AUDIT, RESEARCH AND GUIDELINE UPDATE

'LungGENS': a web-based tool for mapping single-cell gene expression in the developing lung

Yina Du,[1] Minzhe Guo,[1,2] Jeffrey A Whitsett,[1] Yan Xu[1,3]

[1]Department of Perinatal and Pulmonary Biology, The Perinatal Institute and Section of Neonatology, Cincinnati Children's Hospital Medical Center, Cincinnati, Ohio, USA
[2]Department of Electrical Engineering and Computing Systems, University of Cincinnati, Cincinnati, Ohio, USA
[3]Division of Biomedical Informatics, Cincinnati Children's Hospital Medical Center, Cincinnati, Ohio, USA

Correspondence to
Dr Yan Xu, Divisions of Pulmonary Biology, Perinatal Institute and Biomedical Informatics, Cincinnati Children's Hospital Medical Center, 3333 Burnet Avenue, MLC7029, Cincinnati, OH 45229-3039, USA; Yan.Xu@cchmc.org

ABSTRACT
We developed LungGENS (Lung Gene Expression iN Single-cell), a web-based bioinformatics resource for querying single-cell gene expression databases by entering a gene symbol or a list of genes or selecting a cell type of their interest. Gene query provides quantitative RNA expression of the gene of interest in each lung cell type. Cell type query returns associated selective gene signatures and genes encoding cell surface markers and transcription factors in interactive heatmap and tables. LungGENS will be broadly applicable in respiratory research, providing a cell-specific RNA expression resource at single-cell resolution. LungGENS is freely available for non-commercial use at https://research.cchmc.org/pbge/lunggens/default.html.

INTRODUCTION
The lung is a complex multicellular organ composed of a diversity of distinct cell types that interact to accomplish lung morphogenesis and function. Knowledge regarding the proliferation, differentiation and functions of individual cells and the mechanisms by which cells interact to form the lung provides insight into the processes underlying lung morphogenesis, function and repair. Recent advances in single-cell isolation and massive parallel DNA sequencing enable resolution of gene expression in individual cells, providing insight into the diversity of cell types, and gene networks directing cell differentiation, and the complex interactions among diverse cell types. We used massive parallel DNA sequencing and an unbiased analytic approach to identify major cell types and the mRNA signatures in single cells isolated from the entire embryonic mouse lung at the saccular phase of morphogenesis, a time of active proliferation and differentiation (E16.5) before birth. The data provide a rich knowledge base, identifying unique contributions made by multiple pulmonary cell types, and the biological processes mediating formation and function of the lung prior to birth (M Guo, *et al*: SINCERA: A pipeline for single-cell RNA-Seq profiling analysis, submitted). Processing and interpreting the extensive RNA data generated at single-cell level in an entire organ present a major analytical challenge. There are few readily accessible web tools or databases to facilitate query and visualisation of such complex gene expression patterns. In order to facilitate access of single-cell transcriptomic data and visualise the complex data, we developed LungGENS (Lung Gene Expression iN Single-cell), a web tool useful for mapping gene expression patterns in specific pulmonary cells at

single-cell level. The current version of LungGENS was built using the lung single-cell RNA-seq data from mouse lung at embryonic day, E16.5. The programme and website will be extended for expression data from ongoing studies of the mouse and human lung during development as they are completed by the LungMAP consortium and other investigators interested in lung biology.

METHODS
LungGENS was developed in Eclipse (http://www.eclipse.org/), a Java IDE (integrated development environment). Specifically, HTML5, JavaScript and Java programming languages were used for LungGENS web page and server development. We used JSON (JavaScript Object Notation) to support an interchangeable data structure for these programming languages, making data transmission between a server and web application easy and language independent.

MYSQL, a relational database management system, serves as one central component of the web tool by managing data storage and retrieval. In the database, LungGENS data sets were divided into several relational data tables with gene names and cell types as primary keys. This design guarantees the efficiency and accuracy of data querying and table operation. Relational data tables include RNA-Seq gene expression, FPKM (Fragments Per Kilobase of Exon Per Million Fragments Mapped) or raw counts, normalised expression (Z-score transformed), gene summary, gene correlations, cell selective signature genes, surface markers and associated transcription factors.

To provide interactive data visualisation, Highchart (http://www.highcharts.com) was applied during web tool development. Highchart is compatible in both modern mobile and desktop browsers (eg, Safari and Chrome). LungGENS used combinations of heatmap, histogram and bar graph to display gene expression for individual cells and statistical results. Profile charts (line chart) were used to display RNA expression profiles after entry of a query gene and its 20 most highly correlated genes that may share functional similarity with the query gene.

