A Lung Cancer Mouse Model Database

Authors

Ling Cai[1,2,3,#], Ying Gao[1], Ralph J. DeBerardinis[2,3,4], George Acquaah-Mensah[5], Vassilis Aidinis[6], Jennifer E. Beane[7], Shyam Biswal[8], Ting Chen[10], Carla P. Concepcion-Crisol[11], Barbara M. Grüner[12], Deshui Jia[13], Robert Jones[14], Jonathan M. Kurie[15], Min Gyu Lee[16], Per Lindahl[17], Yonathan Lissanu[18], Maria Corina Lorz Lopez[19], Rosanna Martinelli[20], Pawel K. Mazur[21], Sarah A. Mazzilli[7], Shinji Mii[23], Herwig Moll[24], Roger Moorehead[14], Edward E. Morrisey[26], Sheng Rong Ng[27], Matthew G. Oser[28], Arun R. Pandiri[29], Charles A. Powell[30], Giorgio Ramadori[31], Mirentxu Santos Lafuente[19], Eric Snyder[33], Rocio Sotillo[34], Kang-Yi Su[35], Tetsuro Taki[23], Kekoa Taparra[37], Yifeng Xia[38], Ed van Veen[39], Monte M. Winslow[40], Guanghua Xiao[1,3,41], Charles M. Rudin[42], Trudy G. Oliver[43], Yang Xie[1,3,41,#], John D. Minna[3,44,#]

Affiliations

[1]Quantitative Biomedical Research Center, Peter O'Donnell Jr. School of Public Health, UT Southwestern Medical Center, Dallas, TX 75390, USA.
[2]Children's Research Institute, UT Southwestern Medical Center, Dallas, TX 75390, USA.
[3]Simmons Comprehensive Cancer Center, UT Southwestern Medical Center, Dallas, TX 75390, USA.
[4]Howard Hughes Medical Institute, University of Texas Southwestern Medical Center, Dallas, TX 75390, USA.
[5]Massachusetts College of Pharmacy & Health Sciences, 19 Foster Street, Worcester, MA 01608
[6]Institute of Fundamental Biological Research, Biomedical Sciences Research Center Alexander Fleming, 34 Fleming Street, 16672 Athens, Greece
[7]Section of Computational Biomedicine, Boston University School of Medicine, 72 E. Concord Street | Boston, MA 02118
[8]Department of Environmental Health and Engineering, Johns Hopkins University School of Public Health, Baltimore, MD 21205
[10]NYU School of Medicine
[11]Columbia University, 116th and Broadway, New York, NY 10027
[12]Department of Medical Oncology, West German Cancer Center, University Hospital Essen, 45147 Essen, Germany
[13]Institute of Translational Medicine, Shanghai General Hospital, Shanghai Jiao Tong University School of Medicine, No. 650 New Songjiang Road, Shanghai 201620, China
[14]Department of Biomedical Sciences, Ontario Veterinary College, University of Guelph, Guelph, ON, Canada N1G2W1
[15]Department of Thoracic-Head & Neck Med Onc, Division of Cancer Medicine, The University of Texas MD Anderson Cancer Center, 1515 Holcombe Blvd., Houston, TX 77030, USA
[16]Department of Molecular and Cellular Oncology, The University of Texas MD Anderson Cancer Center, 1515 Holcombe Blvd., Houston, TX 77030, USA
[17]Sahlgrenska Center for Cancer Research Institute of Biomedicine | Department of Medical Biochemistry and Cell Biology, University of Gothenburg, Medicanaregatan 1F, 413 90 Gothenburg, Sweden
[18]Department of Thoracic & Cardiovascular Surgery, the University of Texas MD Anderson Cancer Center
[19]Biomedical Innovation Unit. Centro de Investigaciones Energéticas, Medioambientales y Tecnológicas (CIEMAT), 28040 Madrid, Spain
[20]Department of Medicine, Surgery and Dentistry 'Scuola Medica Salernitana', University of Salerno, 84081 Baronissi, SA, Italy
[21]Department of Experimental Radiation Oncology, The University of Texas MD Anderson Cancer Center, Houston, TX, 77030, USA
[23]Department of Pathology, Nagoya University Graduate School of Medicine, Nagoya, Japan.
[24]Medical University of Vienna Center for Physiology and Pharmacology Waehringer Strasse 13a 1090 Vienna, Austria
[26]Penn-CHOP Lung Biology Institute, University of Pennsylvania, Philadelphia, PA 19104
[27]Institute of Molecular and Cell Biology, A*STAR, Singapore 138632
[28]Department of Medical Oncology, Dana-Farber Cancer Institute, Harvard Medical School, Boston, Massachusetts

[29]Cellular and Molecular Pathology Branch, Division of National Toxicology Program (DNTP), National Institute of Environmental Health Sciences (NIEHS), 111 T.W. Alexander Drive, Research Triangle Park, NC, 27709,USA
[30]Icahn School of Medicine at Mount Sinai, One Gustave L. Levy Place, Box 1232, New York, N.Y. 10029
[31]Department of Cell Physiology and Metabolism, University of Geneva; Geneva, 1211, Switzerland
[33]Department of Pathology and Huntsman Cancer Institute, University of Utah, SLC, UT 84112
[34]Molecular Thoracic Oncology, German Cancer Research Center (DKFZ), Im Neuenheimer Feld 280, 69120 Heidelberg
[35]Department of Clinical Laboratory Sciences and Medical Biotechnology, College of Medicine, National Taiwan University
[37]Department of Radiation Oncology, Stanford Health Care, Stanford, CA
[38]Salk Institute for Biological Studies. La Jolla, CA 92037 USA
[39]Department of Integrative Biology and Physiology, University of California Los Angeles
[40]Department of Genetics, Stanford University School of Medicine, Stanford, CA 94305, USA.
[41]Department of Bioinformatics, University of Texas Southwestern Medical Center, Dallas, TX 75390, USA.
[42]Memorial Sloan Kettering Cancer Center, New York, NY 10065, USA.
[43]Department of Pharmacology & Cancer Biology, Duke University, Durham, NC 27710, USA.
[44]Hamon Center for Therapeutic Oncology Research, UT Southwestern Medical Center, Dallas, TX 75390, USA.

[#], co-corresponding author
Ling Cai
5323 Harry Hines Blvd, Dallas, TX 75390; +1-214-697-5362; ling.cai@utsouthwestern.edu
Yang Xie
5323 Harry Hines Blvd, Dallas, TX 75390; +1-214-648-5178; yang.xie@utsouthwestern.edu
John Minna
5323 Harry Hines Blvd, Dallas, TX 75390; +1-214-648-4900; john.minna@utsouthwestern.edu

Abstract

Lung cancer, the leading cause of cancer mortality, exhibits diverse histological 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. Here we established the Lung Cancer Mouse Model Database (LCMMDB), an extensive repository of 1,354 samples from 77 transcriptomic datasets covering 974 samples from genetically engineered mouse models (GEMMs), 368 samples from carcinogen-induced models, and 12 samples from a spontaneous model. Meticulous curation and collaboration with data depositors have produced a robust and comprehensive database, enhancing the fidelity of the genetic landscape it depicts. The LCMMDB aligns 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 GEMMs. Accompanying this resource, we developed a web application that offers researchers intuitive tools for in-depth gene expression analysis. With standardized reprocessing of gene expression data, the LCMMDB 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.

Introduction

Lung cancer remains the most common cause of cancer-related mortality globally, with its complexity reflected in diverse histological subtypes—such as adenocarcinoma (ADC), squamous cell carcinoma (SQCC), large cell carcinoma (LCC), and small cell lung cancer (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 employed [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 (PDXs) [10]. While lung cancer autochthonous

animal models, primarily based on mice, represent a separate but significant line of research, they often lack unified characterization due to independent development across various laboratories.

To address this gap, we conducted a comprehensive review of transcriptomic databases, specifically GEO 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 have culminated in the creation of the Lung Cancer Mouse Model Database (LCMMDB). 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.

Methods

Dataset screening

We performed searches with keywords “lung cancer” refined to species “Mus Musculus” and data type restricted to gene expression profiling by array or high-throughput sequencing in the Gene Expression Omnibus (GEO) and ArrayExpress. Each of these studies was manually inspected to identify gene expression data generated from autochthonous models, including genetically engineered mouse models (GEMMs), chemically induced mouse models, and spontaneous models of lung cancer. R package GEOquery [11] was used to download author-processed gene expression data and sample annotation data from GEO.

LCMMDB data organization and curation process

LCMMDB 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 PMIDs, PMC IDs, and the contact information of data depositors. The sample table contains accession IDs, sample names, types, treatments, strains, sex, age, genotype, histological classification, primary/metastasis status, sources of Affymetrix data, Short 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 genes involved, genetic constructs, zygosity, type of genetic modification (e.g., overexpression, knockout), 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 employed 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, while 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 adenocarcinoma (ADC). For example, in the LSL-K-ras^{G12D} model, tumors can progress from adenoma to adenocarcinoma between 6-16 weeks post-infection [12]. However, not all studies explicitly differentiate between adenoma and adenocarcinoma. Additionally, multiple clonal tumors may present within the same sample, where some may classify as adenomas and others as adenocarcinomas. 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 together under “Adenoma/ADC” to maintain consistency and clarity across the dataset.

