A Lung Cancer Mouse Model Database

Ling Cai[1,2,3,#], Ying Gao[1], Ralph DeBerardinis[2,3,4], Ting Chen[5], Monte Winslow[6], Guanghua Xiao[1,3,7], Charles Rudin[8], Trudy Oliver[9], John Minna[10], Yang Xie[1,3,7,#]

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]Smilow Research Center, NYU School of Medicine, New York, NY 10016, USA.
[6]Department of Genetics, Stanford University School of Medicine, Stanford, CA 94305, USA.
[7]Department of Bioinformatics, University of Texas Southwestern Medical Center, Dallas, TX 75390, USA.
[8]Memorial Sloan Kettering Cancer Center, New York, NY 10065, USA.
[9]Duke University, Durham, NC 27708, USA.
[10]Hamon Cancer Center, 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

Abstract

Lung cancer, the leading cause of cancer mortality, exhibits diverse histological subtypes and genetic complexities. Numerous preclinical 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. 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 genetically engineered mouse models (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 and fostering potential collaborations. 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 and in some cases dictate therapeutic vulnerabilities. To decipher the complexities of tumor biology, high-throughput molecular profiling of patient-derived tumors has been extensively employed. Preclinical models of lung cancer are essential tools for researchers to understand cancer biology and develop therapeutic strategies through experimentation. While there is a concerted effort to aggregate data from cell lines[1, 2] and PDXs[3], lung cancer animal models, often developed and investigated independently by various laboratories, lack a unified characterization.

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 genotype 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 harmonized data supporting a comprehensive understanding of lung cancer biology. 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 genetically engineered mouse models (GEMMs) or chemically induced mouse models. R package GEOquery[4] was used to download author-processed gene expression data and sample annotation data from GEO.

LCMMDB data organization and curation process

In LCMMDB, we organize 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 analysis, primary/metastasis status, sources of Affymetrix data, 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 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 pull out 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 and 16 weeks post-infection [5]. 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 [6], 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 [7]. Gene

expression was quantified using FeatureCounts [8] and GENOCODE [9]. 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[10] (Version 11.0-public) was downloaded from SAGE BIONETWORKS on 04/05/2022 through R package “synapserutils”[11] with Synapse ID “syn26706564”. 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 rearrangement status from “data_fusions.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 [12]. 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 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 syngeneic models to ensure specificity to our research focus. We removed two datasets as we discovered data redundancy from the reprocessed data (Figure S1). Our current data collection includes 77 datasets from 71 unique studies, comprised of 1,354 samples (Figure 1a).

To enhance the accuracy and integrity of our dataset, following a thorough initial harmonization process, we engaged the data depositors by sharing the curated data specific to their studies along with our data schema, soliciting their verification, rectifications, or any insights they could offer. Over 75% 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 reflectively, 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. Complementary to these, seven studies offered 368 samples generated from carcinogen-exposed models, and a unique dataset provided insight into 12 spontaneous tumors in mice aged two years (Figure 1d). Age details were available for 407 samples (30%), with a median age of 34 weeks (Figure 1e). In terms of sex distribution, 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 influences (Figure 1g). Within the 1,101 curated tumor samples, 73 were identified as metastatic, with 53 arising from ADC primary tumors and 20 from SCLC (Figure 1h).

Harmonizing histological data presented a challenge due to the overlapping nature of adenoma and ADC classifications. Despite curating 197 adenomas, 337 ADCs, and 236 cases classified as both based on authors' reports and literature reviews, 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 a mere 5.6% SQCC, indicating an underrepresentation of SQCC when contrasted with the prevalence in human lung cancers (Figure 1i).

The genetic landscape of GEMMs and comparison with patient mutation spectrum

859 precancerous lesions 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*) that were introduced to the GEMMs. We compiled standardized genotypes and simplified genotypes to harmonize genotype curation and identified a total of 122 unique standardized genotypes.

