The Lung Cancer Autochthonous Model Gene Expression Database Enables Cross-Study Comparisons of the Transcriptomic Landscapes Across Mouse Models

Ling Cai[1,2,3], Fangjiang Wu[1], Qinbo Zhou[1], Ying Gao[1], Bo Yao[1], Ralph J. DeBerardinis[2,3,4], George K. Acquaah-Mensah[5], Vassilis Aidinis[6], Jennifer E. Beane[7], Shyam Biswal[8], Ting Chen[9], Carla P. Concepcion-Crisol[10], Barbara M. Grüner[11], Deshui Jia[12], Robert A. Jones[13], Jonathan M. Kurie[14], Min Gyu Lee[15], Per Lindahl[16], Yonathan Lissanu[17], Corina Lorz[18], David MacPherson[19], Rosanna Martinelli[20], Pawel K. Mazur[21], Sarah A. Mazzilli[7], Shinji Mii[22], Herwig P. Moll[23], Roger A. Moorehead[13], Edward E. Morrisey[24], Sheng Rong Ng[25], Matthew G. Oser[26], Arun R. Pandiri[27], Charles A. Powell[28], Giorgio Ramadori[29], Mirentxu Santos[18], Eric L. Snyder[30], Rocio Sotillo[31], Kang-Yi Su[32], Tetsuro Taki[22], Kekoa Taparra[33], Phuoc T. Tran[34], Yifeng Xia[35], J. Edward van Veen[36], Monte M. Winslow[37], Guanghua Xiao[1,3,38], Charles M. Rudin[39], Trudy G. Oliver[40], Yang Xie[1,3,38], and John D. Minna[3,41]

ABSTRACT

Lung cancer, the leading cause of cancer mortality, exhibits diverse histologic subtypes and genetic complexities. Numerous preclinical mouse models have been developed to study lung cancer, but data from these models are disparate, siloed, and difficult to compare in a centralized fashion. In this study, we established the Lung Cancer Autochthonous Model Gene Expression Database (LCAMGDB), an extensive repository of 1,354 samples from 77 transcriptomic datasets covering 974 samples from genetically engineered mouse models (GEMM), 368 samples from carcinogen-induced models, and 12 samples from a spontaneous model. Meticulous curation and collaboration with data depositors produced a robust and comprehensive database, enhancing the fidelity of the genetic landscape it depicts. The LCAMGDB aligned 859 tumors from GEMMs with human lung cancer mutations, enabling comparative analysis and revealing a pressing need to broaden the diversity of genetic aberrations modeled in the GEMMs. To accompany this resource, a web application was developed that offers researchers intuitive tools for in-depth gene expression analysis. With standardized reprocessing of gene expression data, the LCAMGDB serves as a powerful platform for cross-study comparison and lays the groundwork for future research, aiming to bridge the gap between mouse models and human lung cancer for improved translational relevance.

Significance: The Lung Cancer Autochthonous Model Gene Expression Database (LCAMGDB) provides a comprehensive and accessible resource for the research community to investigate lung cancer biology in mouse models.

Introduction

Lung cancer remains the most common cause of cancer-related mortality globally, with its complexity reflected in diverse histologic subtypes, such as adenocarcinoma (ADC), squamous cell carcinoma (SQCC), large cell carcinoma, and small cell lung carcinoma (SCLC), each harboring distinct genetic alterations that drive tumor biology, which in some cases dictates therapeutic vulnerabilities. To decipher the complexities of tumor biology, high-throughput molecular profiling of patient-derived tumors has been extensively used (1–7). Preclinical models of lung cancer are essential tools for researchers to understand cancer biology and develop therapeutic strategies through experimentation. There has been a concerted effort to aggregate data from patient-derived cell lines (8, 9) and patient-derived xenografts (10). Although lung cancer autochthonous animal models, primarily based on mice, represent a separate but significant line of research, they often lack unified characterization because of independent development across various laboratories.

To address this gap, we conducted a comprehensive review of transcriptomic databases, specifically Gene Expression Omnibus and ArrayExpress, collected transcriptomic data from lung cancer mouse models, and standardized associated sample and oncogenotype information. We actively engaged with data depositors to refine our curation process and incorporate their insights. These efforts

[1]Quantitative Biomedical Research Center, Peter O'Donnell Jr. School of Public Health, UT Southwestern Medical Center, Dallas, Texas. [2]Children's Research Institute, UT Southwestern Medical Center, Dallas, Texas. [3]Simmons Comprehensive Cancer Center, UT Southwestern Medical Center, Dallas, Texas. [4]Howard Hughes Medical Institute, University of Texas Southwestern Medical Center, Dallas, Texas. [5]Massachusetts College of Pharmacy & Health Sciences, Worcester, Massachusetts. [6]Institute of Fundamental Biological Research, Biomedical Sciences Research Center Alexander Fleming, Athens, Greece. [7]Section of Computational Biomedicine, Boston University School of Medicine, Boston, Massachusetts. [8]Department of Environmental Health and Engineering, Johns Hopkins University School of Public Health, Baltimore, Maryland. [9]NYU School of Medicine, New York, New York. [10]Columbia University, New York, New York. [11]Department of Medical Oncology, West German Cancer Center, University Hospital Essen Essen, Germany. [12]Institute of Translational Medicine, Shanghai General Hospital, Shanghai Jiao Tong University School of Medicine, Shanghai, China. [13]Department of Biomedical Sciences, Ontario Veterinary College, University of Guelph, Guelph, Canada.

have culminated in the creation of the Lung Cancer Autochthonous Model Gene Expression Database (LCAMGDB). This resource serves as a centralized platform for the research community, providing access to a comprehensive collection of genetically engineered and chemically induced mouse models of lung cancer. We also developed a user-friendly web application populated from this database, offering researchers intuitive tools for dynamic data exploration and sophisticated analysis.

Materials and Methods

Dataset screening

We performed a comprehensive search for transcriptomic datasets in publicly available repositories, specifically the Gene Expression Omnibus and ArrayExpress. Search parameters included the keywords "lung cancer" and were restricted to species "*Mus musculus*," and the data type confined to gene expression profiling by array or high-throughput sequencing. Identified datasets were manually inspected to ensure their relevance and inclusion of data generated from autochthonous models, including genetically engineered mouse models (GEMM), chemically induced mouse models, and spontaneous models of lung cancer. Expression and annotation data were downloaded programmatically using the R package GEOquery (11), and supplementary data files were accessed via the *getGEOSuppFiles* function. Additional metadata were obtained from the associated publications or directly from data depositors. Discrepancies were resolved programmatically and manually. Duplicate entries were flagged and corrected using custom scripts, ensuring no redundancy in the curated database.

LCAMGDB data organization and curation process

LCAMGDB organizes data into three primary tables, for datasets, samples, and genotypes. The dataset table contains data accession IDs, platform IDs, model types, study titles, publication PubMed (PM) IDs, PubMed Central (PMC) IDs, and the contact information of data depositors. The sample table contains accession IDs, sample names, types, treatments, strains, sex, age, genotype, histologic classification, primary/metastasis status, sources of Affymetrix data, Sequence Read Archive (SRA) IDs, and growth protocols. The genotype table was designed to record details of model genetic manipulations. It contains multiple rows for each genotype to specify the genes involved, genetic constructs, zygosity, the type of genetic modification (e.g., overexpression and knockout), the method of genetic manipulation, induction methods, induction systems, promoters used, cell of origin, and additional notes that may provide context or clarifications. This information is further organized to generate both standardized and simplified genotypes, concisely indicating the genetic manipulations and induction methods used in each model.

For data curation, we gathered details from database annotations and carefully reviewed the original publications to extract the necessary information. We standardized terms to ensure consistency across the data. For instance, we categorized sample types into four distinct groups: "bulk tissue," "microdissected," "CD45 depleted," and "sorted cancer cells." We also include data fields for the original curation to preserve the intricacies of the source dataset. For example, although we simplified the primary/metastasis tumor status to "primary" and "metastasis" for consistency, we kept specific details like "liver metastasis" in the "primary/metastasis original" field to capture the full depth of the original classifications. In harmonizing the histology data, we recognized the continuum that exists between mouse tumor classifications of adenoma and ADC. For example, in the LSL-Kras^{G12D} model, tumors can progress from adenoma to ADC between 6 and 16 weeks after infection (12). However, not all studies explicitly differentiate between adenoma and ADC. Additionally, multiple clonal tumors may present within the same sample, in which some may classify as adenomas and others as ADCs. To address this, we carefully reviewed original publications and annotations, assigning the most accurate histology annotation to the "histology.original" field. For cases with clear distinctions, we labeled them as either "Adenoma" or "ADC." For those with ambiguous classifications, we used "Adenoma/ADC." Consequently, in the "histology" field, we grouped these classifications

[14]Department of Thoracic-Head & Neck Med Onc, Division of Cancer Medicine, The University of Texas MD Anderson Cancer Center, Houston, Texas. [15]Department of Molecular and Cellular Oncology, The University of Texas MD Anderson Cancer Center, Houston, Texas. [16]Sahlgrenska Center for Cancer Research Institute of Biomedicine | Department of Medical Biochemistry and Cell Biology, University of Gothenburg Gothenburg, Sweden. [17]Department of Thoracic & Cardiovascular Surgery, The University of Texas MD Anderson Cancer Center, Houston, Texas. [18]Biomedical Innovation Unit. Centro de Investigaciones Energéticas, Medioambientales y Tecnológicas (CIEMAT), Madrid, Spain. [19]Fred Hutchinson Cancer Center, Seattle, Washington. [20]Department of Medicine, Surgery and Dentistry 'Scuola Medica Salernitana', University of Salerno, Baronissi, Italy. [21]Department of Experimental Radiation Oncology, The University of Texas MD Anderson Cancer Center, Houston, Texas. [22]Department of Pathology, Nagoya University Graduate School of Medicine, Nagoya, Japan. [23]Medical University of Vienna, Center for Physiology and Pharmacology, Vienna, Austria. [24]Penn-CHOP Lung Biology Institute, University of Pennsylvania, Philadelphia, Pennsylvania. [25]Institute of Molecular and Cell Biology, A*STAR, Singapore, Singapore. [26]Department of Medical Oncology, Dana-Farber Cancer Institute, Harvard Medical School, Boston, Massachusetts. [27]Cellular and Molecular Pathology Branch, Division of National Toxicology Program (DNTP), National Institute of Environmental Health Sciences (NIEHS), Research Triangle Park, North Carolina. [28]Icahn School of Medicine at Mount Sinai, New York, New York. [29]Department of Cell Physiology and Metabolism, University of Geneva; Geneva, Switzerland. [30]Department of Pathology and Huntsman Cancer Institute, University of Utah, Salt Lake City, Utah. [31]Molecular Thoracic Oncology, German Cancer Research Center (DKFZ), Heidelberg, Germany. [32]Department of Clinical Laboratory Sciences and Medical Biotechnology, College of Medicine, National Taiwan University, Taipei, Taiwan. [33]Department of Radiation Oncology, Stanford Health Care, Stanford, California. [34]Department of Radiation Oncology, University of Maryland School of Medicine, Baltimore, Maryland. [35]Salk Institute for Biological Studies, San Diego, California. [36]Department of Integrative Biology and Physiology, University of California Los Angeles, Los Angeles, California. [37]Department of Genetics, Stanford University School of Medicine, Stanford, California. [38]Department of Bioinformatics, University of Texas Southwestern Medical Center, Dallas, Texas. [39]Memorial Sloan Kettering Cancer Center, New York, New York. [40]Department of Pharmacology & Cancer Biology, Duke University, Durham, North Carolina. [41]Hamon Center for Therapeutic Oncology Research, UT Southwestern Medical Center, Dallas, Texas.

