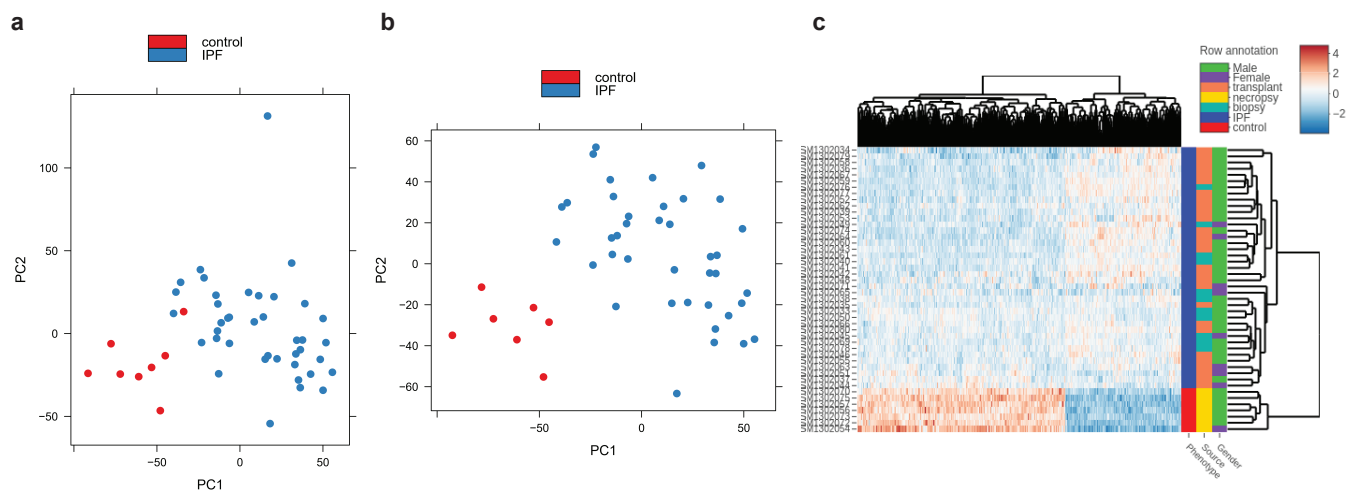


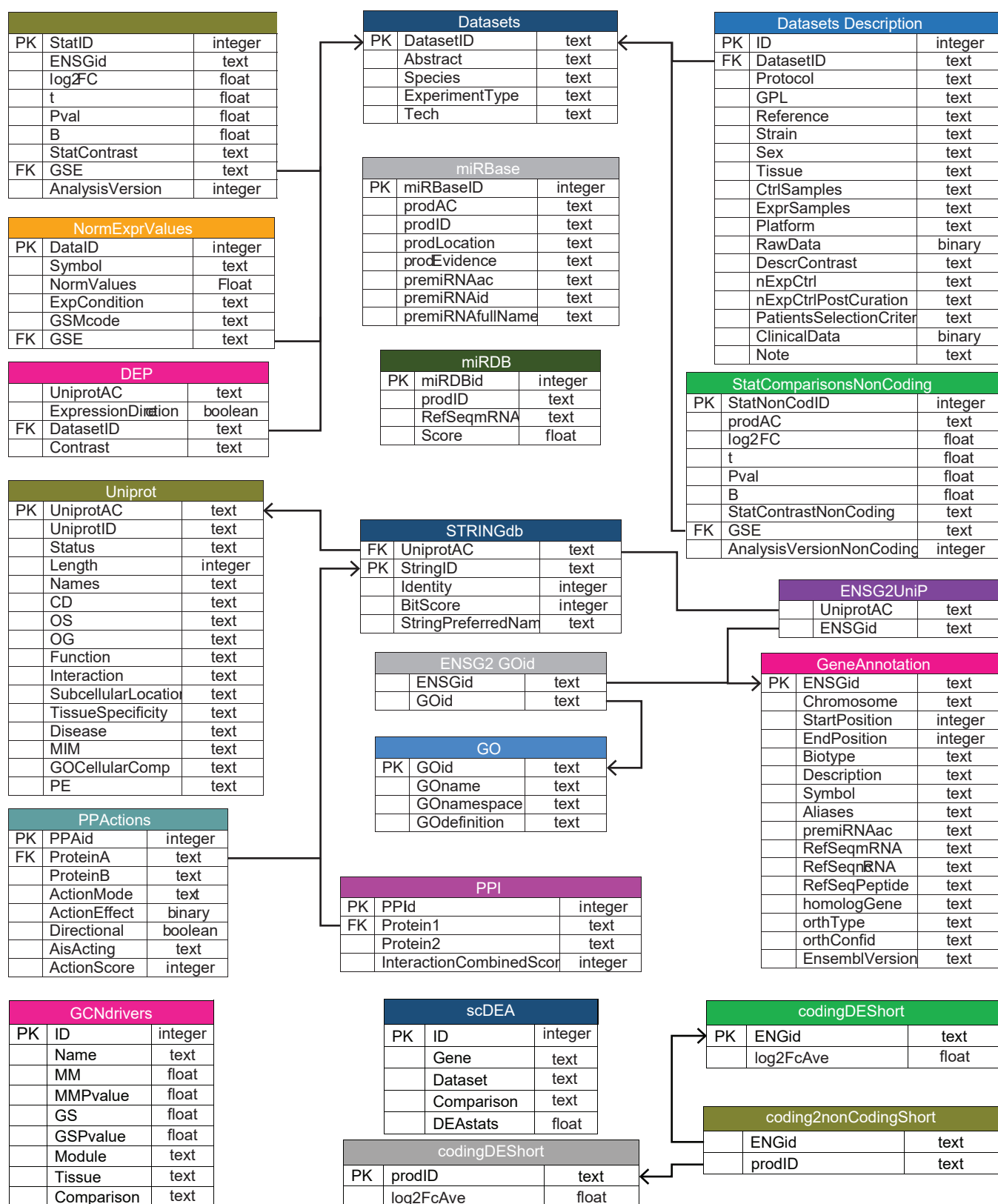
Supplementary information

Fibromine is a multi-omics database and