RESEARCH AND GUIDELINE UPDATE

Lung Gene Expression Analysis (LGEA): an integrative web portal for comprehensive gene expression data analysis in lung development

Yina Du,[1] Joseph A Kitzmiller,[1] Anusha Sridharan,[1] Anne K Perl,[1] James P Bridges,[1] Ravi S Misra,[4] Gloria S Pryhuber,[4] Thomas J Mariani,[4] Soumyaroop Bhattacharya,[4] Minzhe Guo,[1] S Steven Potter,[2] Phillip Dexheimer,[3] Bruce Aronow,[3] Alan H Jobe,[1] Jeffrey A Whitsett,[1] Yan Xu[1,3]

[1]The Perinatal Institute and Section of Neonatology, Perinatal and Pulmonary Biology, Cincinnati Children's Hospital Medical Center, Cincinnati, Ohio, USA
[2]Division of Developmental Biology, Cincinnati Children's Hospital Medical Center, Cincinnati, Ohio, USA
[3]Division of Biomedical Informatics, Cincinnati Children's Hospital Medical Center, Cincinnati, Ohio, USA
[4]Department of Pediatrics, University of Rochester, Rochester, New York, USA

Correspondence to
Dr Yan Xu, The Perinatal Institute and Section of Neonatology, Perinatal and Pulmonary Biology, Cincinnati Children's Hospital Medical Center, Cincinnati, OH 45229, USA; Yan.Xu@cchmc.org

Received 17 October 2016
Revised 30 November 2016
Accepted 8 December 2016
Published Online First
10 February 2017

ABSTRACT

'LungGENS', our previously developed web tool for mapping single-cell gene expression in the developing lung, has been well received by the pulmonary research community. With continued support from the 'LungMAP' consortium, we extended the scope of the LungGENS database to accommodate transcriptomics data from pulmonary tissues and cells from human and mouse at different stages of lung development. Lung Gene Expression Analysis (LGEA) web portal is an extended version of LungGENS useful for the analysis, display and interpretation of gene expression patterns obtained from single cells, sorted cell populations and whole lung tissues. The LGEA web portal is freely available at http://research.cchmc.org/pbge/lunggens/mainportal.html.

INTRODUCTION

We previously developed 'LungGENS',[1] a web tool for mapping single-cell gene expression in the developing lung. LungGENS was visited by approximately 45 institutions in 30 countries during the past year. The initial phase of the LungGENS web tool was based on single-cell RNA sequencing (scRNA-seq) data from normal fetal mouse lung, with our newly developed analytic pipeline 'SINCERA'.[2] With continued support from the 'LungMAP' consortium, transcriptomic data derived from various technical platforms, species and lung developmental stages have become available. Integration and visualisation of these data sets with user-friendly web interfaces will empower investigators to access and interpret data contained in the extended database to better understand lung development and disease. To accommodate heterogeneous data structures and types, we developed Lung Gene Expression Analysis (LGEA) web portal, an extended version of the LungGENS, seeking to identify lung cell types and the dynamic changes in gene expression influencing lung formation and function using RNA-seq from single cells, purified cell populations and whole tissue.

METHODS

The web pages and JavaScript functions of the LGEA web portal were designed and developed using HTML/CSS, JavaScript/jQuery and Java in Eclipse (http://www.eclipse.org/), a Java IDE. Apache Tomcat (http://tomcat.apache.org/) was used as web server. JSON (JavaScript Object Notation) format was adopted as an interchangeable data structure for these programming languages to encode LGEA query results, making downstream data processing and exchange easy and language-independent. When a gene symbol or a cell type is chosen, the client has initiated an HTTP request to the LGEA web server. A Java servlet on the web server handles the request, retrieves the data from database using SQL scripts and prepares retrieved data in JSON format. Finally, a processed HTTP response is returned to the client by displaying a page containing the query data.