Corresponding Authors: Ling Cai, Peter O'Donnell School of Public Health, The University of Texas Southwestern Medical Center, 5323 Harry Hines Blvd, Dallas, TX 75390. E-mail: ling.cai@utsouthwestern.edu; Yang Xie, yang.xie@utsouthwestern.edu; and John D. Minna, john.minna@utsouthwestern.edu

Cancer Res 2025;85:1769–83

doi: 10.1158/0008-5472.CAN-24-1607

together under "Adenoma/ADC" to maintain consistency and clarity across the dataset. Integration of depositor feedback ensured accuracy and completeness, with iterative revisions documented for transparency.

Gene expression reprocessing

Affymetrix raw data were downloaded from the GEO and grouped by platform. For each platform, we downloaded v25 of the gene-level customized chip definition files from the Molecular and Behavioral Neuroscience Institute repository (http://brainarray.mbni.med.umich.edu/Brainarray/Database/CustomCDF/25.0.0/ensg.asp) at the University of Michigan (13), to reprocess the data with the most up-to-date and specific gene annotations. The CEL files were batch-read with the specified platform package and normalized using the Robust Multi-array Average method via the oligo package, yielding an ExpressionSet (eset) from which gene expression matrices were extracted. Entrez IDs were converted to gene symbols based on the NCBI Entrez mapping file.

RNA sequencing (RNA-seq) FASTQ files were downloaded from the SRA through the SRA Toolkit. Paired-end reads were concatenated to be processed as single-end reads. Reads were trimmed to remove adapters and low-quality sequences and subsequently aligned to mouse reference GRCm38 by HISAT2 (14). Gene expression was quantified using featureCounts (15) and GENCODE (16). We retained genes with nonzero values in >10% of samples, normalized their counts to library sizes, and computed log-transformed counts per million (logCPM) values for downstream analyses.

Depositor-processed data were received in various formats, reflecting the diversity of their sources. We applied log-transformation where necessary and performed quantile normalization.

American Association for Cancer Research GENIE data analysis

To compare genetic landscapes between mouse models and human lung cancers, we analyzed the American Association for Cancer Research (AACR) GENIE data (version 15.0-public; ref. 17). Data were downloaded from SAGE BIONETWORKS on March 25, 2024 through R package "synapserutils" (18) with Synapse ID "syn7222066". We used mutation status from "data_mutations_extended.txt", amplification status (value of 2), and deletion status (value of −2) from "data_CNA.txt" and structural variation status from "data_sv.txt" to determine genetic aberrations. Samples of patients with lung cancer were selected from "data_clinical_sample.txt". Cumulative counts of genetic aberration events are summarized at the sample level (note that some patients could have multiple samples in the dataset). Comparisons highlighted disparities in genetic coverage between GEMMs and human tumors, emphasizing the need for models that better recapitulate human lung cancer genetics.

Figure generation and statistical analysis

Statistical software R was used for analyses and web application construction (19). We used principal component analysis (PCA) to summarize gene expression variability across datasets, focusing on primary factors such as genotype, treatment, and tumor status. Gene expression patterns were visualized through scatter plots, heatmaps, and histograms, highlighting variations in sample clustering and pathway activity. Significance testing for group comparisons used the ANOVA and Kruskal–Wallis tests wherever appropriate, with multiple comparison–adjusted P values calculated using the Benjamini–Hochberg method.

For pathway enrichment analysis, the hypergeometric tests were performed to identify significantly altered pathways based on Reactome and Kyoto Encyclopedia of Genes and Genomes annotations. Genes with significant expression differences between groups were analyzed for overrepresented pathways, and the results were visualized using bar and dot plots to convey enrichment significance and gene overlap.

Web application construction

A web application for LCAMGDB was developed using R Shiny, enabling interactive data exploration, visualization, and cross-study comparisons. Functionalities include gene-specific analyses, group comparisons, and PCA visualizations. The application incorporates dynamic filters for metadata fields such as sample type, treatment, and histology. Statistical tools for differential expression and pathway analysis are integrated into the interface. The application is hosted at https://lccl.shinyapps.io/LCAMGDB/ and supports tutorials for new users.

Data availability

The data analyzed in this study were obtained from the GEO and ArrayExpress, with sources listed in Supplementary Table S1. The study, sample, and genotype tables generated in this study are available as supplementary tables and downloadable from https://lccl.shinyapps.io/LCAMGDB/. The processed gene expression data by study or by platform are downloadable from the web app as well. All other raw data are available upon request from the corresponding author.

Results

Construction of the LCAMGDB

An exhaustive search in the GEO and ArrayExpress identified nearly 500 candidate lung cancer autochthonous mouse model datasets. Each of these studies was manually inspected to identify transcriptomic data generated from GEMMs, chemically induced tumors, or spontaneously formed tumors. Additionally, we included control lung samples and those exposed to carcinogenic treatments while excluding mouse cell lines and allografts into syngeneic recipients to ensure specificity to our research focus. We removed two datasets because of data redundancy from the reprocessed data (Supplementary Fig. S1). Our current data collection includes 77 datasets from 71 unique studies (Supplementary Table S1), which comprised 1,354 samples (**Fig. 1A**).

After a thorough data harmonization process, we contacted the data depositors and shared the curated data specific to their studies along with our data schema, soliciting their verification, rectifications, or any insights they could offer. Eighty-nine percent of the data depositors responded to our request to confirm our data curation. More than half of these contributors provided valuable corrections and insights, with some recommending additional datasets for future inclusion (detailed in Supplementary Table S2). The database was updated accordingly, integrating the depositors' revised data and constructive feedback.

Our analysis revealed a general trend toward small sample sizes across the datasets, with a median of 12 samples, ranging from 3 to 143 (**Fig. 1B**). The median number of detected genes per study is 20,942, with older microarray platforms reporting fewer genes (**Fig. 1C**). The majority of the samples originated from 71 GEMM datasets, including

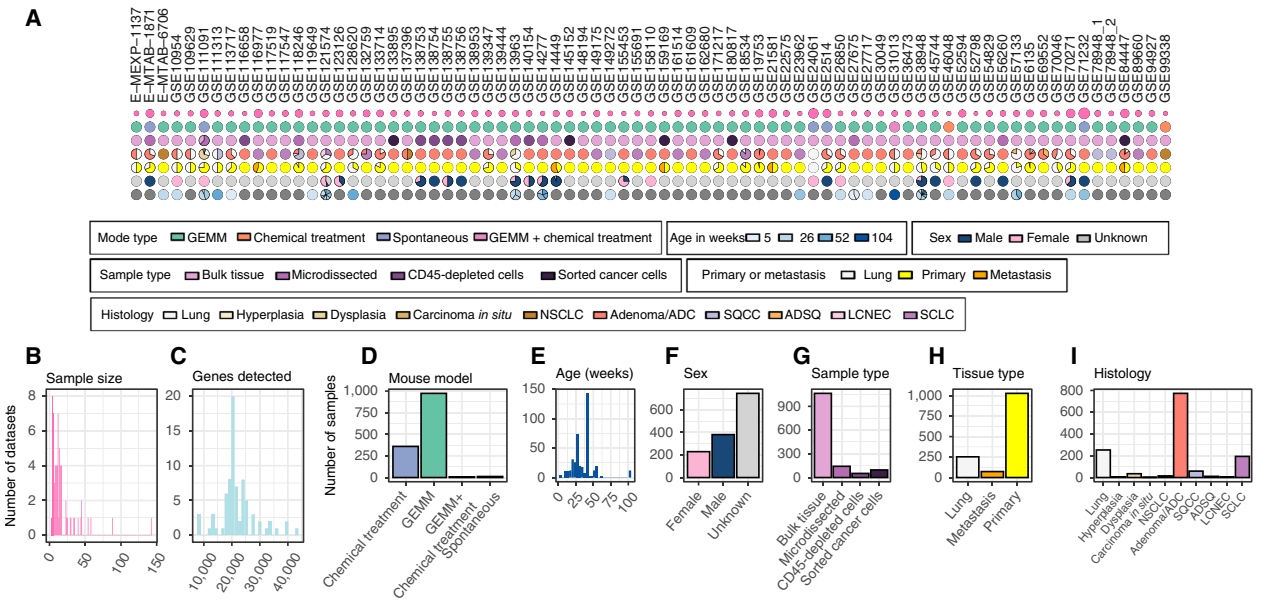

Figure 1.

Overview of sample characteristics and distribution in the LCAMGDB. **A,** Characteristics of individual datasets shown using pie charts. Each column represents a dataset, and each row corresponds to a specific attribute, with color-coding denoting the category. Attributes include model type, age, sex, sample type, histology, and primary or metastasis status. Dark gray denotes missing data. **B** and **C,** Sample size (**B**) and gene feature number (**C**) distribution across all datasets by bar plots. **D–I,** Distribution of samples by model type (**D**), age (**E**), sex (**F**), sample type (**G**), tissue type (**H**), and histology (**I**). Note that "Lung" under tissue type or histology can include normal wild-type lungs and also chemical-treated lungs from toxicology studies, or genetically modified non–wild-type lungs. ADSQ, adenosquamous carcinoma; LCNEC, large cell neuroendocrine carcinoma.

856 cancerous samples and 118 lung samples. Complementary to these, seven studies contained 368 samples generated from carcinogen-exposed models, including 239 cancerous samples and 129 lung samples. One unique dataset covered six spontaneous tumors and six lung samples in mice of 2 years of age (**Fig. 1D**). Age details were available for 407 samples (30%), with a median age of 34 weeks (**Fig. 1E**). Sex annotations were available for 609 samples (45%), encompassing 11 datasets with mixed sexes, six with exclusively female mice, and eight with exclusively male mice (**Fig. 1F**).

The sample types were predominantly from bulk tissue or microdissected specimens. A subset of 146 samples from 11 datasets underwent techniques such as CD45 depletion or fluorescence-based cancer cell sorting to reduce tumor microenvironmental contributions (**Fig. 1G**). Within the 1,101 curated tumor samples, 73 were identified as metastatic, with 53 metastases arising from ADC primary tumors and 20 from SCLC (**Fig. 1H**).