Taking lesions/tumors arising from *Kras* manipulation alone, for example, we identified 10 distinct standardized genotypes, which vary in genetic constructs and induction methods, as shown in Figure 2b. Remarkably, all these genotypes harbor the G12D mutation, which represents only 15% of KRAS mutations in non-small cell lung cancer (NSCLC) patients [13]. This disparity underscores the broader issue of limited genetic variation in GEMM tumors compared to human cancers, exemplified by p53 mutations. Beyond simply inactivating p53, mutations in this gene are known to confer additional gain-of-function properties [14]. 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. 28 genes are predominantly activated while 24 are primarily inactivated. Two genes, *Nfe2l2* and *Stat3* were subjected 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 out 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, while manipulations of 11 different single genes were also able to generate GEMM tumors (Figure 2d).

Figure 2e provides a comparative analysis of the frequency of genetic alterations in mouse lung cancer GEMM-derived tumors from the LCMMDB with those identified in human lung cancers, as recorded in the AACR GENIE v7 database, which is based on clinical sequencing data from real-world patient populations. It is evident that mutations in p53 and KRAS, which are among the most prevalent in human lung cancer patients, 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 issue of 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, specifically 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), while clear

distinctions are 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 effectively consolidate 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 2-4 in our paper. 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 provided 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* (Lkb1) knockout, where Cd274 expression is markedly downregulated, corroborating with clinical findings that STK11 mutations are significantly enriched among PD-L1-negative lung tumors [15]. On the other hand, treatment of oxaliplatin and cyclophosphamide (Ox/Cy)[16], 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 [17]. 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) [18]. 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 [19]. 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[20], we observe *Ezh2* expression increase with tumor invasiveness, which corroborates clinical findings that this chromatin modifier is associated with cancer progression and metastasis [21].

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 (like 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 long plot. An example is given in Figure 6. 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). Bulk tissue samples with the lowest *Cd19* expression are from SCLC, consistent with the immune cold nature of this histological subtype. The highest expression of Cd19 is found in dysplasia samples derived from treatments 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. Figure S6 showcases an example of *Ezh2* expression in reprocessed data of Mouse430_2, revealing its much higher in SCLC compared to NSCLC samples.

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. 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 revealing sample details, indicate that these outliers represent ADCs from a model with constitutive-active Fgfr1 in an Rb1/p53-deficient background, typically used to study classic SCLC (Figure 3b). While Fgfr1 activation has downregulated *Ascl1* expression in this model [22], the *Ascl1* levels are still higher than ADC tumors from other models, implying that Fgfr1-induced signaling can dominate and suppress neuroendocrine program driven by Ascl1, even when Ascl1 is still present. Figure 7b showcases another example of a PCA plot with histology color mapping. Within this visual, a few notable outliers among the NSCLC samples were identified as SCLC samples. This particular discrepancy is clarified upon discovering that these SCLC samples have undergone *Ascl1* knockout, leading to a significant reduction in neuroendocrine gene expression. Consequently, this alteration in gene expression profiles 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 S7).

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 elucidates 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 also need to consider the limitations inherent to the database's scope. The LCMMDB, founded on transcriptomic data, may omit mouse models lacking such characterization, potentially skewing the genetic landscape it portrays. This limitation underscores the need for an inclusive approach that considers unpublished or less-publicized models to achieve a truly representative overview. 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. 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 also 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 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.

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

Reference

1. Ghandi, M., et al., *Next-generation characterization of the Cancer Cell Line Encyclopedia.* Nature, 2019. **569**(7757): p. 503-508.
2. 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.
3. Sun, H., et al., *Comprehensive characterization of 536 patient-derived xenograft models prioritizes candidatesfor targeted treatment.* Nat Commun, 2021. **12**(1): p. 5086.
4. Davis, S. and P.S. Meltzer, *GEOquery: a bridge between the Gene Expression Omnibus (GEO) and BioConductor.* Bioinformatics, 2007. **23**(14): p. 1846-7.
5. 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.
6. Dai, M., et al., *Evolving gene/transcript definitions significantly alter the interpretation of GeneChip data.* Nucleic Acids Res, 2005. **33**(20): p. e175.
7. Kim, D., et al., *Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype.* Nat Biotechnol, 2019. **37**(8): p. 907-915.
8. 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.
9. Frankish, A., et al., *GENCODE reference annotation for the human and mouse genomes.* Nucleic Acids Res, 2019. **47**(D1): p. D766-D773.
10. Consortium, A.P.G., *AACR Project GENIE: Powering Precision Medicine through an International Consortium.* Cancer Discov, 2017. **7**(8): p. 818-831.
11. Kimyen Ladia, *synapserutils: Collection of utilities building on top of synapser*. 2019.
12. R Development Core Team, *R: A language and environment for statistical computing*, in *Vienna, Austria*. 2020, R Foundation for Statistical Computinng.
13. Judd, J., et al., *Characterization of KRAS Mutation Subtypes in Non-small Cell Lung Cancer.* Mol Cancer Ther, 2021. **20**(12): p. 2577-2584.
14. 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.
15. 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.
16. 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.