RESULTS
LungGENS is an open access tool that can be freely accessed at https://research.cchmc.org/pbge/lunggens/default.html. The search tool is easy to use and offers two distinct interfaces that facilitate user-defined queries: 'Query by gene' and 'Query by cell type' as depicted in figure 1.
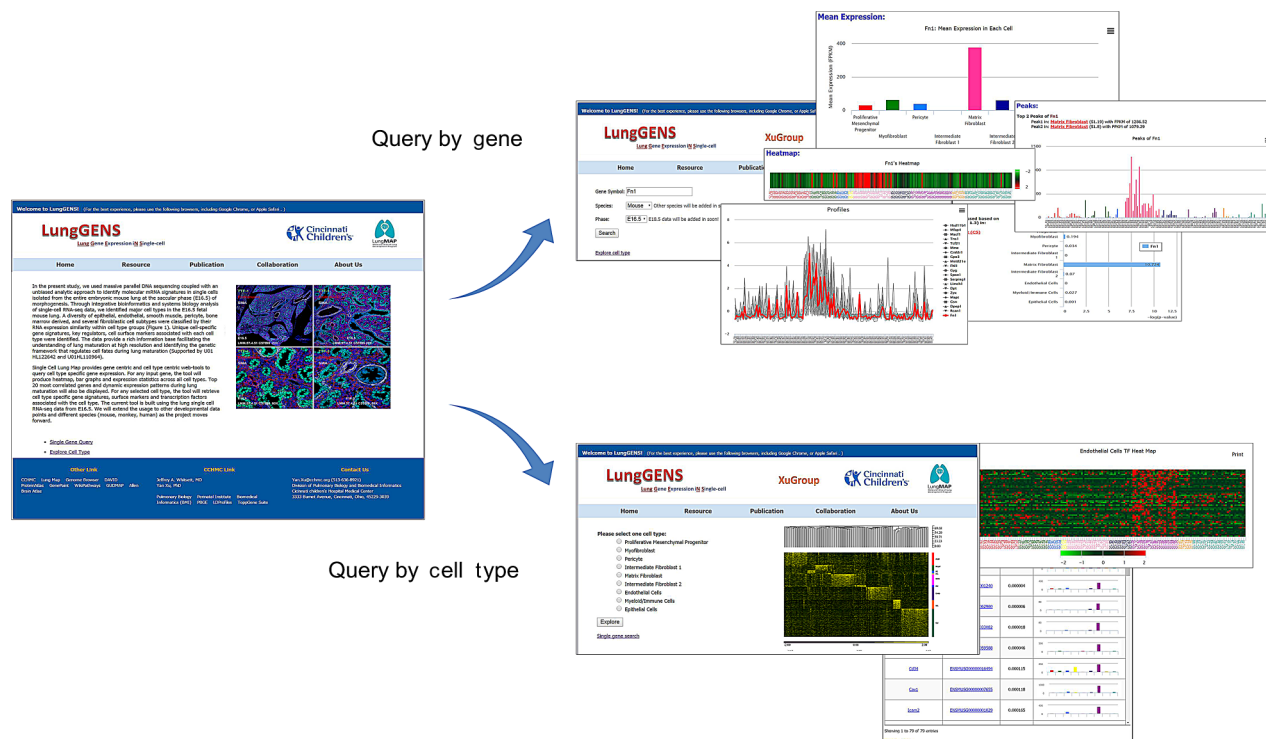
Figure 1 LungGENS (Lung Gene Expression iN Single-cell) provides two main queries: 'Query by gene' and 'Query by cell type'.

Query by gene

'Query by gene' allows users to input a gene symbol of interest. Query output consists of five sections:

'Identifications' ('IDs') section: LungGENS provides external links by mapping query gene to benchmark databases and resources (NCBI, GeneCard, MGI, Ensembl and Protein Atlas). By redirecting users to these authoritative databases, users can readily incorporate external knowledge to gain a more comprehensive understanding of the gene of interest.

'Summary' section: LungGENS displays query gene expression from 148 single cells isolated from whole lung digests from fetal mice at E16.5 using heatmap and individual bar graphs. The programme provides mean expression and t test p values (negative log-transformed) of the query gene in nine distinct cell types using histograms. LungGENS highlights the cell type in which the query gene is significantly more highly expressed in red colour. The highlighted cell type is a 'clickable' hyperlink which will redirect users to 'Query by cell type' page. Cells with highest and the second highest levels of expression are identified.

'Cell Type Statistics' section: LungGENS summarises cell type statistics for the queried gene, including maximum, minimum, average expression, the number and identity of cells expressing the RNA and the percentage of cells expressing it in each cell type.

'Top 20 Correlated Genes' section: LungGENS retrieves 20 most closely correlated genes across 148 cells from precompiled correlation tables for each query gene. Data are presented as a group of profiles in charts, the query gene being highlighted in red. A table lists the 20 most correlated genes, their correlation coefficient and their mean expression across nine cell types is shown using histograms. Users can further examine the closely correlated genes by clicking the gene symbol/ID or after downloading the table for further analysis.