We curated 197 adenomas, 337 ADCs, and 236 cases classified as both based on authors' reports and literature reviews. Due to the overlapping lineage relationship of adenoma and ADC classifications, we opted to aggregate these under the "Adenoma/ADC" category for standardized histologic classification. This aggregation highlighted a dataset composition with 73% Adenoma/ADC, 18.3% SCLC, and only 5.6% SQCC, indicating an underrepresentation of SQCC when contrasted with its prevalence in human lung cancer (**Fig. 1I**).

The genetic landscape of GEMMs and comparison with patient mutation spectrum

About 859 precancerous lesions (hyperplasia, dysplasia, and carcinoma *in situ*) and tumor samples in the LCAMGDB collection were developed from the GEMMs. We curated genotype tables to record the involved genes, allele zygosity, genetic modifications, manipulative techniques, induction methods, and cells of origin. We illustrate the genetic alteration landscape, involving either single or combined manipulations of 54 genes in these GEMM samples in **Fig. 2A**. These include six human genes (*EGFR*, *IGF1R*, *EZH2*, *MYCN*, *CCNE1*, and *SNAI1*) and two viral genes (human papillomavirus *E6* and *E7*) introduced to the GEMMs. We compiled standardized genotypes and simplified them to harmonize genotype curation and identified a total of 122 unique standardized genotypes.

Considering lesions/tumors arising from *Kras* manipulation alone, for example, 10 distinct standardized genotypes were identified, which vary in genetic constructs and induction methods (**Fig. 2B**). Remarkably, all these genotypes harbor the G12D mutation, representing only ~15% of *KRAS* mutations in patients with non–small cell lung cancer (NSCLC; ref. 20). This disparity underscores the broader issue of limited genetic variation in GEMM tumors when compared with human cancers, which is also exemplified by *Trp53* mutations. Beyond simply inactivating p53, mutations in this gene are known to confer additional gain-of-function properties (21). However, in our current LCAMGDB database, out of 404 GEMM tumors with *Trp53* manipulation, only 16 samples originate from a single study using a $Trp53^{R172H}$ mutant model, with the remainder predominantly involving knockouts or knockdowns. In our more recent search (Supplementary Table S3), we identified newer models beginning to address this diversity, such as two new G12C datasets (22, 23). This highlights the importance of providing timely updates to reflect advancements as the field evolves.

The gene-centric distribution of genetic alterations in the tumor samples is detailed in **Fig. 2C**. Twenty-eight genes are predominantly activated, whereas 24 are primarily inactivated. Two genes, *Nfe2l2* and *Stat3*, were subject to both activation and inactivation

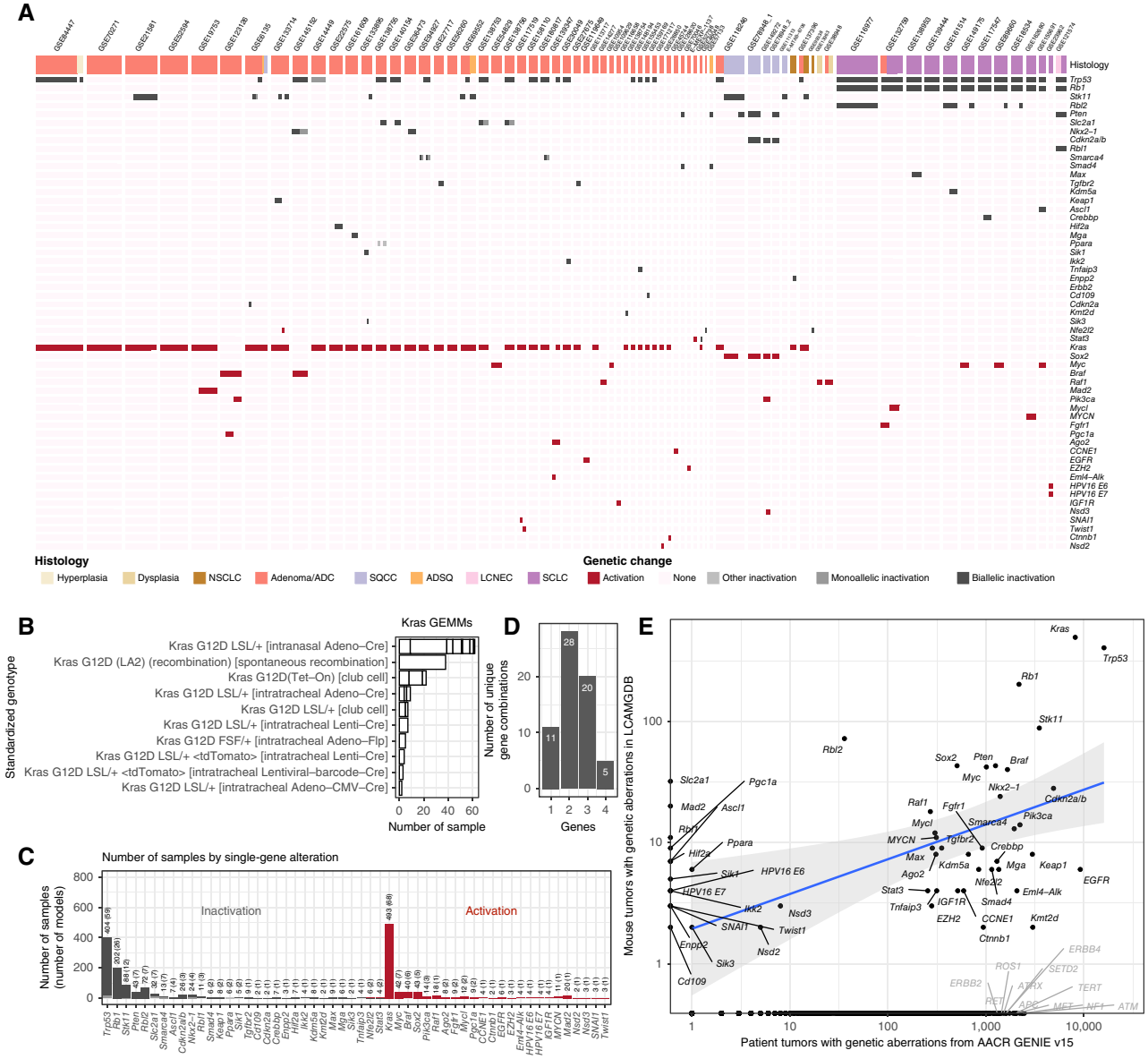

Figure 2.

Summary of GEMM genotypes in LCAMGDB. **A,** Landscape of genetic modifications in LCAMGDB GEMM tumors by dataset and histology. ADSQ, adenosquamous carcinoma; LCNEC, large cell neuroendocrine carcinoma. **B,** Sample count by standardized genotype in GEMM tumors with *Kras* mutation alone. Small boxes within the bars represent samples from different datasets. **C,** Sample count in GEMM tumors by single-gene alteration. The *y*-axis values are printed on the top of each bar to indicate the total number of tumors with specific genes altered, and the number of unique standardized genotypes involving the specified genes is given in parentheses. **D,** Count of GEMMs by the number of altered genes. Bars represent the number of GEMMs with one to four manipulated genes, irrespective of the manipulation method used or mutation. **E,** GEMM tumor alterations in LCAMGDB vs. human lung cancer genetic aberrations in the AACR GENIE version 15 database by gene. Selected human oncogene and tumor suppressor genes not represented in LCAMGDB are highlighted in gray.

studies within the GEMMs. This figure also denotes the number of standardized genotypes associated with each gene, represented in parentheses next to the total sample count. Notably, 29 of the 54 genes were exclusive to a single model. When considering the unique gene combinations, dual-gene manipulations emerged as the most common scenario, presented in 28 distinct instances. By contrast, manipulations of 11 different single genes were adequate to generate GEMM tumors (**Fig. 2D**). Only five models contained alterations in four genes (**Fig. 2D**), likely reflecting the inherent challenges associated with the time and expense required to generate mice with quadruple-modified alleles.

We next performed a comparative analysis of the frequency of genetic alterations in mouse lung cancer GEMM–derived tumors with those identified in human lung cancers (**Fig. 2E**), as recorded in the AACR GENIE v15 database, based on clinical sequencing data from real-world patient populations (17). Mutations in *TP53* and *KRAS*, among the most prevalent mutations in human lung cancer, are adequately represented in the GEMM tumors. The

observed positive correlation in gene alteration frequencies between mouse tumors and patient tumors suggests that GEMMs frequently incorporate genes commonly mutated in human lung cancer. However, our review indicates that some genes implicated in human lung cancers are understudied within the GEMM framework. For instance, the *Eml4–Alk* translocation and *Kmt2d* inactivation have each been characterized in only one study in our database. Moreover, pivotal oncogenes such as *ROS1*, *MET*, *RET*, and *ERBB4* and critical tumor suppressor genes like *NF1*, *ATM*, and *APC* are currently absent from the LCAMGDB (**Fig. 2E**). Supplementary Figure S2 details the frequency and types of genetic alterations for the top 100 genes most frequently altered in patients with lung cancer according to the AACR GENIE data, with an emphasis on 78 genes that are not yet included in the LCAMGDB. These findings underscore the need to broaden the scope of lung cancer GEMM development and characterization to cover a more extensive array of genetic drivers of the disease.

Harmonization of gene expression data

To address the limited sample size within individual datasets, we acquired raw data wherever possible and reprocessed them using standardized pipelines organized by platforms, each with the latest probe and gene annotations. This standardization effort enabled us to make reprocessed data available for 85% of the samples (**Fig. 3A**). Notably, approximately half of these samples ($n = 563$) is derived from RNA-seq and encompasses 38 distinct datasets (**Fig. 3B**). PCA conducted on the top 1,000 variable genes from the reprocessed RNA-seq data revealed that the first two principal components (PC) capture 62% of the total variance, indicating a strong structuring of the data (**Fig. 3C**). Despite potential batch effects, the PCA demonstrates that different datasets exhibit substantial overlap (**Fig. 3D**), with clear distinctions observed between SCLC and NSCLC samples along PC1 and between primary and metastatic samples along PC2 (**Fig. 3E** and **F**, respectively). For microarray datasets, we used a similar processing strategy. As an example, PCA on data from the Mouse430_2 platform (the most represented microarray platform with 283 samples across 15 datasets) demonstrated a comparable success in data integration (Supplementary Fig. S3). Although batches from various experimental conditions, sample types, and biological differences such as mouse age, sex, and strain may still be present, our reprocessing method seems to have effectively consolidated the datasets, thereby facilitating cross-dataset comparisons and potentially uncovering broader trends within the merged data.

A user-friendly web application for LCAMGDB

To facilitate the exploration and analysis of the LCAMGDB data, we constructed a web application that can be accessed at https://lccl.shinyapps.io/LCAMGDB/. This application is structured into two primary sections: a data review panel and an analysis panel. We have included eight sets of step-by-step tutorials for these functionalities providing guidance for users to navigate and analyze the datasets.