17. 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**.
18. Sayin, V.I., et al., *Antioxidants accelerate lung cancer progression in mice.* Sci Transl Med, 2014. **6**(221): p. 221ra15.
19. 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).
20. 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.
21. 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.
22. 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.

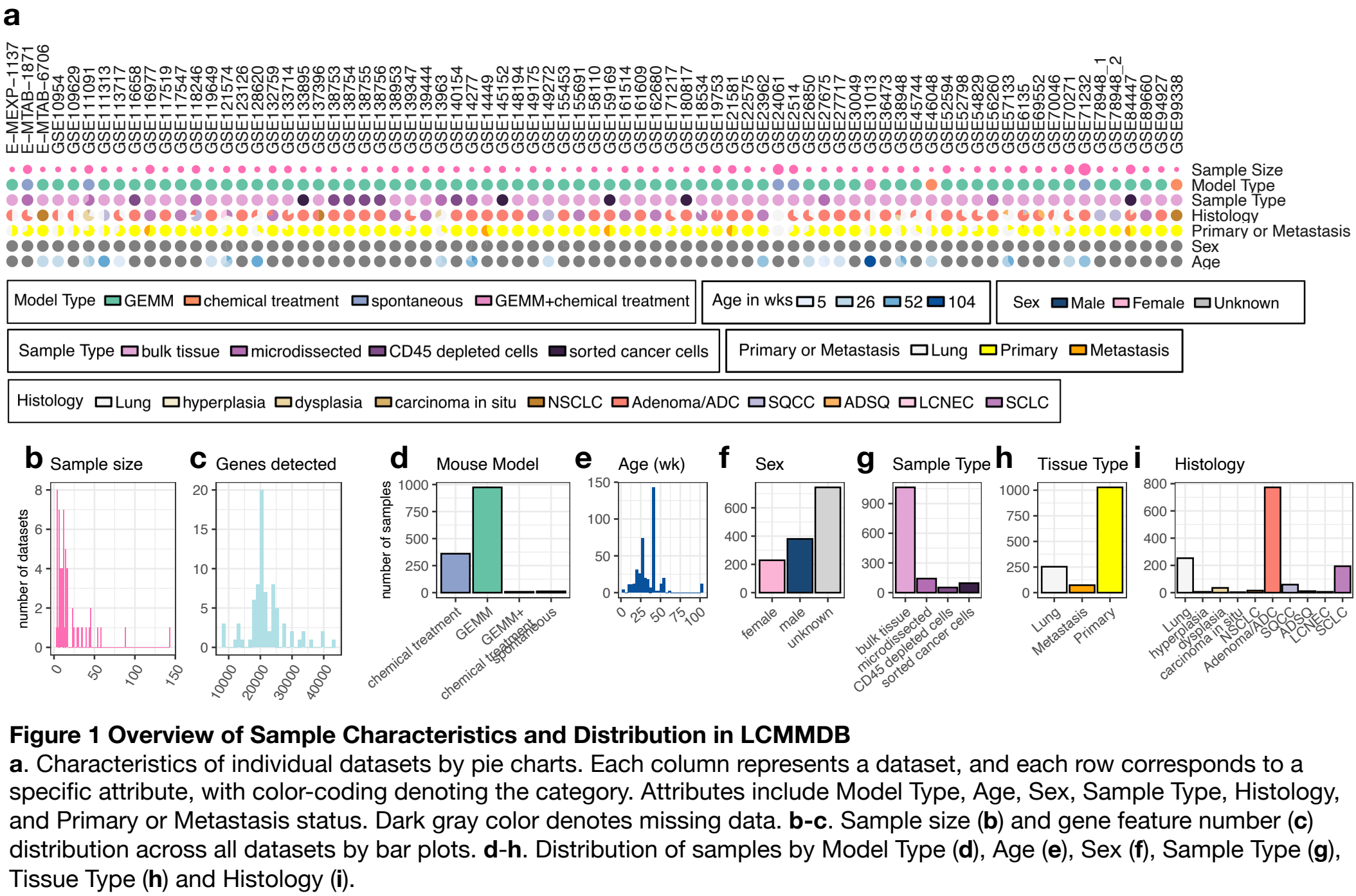

Figure 1 Overview of Sample Characteristics and Distribution in LCMMDB

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**).