Gene expression reprocessing

Affymetrix raw data were downloaded from GEO and grouped by platform. For each platform, we downloaded v25 of the gene-level customized Chip Definition Files (CDFs) from the Molecular & Behavioral Neuroscience Institute (MBNI) 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 (RMA) method via the oligo package, yielding an expression set (eset) from which gene expression matrices were extracted. Entrez IDs were converted to gene symbols based on the NCBI Entrez mapping file.

RNA-seq fastq files were downloaded from Sequence Read Archive (SRA) through 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 GRCh38m by hisat2 [14]. Gene expression was quantified using FeatureCounts [15] and GENOCODE [16]. We retained genes with non-zero values in more than 10% of samples, normalized their counts to library sizes, and computed log-transformed counts per million (logCPM) values for downstream analyses.

AACR GENIE data analysis

AACR GENIE data [17] (Version 15.0-public) was downloaded from SAGE BIONETWORKS on 03/25/2024 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 (sv) status from “data_sv.txt” to determine genetic aberrations. Lung cancer patient samples 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).

Web application construction

Statistical software R was used for analyses and web application construction [19]. The web application https://lccl.shinyapps.io/LCMMDB/ is a shiny app deployed at the shinyapps.io servers. It is implemented through the following R packages: 'data.table', 'reshape2', 'digest', 'stringr', 'dplyr', 'tidyverse', 'Hmisc', ‘ggplot2’, 'RColorBrewer', 'ggpubr', 'patchwork', 'shinyjs', 'shinycssloaders', 'bslib', 'htmltools', 'shinyWidgets', 'shinyTree', 'DT', and 'plotly'.

Results

Construction of the Lung Cancer Mouse Model Database (LCMMDB)

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 due to data redundancy from the reprocessed data (**Figure S1**). Our current data collection includes 77 datasets from 71 unique studies, comprised of 1,354 samples (**Table 1** and **Figure 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. 89% 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 **Table S1**). The database was updated accordingly, integrating the depositors' revised data and constructive feedback.

Our analysis revealed a general trend towards small sample sizes across the datasets, with a median of 12 samples, ranging from 3 to 143 (**Figure 1b**). The median number of detected genes per study is 20,942, with older microarray platforms reporting fewer genes (**Figure 1c**). The majority of the samples originated from 71 GEMM datasets, including 856 cancerous samples and 118 lung samples. Complementary to these, seven studies offered 368 samples generated from carcinogen-exposed models, including 239 cancerous samples

and 129 lung samples. One unique dataset provided insight into 6 spontaneous tumors and 6 lung samples in mice aged two years (**Figure 1d**). Age details were available for 407 samples (30%), with a median age of 34 weeks (**Figure 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 (**Figure 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 (**Figure 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 (**Figure 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 histological 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 (**Figure 1i**).

<u>The genetic landscape of GEMMs and comparison with patient mutation spectrum</u>
859 precancerous lesions (hyperplasia, dysplasia, carcinoma *in situ*) and tumor samples in the LCMMDB collection were developed from 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 **Figure 2a**. These include six human genes (*EGFR*, *IGF1R*, *EZH2*, *MYCN*, *CCNE1*, and *SNAI1*) and two viral genes (HPV *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 (**Figure 2b**). Remarkably, all these genotypes harbor the G12D mutation, representing only ~15% of *KRAS* mutations in non-small cell lung cancer (NSCLC) patients [20]. This disparity underscores the broader issue of limited genetic variation in GEMM tumors compared to 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 LCMMDB 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.

The gene-centric distribution of genetic alterations in the tumor samples is detailed in **Figure 2c**. Twenty-eight genes are predominantly activated while 24 are primarily inactivated. Two genes, *Nfe2l2* and *Stat3*, were subject to both activation and inactivation 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. In contrast, manipulations of 11 different single genes were adequate to generate GEMM tumors (**Figure 2d**). Only five models contained alterations in 4 genes (**Figure 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 (**Figure 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*, *ERBB4*, and critical tumor suppressor genes like *NF1*, *ATM*, and *APC* are currently absent from the LCMMDB (**Figure 2e**). **Figure S2** details the frequency and types of genetic alterations for the top 100 genes most frequently altered in lung cancer patients according to the AACR GENIE data, with an emphasis on 78 genes that are not yet included in the LCMMDB. 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 where possible and reprocessed them through standardized pipelines by platform, each with the latest probe and gene annotations. This standardization effort enabled us to make reprocessed data available for 85% of the samples (**Figure 3a**). Notably, approximately half of these samples (n = 563) are derived from RNA-seq and encompass 38 distinct datasets (**Figure 3b**). Principal component analysis (PCA) conducted on the top 1000 variable genes from the reprocessed RNA-seq data revealed that the first two principal components (PCs) capture 62% of the total variance, indicating a strong structuring of the data (**Figure 3c**). Despite potential batch effects, the PCA demonstrates that different datasets exhibit substantial overlap (**Figure 3d**), with clear distinctions observed between SCLC and NSCLC samples along PC1, and between primary and metastatic samples along PC2 (**Figures 3e** and **3f**, respectively). For microarray datasets, we executed a parallel processing strategy on data from the Mouse430_2 platform—the most represented microarray platform with 283 samples across 15 datasets—and noted a comparable success in data integration (**Figure 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 appears 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 LCMMDB

To facilitate the exploration and analysis of the LCMMDB data, we constructed a web application that can be accessed at https://lccl.shinyapps.io/LCMMDB/. This application is structured into two primary sections: a data review panel and an analysis panel. Within the data review panel, the "Overview" tab presents graphical summaries of the LCMMDB, while the "Studies," "Samples," and "GEMMs" tabs allow users to navigate and refine detailed data tables. These tables correspond to **Supplementary Tables 2-5** in this manuscript. Specifically, the "GEMMs" tab displays a table where genetic alterations are recorded with one gene per line. Each genotype within a study is distinctively highlighted to ensure clear 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 in **Figure S4**.

The analysis panel offers users an interactive environment to delve deeper into the gene expression profiles across multiple datasets. 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.

Comparisons in individual datasets

We offer three options for 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 utilizing the available filters within the dropdown menu, tailoring the analysis to their specific research interests (**Figure S5**). As exemplified in **Figure 4**, where 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, where *Cd274* expression is markedly downregulated, corroborating clinical findings that *STK11* mutations are significantly enriched among PD-L1-negative lung tumors [22]. On the other hand, treatment of oxaliplatin and cyclophosphamide (Ox/Cy)[23], known to induce immunogenic cell death increased the expression of *Cd274*.

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 ten 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 (**Figure 5a**). We observed a pronounced increase in *Cd19* expression, indicative of B cell infiltration, in a Braf-driven GEMM under MAPK inhibitor treatment (GSE145152 dataset), which aligns with tumor regression [24]. Conversely, a significant decrease in *Cd19* expression was noted in samples with tumor progression, such as in Kras-driven GEMMs treated with antioxidants (GSE52594 dataset) [25]. *Cd19* also increased in Egfr-driven GEMMs subjected to a high-fat diet (GSE119649) and a Kras-driven GEMM under a high-caloric diet (GSE56260), in line with previous findings that obesity creates a more inflammatory tumor microenvironment in mouse [26], as well as observation in patients that high body mass index (BMI) is independently associated with overall survival benefit from immune checkpoint inhibitor therapy in advanced NSCLC [27].

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 (**Figure 5b**). With the fine curation of metastatic status in samples from GSE84447 [28], 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 [29].

Comparisons in merged reprocessed data

Analysis of reprocessed data merged by 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 **Figure S5**). We provide two visualization approaches for analyses. The first approach is to generate a dotplot 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. 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 (**Figure 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 histological subtype [30-32]. The highest expression of *Cd19* is found in dysplasia samples derived following treatment with the alkylating agent N-nitroso-tris-chloroethylurea (NCTU), potentially due to abundant neoantigen resulting from carcinogen treatment (**Figure 6**, top). Users may select from additional profiling microarray platforms. For example, analysis of *Ezh2* expression in reprocessed data of Mouse430_2 reveals its expression is much higher in SCLC compared to NSCLC samples (**Figure S6**), as previously established [33, 34].

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. In **Figure 7a**, we demonstrate this with RNA-seq samples colored to reflect *Ascl1* expression—a neuroendocrine lineage transcription factor instrumental in SCLC pathogenesis [35, 36]. Corroborating the histological segregation observed in **Figure 3e**, we found that samples characterized by lower PC1 scores—typical of SCLC—exhibited elevated *Ascl1* expression. Notably, some adenocarcinoma (ADC) samples, despite having higher PC1 scores, also showed high *Ascl1* levels. Our interactive plots equipped with informative tooltips reveal sample details, indicating that these outliers are ADCs from a model with constitutively-active *Fgfr1*^{K656E} in an *Rb1/Trp53*-deficient background [37], typically used to study classic SCLC (**Figure 3b**). While FGFR1 activation in this model has reduced *Ascl1* expression compared to classic SCLC tumors [37], the *Ascl1* levels are still higher than ADC tumors from other models (**Figure 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 (**Figure 7b**). This particular discrepancy is clarified upon recognizing that these SCLC samples have undergone *Ascl1* knockout, leading to a complete loss of neuroendocrine cell fate [36] Consequently, ASCL1 knockout in a SCLC model makes the transcriptomic landscape of the SCLC sample more akin to that of NSCLC samples, explaining its outlier position in the PCA

plot. 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 (**Figure S8**).