Oracle Database 11g (https://www.oracle.com/database/index.html) is used as a central component of LGEA web portal to improve data storage and efficient database management. The relational database of LGEA web portal is designed in compliance with the design structure in the previous LungGENS relational database using gene symbols and their associated cell types as primary keys within the relational data tables.

The interaction and visualisation of LGEA web portal is supported by Highcharts (http://www.highcharts.com/), an interactive charting library. Highcharts is compatible with modern mobile and desktop browsers (eg, Safari, Firefox and Chrome). In addition to using interactive heatmaps, histograms, bar graphs and profile charts to display gene expression data from individual cells, we implemented new graphical and statistical presentations including principal component analysis (PCA), scatter plot, box plot and Venn diagram into the LGEA web page design.

RESULTS

Lung development is a highly regulated and coordinated process typified by stage-specific changes in structure and function including branching morphogenesis, angiogenesis, sacculation, alveologenesis and cytodifferentiation.[3] In mice, formation and maturation of the gas exchange region of the lung begins at approximately embryonic day 15 (E15) and ends at postnatal day 30 (PN30). In addition to mouse lung E16.5 single-cell RNA-seq data previously published in LungGENS, the LGEA database has been extended to include single cell,

To cite: Du Y, Kitzmiller JA, Sridharan A, *et al*. *Thorax* 2017;**72**:481–484.

sorted cell and developmental time course data from whole lung tissues from E16.5 to PN28 and adults. The database is synchronised with ongoing studies from the research centres of 'LungMAP' consortium. The LGEA web portal provides three major types of analyses using the extended database: (1) single-cell transcriptome analysis using 'LungGENS' (2) sorted lung cell populations analysis using 'LungSortedCells' and (3) Lung Developmental Time Course analysis using 'LungDTC' as depicted in figure 1A.

LungGENS

The initial release of LungGENS was hosted using scRNA-seq data obtained from fetal mouse lung at E16.5 (148 cells). The current version of LungGENS database contains additional cells sequenced from E16.5 and E18.5 mouse lung, processed using Fluidigm C1 microfluidics technology. 'Gene Query' and 'Cell Type Query' retrieve data from the expanded database to provide cell-specific gene expression patterns for each lung cell type and associated gene signatures, surface markers and transcription factors for cell types of interest. 'Gene list query' has been expanded to all data sets in the LGEA web portal. Users can input a list of gene symbols and retrieve predicted cell types co-expressing their gene list of interest.

LungSortedCells

The LungSortedCells database includes fluorescence-activated cell sorting (FACS) sorted cell populations enriched for endothelial, mesenchymal, immune and epithelial cells from human lung (processed by Human Tissue Core (HTC) at University of Rochester supported by the LungMAP consortium) at day 1 and 20 months; sorted mouse alveolar type 2 cells at PN7 and PN28, sorted mouse mesenchymal, immune and epithelial cells at PN7 and PN28, and *Pdgfra* expressing fibroblasts at E16.5, E18.5, PN7 and PN28 (processed by 'CCHMC'" Mouse Hub supported by the LungMAP consortium). 'Gene Query' allows users to input a gene symbol of interest. Query output uses a bar graph to display its expression levels across all cell types and a monocolor heatmap to provide an overview of the levels of expression of the queried gene expression across all data sets in LGEA database (figure 1D). 'Cell type query' identifies a list of signature genes for a given cell type, displays them using an interactive heatmap and provides downloadable data table for the query. Transcription factors, cell surface markers and signature genes identified from scRNA-seq analysis were also listed in tabular form as cross-reference (figure 1D). Gene symbols in the data tables are designed as a pop-up query panel, enabling users to redirect the query gene to any data set within LGEA database (figure 1D). Signature genes are identified using the following criteria: (1) gene A is expressed in cell B with an expression level >0.6 quantile in the whole-genome distribution; (2) gene A is expressed in cell B at least fivefold higher than the average expression of gene A in all other cell types; (3) gene A is most highly expressed in cell B with at least 1.5-fold higher expression than the cell type expressing the next highest level of gene A and (4) the coefficient of variation of gene A in cell B among biological replicates is <0.5.