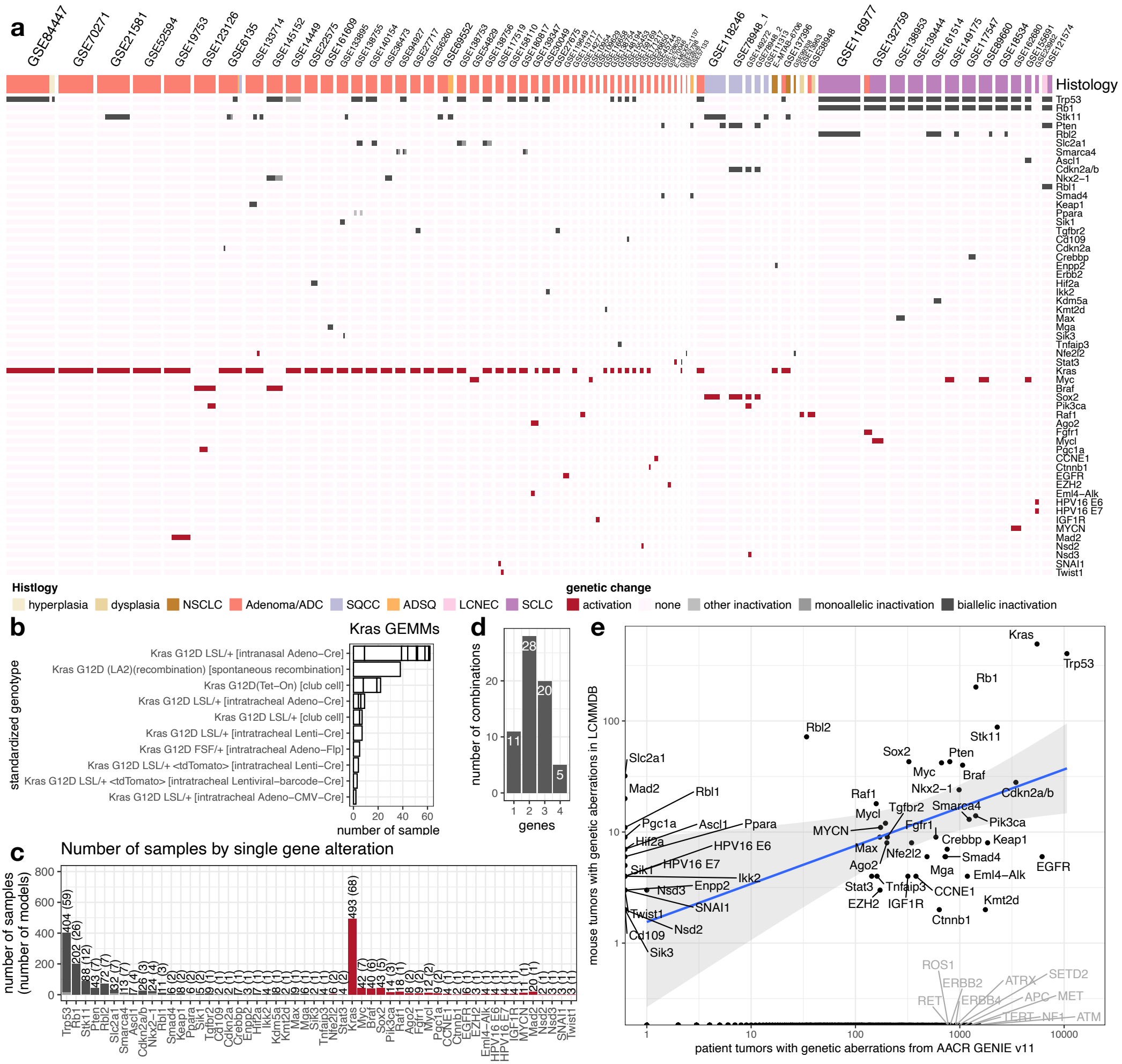

Figure 2 Summary of GEMM genotypes in LCMMDB

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 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 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 v7 database by gene. Selected human oncogene and tumor suppressor genes not represented in LCMMDB are highlighted in gray.

