
- Xie Y, Lu W, Wang S, Tang X, Tang H, Zhou Y, Moran C, Behrens C, Roth JA, Zhou Q, et al. 2019. Validation of the 12-gene predictive signature for adjuvant chemotherapy response in lung cancer. *Clin Cancer Res* **25**: 150–157. doi:10.1158/1078-0432.CCR-17-2543
- Yanagawa J, Tran LM, Fung E, Wallace WD, Prosper AE, Fishbein GA, Shea C, Hong R, Liu B, Salehi-Rad R, et al. 2020. Single-cell characterization of subsolid and solid lesions in the lung adenocarcinoma spectrum. *bioRxiv* doi:10.1101/2020.12.25.424416
- Yang L, Wang S, Gerber DE, Zhou Y, Xu F, Liu J, Liang H, Xiao G, Zhou Q, Gazdar A, et al. 2018. Main bronchus location is a predictor for metastasis and prognosis in lung adenocarcinoma: a large cohort analysis. *Lung Cancer* **120**: 22–26. doi:10.1016/j.lungcan.2018.03.011
- Yu C, Mannan AM, Yvone GM, Ross KN, Zhang YL, Marton MA, Taylor BR, Crenshaw A, Gould JZ, Tamayo P, et al. 2016a. High-throughput identification of genotype-specific cancer vulnerabilities in mixtures of barcoded tumor cell lines. *Nat Biotechnol* **34**: 419–423. doi:10.1038/nbt.3460
- Yu KH, Zhang C, Berry GJ, Altman RB, Ré C, Rubin DL, Snyder M. 2016b. Predicting non-small cell lung cancer prognosis by fully automated microscopic pathology image features. *Nat Commun* **7**: 12474. doi:10.1038/ncomms12474
- Zhang Z, Li H, Jiang S, Li R, Li W, Chen H, Bo X. 2019. A survey and evaluation of Web-based tools/databases for variant analysis of TCGA data. *Brief Bioinform* **20**: 1524–1541. doi:10.1093/bib/bby023
- Zhao Y, Gao Y, Xu X, Zhou J, Wang H. 2021. Multi-omics analysis of genomics, epigenomics and transcriptomics for molecular subtypes and core genes for lung adenocarcinoma. *BMC Cancer* **21**: 257. doi:10.1186/s12885-021-07888-4

Cold Spring Harbor Perspectives in Medicine

Lung Cancer Computational Biology and Resources

Ling Cai, Guanghua Xiao, David Gerber, John D. Minna and Yang Xie

Cold Spring Harb Perspect Med 2022; doi: 10.1101/cshperspect.a038273 originally published online November 8, 2021

Subject Collection [Lung Cancer: Disease Biology and Its Potential for Clinical Translation](#)

Tumor Immunology and Immunotherapy of Non-Small-Cell Lung Cancer

Tina Cascone, Jared Fradette, Monika Pradhan, et al.

Molecular Pathology of Lung Cancer

James J. Saller and Theresa A. Boyle

Preclinical Models for the Study of Lung Cancer Pathogenesis and Therapy Development

Anna Arnal-Estapé, Giorgia Foggetti, Jacqueline H. Starrett, et al.

Radiation Therapy in Non-Small-Cell Lung Cancer

Michael Dohopolski, Sujana Gottumukkala, Daniel Gomez, et al.

Application of Radiomics and Artificial Intelligence for Lung Cancer Precision Medicine

Ilke Tunali, Robert J. Gillies and Matthew B. Schabath

Advances in Small-Cell Lung Cancer (SCLC) Translational Research

Benjamin J. Drapkin and Charles M. Rudin

Lung Cancer Stem Cells and Their Clinical Implications

Samuel P. Rowbotham, Mounika U.L. Goruganthu, Rajeswara R. Arasada, et al.

Lung Cancer Computational Biology and Resources

Ling Cai, Guanghua Xiao, David Gerber, et al.

Metabolic Phenotypes, Dependencies, and Adaptation in Lung Cancer

Gina M. DeNicola and David B. Shackelford

Early Diagnosis and Screening for Lung Cancer

Humam Kadara, Linh M. Tran, Bin Liu, et al.

Targeting Epigenetics in Lung Cancer

Yvonne L. Chao and Chad V. Pecot

For additional articles in this collection, see <http://perspectivesinmedicine.cshlp.org/cgi/collection/>