mining tool for target discovery in pulmonary fibrosis

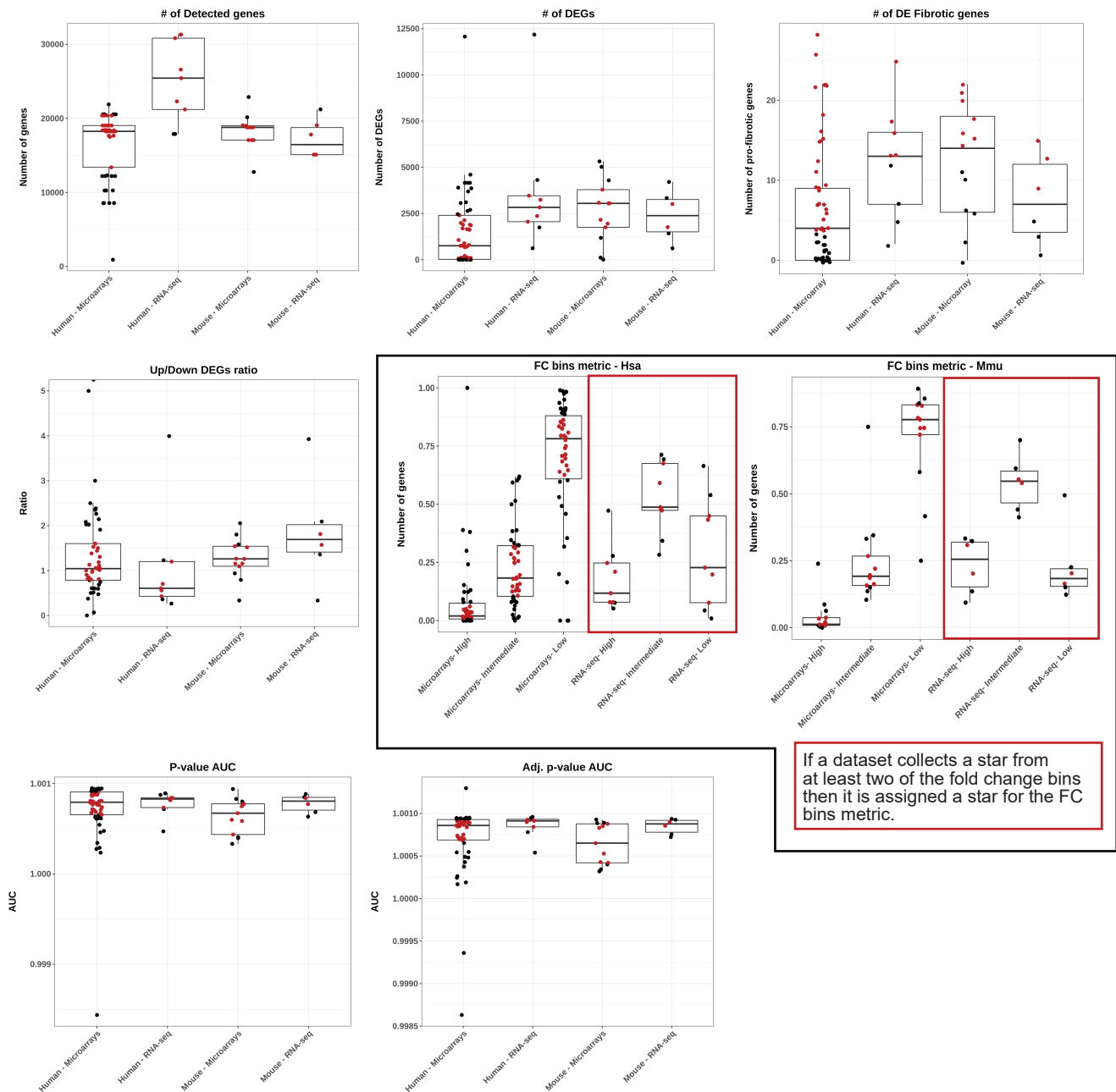
Dionysios Fanidis, Panagiotis Moulos and Vassilis Aidinis