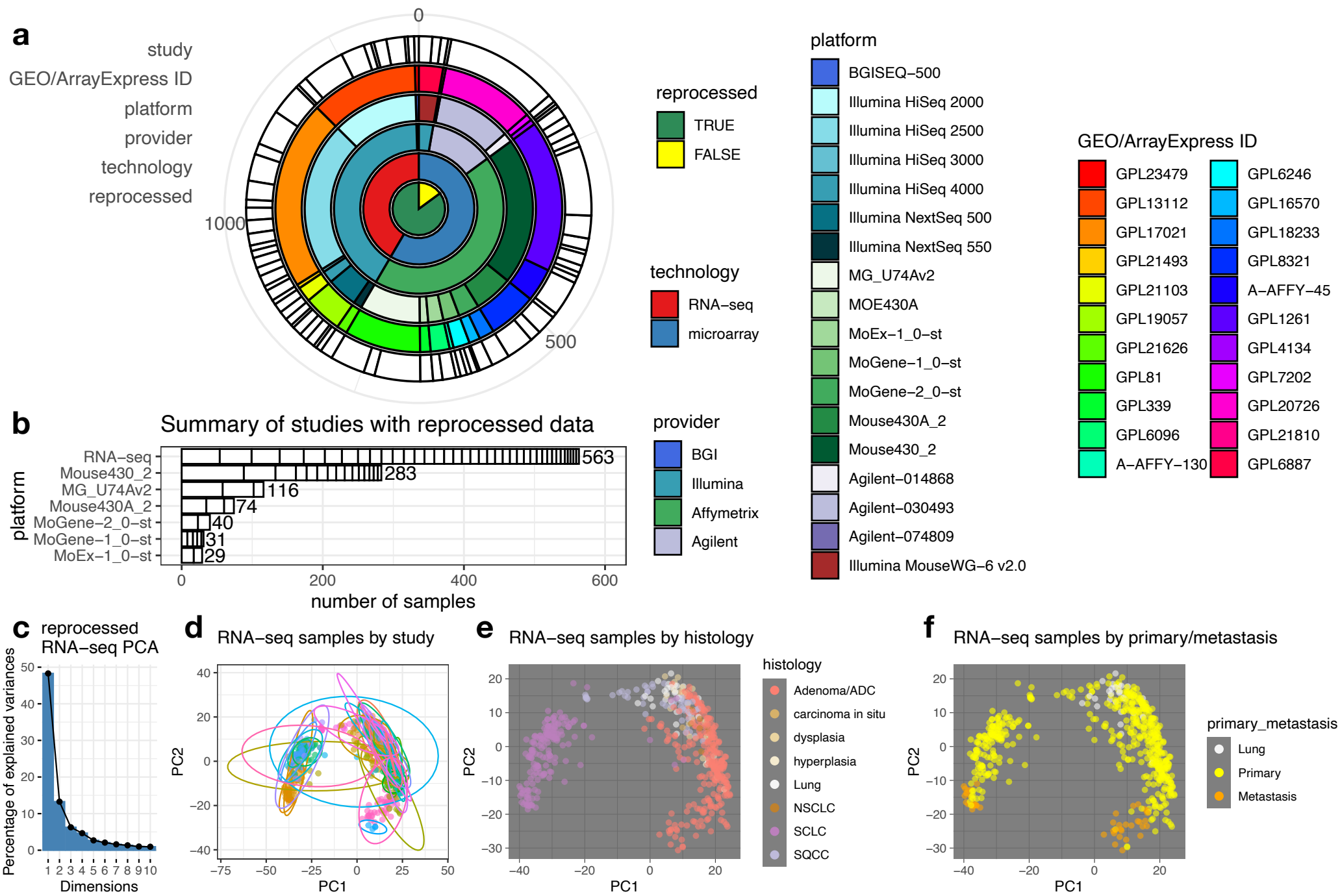

Figure 3. Data reprocessing by platform

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**).