Within the data review panel, the "Overview" tab presents graphical summaries of the LCAMGDB, and the "Studies," "Samples," and "GEMMs" tabs allow users to review and refine detailed data tables. These tables correspond to Supplementary Tables S4–S7 in this article, and they are also downloadable from the web app. Specifically, the "GEMMs" tab displays a table in which genetic alterations are recorded with one gene per line. Each genotype within a study is distinctively highlighted to facilitate visual separation. Users can customize their view, choosing which columns to display and applying filters to refine row entries, such as querying specific gene combinations, with an illustrative example as shown in Supplementary Fig. S4. Additionally, we provide data download links organized by dataset for author-processed data and by platform for reprocessed data, allowing users to conduct their own analyses offline.

The analysis panel offers users an interactive environment to delve deeper into the gene expression profiles across multiple datasets. To examine a single gene of interest, the "Depositor-processed" data option allows researchers to analyze the expression data as originally submitted, maintaining consistency within datasets and enabling reliable within-dataset comparisons. The results are visualized as a series of dot plots, ranked by the statistical significance of expression differences determined by one-way ANOVA. The "Merged by platform" data option allows users to examine the reprocessed data by platform, leveraging the harmonized datasets to discern patterns and insights across different studies. Additionally, researchers may perform two-group comparisons across the transcriptome to screen for genes and pathways with differential expression patterns. The following sections illustrate these tools with practical examples.

Comparisons of a single gene in individual datasets

We offer three options for single-gene expression comparison using depositor-processed data. The first, which compares expression by genotype and/or treatment, provides the broadest dataset range and includes versatile sample filtering capabilities. Users can refine the analysis parameters by using the available filters within the dropdown menu, tailoring the analysis to their specific research interests (Supplementary Fig. S5). As exemplified in **Fig. 4**, in which the top 6 of 44 datasets qualified from the specified criteria are shown, we identified several genetic and treatment conditions that induced the most prominent *Cd274* (PD-L1) expression changes in NSCLC bulk tissue samples. This is particularly notable in models with *Stk11* (also known as *Lkb1*) knockout, wherein *Cd274* expression is markedly downregulated, corroborating clinical findings that *STK11* mutations are significantly enriched among PD-L1–negative lung tumors (24). On the other hand, treatment with oxaliplatin and cyclophosphamide (25) known to induce immunogenic cell death increased the expression of *Cd274* (**Fig. 4**, bottom).

To enable more focused analyses on treatment/carcinogenesis response and cancer progression, we devised two additional comparison options for analyzing gene expression: one for treatment comparisons from 10 studies and another for examining differences between primary tumors and metastatic lesions from five studies. The treatment comparison tool is showcased by analysis of the B-cell marker *Cd19* to reveal distinct trends in tumor microenvironments (**Fig. 5A**). B cells play dual roles in the tumor microenvironment, contributing to both antitumor and protumorigenic processes. On one hand, they support antitumor immunity by forming tertiary lymphoid structures and facilitating immunoresponses, which correlate with improved prognosis (26) and enhanced immunotherapy outcomes (27). On the other hand, immunosuppressive regulatory B cells can promote tumor development by secreting anti-inflammatory cytokines (28). In our analysis, we observed *Cd19* upregulation indicative of B-cell infiltration in a Braf-driven GEMM under MAPK inhibitor treatment (GSE145152 dataset), which coincides with tumor regression in this study (29). *Cd19* downregulation, by contrast, was noted in Kras-driven GEMMs treated with antioxidants (GSE52594 dataset), which coincides with accelerated tumor progression in this study

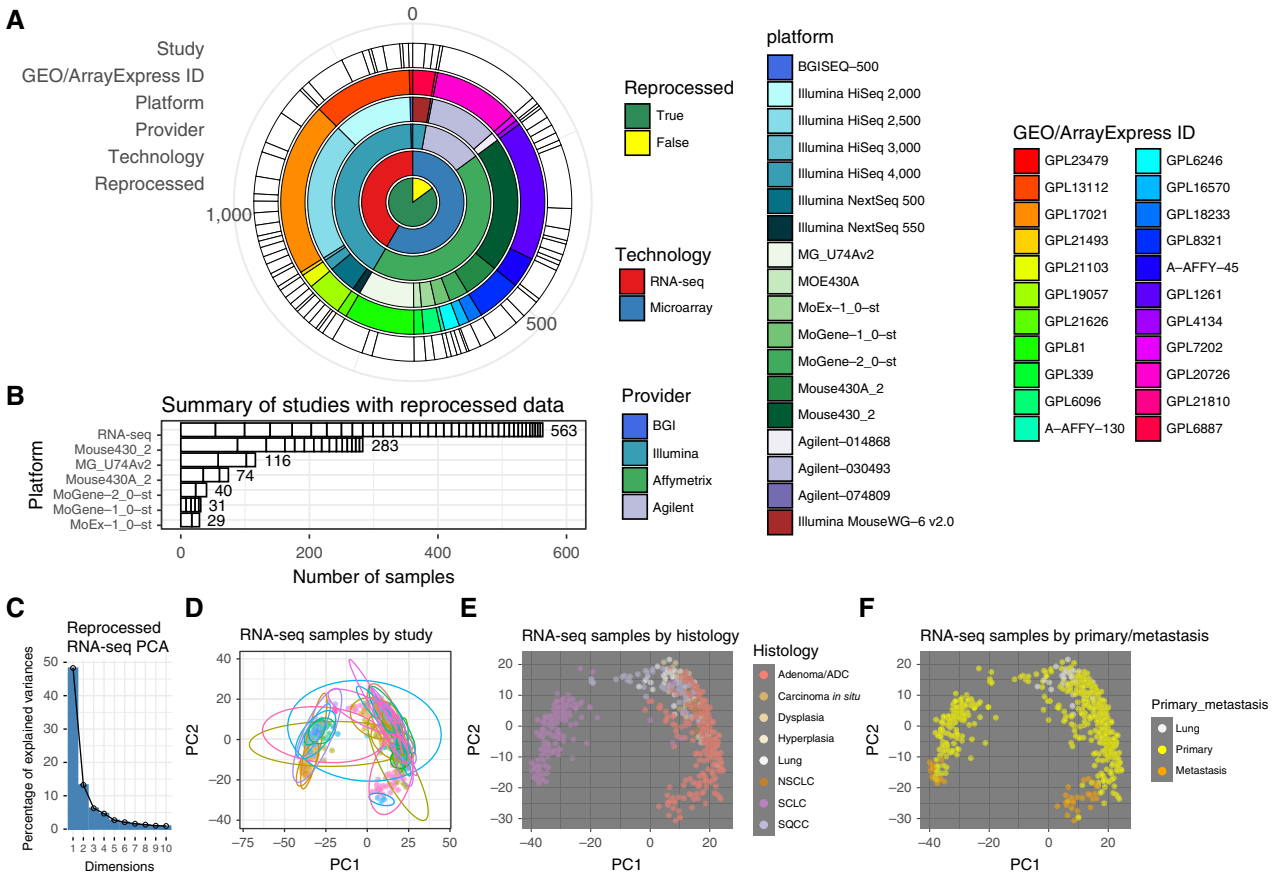

Figure 3.

Data reprocessing by platform. **A,** Hierarchical relationship of transcriptomic profiling technology and platforms. Eighty-five percent (1,152 samples) of the LCAMGDB gene expression data were reprocessed. **B,** Platforms with multiple studies reprocessed through standardized workflow. Each box within the bars represents a single dataset. **C,** In PCA using the 1,000 most variable genes from reprocessed RNA-seq data, the top two PCs account for 62% of the total variance. **D–F,** Distribution of 563 RNA-seq samples by source dataset (**D**), histology (**E**), and primary/metastasis status (**F**). BGI, Beijing Genomics Institute.

(30). These findings suggest that B-cell infiltration in these treatment contexts supports antitumor immunity. Additionally, we observe *Cd19* upregulation in Egfr-driven GEMMs subjected to a high-fat diet (GSE119649) and a Kras-driven GEMM under a high-caloric diet (GSE56260), consistent with findings that obesity creates a more inflammatory tumor microenvironment in mouse models (31). Similar observations in human studies have associated high body mass index with improved overall survival after immune checkpoint inhibitor therapy in advanced NSCLC (32).

The primary/metastasis comparison tool is exemplified by the examination of *Ezh2* expression, a component of the polycomb repressive complex 2, which is implicated in gene silencing (**Fig. 5B**). With the fine curation of metastatic status in samples from GSE84447 (33), we observe *Ezh2* expression increase with tumor invasiveness in Kras-driven models, which corroborates clinical findings that this chromatin modifier is associated with cancer progression and metastasis (34).

Comparisons of a single gene in merged reprocessed data

Analysis of reprocessed data merged by profiling platform enables cross-study comparison. Users may select from six platforms with two or more merged datasets and further filter the input sample (as in Supplementary Fig. S5). We provide two visualization approaches for analyses. The first approach is to generate a dot plot with samples colored by histology and ordered by the median expression of the user-defined gene in groups stratified by a combination of data source, genotype, treatment, primary/metastasis status, and sample type. In the reprocessed RNA-seq data, this gives rise to 115 unique groups and creates a very extensive plot. To demonstrate this tool, we refined our selection to primary tumors from the RNA-seq reprocessed data and examined the expression of *Cd19*. The lowest expression is found in sorted cancer cells and samples with CD45 depletion (**Fig. 6**, bottom), as expected from the depletion of immune cells. Bulk tissue samples with the lowest *Cd19* expression are from SCLC, consistent with the immune cold nature of this histologic subtype (35–37). The highest expression of *Cd19* is found in dysplasia samples derived after treatment with the alkylating agent N-nitroso-tris-chloroethylurea, potentially due to abundant neoantigen resulting from carcinogen treatment (**Fig. 6**, top). Users may select from additional profiling microarray platforms. For example, the analysis of *Ezh2* expression in reprocessed data of Mouse430_2 reveals that its expression is much higher in SCLC than in NSCLC samples (Supplementary Fig. S6), as previously established (38, 39).