LungDTC

We collected Developmental Time Course (DTC) data sets from whole mouse lung RNA microarray experiments from three mouse strains of (E15.5 to PN30),[3 4] whole mouse lung RNA-seq at E16.5, E18.5, PN1, PN3, PN7, PN14 and PN28 (processed by 'CCHMC' Mouse Hub) and whole Rhesus macaque lung RNA-seq (GA100, 130 and 150). We present a combination of PCA, scatter chart, line chart and downloadable differentially expressed gene tables to display dynamic gene expression patterns across important developmental time periods (figure 1E). Users can compare the expression data from distinct mouse strains and different technical platforms including RNA microarray and RNA-seq. Dynamic profile patterns are displayed in line charts and downloadable gene tables (figure 1E) from which users can explore the expression of an individual gene profile and redirect sets of genes sharing similar expression patterns to ToppGene (https://toppgene.cchmc.org/enrichment.jsp) for gene set enrichment analysis.

LGEA Tools

To facilitate comparison and integration analyses, we developed new tools including 'Gene at a Glance' (figure 1B) and Signature Comparison ('SigComparison') (figure 1C). 'Gene at a Glance' enables users to input any gene of interest and displays the given gene expression information across developmental times and conditions within the LGEA database (figure 1B). 'SigComparison' compares signature genes between two experimental conditions within the LGEA database, displays the results using a Venn diagram and calculates the correlation of the overlapping data sets. Alternatively, users can input and compare their gene lists with the signature genes identified for specific cell types in LGEA database or compare two gene lists independently of the LGEA database (figure 1C). In addition to hosting analytic tools developed for LungMAP, LGEA provides URL links to >60 commonly used internal and external resources. For example, by clicking the tab for 'Lung Image' from the LGEA homepage, users will be redirected to Lung Image web collection (https://research.cchmc.org/lungimage/) hosted by Dr Whitsett's laboratory. The Lung Image gallery contains >2000 immunofluorescence confocal microscopies images obtained from embryonic (E16.5 and E18.5) and postnatal mouse (PN1, PN3, PN7, PN10, PN14 and PN28) lungs, with protein markers representing major pulmonary cell types. The gallery also contains >1000 images of postnatal human lung from 4 months to 4 years of age. A link between each protein marker and available single-cell RNA-seq data in LungGENS is provided.

LIMITATIONS AND FUTURE DIRECTIONS

The first phase of LGEA is aimed to develop user-friendly tools for the lung research community for quick and easy transcriptomic data access. Some areas are still needed for further development in order to improve the functionality and to better meet different levels of data analysis. Some of the limitations are described below. (1) Currently, LGEA only covers transcriptomic data from normal mouse and human lung. Other omics data types including proteomic, metabolomics and lipidomic and data related to lung diseases are not yet included. We are actively working to expand LGEA database to include single-cell data from idiopathic pulmonary fibrosis, cystic fibrosis and other chronic lung diseases. (2) The current version of LGEA contains single-cell data processed using Fluidigm C1 microfluidics technology, limiting the number of cells being captured, in turn influencing the power of statistical analysis. At present, RNA data is being produced from thousands of individual lung cells using 'Drop-seq'.[5] We are actively working on analytic pipeline to facilitate the complex data mining of the 'Drop-seq' RNA-sequencing data. LungGENS will be expanded to include single-cell RNA-seq data from this new platform. Transcriptomic data from increasing numbers of single cells will increase statistical power used for cell-type characterisation and