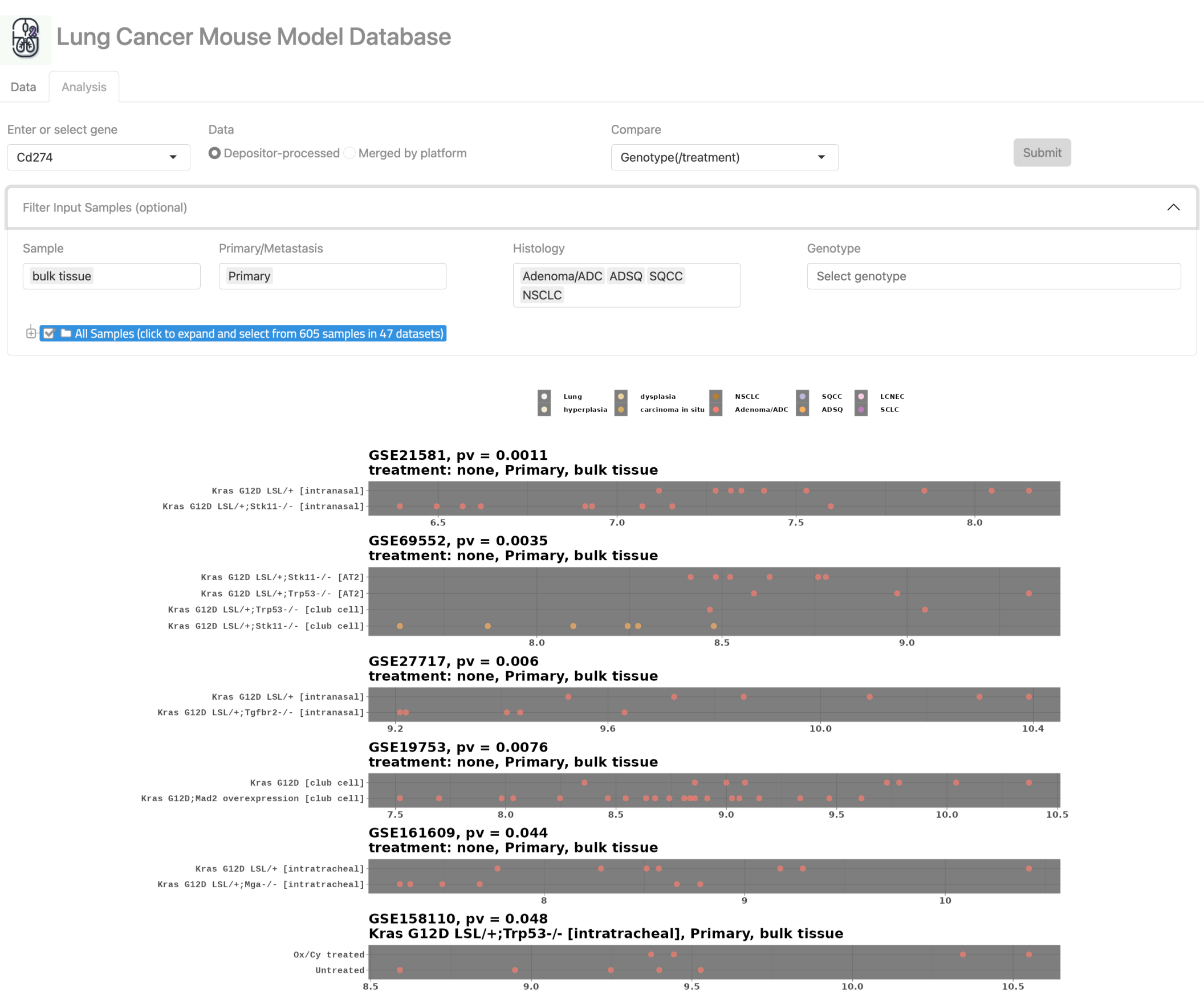

Figure 4: Interactive visualization of gene expression across multiple datasets
This figure features the web application's capability for users to interrogate the expression of a selected gene, *Cd274* (PD-L1), across a range of datasets. Utilizing the "Depositor-processed" option leverages the original data as 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 out of a total of 36 available, giving users a snapshot of the gene expression landscape within the application's extensive repository.