Discussion

Our LCMMDB 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 need to consider the limitations inherent to the database's scope. The LCMMDB 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. The dynamic nature of scientific research also necessitates the LCMMDB 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. The candidate datasets to be included in the next update are listed in **Supplementary Table S6**. 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. While not expounded upon in this manuscript, 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. While standardization efforts have been rigorous, batch effects from diverse experimental and genetic backgrounds may still be present. Future updates will aim to support meta-analytical capabilities and provide insights from comprehensive cross-transcriptomic evaluations. Central to the LCMMDB's utility 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 LCMMDB 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.

Acknowledgments

This study is supported by funding from UTSW ACS-IRG (IRG-21-142-16), P50CA70907, U24CA213274, R01GM140012, R01GM141519, R01DE030656, R01CA244841 (TGO), U01AI169298, U01CA249245, U01AI156189, R35CA22044901, U01CA213338, R35GM136375, and R35CA263816. The authors declare no competing interests.

L.C. designed the study, performed the analyses, and wrote the manuscript. Y.G. reviewed additional datasets for inclusion in a future update. T.G.O. and J.D.M. provided critical inputs for the manuscript. T.G.O., R.J.D., J.D.M., and Y.G. helped edit the manuscript. L.C., R.J.D., G.X., T.G.O., C.M.R., and Y.X. obtained funding to support this work. We thank the lung cancer mouse model community for their efforts in helping to curate this resource. Among the authors, G.A., V.A., J.E.B., S.B., J.B., T.C., C.P.C., B.M.G., D.J., R.J., J.M.K., M.L. (Lee), P.L., Y.L., M.L. (Lopez), R.M. (Martinelli), P.K.M., S.A.M., S.M., H.M., R.M. (Moorehead), E.E.M., S.N., T.G.O., M.G.O., A.R.P., C.A.P., G.R., M.S., E.S., R.S., K.S., T.T., K.T., Y.X., E.V., and M.W. provided dataset

confirmation, corrected or improved data curation, and feedback; M.G.O., M.W., T.G.O., and J.D.M. suggested additional datasets to include. In addition to the listed authors, we want to acknowledge additional data curation confirmation/correction help from Drs. Andrea Ventura, Anneleen Daemen, Anton Berns, Bob Stearman, Casanova Emilio, Celeste Simon, Cong Yan, Emmy Verschuren, Francesco Demayo, Julien Sage, Karsta Luettich, Kwok-Kin Wong, Meylan Etienne, Shivani Srivastava, and Thomas Russell; and dataset suggestions from Drs. Karsta Luettich, Julien Sage, and Thomas Russell.

Tables

Accession	Model	Title in data repository	N	Ref
E-MEXP-1137	GEMM	Transcription profiling of lung alveolar type II epithelial cells from CCSP-rtTA/(teto)7Stat3C bitransgenic mice treated with doxycycline and with spontaneous lung bronchoalveolar adenocarcinoma	6	[38]
E-MTAB-6706	GEMM	Expression profiling of K-Ras-driven lung carcinogenesis upon the genetic absence of Enpp2 from pulmonary cells	6	[39]
GSE10954	GEMM	Transcription Profiling of Lung Adenocarcinomas of c-Myc-Transgenic Mice	8	[40]
GSE109629	GEMM	RNA and miRNA sequencing of an IGF1R transgenic mouse model of lung cancer	8	[41]
GSE111313	GEMM	The Role of Lkb1 in Mouse Lung Squamous Cell Carcinoma (SCC) Development	10	[42]
GSE113717	GEMM	De novo lipogenesis represents a therapeutic target in Kras mutant NSCLC	8	[43]
GSE116658	GEMM	Global gene expression profiling of K-RasLSL-G12D and K-RasLSL-G12D;Mll4fl/fl lung tumors using RNA sequencing	4	[44]
GSE116977	GEMM	Inter-tumoral heterogeneity in SCLC is influenced by the cell-type of origin	45	[45]
GSE117519	GEMM	Genome-wide analysis of EMT induced gene expression in accelerating mouse KrasG12D lung tumorigenesis	9	[46]
GSE117547	GEMM	Crebbp loss drives small cell lung cancer and increases sensitivity to HDAC inhibition[SCLC]	14	[47]
GSE118246	GEMM	The Lineage-Defining Transcription Factors SOX2 and NKX2-1 Determine Lung Cancer Cell Fate and Shape the Tumor Immune Microenvironment	34	[48]
GSE119649	GEMM	Gene expression profiling from high-fat diet (HFD)-treated and regular diet (RD)-treated lung cancer	6	[26]
GSE121574	GEMM	Wide versus cell type restricted deletion of four tumor suppressor genes determines the type of high grade neuroendocrine lung cancer	17	[49]
GSE123126	GEMM	Mutationally-activated PI3-kinase-a promotes de-differentiation of lung tumors initiated by the BRAFV600E oncoprotein kinase	23	[50]
GSE128620	GEMM	ERBB2 Regulates MED24 during Cancer Progression in Mice with Pten and Smad4 Deletion in the Pulmonary Epithelium	9	[51]
GSE132759	GEMM	The contribution of FGFR1 to the development of SCLC is dictated by the cell-of-origin	24	[37]
GSE133714	GEMM	Transcriptional profiling of BL/6 mice with alterations in Kras, Stk11, and/or Keap1/Nrf2.	23	[52]
GSE133895	GEMM	An Lkb1-Sik axis suppresses tumor growth and controls differentiation [RNA-Seq]	12	[53]
GSE137396	GEMM	Characterization of transcriptional analysis of LKB1 mutant lung nodules	10	[54]
GSE138753	GEMM	CC10+ KP vs CC10+ KPG1 tumours	10	[55]
GSE138754	GEMM	KP lung tumours	4	[55]

GSE138755	GEMM	PPARaDN KP vs PPARaDN KPG1 tumours	12	[55]
GSE138756	GEMM	SPC+ KP vs SPC+ KPG1 tumours	10	[55]
GSE138953	GEMM	Transcriptional analyses of RPMax-/- versus RPMaxwt mouse lung tumors	16	[56]
GSE139347	GEMM	RNA-seq to measure mRNA expression in normal lung tissues, KRasG12D; p53-null(KP)-driven and Eml4-Alk(EA)-driven lung adenocarcinomas	11	[57]
GSE139444	GEMM	CRISPR-mediated modeling and functional validation of candidate tumor suppressor genes in small cell lung cancer	16	[58]
GSE13963	GEMM	Molecular characterization of lung dysplasia induced by c-raf-1	15	[59]
GSE140154	GEMM	Expression profiling of CD45- lung tumor cells	12	[60]
GSE14277	GEMM	Cancer genomics identifies regulatory gene networks associated with the transition from dysplasia to adenocarcinomas	5	[61]
GSE14449	GEMM	Gene expression profiles of spontaneous metastasis in a K-ras/p53 mutant mouse model	16	[62]
GSE145152	GEMM	An NKX2-1/ERK/WNT feedback loop modulates gastric identity and response to targeted therapy in lung adenocarcinoma.	17	[24]
GSE148194	GEMM	Downregulation of anti-inflammatory A20 promotes immune escape of lung adenocarcinomas	4	[63]
GSE149175	GEMM	MYC drives temporal evolution of small cell lung cancer subtypes by reprogramming neuroendocrine fate [Bulk RNA-seq of RPM and RPR2 tumors]	15	[36]
GSE149272	GEMM	Elevated NSD3 Histone Methylation Activity Drives Squamous Cell Lung Cancer	6	[64]
GSE155453	GEMM	CD109 regulates in vivo tumor invasion in lung adenocarcinoma through TGF-β signaling	4	[65]
GSE155691	GEMM	ASCL1 represses a latent osteogenic program in small cell lung cancer in multiple cells of origin [RNA-Seq]	7	[36]
GSE158110	GEMM	Immunogenic chemotherapy enhances recruitment of CAR-T cells to solid tumors and improves anti-tumor efficacy when combined with checkpoint blockade [RNA-Seq]	9	[23]
GSE159169	GEMM	Unbiased genomic analysis of murine lung cancer cell lines	4	[66]
GSE161514	GEMM	RNA-sequencing of bulk small cell lung cancer lung tumors generating by either injecting LSL-Cas9 mice with a sgControl RPP CMV Cre adenovirus or a sgKdm5a RPP CMV Cre adenovirus	16	[67]
GSE161609	GEMM	RNA sequencing comparing mouse Kras G12D lung tumors that have WT or inactivated Mga via CRISPR	13	[68]
GSE162680	GEMM	MYCN drives chemoresistance in small cell lung cancer while USP7 inhibition can restore chemosensitivity [tumor]	11	[69]
GSE171217	GEMM	NSD2 amplifies oncogenic transcriptional output to promote lung adenocarcinoma pathogenesis [RNA-seq]	6	[70]
GSE180817	GEMM	Smarca4 inactivation promotes lineage-specific transformation and early metastatic features in the lung [Bulk RNA-seq]	9	[71]
GSE18534	GEMM	Mouse small cell lung cancer model	15	[72]
GSE19753	GEMM	Mad2-induced chromosome instability leads to lung tumor relapse after oncogene withdrawal	29	[73]
GSE21581	GEMM	Expression data from lung adenocarcinoma mouse tumors	35	[74]
GSE22575	GEMM	Expression data from Hif 2alpha Knockdown study	14	[75]
GSE23962	GEMM	Transgenic SCLC induced by E6/E7 oncoproteins vs normal controls	4	[76]