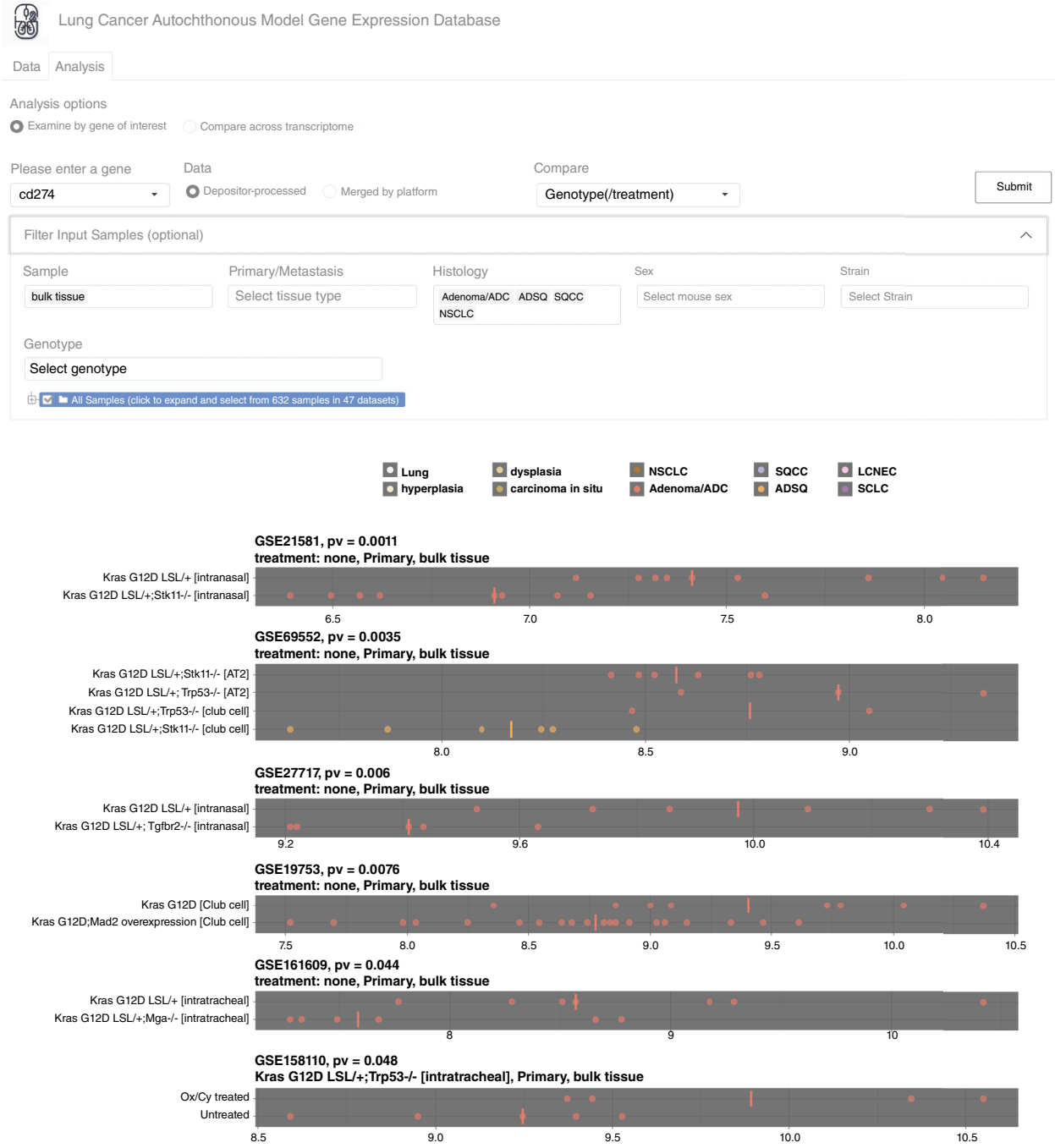

Figure 4.

Interactive visualization of gene expression across multiple datasets. This figure features the web application's capability for users to interrogate the expression of a selected gene, *Cd274* (PD-L1), across a range of datasets. The "Depositor-processed" option leverages the original data processed in the deposited datasets, optimizing the within-dataset comparisons. Users can tailor the analysis by applying filters via the dropdown menu. After selecting the appropriate parameters and clicking "Submit," the application generates dot plots arrayed by the statistical significance of their expression differences, as assessed by one-way ANOVA. Displayed here are the top six datasets from the full results, giving users a snapshot of the gene expression landscape within the application's extensive repository. Bars in each plot denote the group median. ADSQ, adenosquamous carcinoma; LCNEC, large cell neuroendocrine carcinoma; Ox/Cy, oxaliplatin and cyclophosphamide.

The second visualization option generates a two-dimensional PCA plot, with sample points colored based on variables such as gene expression, histology, primary/metastasis status, sample type, or data source. Precomputed and *ad hoc* PCA computation is available, and the latter allows users to dynamically recompute PCs based on their selected subsets of data. In **Fig. 7A**, we demonstrate

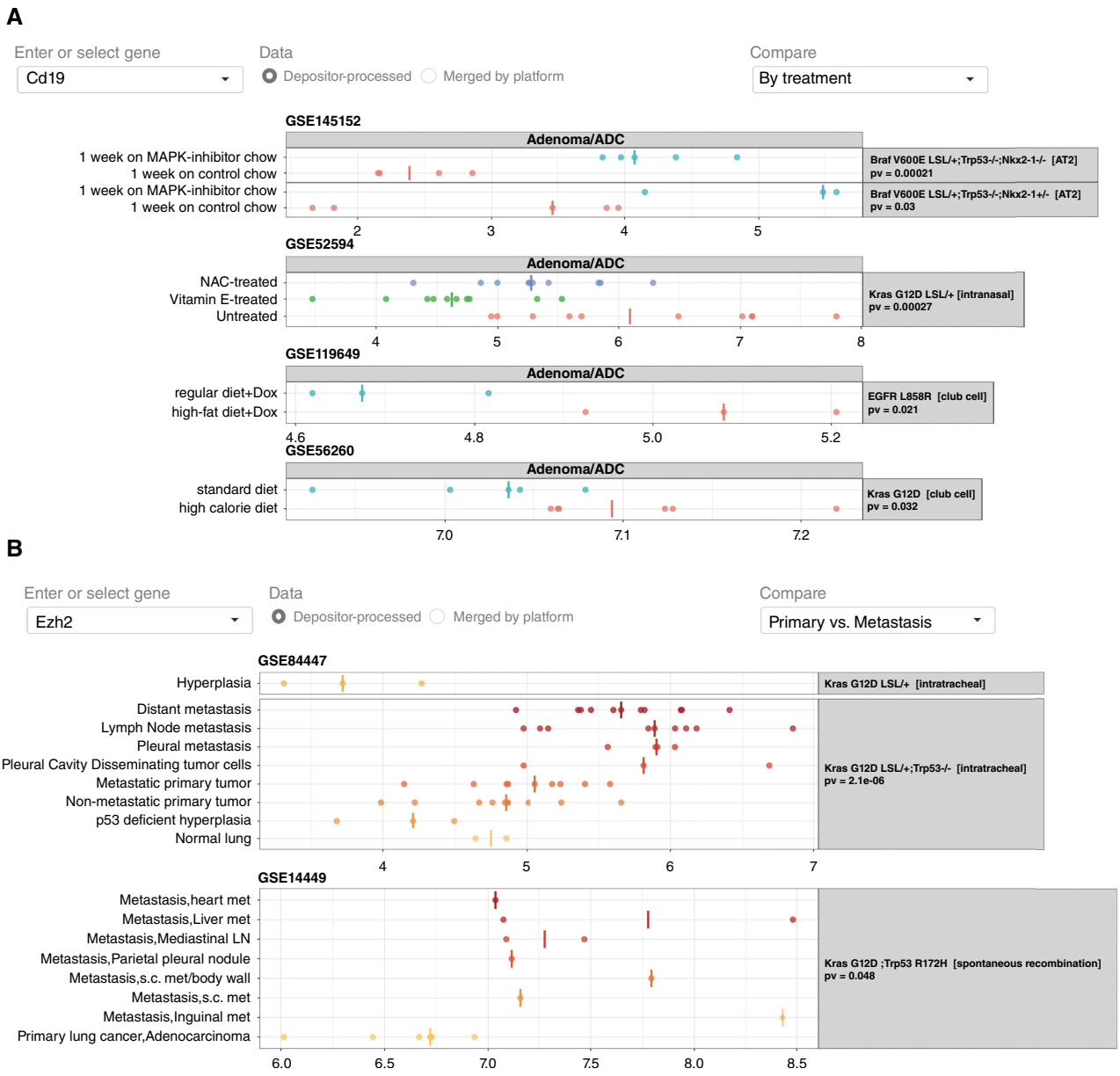

Figure 5.

Gene expression comparison by treatment and primary/metastasis status. **A,** Expression of *Cd19* revealing B-cell infiltration in various treatment contexts. Data points are categorized by treatment conditions under each genotype. **B,** Expression of *Ezh2* in primary and metastatic tumor samples. Color gradient signifying the spectrum of metastatic progression stages. *P* values from one-way ANOVA are indicated, and results were ordered by statistical significance. For conciseness, only the top four datasets of the 10 for *Cd19* (**A**) and the top two of the five for *Ezh2* (**B**) have been included in the snapshots. Bars in each plot denote group median.

this with RNA-seq samples colored to reflect *Ascl1* expression—a neuroendocrine lineage transcription factor instrumental in SCLC pathogenesis (40, 41). Consistent with the histologic segregation observed in **Fig. 3E**, samples with lower PC1 scores, predominantly of SCLC histology, exhibit elevated *Ascl1* expression.

However, some outliers were identified: a few ADC samples with higher PC1 scores also showed high *Ascl1* levels. Using the interactive tooltips, users can examine these outliers in detail to understand their context. In this example, we found that ADC samples are derived from a model with constitutively active $Fgfr1^{K656E}$ in an *Rb1/Trp53*–deficient background (**Fig. 7A**; ref. 42). Although *Rb1/Trp53* models generate classic SCLC, Fgfr1 activation in this model has reduced *Ascl1* expression (42), making it lower than classic SCLC but still higher than classic ADC tumors from other models (Supplementary Fig. S7).