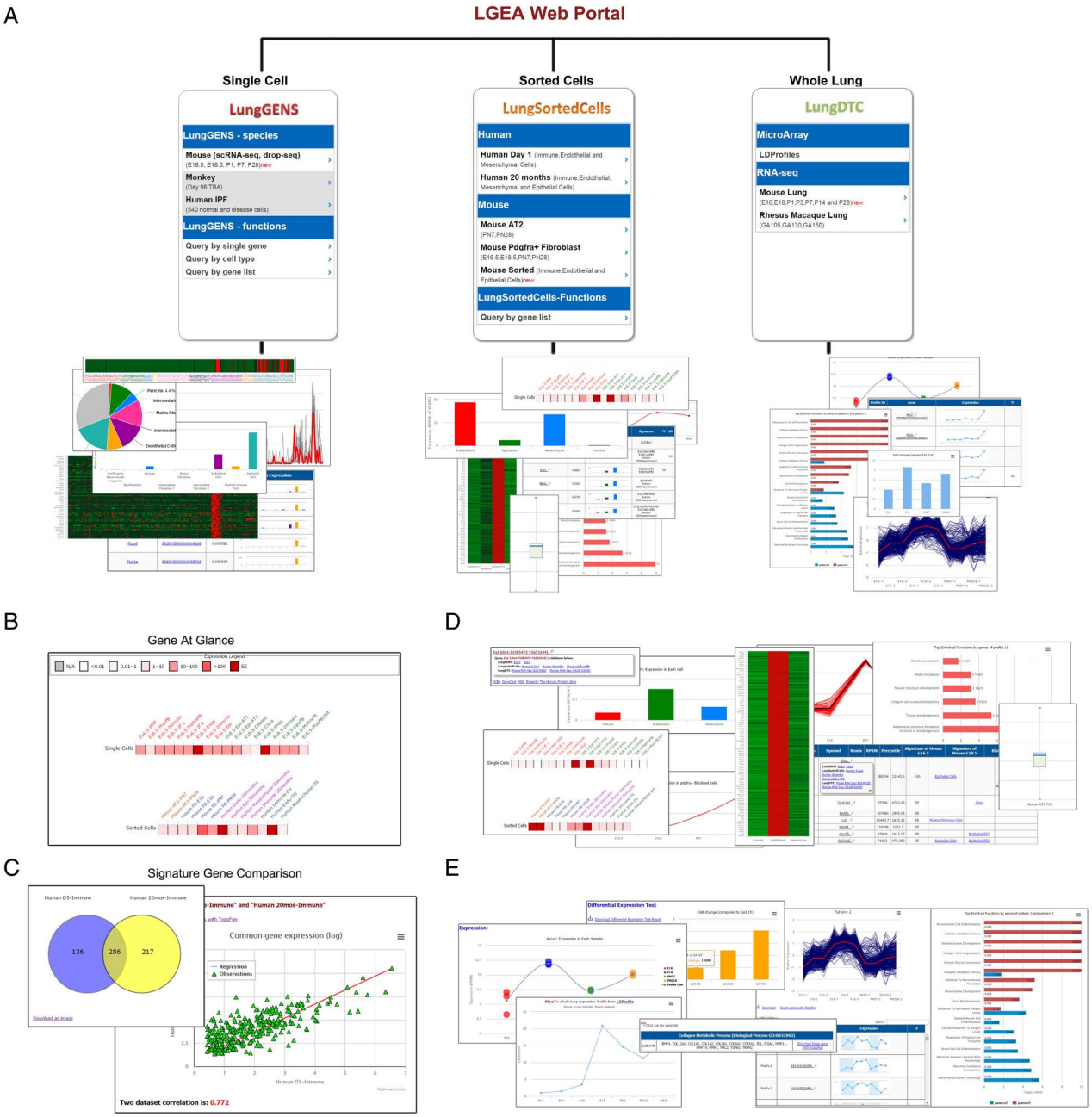

Figure 1 (A) The home page of the Lung Gene Expression Analysis (LGEA) web portal provides access to data and to query results. Two integrative analytical tools 'Gene At Glance' (B) and 'SigComparison' (C) in LGEA are shown. LungSortedCells and LungDTC query functions are shown (D and E).