GSE26850	GEMM	Promotion of Lung Tumorigenesis By Beta-catenin	6	[77]
GSE27675	GEMM	Expression data from lung tumor and stromal cells of KrasTgfbr2 -/- mouse model	7	[78]
GSE27717	GEMM	Expression data from lung tumors of KrasTgfbr2 -/- mouse model	11	[78]
GSE30049	GEMM	Expression data comparing the effect of IKK2 KO on a KRAS mouse model	8	[79]
GSE36473	GEMM	Nkx2.1 represses a latent gastric differentiation program in lung adenocarcinoma	12	[80]
GSE38948	GEMM	A whole genome map of alternative splicing events in precancerous lung dysplasia and adenocarcinomas of c-Raf mice	17	
GSE45744	GEMM	Whole-genome expression data from normal FVB mouse lung tissue, transgenic cyclin E overexpressing (CEO) normal mouse lung tissue, and transgenic CEO lung adenocarcinomas	12	[81]
GSE46048	GEMM + chemical treatment	High expression genes in urethane-induced lung tumor	4	[82]
GSE52594	GEMM	Antioxidants Accelerate Lung Cancer Progression in Mice	30	[25]
GSE52798	GEMM	Disruption of STAT3 signaling promotes KRAS induced lung tumorigenesis	3	[83]
GSE54829	GEMM	Oncogenomics of c-Myc-induced papillary lung adenocarcinoma	14	[84]
GSE56260	GEMM	Gene expression in Kras-driven lung tumors in mice fed with standard or high calorie diet	11	[85]
GSE57133	GEMM	ErbB2 Pathway Activation upon Smad4 Loss Promotes Lung Tumor Growth and Metastasis [expression]	15	[86]
GSE6135	GEMM	LKB1 modulates lung cancer differentiation and metastasis	25	[87, 88]
GSE69552	GEMM	Cell-of-origin links lung tumor histotype spectrum to immune microenvironment diversity	17	[89]
GSE70046	GEMM	Oncogenic deregulation of EZH2 as an opportunity for targeted therapy in lung cancer [RNA-Seq]	7	[90]
GSE70271	GEMM	caArray_jacks-00113: Murine KRASLA lung cancer gene expression	58	[91]
GSE78948_1	GEMM	Sox2 is the determining oncogenic switch in promoting lung squamous cell carcinoma from different cells-of-origin	14	[92]
GSE78948_2	GEMM	Sox2 is the determining oncogenic switch in promoting lung squamous cell carcinoma from different cells-of-origin	6	[92]
GSE84447	GEMM	Unbiased genomic analysis of multiple stages of lung cancer development	54	[28]
GSE89660	GEMM	Comparison of gene expression patterns of two SCLC genetically-engineered mouse models; Rb1 floxed, Trp53 floxed, LSL-Myc T58A-IRES-Luc vs. Rb1 floxed, Trp53 floxed, Rbl2 (p130) floxed	14	[93]
GSE94927	GEMM	RNAseq of genetically engineered mouse model (GEMM) tumors	12	[94]
GSE99338	GEMM + chemical treatment	RNAseq analysis of wildtype and Nrf2-/- advanced lung tumors	4	[95]
E-MTAB-1871	chemical treatment	Mechanisms of CS-related A/J lung tumorigenesis	45	[96]
GSE111091	chemical treatment	Establishing a preclinical model of squamous lung cancer to investigate of novel chemopreventive approaches in high risk individuals with bronchial premalignant lesion	40	[97]
GSE24061	chemical	MAQC-II Project: Hamner data set	88	[98,

	treatment			99]
GSE2514	chemical treatment	Lung tumors	44	[100]
GSE31013	spontaneous	Global Differential Gene Expression Analysis of Spontaneous Lung Tumors in B6C3F1 Mice: Comparison to Human Non-Small Cell Lung Cancer	12	[101]
GSE71232	chemical treatment	Expression quantitative trait loci (eQTL) analysis of mouse lung tumors	143	[102]

Table 1. Current collection of LCMMDB datasets

References

1. Al Bakir, M., et al., *The evolution of non-small cell lung cancer metastases in TRACERx.* Nature, 2023. **616**(7957): p. 534-542.
2. Campbell, J.D., et al., *Distinct patterns of somatic genome alterations in lung adenocarcinomas and squamous cell carcinomas.* Nat Genet, 2016. **48**(6): p. 607-16.
3. Cancer Genome Atlas Research, N., *Comprehensive molecular profiling of lung adenocarcinoma.* Nature, 2014. **511**(7511): p. 543-50.
4. Frankell, A.M., et al., *The evolution of lung cancer and impact of subclonal selection in TRACERx.* Nature, 2023. **616**(7957): p. 525-533.
5. George, J., et al., *Comprehensive genomic profiles of small cell lung cancer.* Nature, 2015. **524**(7563): p. 47-53.
6. George, J., et al., *Evolutionary trajectories of small cell lung cancer under therapy.* Nature, 2024.
7. Martinez-Ruiz, C., et al., *Genomic-transcriptomic evolution in lung cancer and metastasis.* Nature, 2023. **616**(7957): p. 543-552.
8. Ghandi, M., et al., *Next-generation characterization of the Cancer Cell Line Encyclopedia.* Nature, 2019. **569**(7757): p. 503-508.
9. Gazdar, A.F., et al., *Lung cancer cell lines as tools for biomedical discovery and research.* J Natl Cancer Inst, 2010. **102**(17): p. 1310-21.
10. Sun, H., et al., *Comprehensive characterization of 536 patient-derived xenograft models prioritizes candidatesfor targeted treatment.* Nat Commun, 2021. **12**(1): p. 5086.
11. Davis, S. and P.S. Meltzer, *GEOquery: a bridge between the Gene Expression Omnibus (GEO) and BioConductor.* Bioinformatics, 2007. **23**(14): p. 1846-7.
12. Jackson, E.L., et al., *Analysis of lung tumor initiation and progression using conditional expression of oncogenic K-ras.* Genes Dev, 2001. **15**(24): p. 3243-8.
13. Dai, M., et al., *Evolving gene/transcript definitions significantly alter the interpretation of GeneChip data.* Nucleic Acids Res, 2005. **33**(20): p. e175.
14. Kim, D., et al., *Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype.* Nat Biotechnol, 2019. **37**(8): p. 907-915.
15. Liao, Y., G.K. Smyth, and W. Shi, *featureCounts: an efficient general purpose program for assigning sequence reads to genomic features.* Bioinformatics, 2014. **30**(7): p. 923-30.
16. Frankish, A., et al., *GENCODE reference annotation for the human and mouse genomes.* Nucleic Acids Res, 2019. **47**(D1): p. D766-D773.
17. Consortium, A.P.G., *AACR Project GENIE: Powering Precision Medicine through an International Consortium.* Cancer Discov, 2017. **7**(8): p. 818-831.
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*, in *Vienna, Austria*. 2020, R Foundation for Statistical Computinng.