In another example using the PCA plot with histology color mapping, there are a few notable outliers among the NSCLC samples that are identified as SCLC (**Fig. 7B**). This particular discrepancy is clarified upon recognizing that these SCLC samples have

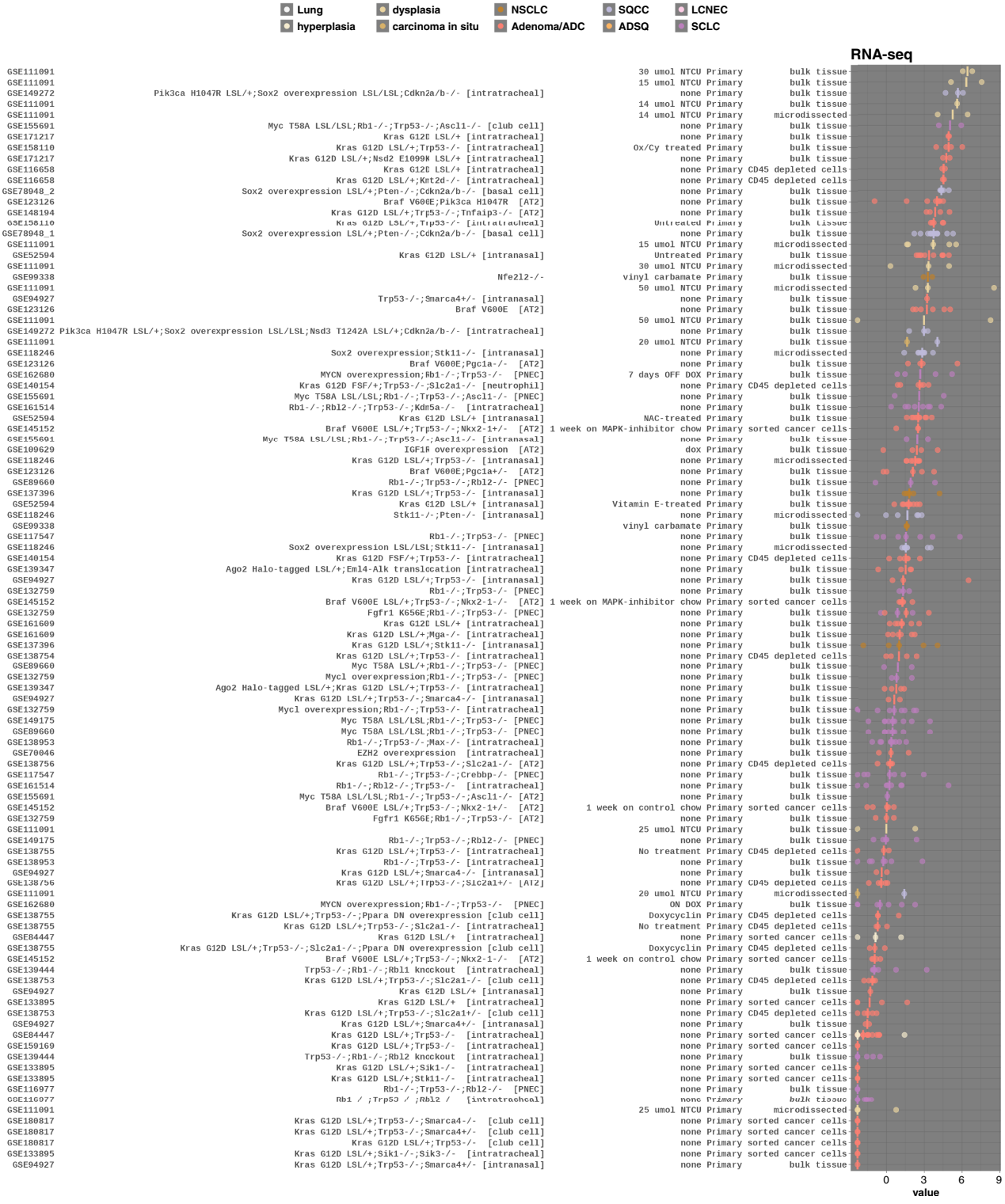

Figure 6.

Expression of *Cd19* in reprocessed RNA-seq data. Each dot represents a unique sample, colored according to histology and ordered by the median expression of *Cd19*. The inputs are filtered to display primary tumors only. The median of the group is shown as a bar for each row. ADSQ, adenosquamous carcinoma; LCNEC, large cell neuroendocrine carcinoma; NTCU, N-nitroso-tris-chloroethylurea; Ox/Cy, oxaliplatin and cyclophosphamide.

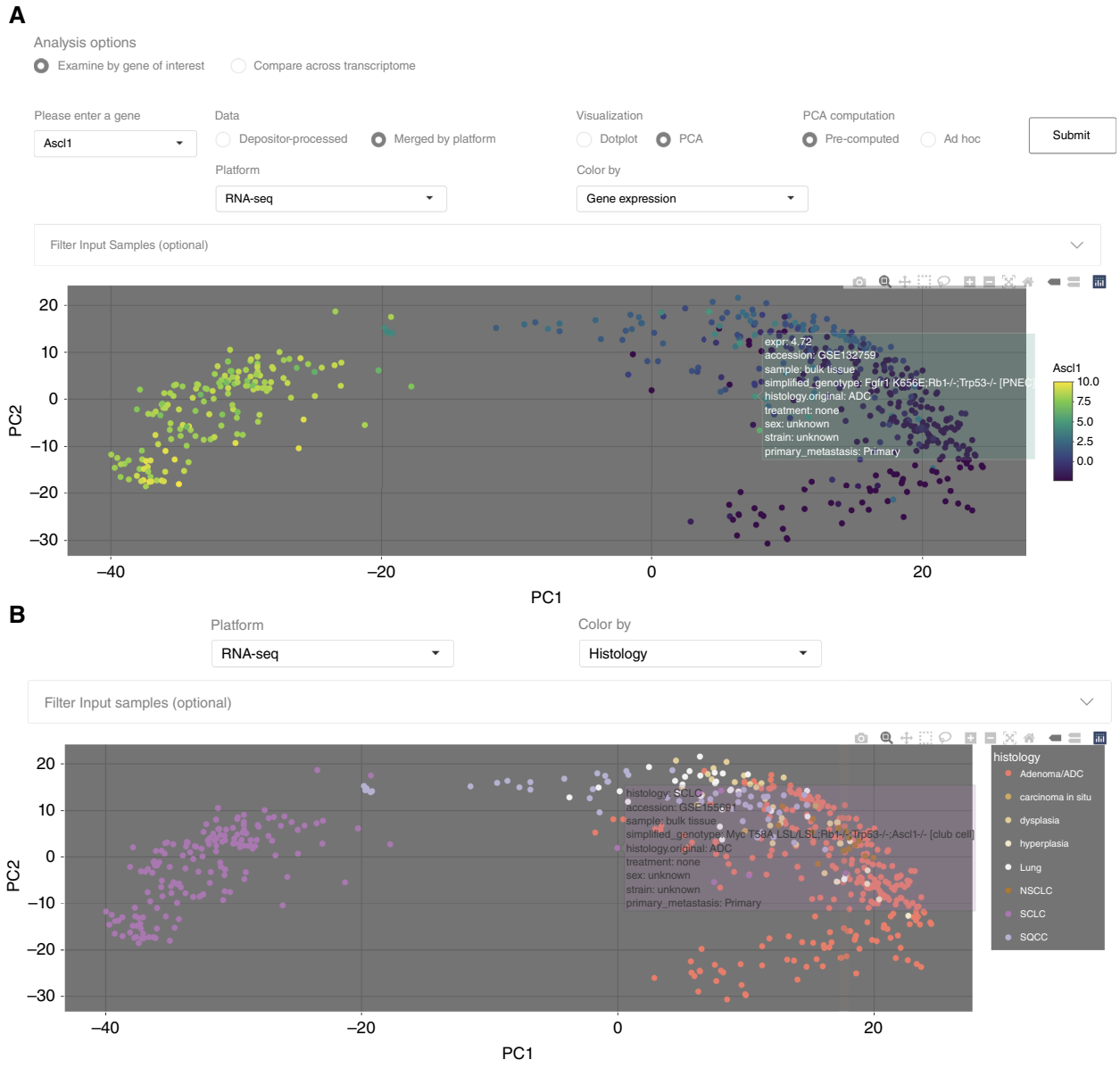

Figure 7.
Interactive visualization of gene expression in reprocessed data merged by platform. **A,** PCA plot of reprocessed RNA-seq samples, color-coded by the expression of *Ascl1*, a neuroendocrine lineage transcription factor highly expressed in SCLC. The interactive tooltip uncovers the origin of an outlier sample with elevated *Ascl1* levels as an ADC sample from an Rb1/p53-deficient model featuring Fgfr1 activation. **B,** PCA plot colored by histology. Details of an SCLC sample located near the NSCLC samples are displayed. This outlier sample has *Ascl1* knocked out, which explains the loss of neuroendocrine gene expression that renders the transcriptomic profile more similar to NSCLC.

undergone *Ascl1* knockout, leading to a complete loss of neuroendocrine cell fate (41) that rendered the transcriptomic landscape of the SCLC sample more akin to that of NSCLC samples, explaining its outlier position in the PCA plot. These examples underscore how interactive plots equipped with tooltips provide critical insights into sample-specific details, enabling deeper exploration of unexpected patterns.

Users also have the option to visualize data points according to the data source; the interactive plots enable users to selectively focus on, or exclude, samples from specific sources by clicking or double-clicking on dataset identifiers, thereby providing a clearer understanding of the underlying data distribution across studies (Supplementary Fig. S8).

Transcriptome-wide group comparison

The transcriptome-wide group comparison feature in LCAMGDB provides users with the ability to conduct differential expression analyses between two distinct sample groups. This

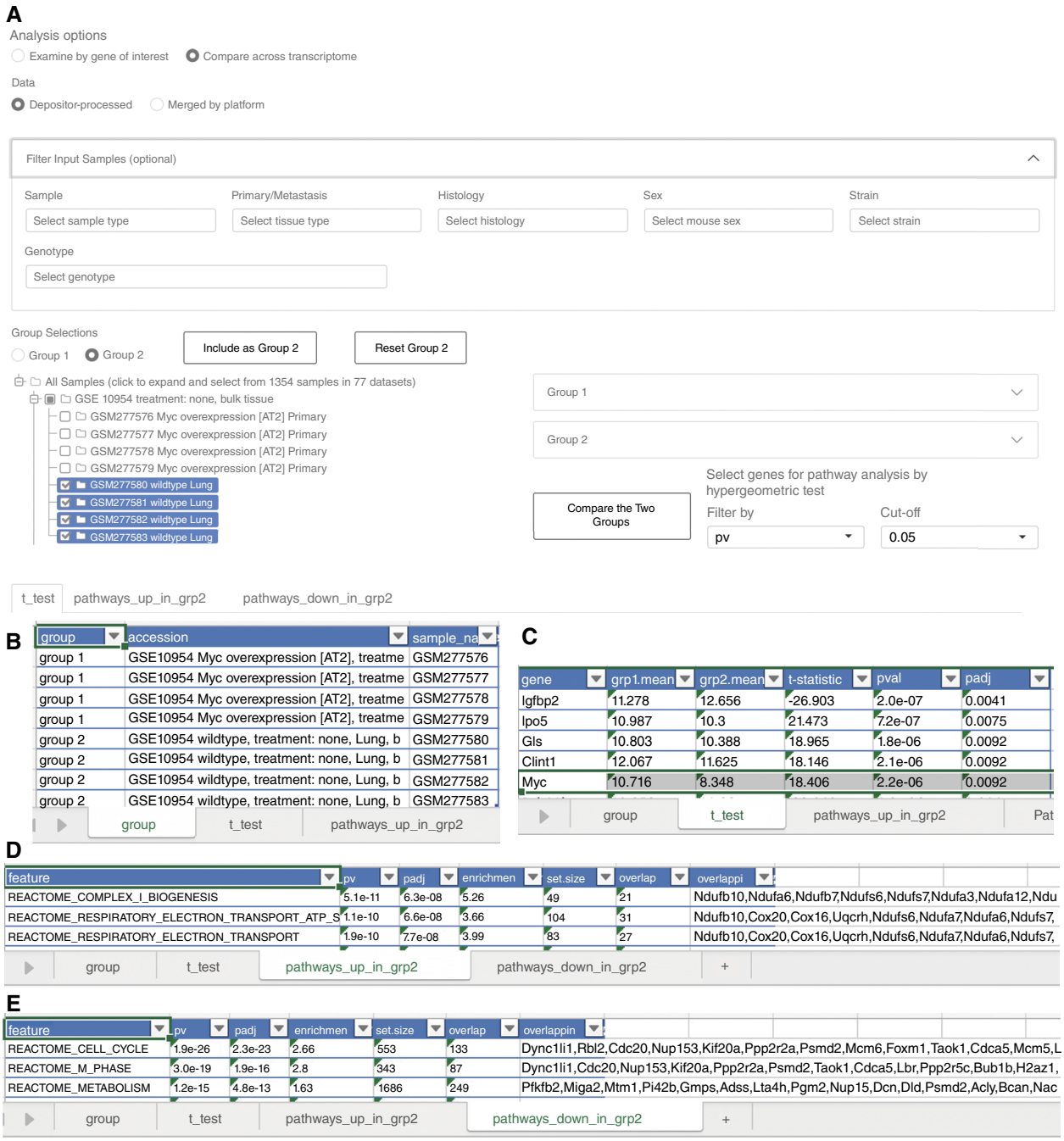

Figure 8.