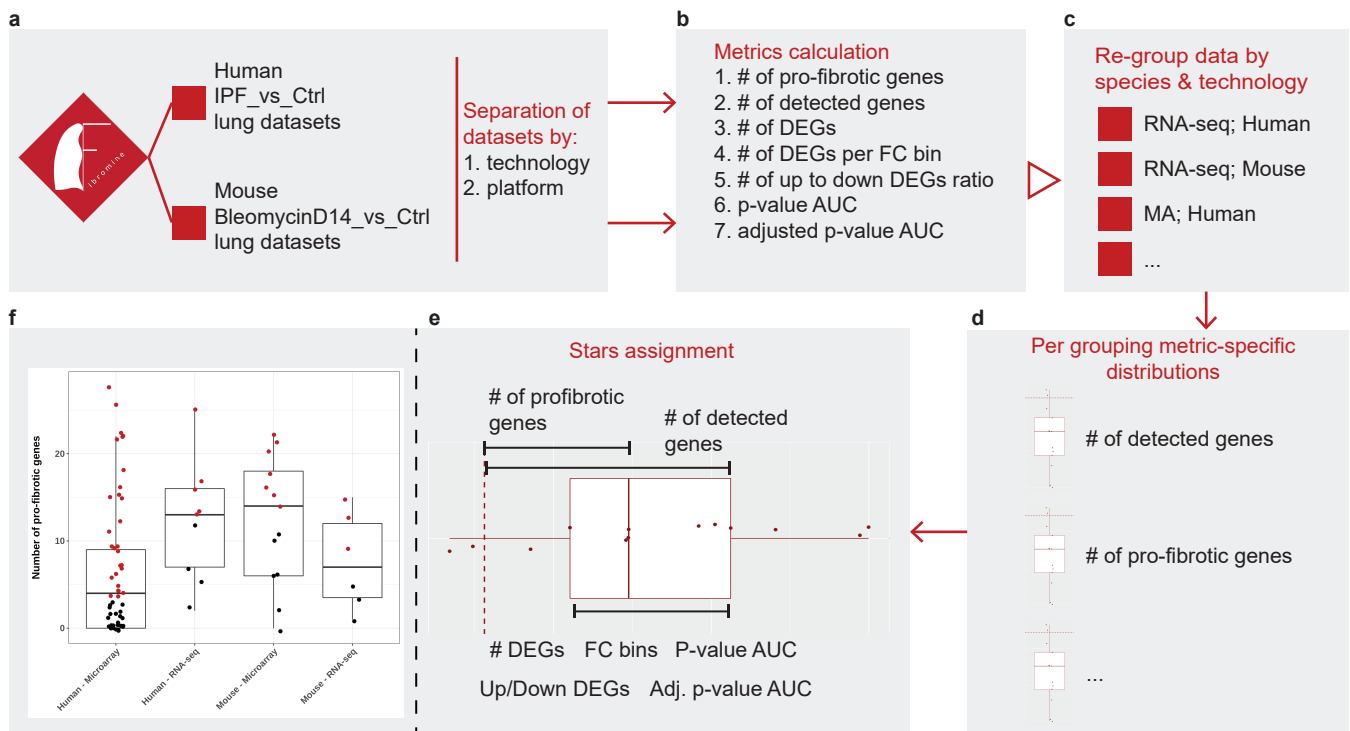
Supplementary Figure 1: Datasets manual curation example. All transcriptomic datasets included into Fibromine were manually curated using PCA/MDS and samples hierarchical clustering plots. **(a)** PCA plot of the GSE53845 dataset prior to manual curation. **(b-c)** PCA and hierarchical clustering plot of the same dataset post to manual curation. *PCA: principal component analysis; MDS: multi-dimensional scaling.* Figure created using ggplot2 v3.3.5, heatmaply v1.2.1 and Google slides.