20. Judd, J., et al., *Characterization of KRAS Mutation Subtypes in Non-small Cell Lung Cancer.* Mol Cancer Ther, 2021. **20**(12): p. 2577-2584.
21. Kennedy, M.C. and S.W. Lowe, *Mutant p53: it's not all one and the same.* Cell Death Differ, 2022. **29**(5): p. 983-987.
22. Skoulidis, F., et al., *STK11/LKB1 Mutations and PD-1 Inhibitor Resistance in KRAS-Mutant Lung Adenocarcinoma.* Cancer Discov, 2018. **8**(7): p. 822-835.
23. Srivastava, S., 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**(2): p. 193-208 e10.
24. Zewdu, R., et al., *An NKX2-1/ERK/WNT feedback loop modulates gastric identity and response to targeted therapy in lung adenocarcinoma.* Elife, 2021. **10**.
25. Sayin, V.I., et al., *Antioxidants accelerate lung cancer progression in mice.* Sci Transl Med, 2014. **6**(221): p. 221ra15.
26. Hsu, W.L., et al., *High-fat diet induces C-reactive protein secretion, promoting lung adenocarcinoma via immune microenvironment modulation.* Dis Model Mech, 2023. **16**(11).
27. Kichenadasse, G., et al., *Association Between Body Mass Index and Overall Survival With Immune Checkpoint Inhibitor Therapy for Advanced Non-Small Cell Lung Cancer.* JAMA Oncol, 2020. **6**(4): p. 512-518.
28. Chuang, C.H., et al., *Molecular definition of a metastatic lung cancer state reveals a targetable CD109-Janus kinase-Stat axis.* Nat Med, 2017. **23**(3): p. 291-300.
29. Behrens, C., 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**(23): p. 6556-65.
30. Busch, S.E., et al., *Lung Cancer Subtypes Generate Unique Immune Responses.* J Immunol, 2016. **197**(11): p. 4493-4503.
31. Sutherland, K.D., A.S. Ireland, and T.G. Oliver, *Killing SCLC: insights into how to target a shapeshifting tumor.* Genes Dev, 2022. **36**(5-6): p. 241-258.
32. Doyle, A., et al., *Markedly decreased expression of class I histocompatibility antigens, protein, and mRNA in human small-cell lung cancer.* J Exp Med, 1985. **161**(5): p. 1135-51.
33. Murai, F., et al., *EZH2 promotes progression of small cell lung cancer by suppressing the TGF-beta-Smad-ASCL1 pathway.* Cell Discov, 2015. **1**: p. 15026.
34. Poirier, J.T., et al., *DNA methylation in small cell lung cancer defines distinct disease subtypes and correlates with high expression of EZH2.* Oncogene, 2015. **34**(48): p. 5869-78.
35. Borromeo, M.D., et al., *ASCL1 and NEUROD1 Reveal Heterogeneity in Pulmonary Neuroendocrine Tumors and Regulate Distinct Genetic Programs.* Cell Rep, 2016. **16**(5): p. 1259-1272.
36. Olsen, R.R., et al., *ASCL1 represses a SOX9(+) neural crest stem-like state in small cell lung cancer.* Genes Dev, 2021. **35**(11-12): p. 847-869.
37. Ferone, G., et al., *FGFR1 Oncogenic Activation Reveals an Alternative Cell of Origin of SCLC in Rb1/p53 Mice.* Cell Rep, 2020. **30**(11): p. 3837-3850 e3.
38. Li, Y., et al., *Activation of the signal transducers and activators of the transcription 3 pathway in alveolar epithelial cells induces inflammation and adenocarcinomas in mouse lung.* Cancer Res, 2007. **67**(18): p. 8494-503.
39. Magkrioti, C., et al., *The Autotaxin-Lysophosphatidic Acid Axis Promotes Lung Carcinogenesis.* Cancer Res, 2018. **78**(13): p. 3634-3644.
40. Reymann, S. and J. Borlak, *Transcription profiling of lung adenocarcinomas of c-myc-transgenic mice: identification of the c-myc regulatory gene network.* BMC Syst Biol, 2008. **2**: p. 46.
41. Jones, R.A., S.E. Franks, and R.A. Moorehead, *Comparative mRNA and miRNA transcriptome analysis of a mouse model of IGFIR-driven lung cancer.* PLoS One, 2018. **13**(11): p. e0206948.

42. Liu, J., et al., *JNK(1/2) represses Lkb(1)-deficiency-induced lung squamous cell carcinoma progression.* Nat Commun, 2019. **10**(1): p. 2148.
43. Singh, A., et al., *De novo lipogenesis represents a therapeutic target in mutant Kras non-small cell lung cancer.* FASEB J, 2018. **32**(12): p. fj201800204.
44. Alam, H., et al., *KMT2D Deficiency Impairs Super-Enhancers to Confer a Glycolytic Vulnerability in Lung Cancer.* Cancer Cell, 2020. **37**(4): p. 599-617 e7.
45. Yang, D., et al., *Axon-like protrusions promote small cell lung cancer migration and metastasis.* Elife, 2019. **8**.
46. Taparra, K., et al., *O-GlcNAcylation is required for mutant KRAS-induced lung tumorigenesis.* J Clin Invest, 2018. **128**(11): p. 4924-4937.
47. Jia, D., et al., *Crebbp Loss Drives Small Cell Lung Cancer and Increases Sensitivity to HDAC Inhibition.* Cancer Discov, 2018. **8**(11): p. 1422-1437.
48. Mollaoglu, G., et al., *The Lineage-Defining Transcription Factors SOX2 and NKX2-1 Determine Lung Cancer Cell Fate and Shape the Tumor Immune Microenvironment.* Immunity, 2018. **49**(4): p. 764-779 e9.
49. Lazaro, S., et al., *Differential development of large-cell neuroendocrine or small-cell lung carcinoma upon inactivation of 4 tumor suppressor genes.* Proc Natl Acad Sci U S A, 2019. **116**(44): p. 22300-22306.
50. van Veen, J.E., et al., *Mutationally-activated PI3'-kinase-alpha promotes de-differentiation of lung tumors initiated by the BRAF(V600E) oncoprotein kinase.* Elife, 2019. **8**.
51. Liu, J., et al., *ERBB2 Regulates MED24 during Cancer Progression in Mice with Pten and Smad4 Deletion in the Pulmonary Epithelium.* Cells, 2019. **8**(6).
52. Singh, A., et al., *NRF2 Activation Promotes Aggressive Lung Cancer and Associates with Poor Clinical Outcomes.* Clin Cancer Res, 2021. **27**(3): p. 877-888.
53. Murray, C.W., et al., *An LKB1-SIK Axis Suppresses Lung Tumor Growth and Controls Differentiation.* Cancer Discov, 2019. **9**(11): p. 1590-1605.
54. Deng, J., et al., *ULK1 inhibition overcomes compromised antigen presentation and restores antitumor immunity in LKB1 mutant lung cancer.* Nat Cancer, 2021. **2**(5): p. 503-514.
55. Contat, C., et al., *Combined deletion of Glut1 and Glut3 impairs lung adenocarcinoma growth.* Elife, 2020. **9**.
56. Augert, A., et al., *MAX Functions as a Tumor Suppressor and Rewires Metabolism in Small Cell Lung Cancer.* Cancer Cell, 2020. **38**(1): p. 97-114 e7.
57. Li, X., et al., *High-Resolution In Vivo Identification of miRNA Targets by Halo-Enhanced Ago2 Pull-Down.* Mol Cell, 2020. **79**(1): p. 167-179 e11.
58. Ng, S.R., et al., *CRISPR-mediated modeling and functional validation of candidate tumor suppressor genes in small cell lung cancer.* Proc Natl Acad Sci U S A, 2020. **117**(1): p. 513-521.
59. Rohrbeck, A., V.S. Muller, and J. Borlak, *Molecular characterization of lung dysplasia induced by c-Raf-1.* PLoS One, 2009. **4**(5): p. e5637.
60. Ancey, P.B., et al., *GLUT1 Expression in Tumor-Associated Neutrophils Promotes Lung Cancer Growth and Resistance to Radiotherapy.* Cancer Res, 2021. **81**(9): p. 2345-2357.
61. Rohrbeck, A. and J. Borlak, *Cancer genomics identifies regulatory gene networks associated with the transition from dysplasia to advanced lung adenocarcinomas induced by c-Raf-1.* PLoS One, 2009. **4**(10): p. e7315.
62. Gibbons, D.L., et al., *Expression signatures of metastatic capacity in a genetic mouse model of lung adenocarcinoma.* PLoS One, 2009. **4**(4): p. e5401.
63. Breitenecker, K., et al., *Down-regulation of A20 promotes immune escape of lung adenocarcinomas.* Sci Transl Med, 2021. **13**(601).
64. Yuan, G., et al., *Elevated NSD3 histone methylation activity drives squamous cell lung cancer.* Nature, 2021. **590**(7846): p. 504-508.