Transcriptome-wide group comparison of Myc-overexpressed tumors vs. wild-type lung samples. **A,** User interface for defining group 1 (Myc-overexpressed tumors) and group 2 (wild-type lung samples) using depositor-processed data. The panel allows users to refine input datasets and samples based on variables such as genotype, histology, and treatment, with a sample tree for finalizing sample selection. **B–E,** Example outputs from the downloadable Excel file, including the group stratification (**B**), the gene-level comparison from the two-sample *t* test (**C**), and pathway enrichment analysis with hypergeometric test from Reactome pathway library for genes higher in the lungs (group 2; **D**) or higher in tumors (**E**).

functionality is available for both depositor-processed data and reprocessed data.

In **Fig. 8**, we provide an example comparing Myc-overexpressed tumors with wild-type lung samples from GSE10954 (43). The interactive filters allow the selection of groups based on variables such as genotype, treatment, or sample type (**Fig. 8A**). After defining the groups, users may explore the analysis outputs on the web app or download them as an Excel file for further examination. **Figure 8B–E** are examples of the downloadable results, including the user-defined group stratification (**Fig. 8B**), differentially expressed genes

from the two-sample *t* test, identifying *Myc* among the most upregulated genes in tumor samples, along with other well-known Myc targets such as *Gls* (encoding glutaminase; **Fig. 8C**; ref. 44). Furthermore, pathway enrichment analyses revealed higher expression of oxidative phosphorylation genes in the lungs (**Fig. 8D**) and upregulation of cell-cycle genes in tumors (**Fig. 8E**). This example analysis highlights how this tool uncovers biologically meaningful patterns, leveraging the high fidelity of depositor-processed data.

Reprocessed data analyses expand the scope to include samples across multiple datasets. Users can select groups from harmonized data but should be mindful of potential batch effects. Researchers may reference our online tutorial with an example that compared SCLC and NSCLC samples from different studies, to better understand how to use this tool. Although this approach is ideal for identifying trends and validating findings, depositor-processed data remain more reliable for finer, exploratory analyses, minimizing the risk of confounding factors. We welcome collaborations with researchers interested in performing more sophisticated statistical analyses or customized study designs.

Discussion

Our LCAMGDB presents a curated compendium of transcriptomic data covering 1,354 samples from 71 studies, summarizing a vast array of lung cancer mouse models. This resource interrogates the genetic aberration landscape across 859 GEMM tumors, providing an unprecedented platform for cross-study comparison. Our collaborative approach, engaging with data depositors, has ensured the integrity and enhancement of the database, leading to its current comprehensive state.

However, we have to consider the limitations inherent to the database's scope. The LCAMGDB is founded on transcriptomic data, which excludes mouse models lacking such characterization. This limitation underscores the need for an inclusive approach that considers unpublished or less-publicized models to achieve a comprehensive and representative overview of genetic alterations in lung cancer autochthonous models. The dynamic nature of scientific research also necessitates the LCAMGDB to be a living database, with ongoing updates and expansions informed by both community feedback and continual data discovery. Future versions will integrate additional datasets, reflecting the latest advancements and filling in gaps identified through collaborative suggestions and our active searches. These updates will include datasets featuring key *Kras* mutant models, such as *Kras* G12C (e.g., *Kras/Trp53* and *Kras/Stk11* models), which are not represented in the current release. The candidate datasets to be included in the next update are listed in Supplementary Table S3. These include datasets from a more recent GEO/ArrayExpress screen and datasets suggested by the community, such as toxicology studies of mouse lungs treated with carcinogens. We will also continue to develop the collaborator login system, enabling researchers to privately assess their data alongside public datasets. Although not expounded upon in this article, this function highlights the platform's potential for fostering collaborative research endeavors.

It is important to note the caution required in interpreting the reprocessed data. Although standardization efforts have been rigorous, batch effects from diverse experimental and genetic backgrounds may still be present. Users should not attribute the expression variation solely to genotype. Future updates will aim to support meta-analytical capabilities and provide insights from comprehensive cross-transcriptomic evaluations. Central to the LCAMGDB use is its facilitation of comprehensive comparisons between mouse models, additional preclinical model data, such as patient-derived cell lines, patient-derived xenografts, syngeneic mouse models, and human lung cancer data. This alignment is crucial for translating preclinical findings to clinical relevance, aiding in the development of personalized therapies. The database's current iteration lays the groundwork for such comparative studies, which we plan to explore in-depth in subsequent analyses.

In sum, the LCAMGDB offers a robust framework for the exploration of gene expression data within mouse models, setting the stage for additional comprehensive analyses that have the potential to unveil new discoveries and guide the design of future models for a more accurate reflection of human lung cancer.

Authors' Disclosures

L. Cai reports grants from the NCI and American Cancer Society during the conduct of the study. R.J. DeBerardinis reports personal fees from Agios Pharmaceuticals, Vida Ventures, and Atavistik Bio outside the submitted work. J.E. Beane reports grants from American Cancer Society during the conduct of the study, as well as grants from Johnson & Johnson, NCI, LUNGevity Foundation, and American Lung Association outside the submitted work. C.P. Concepcion-Crisol reports grants from American Cancer Society during the conduct of the study, as well as grants from the Lung Cancer Research Foundation and Prelude Therapeutics and other support from Genentech outside the submitted work. B.M. Grüner reports grants from German Research Foundation (Deutsche Forschungsgemeinschaft), German Cancer Aid, the German Federal Ministry of Education and Research, Wilhelm Sander-Stiftung, and the German Federal Ministry of Economic Affairs and Climate Action outside the submitted work. Y. Lissanu reports personal fees from aMoon Fund outside the submitted work, as well as a patent for WO 2023/129506 A1 pending. R. Martinelli reports grants from Ministero della Salute (Roma), Convenzione CEINGE-MIUR (2000) art 5.2 during the conduct of the study. P.K. Mazur reports being a consultant and stockholder of Ikena Oncology, Inc., and Alternative Bio, Inc. S.A. Mazzilli reports grants from American Cancer Society during the conduct of the study, as well as grants from Johnson & Johnson, NCI, LUNGevity Foundation, and American Lung Association outside the submitted work. S. Mii reports grants from JSPS KAKENHI during the conduct of the study. M.G. Oser reports grants from Novartis, Circle Pharma, Bristol Myers Squibb, and Eli Lilly and Company outside the submitted work. C.A. Powell reports grants from NIH during the conduct of the study. M. Santos reports grants from Instituto de Salud Carlos III during the conduct of the study, as well as grants from Instituto de Salud Carlos III outside the submitted work. P.T. Tran reports grants from NCI during the conduct of the study; personal fees from Natsar Pharmaceuticals, RefleXion Medical, Bayer, Janssen, Regeneron, Lantheus, Pfizer, and Amgen outside the submitted work; and a patent for Compounds and Methods of Use in Ablative Radiotherapy, patent number 9114158, with royalties paid from Natsar Pharmaceuticals. C.M. Rudin reports personal fees from AbbVie, Amgen, Boehringer Ingelheim, Daiichi Sankyo, Hoffmann-La Roche, Jazz Pharmaceuticals, Eli Lilly and Company, Puma Biotechnology, and Treeline Biosciences outside the submitted work, as well as being on the scientific advisory boards for Auron Therapeutics, DISCO Pharmaceuticals, and Earli. T.G. Oliver reports grants from NCI (5U24-CA213274-08, 1R01-CA262134-01A1, 5R01-CA244841-06, and 5R01-CA251147-04) during the conduct of the study; other support from Auron Therapeutics and personal fees from Light Horse Therapeutics and Nuage Therapeutics outside the submitted work; a patent for SCLC subtyping (US11,124,841B2) issued; and being a senior editor at *Cancer Research* and on the editorial board for *Genes & Development*. Y. Xie reports grants from NIH and Cancer Prevention and Research Institute of Texas (CPRIT) during the conduct of the study. J.D. Minna reports grants from NCI and personal fees from NIH and University of Texas Southwestern Medical Center during the conduct of the study. No disclosures were reported by the other authors.

Authors' Contributions

L. Cai: Conceptualization, resources, data curation, software, formal analysis, supervision, funding acquisition, validation, investigation, visualization, methodology, writing–original draft, project administration, writing–

review and editing. **F. Wu:** Software. **Q. Zhou:** Data curation. **Y. Gao:** Data curation. **B. Yao:** Resources. **R.J. DeBerardinis:** Funding acquisition, writing–review and editing. **G.K. Acquaah-Mensah:** Data curation, validation. **V. Aidinis:** Data curation, validation. **J.E. Beane:** Data curation, validation. **S. Biswal:** Data curation, validation. **T. Chen:** Data curation, validation. **C.P. Concepcion-Crisol:** Data curation, validation. **B.M. Grüner:** Data curation, validation. **D. Jia:** Data curation, validation. **R.A. Jones:** Data curation, validation. **J.M. Kurie:** Data curation, validation. **M.G. Lee:** Data curation, validation. **P. Lindahl:** Data curation, validation. **Y. Lissanu:** Data curation, validation. **C. Lorz:** Data curation, validation. **D. MacPherson:** Data curation. **R. Martinelli:** Data curation, validation. **P.K. Mazur:** Data curation, validation. **S.A. Mazzilli:** Data curation, validation. **S. Mii:** Data curation, validation. **H.P. Moll:** Data curation, validation. **R.A. Moorehead:** Data curation, validation. **E.E. Morrisey:** Data curation. **S.R. Ng:** Data curation, validation. **M.G. Oser:** Data curation, validation. **A.R. Pandiri:** Data curation, validation. **C.A. Powell:** Data curation, validation. **G. Ramadori:** Data curation, validation. **M. Santos:** Data curation, validation. **E.L. Snyder:** Data curation, validation. **R. Sotillo:** Data curation, validation. **K.-Y. Su:** Data curation, validation. **T. Taki:** Data curation, validation. **K. Taparra:** Data curation, validation. **P.T. Tran:** Data curation. **Y. Xia:** Data curation, validation. **J.E. van Veen:** Data curation, validation. **M.M. Winslow:** Data curation, validation, methodology. **G. Xiao:** Supervision, funding acquisition. **C.M. Rudin:** Supervision, funding acquisition. **T.G. Oliver:** Data curation, supervision, funding acquisition, validation, methodology, writing–review and editing. **Y. Xie:** Supervision, funding acquisition. **J.D. Minna:** Conceptualization, supervision, funding acquisition, project administration, writing–review and editing.