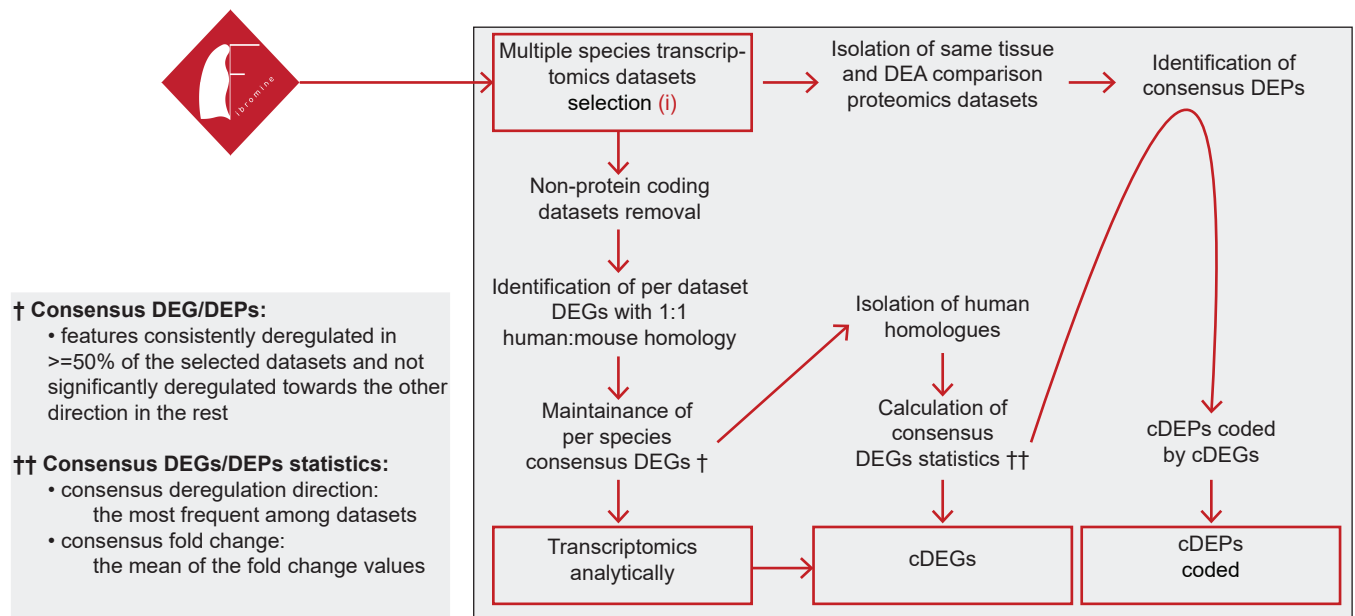
Supplementary Figure S2: Fibromine schema. Figure created using Google slides.