65. Taki, T., et al., *CD109 regulates in vivo tumor invasion in lung adenocarcinoma through TGF-beta signaling.* Cancer Sci, 2020. **111**(12): p. 4616-4628.
66. Chuang, C.H., et al., *Altered Mitochondria Functionality Defines a Metastatic Cell State in Lung Cancer and Creates an Exploitable Vulnerability.* Cancer Res, 2021. **81**(3): p. 567-579.
67. Oser, M.G., et al., *The KDM5A/RBP2 histone demethylase represses NOTCH signaling to sustain neuroendocrine differentiation and promote small cell lung cancer tumorigenesis.* Genes Dev, 2019. **33**(23-24): p. 1718-1738.
68. Mathsyaraja, H., et al., *Loss of MGA repression mediated by an atypical polycomb complex promotes tumor progression and invasiveness.* Elife, 2021. **10**.
69. Grunblatt, E., et al., *MYCN drives chemoresistance in small cell lung cancer while USP7 inhibition can restore chemosensitivity.* Genes Dev, 2020. **34**(17-18): p. 1210-1226.
70. Sengupta, D., et al., *NSD2 dimethylation at H3K36 promotes lung adenocarcinoma pathogenesis.* Mol Cell, 2021. **81**(21): p. 4481-4492 e9.
71. Concepcion, C.P., et al., *Smarca4 Inactivation Promotes Lineage-Specific Transformation and Early Metastatic Features in the Lung.* Cancer Discov, 2022. **12**(2): p. 562-585.
72. Schaffer, B.E., et al., *Loss of p130 accelerates tumor development in a mouse model for human small-cell lung carcinoma.* Cancer Res, 2010. **70**(10): p. 3877-83.
73. Sotillo, R., et al., *Mad2-induced chromosome instability leads to lung tumour relapse after oncogene withdrawal.* Nature, 2010. **464**(7287): p. 436-40.
74. Carretero, J., et al., *Integrative genomic and proteomic analyses identify targets for Lkb1-deficient metastatic lung tumors.* Cancer Cell, 2010. **17**(6): p. 547-59.
75. Mazumdar, J., et al., *HIF-2alpha deletion promotes Kras-driven lung tumor development.* Proc Natl Acad Sci U S A, 2010. **107**(32): p. 14182-7.
76. Buonomo, T., et al., *Involvement of aryl hydrocarbon receptor signaling in the development of small cell lung cancer induced by HPV E6/E7 oncoproteins.* J Transl Med, 2011. **9**: p. 2.
77. Pacheco-Pinedo, E.C., et al., *Wnt/beta-catenin signaling accelerates mouse lung tumorigenesis by imposing an embryonic distal progenitor phenotype on lung epithelium.* J Clin Invest, 2011. **121**(5): p. 1935-45.
78. Borczuk, A.C., et al., *Progression of human bronchioloalveolar carcinoma to invasive adenocarcinoma is modeled in a transgenic mouse model of K-ras-induced lung cancer by loss of the TGF-beta type II receptor.* Cancer Res, 2011. **71**(21): p. 6665-75.
79. Xia, Y., et al., *Reduced cell proliferation by IKK2 depletion in a mouse lung-cancer model.* Nat Cell Biol, 2012. **14**(3): p. 257-65.
80. Snyder, E.L., et al., *Nkx2-1 represses a latent gastric differentiation program in lung adenocarcinoma.* Mol Cell, 2013. **50**(2): p. 185-99.
81. Busch, A.M., et al., *Evidence for tankyrases as antineoplastic targets in lung cancer.* BMC Cancer, 2013. **13**: p. 211.
82. Satoh, H., et al., *Nrf2 prevents initiation but accelerates progression through the Kras signaling pathway during lung carcinogenesis.* Cancer Res, 2013. **73**(13): p. 4158-68.
83. Grabner, B., et al., *Disruption of STAT3 signalling promotes KRAS-induced lung tumorigenesis.* Nat Commun, 2015. **6**: p. 6285.
84. Ciribilli, Y., et al., *Decoding c-Myc networks of cell cycle and apoptosis regulated genes in a transgenic mouse model of papillary lung adenocarcinomas.* Oncotarget, 2015. **6**(31): p. 31569-92.
85. Ramadori, G., et al., *Diet-induced unresolved ER stress hinders KRAS-driven lung tumorigenesis.* Cell Metab, 2015. **21**(1): p. 117-25.
86. Liu, J., et al., *ErbB2 Pathway Activation upon Smad4 Loss Promotes Lung Tumor Growth and Metastasis.* Cell Rep, 2015. **10**(9): p. 1599-1613.

87. Ji, H., et al., *LKB1 modulates lung cancer differentiation and metastasis.* Nature, 2007. **448**(7155): p. 807-10.
88. Han, X., et al., *Transdifferentiation of lung adenocarcinoma in mice with Lkb1 deficiency to squamous cell carcinoma.* Nat Commun, 2014. **5**: p. 3261.
89. Nagaraj, A.S., et al., *Cell of Origin Links Histotype Spectrum to Immune Microenvironment Diversity in Non-small-Cell Lung Cancer Driven by Mutant Kras and Loss of Lkb1.* Cell Rep, 2017. **18**(3): p. 673-684.
90. Zhang, H., et al., *Oncogenic Deregulation of EZH2 as an Opportunity for Targeted Therapy in Lung Cancer.* Cancer Discov, 2016. **6**(9): p. 1006-21.
91. Sweet-Cordero, A., et al., *Comparison of gene expression and DNA copy number changes in a murine model of lung cancer.* Genes Chromosomes Cancer, 2006. **45**(4): p. 338-48.
92. Ferone, G., et al., *SOX2 Is the Determining Oncogenic Switch in Promoting Lung Squamous Cell Carcinoma from Different Cells of Origin.* Cancer Cell, 2016. **30**(4): p. 519-532.
93. Mollaoglu, G., et al., *MYC Drives Progression of Small Cell Lung Cancer to a Variant Neuroendocrine Subtype with Vulnerability to Aurora Kinase Inhibition.* Cancer Cell, 2017. **31**(2): p. 270-285.
94. Lissanu Deribe, Y., et al., *Mutations in the SWI/SNF complex induce a targetable dependence on oxidative phosphorylation in lung cancer.* Nat Med, 2018. **24**(7): p. 1047-1057.
95. Zhang, D., et al., *Identification of an Unfavorable Immune Signature in Advanced Lung Tumors from Nrf2-Deficient Mice.* Antioxid Redox Signal, 2018. **29**(16): p. 1535-1552.
96. Luettich, K., et al., *Systems toxicology approaches enable mechanistic comparison of spontaneous and cigarette smoke-related lung tumor development in the A/J mouse model.* Interdiscip Toxicol, 2014. **7**(2): p. 73-84.
97. Beane, J.E., et al., *Molecular subtyping reveals immune alterations associated with progression of bronchial premalignant lesions.* Nat Commun, 2019. **10**(1): p. 1856.
98. Shi, L., et al., *The MicroArray Quality Control (MAQC)-II study of common practices for the development and validation of microarray-based predictive models.* Nat Biotechnol, 2010. **28**(8): p. 827-38.
99. Thomas, R.S., et al., *Use of short-term transcriptional profiles to assess the long-term cancer-related safety of environmental and industrial chemicals.* Toxicol Sci, 2009. **112**(2): p. 311-21.
100. Stearman, R.S., et al., *Analysis of orthologous gene expression between human pulmonary adenocarcinoma and a carcinogen-induced murine model.* Am J Pathol, 2005. **167**(6): p. 1763-75.
101. Pandiri, A.R., et al., *Differential transcriptomic analysis of spontaneous lung tumors in B6C3F1 mice: comparison to human non-small cell lung cancer.* Toxicol Pathol, 2012. **40**(8): p. 1141-59.
102. Cotroneo, C.E., et al., *Expression quantitative trait analysis reveals fine germline transcript regulation in mouse lung tumors.* Cancer Lett, 2016. **375**(2): p. 221-230.

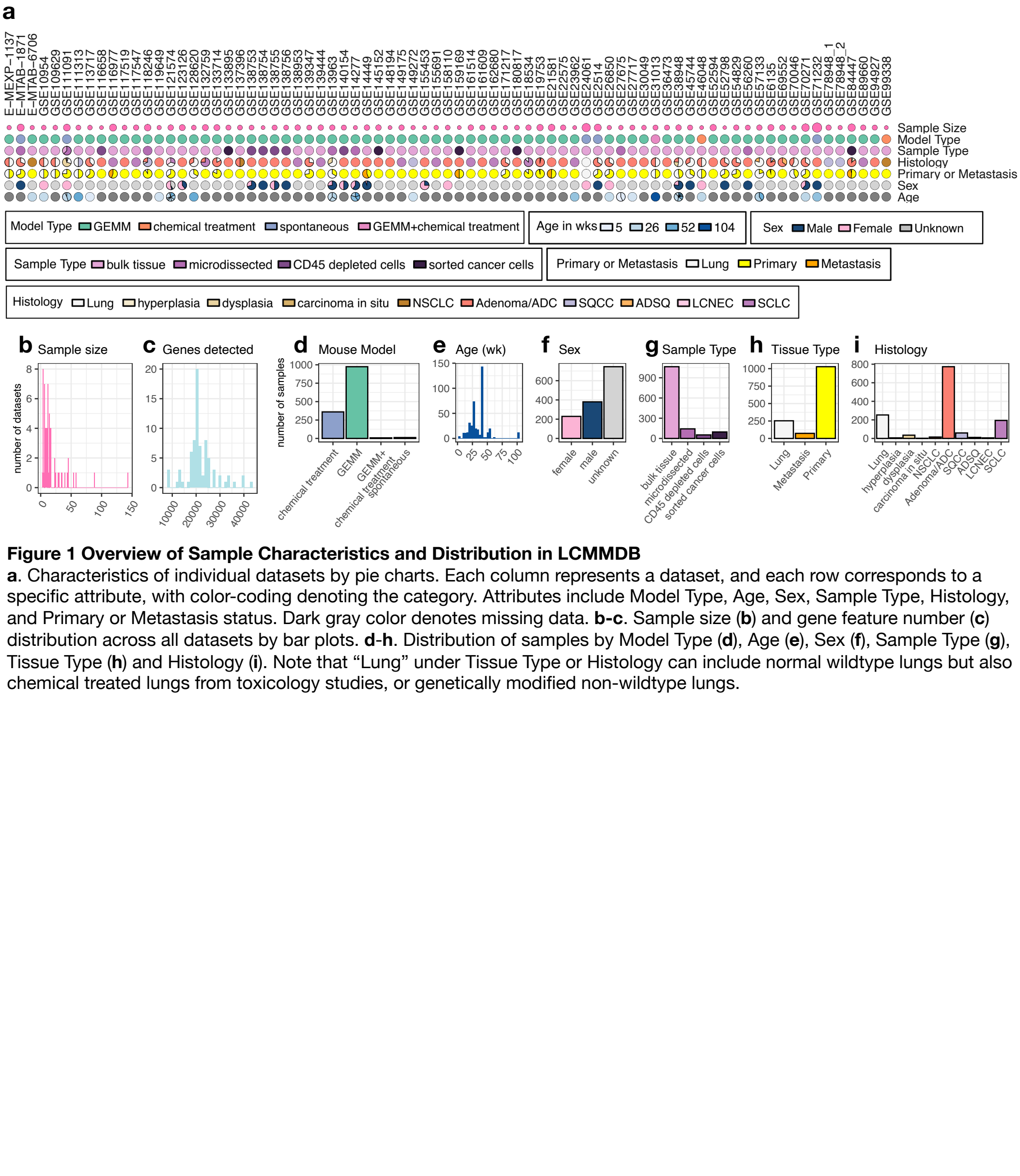

Figure 1 Overview of Sample Characteristics and Distribution in LCMMDB

a. Characteristics of individual datasets by 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 color denotes missing data. **b-c**. Sample size (**b**) and gene feature number (**c**) distribution across all datasets by bar plots. **d-h**. 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 wildtype lungs but also chemical treated lungs from toxicology studies, or genetically modified non-wildtype lungs.