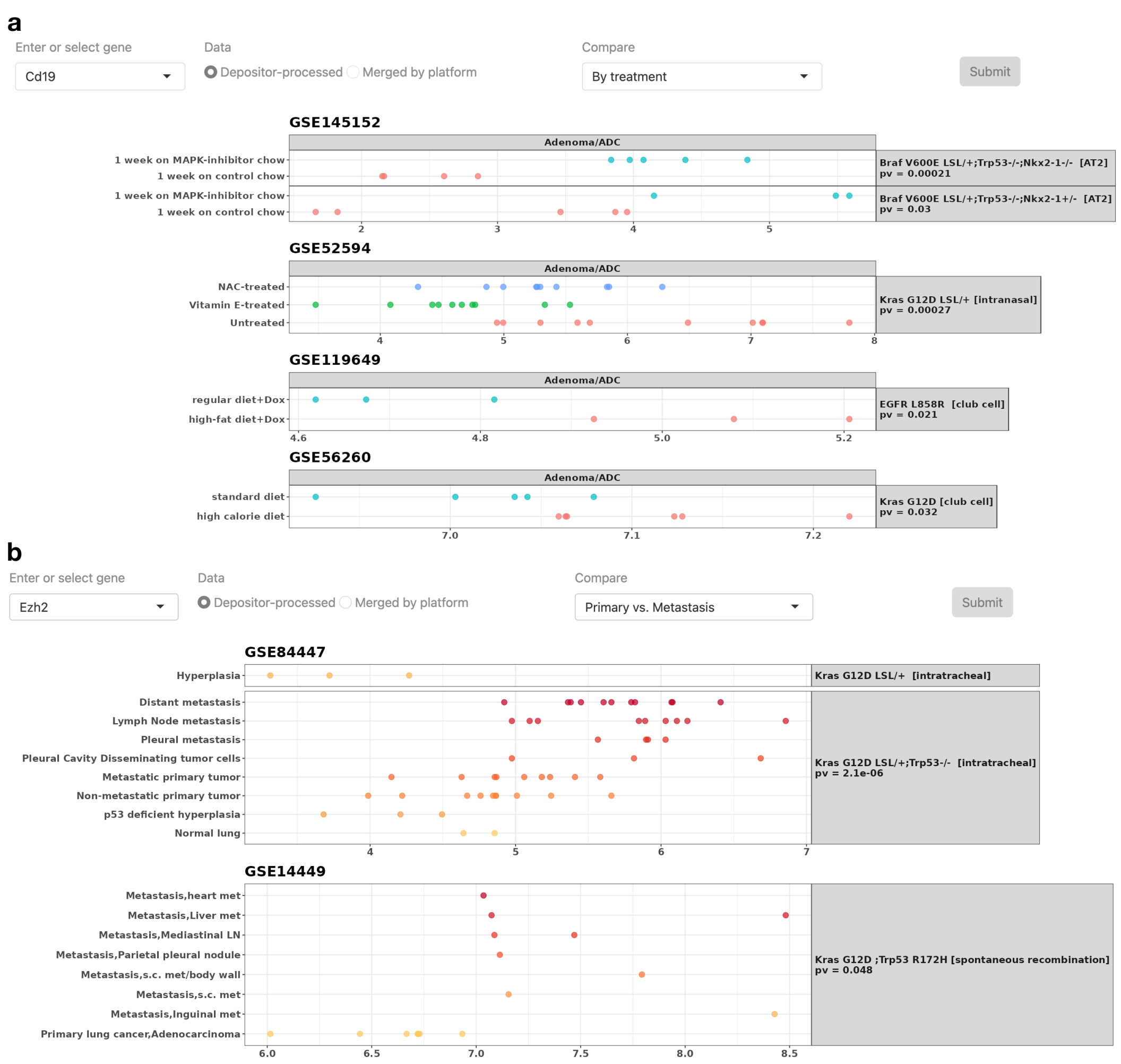

Figure 5: Gene expression comparison by treatment and primary/metastasis status
a. Expression of *Cd19* revealing B cell infiltration in various treatment contexts. Data points are categorized by treatment conditions under each genotype. **b**. Expression of Ezh2 in primary and metastatic tumor samples. 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.