Supplementary Figure S3: Fibromine datasets benchmarking analytically. All coding transcriptomic datasets were assessed using seven metrics in an attempt to help the user choose those of interest. After separating datasets/comparisons as depicted in Figure 2a, we have created a per-metric distribution for each of the dataset groups. A dataset was starred for a certain metric (red dots), if the latter's value was within the interquartile range of the respective distribution, as depicted above. Exception to the latter is the Detected genes and Pro-fibrotic genes metrics, where the median to 90th percentile and the above the median ranges were utilized. ● *Starred dataset/comparison* ● *Non-starred dataset/comparison*. Figure created using ggplot2 v3.3.5 and Google slides.



Supplementary Figure S4. Datasets benchmarking for major Fibromine DEA comparisons. In order to reveal the most homogeneous transcriptomic datasets for two major Fibromine DEA comparisons (**a**) seven metrics were calculated (**b**). Then, data were re-grouped per species and technology (**c**) and metric-specific distributions were created for each of the final groupings (**d**). Last, every dataset was evaluated with respect to its group and assigned a per-metric star if the calculated criterion value lied within a pre-specified interval of the respective distribution (**e**). Each dataset received at maximum seven and at minimum zero stars. Datasets with many stars are more closely related to each other than to the rest of the group. (**f**) A real example of stars assignment: red dots are datasets starred for the *Number of pro-fibrotic genes* criterion. Boxplots depict the interquartile range and median of the data; whiskers extend no longer than 1.5 times the length of the boxplot. *DEGs*: differentially expressed genes; *FC*: fold change; *AUC*: area under the curve. Figure created using ggplot2 v3.3.5 and Google slides.



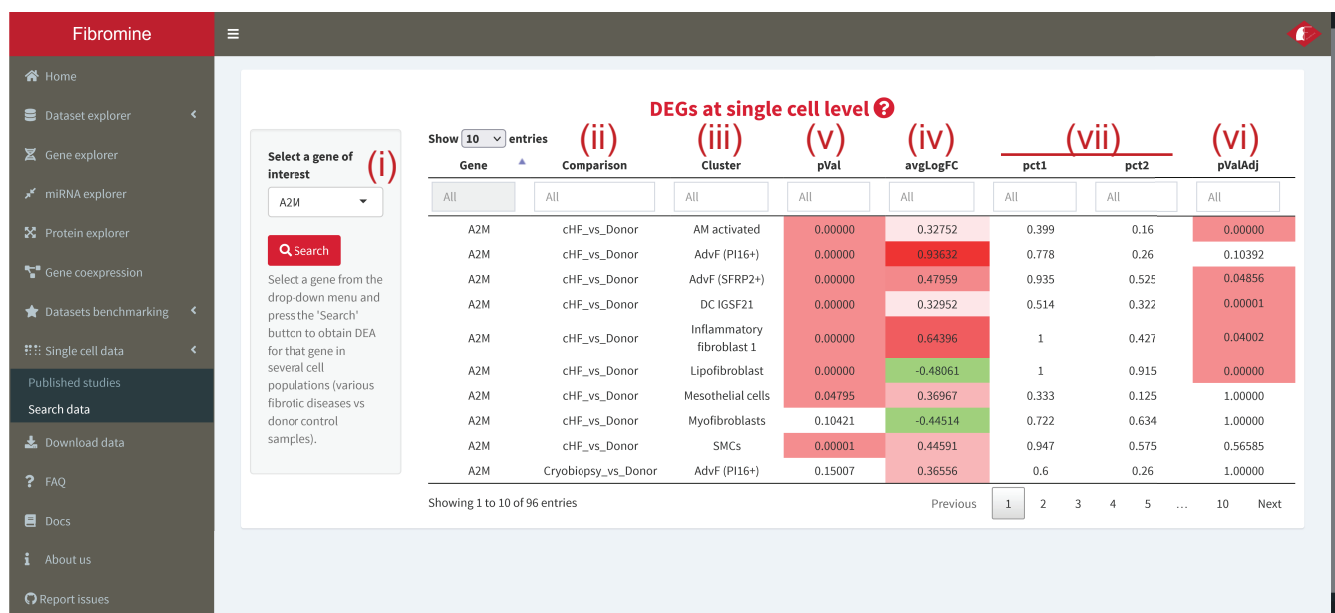
Supplementary Figure S5. Across species datasets integration workflow. Across species transcriptomic datasets *Dataset explorer* back-end integration workflow. If datasets from more than one species is selected, then datasets integration workflow is slightly modified and finally, for clarity reasons, focus is given to human consensus DEGs. Consensus DEPs returned must share the same direction of deregulation with their coding gene. Boxes of red correspond to front-end elements, while the rest account for back-end processes. *cDEGs*: consensus differentially expressed genes; *cDEPs*: consensus differentially expressed proteins. Figure created using Google slides.