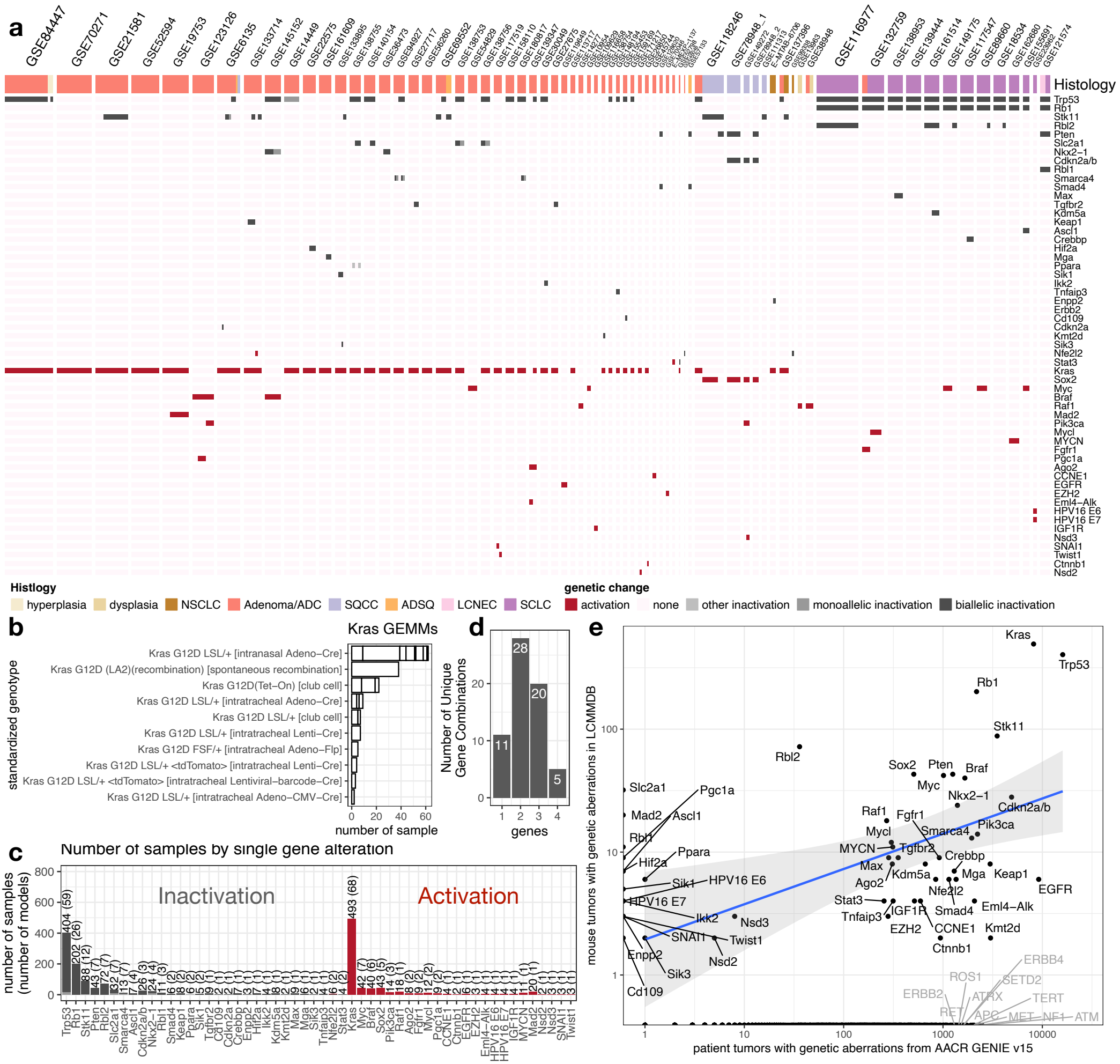

Figure 2 Summary of GEMM genotypes in LCMMDB

a. Landscape of genetic modifications in LCMMDB GEMM tumors by dataset and histology. **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. Y-axis labels indicate the total number of tumors with the specific gene altered and the number of unique standardized genotypes with the specified gene altered 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 manipulation method or mutation. **e**. GEMM tumor alterations in LCMMDB vs. human lung cancer genetic aberrations in the AACR GENIE v15 database by gene. Selected human oncogene and tumor suppressor genes not represented in LCMMDB are highlighted in gray.

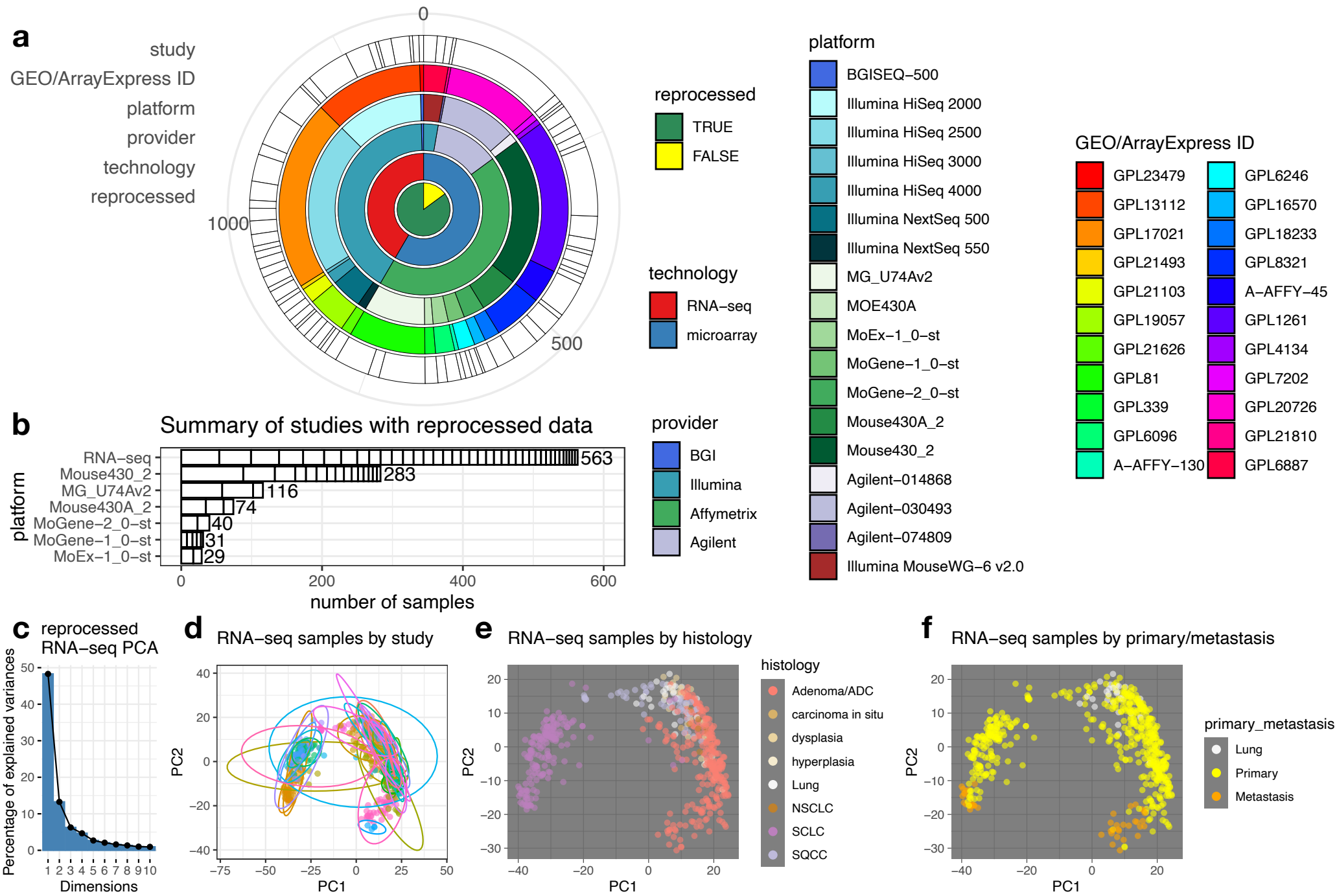

Figure 3. Data reprocessing by platform

a. Hierarchical relationship of technology and platforms. 85% (1152 samples) of the LCMMDB gene expression data was reprocessed. **b**. Platforms with multiple studies reprocesed through standardized workflow. Each box within the bars represents a single dataset. **c**. In principal component analysis using 1000 most variable genes from reprocessed RNA-seq data, the top two principal components accounts 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**).