'Dynamic Gene Expression' section: LungGENS redirects users to 'Lung Developmental Gene Expression Profiles', a web tool we built using lung developmental time course RNA microarray data. This database enables users to query dynamic RNA expression profiles during prenatal–perinatal lung development in three mouse strains A/J, C57BL/6J (B6) and Swiss-Webster Strain, visualised as interactive line graphs.[1][2] The relative expression of the query gene across 52 different tissues was displayed as bar graphs.[3]

The current version of LungGENS supports both 'Query by single gene' and 'Query by gene list'. 'Query by gene list' allows the user to type in or paste their list of genes of interest to identify cell types that selectively express in that set of gene. The search engine will retrieve the expression profile of each gene in the list across all cell types in the database (E16.5 data in current version) and use t test comparison of the gene among various cell types. The cell type distribution is then depicted using a pie chart that represents percentage of genes selectively expressed in each cell type with a t test p value <0.05. Genes with a t test p value>0.05 are defined as unselectively expressed. The corresponding heatmaps and gene lists associated with each cell type are provided. Functional enrichment of the gene set can be performed by submitting the list to 'Toppgene Suite', another Cincinnati Children's Hospital Medical Center (CCHMC) developed web tool (https://toppgene.cchmc.org/).[4] LungGENS provides the interaction and redirection functions for users to select signature genes or correlated genes from LungGENS and directly import them into Toppgene Suite for analysis.

Query by cell type

'Query by cell type' enables users to select one of nine predefined cell types (identified by unsupervised hierarchical clustering followed by functional analysis and biomarker validation) (M Guo, et al, submitted) to query the database. For each cell type of interest, the web tool will provide relatively selective

gene signatures, associated cell surface markers and transcription factors. Featured genes are visualised in an interactive heatmap which enables users to compare gene expression across all individual cells in the nine major cell clusters. Each heatmap is associated with a data table containing Gene symbol, Ensemble ID, t test p value (the lower the p value, the more selective is the associated RNA with the query cell type) and mean expression across nine cell types. Gene symbol and Ensemble ID columns were designed with hyperlinks to enable users to redirect the cell type query back to the gene query page in LungGENS. LungGENS provides a search box associated with each table, which enables users to identify if their gene of interest is a 'signature gene' that is selective for a given cell type. Both heatmap and tables can be downloaded.

In addition to the E16.5 single-cell RNA data generated at CCHMC, single-cell RNA-seq from mouse lung epithelial cells at E18.5 published by Treutlein et al,[5] can be searched in LungGENS. Users can select 'Explore cell type (Mouse E18.5 Epithelium)' under 'New Query Functions' or select 'E18.5 (Mouse-Epi)' in the drop down menu listed under the 'Explore cell type' function. After selection of one of four epithelial subtypes (ie, Club Cells, Ciliated Cells, Alveolar Type I and Type II cells), LungGENS will display a heatmap and bar graphs representing the relative expression of each signature gene for each of the four epithelial cell types. Since we expect that more lung single-cell data will be produced by the research community, we will add new expression data sets as they are published and accessible.

CONCLUSION

Based on our knowledge, there is no readily available web tool to enable users to (1) input a gene of their interest for identification of lung cell types expressing the gene or (2) input a cell type of interest to identify 'signature' genes, surface markers and transcription factors that are selectively expressed in various lung cell types. LungGENS is designed to facilitate the retrieval of lung cell-specific gene expression information from extensive data sets derived from RNA sequencing of single cells and to integrate to the data with previous RNA expression studies from mouse lung at various developmental times.

LungGENS was developed on behalf of the National Institutes of Health, Heart Lung Blood Institute 'LungMAP' research project. The initial phase of the web development was based on RNA-seq data obtained from single cells isolated from fetal mouse lung at E16.5. The current LungGENS database will be naturally extended to ongoing data generated from normal and abnormal lung tissues and cells from additional species, including human, at additional developmental time points.

REFERENCES

1 Xu Y, Wang Y, Besnard V, et al. Transcriptional programs controlling perinatal lung maturation. PLoS ONE 2012;7:e37046.
2 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.
3 Su AI, Wiltshire T, Batalov S, et al. A gene atlas of the mouse and human protein-encoding transcriptomes. Proc Natl Acad Sci USA 2004;101:6062–7.
4 Chen J, Bardes EE, Aronow BJ, et al. ToppGene Suite for gene list enrichment analysis and candidate gene prioritization. Nucleic Acids Res 2009;37(Web Server issue):W305–11.
5 Treutlein B, Brownfield DG, Wu AR, et al. Reconstructing lineage hierarchies of the distal lung epithelium using single-cell RNA-seq. Nature 2014;509:371–5.