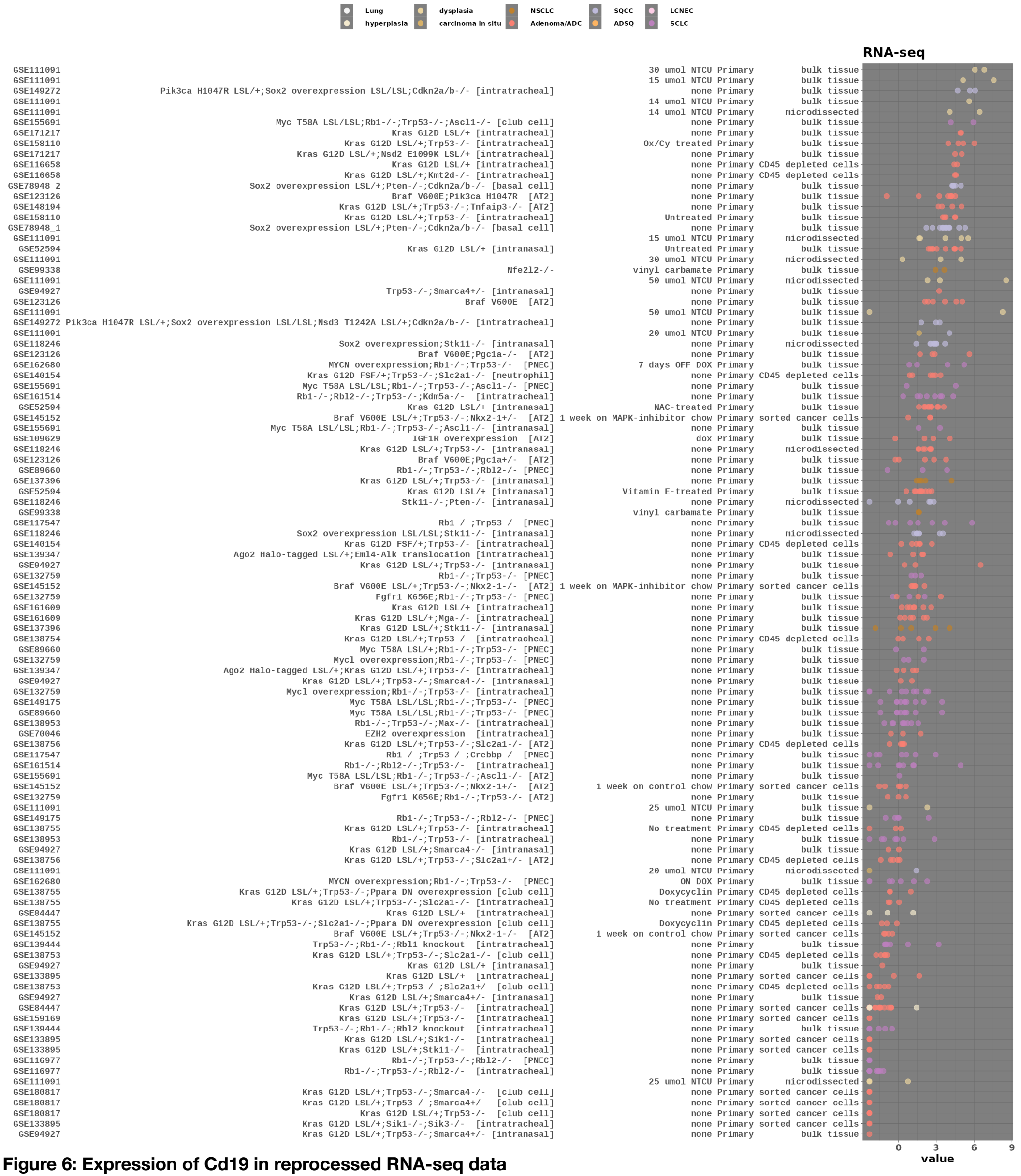

Figure 6: Expression of Cd19 in reprocessed RNA-seq data

Each dot represents a unique sample, colored according to histology, and ordered by the median expression of Cd19. The plot is refined to display primary tumors only.

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.