signature gene identification, enabling identification of rare or novel cell types. (3) Current LGEA web queries are performed on data for one gene/cell at a time or for gene lists containing <500 genes at a time. (4).The current version of LGEA query only accepts official gene symbols from annotated human and mouse genomes. Since there are many online tools for the gene ID conversion; including Biomart (http://central.biomart.org/), DAVID (http://david.abcc.ncifcrf.gov/) and biological DataBase network (http://biodbnet.abcc.ncifcrf.gov/db/db2db.php), we recommend users to convert different types of IDs to official gene symbols prior to LGEA applications. (5) The current version of LGEA does not provide customised application program interfaces (API) for programmatic data access. Nevertheless, at present, a portion of the LGEA functions can be directly accessed using programming language, such as R, Python or Java, by using appropriate network (HTTP) client-side APIs since the query results are encoded in JSON format.

CONCLUSIONS

The new LGEA web portal is designed to implement new features and analytical methods to provide an extended database enabling rapid analysis of (1) scRNA-seq using 'LungGENS' (2) sorted lung cell populations using 'LungSortedCells' and (3) lung developmental time course data using 'LungDTC'. LGEA provides useful graphical interfaces with new interactive options to uses increasingly comprehensive RNA expression data sets.

The new LGEA web portal database will be naturally extended to new data generated from normal and abnormal lung tissues and cells from additional species, developmental times and experimental protocols. The LGEA will be broadly applicable for lung research and freely available at http://research.cchmc.org/pbge/lunggens/mainportal.html and LungMAP research Consortium website (http://www.lungmap.net/) supported by National Heart, Lung, and Blood Institute (NHLBI).

Twitter Follow Yan Xu @YanXu_Cincy

Acknowledgements The authors thank Dr Sara Lin (Program Director) and all members of LungMAP research consortium. The authors acknowledge the technical contributions from Mehari Endale for mouse Pdgfra+ cell preparation, Dr Susan Wert for coordinating monkey tissue and cells. Charleen Slaunwhite and Terry Wightman from the University of Rochester Flow Cytometry Core; John Ashton, Ben Smith, Michelle Zanche, Kelly Schooping and Jason Myers from the University of Rochester Genomics Research Core, for their contribution to study design, cell sorting and RNA sequencing for the HTC.

Contributors YX and YD conceived and designed the web application. YD developed the database and web application. JAK, AS, AKP, JPB and JAW designed and conducted mouse single-cell and sorted cell RNA-seq experiments. AHJ and JAW designed and conducted rhesus macaque RNA-seq experiments. RSM, GSP, TJM and SB designed and conducted human sorted cell RNA-seq experiments. SSP, JAW, PD and BA conceived and performed the single-cell experiments, helped generate and interpret the lung single-cell data. MG, YD and YX contributed to data analysis and interpretation. YD, JAW and YX contributed to the writing of the manuscript. All authors have read and provided inputs to the manuscript.

Funding NHLBI U01 HL122642 (LungMAP).

Competing interests None declared.

Provenance and peer review Not commissioned; externally peer reviewed.

Data sharing statement LGEA web portal and LungGENS are freely available for non-commercial use at http://research.cchmc.org/pbge/lunggens/mainportal.html and data will be integrated with other omics data and lung image data at 'BREATH' database and display on the LungMAP website (http://www.lungmap.net/).

REFERENCES

1 Du Y, Guo M, Whitsett JA, *et al*. 'LungGENS': a web-based tool for mapping single-cell gene expression in the developing lung. *Thorax* 2015;70:1092–4.
2 Guo M, Wang H, Potter SS, *et al*. SINCERA: a Pipeline for Single-Cell RNA-Seq Profiling Analysis. *PLoS Comput Biol* 2015;11:e1004575.
3 Xu Y, Wang Y, Besnard V, *et al*. Transcriptional programs controlling perinatal lung maturation. *PLoS ONE* 2012;7:e37046.
4 Mariani TJ, Reed JJ, Shapiro SD. Expression profiling of the developing mouse lung: insights into the establishment of the extracellular matrix. *Am J Respir Cell Mol Biol* 2002;26:541–8.
5 Macosko EZ, Basu A, Satija R, *et al*. Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. *Cell* 2015;161:1202–14.