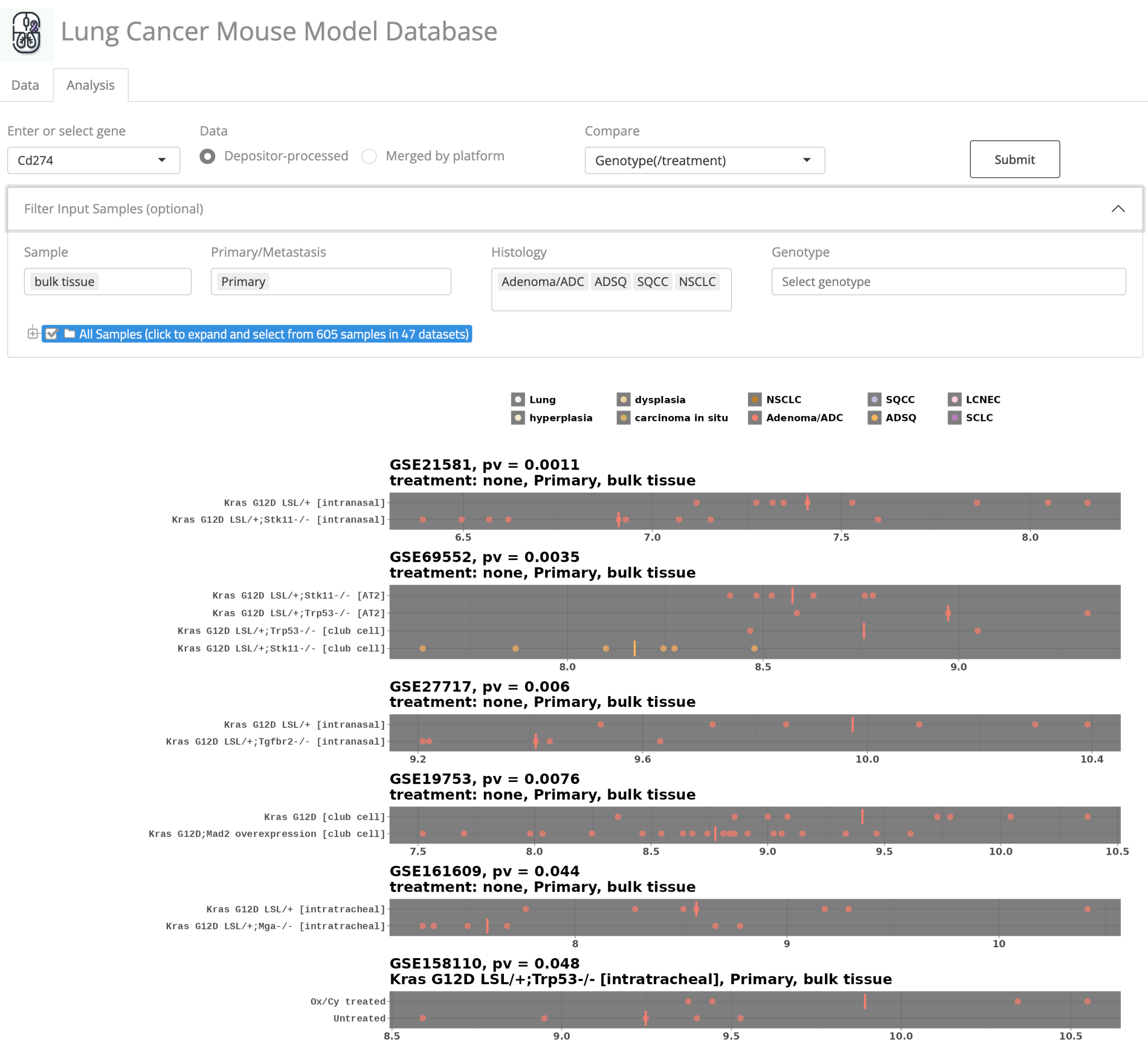

Figure 4: Interactive visualization of gene expression across multiple datasets
This figure features the web application's capability for users to interrogate the expression of a selected gene, *Cd274* (PD-L1), across a range of datasets. The "Depositor-processed" option leverages the original data processed 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 6 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.

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 4 datasets out of 10 for *Cd19* (a) and the top 2 out of 5 for *Ezh2* (b) included in the snapshots. Bars in each plot denote group median.

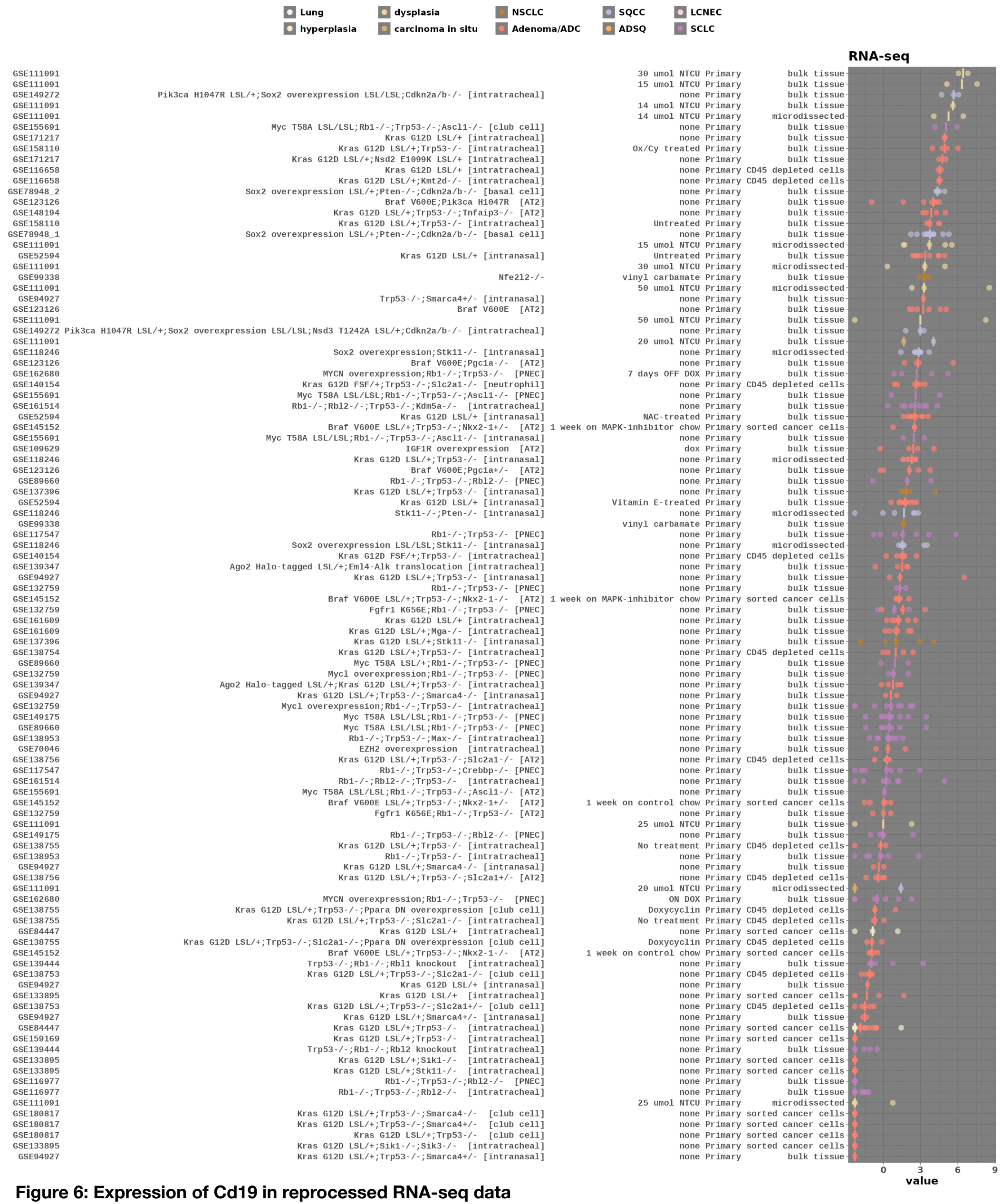

Figure 6: Expression of Cd19 in reprocessed RNA-seq data

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. Median of the group is shown as a bar for each row.

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 a Rb1/p53 deficient model featuring Fgfr1 activation. **b**. PCA plot colored by histology. Details of a SCLC sample located near the NSCLC samples are read. This outlier sample has *Ascl1* knocked out, which explains the loss of neuroendocrine gene expression that renders the transcriptomic profile more similar to NSCLC.