Acknowledgments

This study is supported by funding from R01CA285336, U24CA213274, ACS-IRG (IRG-21-142-16), P50CA070907, R01GM140012, R01GM141519, R01DE030656, R35CA220449, R35CA263816, R01CA271540, R01CA244841, U01CA249245, U01AI156189, U01CA213338, R35GM136375, R01CA212415, R01CA240317, R01CA272945, R37CA251629, P50CA228944, as well as Howard Hughes Medical Institute (to R.J. DeBerardinis); the German Research Foundation (Deutsche Forschungsgemeinschaft) GR4575/1-2 and SFB1430 project-ID 424228829 (to B.M. Grüner); and Instituto de Salud Carlos III Project PI21/00764 (to M. Santos), co-funded by the Fondo Europeo de Desarrollo Regional and the European Union. The authors declare no competing interests.

Note

Supplementary data for this article are available at Cancer Research Online (http://cancerres.aacrjournals.org/).

Received May 15, 2024; revised December 23, 2024; accepted February 26, 2025; posted first April 29, 2025.

References

1. Al Bakir M, Huebner A, Martínez-Ruiz C, Grigoriadis K, Watkins TBK, Pich O, et al. The evolution of non-small cell lung cancer metastases in TRACERx. Nature 2023;616:534–42.
2. Campbell JD, Alexandrov A, Kim J, Wala J, Berger AH, Pedamallu CS, et al. Distinct patterns of somatic genome alterations in lung adenocarcinomas and squamous cell carcinomas. Nat Genet 2016;48:607–16.
3. The Cancer Genome Atlas Research Network. Comprehensive molecular profiling of lung adenocarcinoma. Nature 2014;511:543–50.
4. Frankell AM, Dietzen M, Al Bakir M, Lim EL, Karasaki T, Ward S, et al. The evolution of lung cancer and impact of subclonal selection in TRACERx. Nature 2023;616:525–33.
5. George J, Lim JS, Jang SJ, Cun Y, Ozretić L, Kong G, et al. Comprehensive genomic profiles of small cell lung cancer. Nature 2015;524:47–53.
6. George J, Maas L, Abedpour N, Cartolano M, Kaiser L, Fischer RN, et al. Evolutionary trajectories of small cell lung cancer under therapy. Nature 2024;627:880–9.
7. Martínez-Ruiz C, Black JRM, Puttick C, Hill MS, Demeulemeester J, Larose Cadieux E, et al. Genomic-transcriptomic evolution in lung cancer and metastasis. Nature 2023;616:543–52.
8. Ghandi M, Huang FW, Jané-Valbuena J, Kryukov GV, Lo CC, McDonald ER III, et al. Next-generation characterization of the cancer cell line encyclopedia. Nature 2019;569:503–8.
9. Gazdar AF, Girard L, Lockwood WW, Lam WL, Minna JD. Lung cancer cell lines as tools for biomedical discovery and research. J Natl Cancer Inst 2010;102:1310–21.
10. Sun H, Cao S, Mashl RJ, Mo CK, Zaccaria S, Wendl MC, et al. Comprehensive characterization of 536 patient-derived xenograft models prioritizes candidatesfor targeted treatment. Nat Commun 2021;12:5086.
11. Davis S, Meltzer PS. GEOquery: a bridge between the gene expression Omnibus (GEO) and BioConductor. Bioinformatics 2007;23:1846–7.
12. Jackson EL, Willis N, Mercer K, Bronson RT, Crowley D, Montoya R, et al. Analysis of lung tumor initiation and progression using conditional expression of oncogenic K-ras. Genes Dev 2001;15:3243–8.
13. Dai M, Wang P, Boyd AD, Kostov G, Athey B, Jones EG, et al. Evolving gene/transcript definitions significantly alter the interpretation of GeneChip data. Nucleic Acids Res 2005;33:e175.
14. Kim D, Paggi JM, Park C, Bennett C, Salzberg SL. Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. Nat Biotechnol 2019;37:907–15.
15. Liao Y, Smyth GK, Shi W. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. Bioinformatics 2014;30:923–30.
16. Frankish A, Diekhans M, Ferreira AM, Johnson R, Jungreis I, Loveland J, et al. GENCODE reference annotation for the human and mouse genomes. Nucleic Acids Res 2019;47:D766–73.
17. AACR Project GENIE Consortium. AACR project GENIE: powering precision medicine through an international consortium. Cancer Discov 2017;7:818–31.
18. KimYen Ladia. Synapserutils: collection of utilities building on top of synapser. 2019.
19. R Development Core Team. R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computinng; 2020.
20. Judd J, Abdel Karim N, Khan H, Naqash AR, Baca Y, Xiu J, et al. Characterization of KRAS mutation subtypes in non-small cell lung cancer. Mol Cancer Ther 2021;20:2577–84.
21. Kennedy MC, Lowe SW. Mutant p53: it's not all one and the same. Cell Death Differ 2022;29:983–7.
22. Salmón M, Álvarez-Díaz R, Fustero-Torre C, Brehey O, Lechuga CG, Sanclemente M, et al. Kras oncogene ablation prevents resistance in advanced lung adenocarcinomas. J Clin Invest 2023;133:e164413.
23. Mukhopadhyay S, Huang HY, Lin Z, Ranieri M, Li S, Sahu S, et al. Genome-wide CRISPR screens identify multiple synthetic lethal targets that enhance KRASG12C inhibitor efficacy. Cancer Res 2023;83:4095–111.
24. Skoulidis F, Goldberg ME, Greenawalt DM, Hellmann MD, Awad MM, Gainor JF, et al. STK11/LKB1 mutations and PD-1 inhibitor resistance in KRAS-mutant lung adenocarcinoma. Cancer Discov 2018;8:822–35.
25. Srivastava S, Furlan SN, Jaeger-Ruckstuhl CA, Sarvothama M, Berger C, Smythe KS, et al. Immunogenic chemotherapy enhances recruitment of CAR-T cells to lung tumors and improves antitumor efficacy when combined with checkpoint blockade. Cancer Cell 2021;39:193–208.e10.
26. Tsou P, Katayama H, Ostrin EJ, Hanash SM. The emerging role of B cells in tumor immunity. Cancer Res 2016;76:5597–601.
27. Lauss M, Donia M, Svane IM, Jönsson G. B cells and tertiary lymphoid structures: friends or foes in cancer immunotherapy? Clin Cancer Res 2022;28:1751–8.
28. Tan R, Nie M, Long W. The role of B cells in cancer development. Front Oncol 2022;12:958756.
29. Zewdu R, Mehrabad EM, Ingram K, Fang P, Gillis KL, Camolotto SA, et al. An NKX2-1/ERK/WNT feedback loop modulates gastric identity and response to targeted therapy in lung adenocarcinoma. Elife 2021;10:e66788.
30. Sayin VI, Ibrahim MX, Larsson E, Nilsson JA, Lindahl P, Bergo MO. Antioxidants accelerate lung cancer progression in mice. Sci Transl Med 2014;6:221ra15.
31. Hsu WL, Hsieh YT, Chen WM, Chien MH, Luo WJ, Chang JH, et al. High-fat diet induces C-reactive protein secretion, promoting lung adenocarcinoma via immune microenvironment modulation. Dis Model Mech 2023;16:dmm050360.
32. Kichenadasse G, Miners JO, Mangoni AA, Rowland A, Hopkins AM, Sorich MJ. Association between body mass index and overall survival with immune checkpoint inhibitor therapy for advanced non-small cell lung cancer. JAMA Oncol 2020;6:512–18.

33. Chuang CH, Greenside PG, Rogers ZN, Brady JJ, Yang D, Ma RK, et al. Molecular definition of a metastatic lung cancer state reveals a targetable CD109-Janus kinase-Stat axis. Nat Med 2017;23:291–300.
34. Behrens C, Solis LM, Lin H, Yuan P, Tang X, Kadara H, et al. EZH2 protein expression associates with the early pathogenesis, tumor progression, and prognosis of non-small cell lung carcinoma. Clin Cancer Res 2013;19:6556–65.
35. Busch SE, Hanke ML, Kargl J, Metz HE, MacPherson D, Houghton AM. Lung cancer subtypes generate unique immune responses. J Immunol 2016;197: 4493–503.
36. Sutherland KD, Ireland AS, Oliver TG. Killing SCLC: insights into how to target a shapeshifting tumor. Genes Dev 2022;36:241–58.
37. Doyle A, Martin WJ, Funa K, Gazdar A, Carney D, Martin SE, et al. Markedly decreased expression of class I histocompatibility antigens, protein, and mRNA in human small-cell lung cancer. J Exp Med 1985;161:1135–51.
38. Murai F, Koinuma D, Shinozaki-Ushiku A, Fukayama M, Miyaozono K, Ehata S. EZH2 promotes progression of small cell lung cancer by suppressing the TGF-beta-Smad-ASCL1 pathway. Cell Discov 2015;1:15026.
39. Poirier JT, Gardner EE, Connis N, Moreira AL, de Stanchina E, Hann CL, et al. DNA methylation in small cell lung cancer defines distinct disease subtypes and correlates with high expression of EZH2. Oncogene 2015;34: 5869–78.
40. Borromeo MD, Savage TK, Kollipara RK, He M, Augustyn A, Osborne JK, et al. ASCL1 and NEUROD1 reveal heterogeneity in pulmonary neuroendocrine tumors and regulate distinct genetic programs. Cell Rep 2016;16: 1259–72.
41. Olsen RR, Ireland AS, Kastner DW, Groves SM, Spainhower KB, Pozo K, et al. ASCL1 represses a SOX9^{+} neural crest stem-like state in small cell lung cancer. Genes Dev 2021;35:847–69.
42. Ferone G, Song JY, Krijgsman O, van der Vliet J, Cozijnsen M, Semenova EA, et al. FGFR1 oncogenic activation reveals an alternative cell of origin of SCLC in Rb1/p53 mice. Cell Rep 2020;30:3837–50.e3.
43. Reymann S, Borlak J. Transcription profiling of lung adenocarcinomas of c-myc-transgenic mice: identification of the c-myc regulatory gene network. BMC Syst Biol 2008;2:46.
44. Wise DR, DeBerardinis RJ, Mancuso A, Sayed N, Zhang XY, Pfeiffer HK, et al. Myc regulates a transcriptional program that stimulates mitochondrial glutaminolysis and leads to glutamine addiction. Proc Natl Acad Sci U S A 2008; 105:18782–7.