The screenshot displays the Fibromine Gene Explorer interface. On the left is a dark sidebar with navigation links: Home, Dataset explorer, Gene explorer (selected), Protein explorer, Datasets benchmarking, Single cell data, Download data, How to, Docs, About us, and Report issues.

The main content area is divided into several sections:

- Top Left:** A search bar with "MAP3K8" entered. Below it are filters for species (Both, Human, Mouse) and buttons for Search, Example (i), and About.
- Top Right:** Section (iii) General information. It shows a table with columns: Name, Aliases, Code, Chromosome, StartPosition, EndPosition, Biotype, and IsTF. The entry for MAP3K8 is displayed.
- Middle Right:** Section (viii) Map to single cell data. It shows a table with columns: Name, Code, and Single cell dataset mapping. The entry for MAP3K8 is displayed.
- Bottom Left:** Section (vi) Expression data and (vii) Gene ontology. The Expression data tab is active, showing a table with columns: Name, Species, Tissue, DatasetID, Comparison, Stars count, #Exp, #Ctrl, Tech, log2FC, Pval, and FDR. The table is filtered for "Lung tissue" and "IPF_vs_Ctrl" comparisons.
- Bottom Right:** Section (iv) DEG statistics. It shows a table with columns: Name, Species, Tissue, DatasetID, Comparison, Stars count, #Exp, #Ctrl, Tech, log2FC, Pval, and FDR. The table is filtered for "Lung tissue" and "IPF_vs_Ctrl" comparisons.
- Bottom Center:** Section (v) DEP data. It shows a table with columns: Name, UniprotAC, ExpressionDirection, DatasetID, Tissue, and Contrast. The table is empty, with a message "No data available in table".

Supplementary Figure S6. Gene explorer interface example. Through Gene explorer the user can interrogate Fibromine data for one or multiple genes of interest simultaneously. Here the output of the *Example* button (i) is displayed. *DEG statistics* table results are filtered to present lung tissue IPF_vs_Ctrl comparisons (ii). Initially, some general information is displayed (iii), while DEG statistics table (iv) records in detail the respective expression patterns. Differentially expression analysis data from the proteomic datasets corresponding to the genomic feature queried are presented at the DEP data table (v), if any. Gene ontology (vi) and RefSeq - miRNA tabs (vii) present the GO terms matched with the query elements, RefSeq data and miRNA-mRNA potential interactions sourced from miRDB, respectively. Finally, the user has the option of exploring gene expression at the single cell level via dynamic redirection to the NU-Pulmonary online resource (viii) in a species-specific manner. *Figure created using Google slides.*



Supplementary Figure S9. Single cell data exploration interface example. Differential expression analysis results from the single cell dataset of Mayr et al. (2021) can be investigated via *Single cell data* > *Search data* tab of Fibromine. The user can query the expression profile of the most variable features (i) as it changes between any of the ten included pathologies relative to control donor samples (ii) in a set of 45 cellular types (iii). Direction of deregulation is color coded in the Average log Fold Change (avgLogFC; iv) column: positive fold changes are marked with shades of red and negative fold changes with shades of green. P-value (pVal; v) and p-value adjusted (pValAdj; vi) column cells are colored in red every time the respective value is smaller than 0.05. Columns pct1 and pct2 (vii) record the percentage of cells in each comparison part that expresses the queried gene. *Figure created using Google slides.*