Complex Sources of Variation in Tissue Expression Data: Analysis of the GTEx Lung Transcriptome

Matthew N. McCall,[1,*] Peter B. Illei,[2] and Marc K. Halushka[2,*]

The sources of gene expression variability in human tissues are thought to be a complex interplay of technical, compositional, and disease-related factors. To better understand these contributions, we investigated expression variability in a relatively homogeneous tissue expression dataset from the Genotype-Tissue Expression (GTEx) resource. In addition to identifying technical sources, such as sequencing date and post-mortem interval, we also identified several biological sources of variation. An in-depth analysis of the 175 genes with the greatest variation among 133 lung tissue samples identified five distinct clusters of highly correlated genes. One large cluster included surfactant genes (*SFTPA1*, *SFTPA2*, and *SFTPC*), which are expressed exclusively in type II pneumocytes, cells that proliferate in ventilator associated lung injury. High surfactant expression was strongly associated with death on a ventilator and type II pneumocyte hyperplasia. A second large cluster included dynein (*DNAH9* and *DNAH12*) and mucin (*MUC5B* and *MUC16*) genes, which are exclusive to the respiratory epithelium and goblet cells of bronchial structures. This indicates heterogeneous bronchiole sampling due to the harvesting location in the lung. A small cluster included acute-phase reactant genes (*SAA1*, *SAA2*, and *SAA2–SAA4*). The final two small clusters were technical and gender related. To summarize, in a collection of normal lung samples, we found that tissue heterogeneity caused by harvesting location (medial or lateral lung) and late therapeutic intervention (mechanical ventilation) were major contributors to expression variation. These unexpected sources of variation were the result of altered cell ratios in the tissue samples, an underappreciated source of expression variation.

Introduction

Evaluating tissues for patterns of expression is a central tenet of genomic-era studies. These tissue studies have been pivotal to our understanding of diseases and biologic processes, including malignancies, inflammatory disorders, and developmental changes.[1–4] As more high-throughput -omic tools develop, there will be even more desire to assay all manner of RNA, DNA, protein, and epigenetic signals in human and animal tissues. Although obtaining and using human tissue is fundamental to these processes, the sources of heterogeneity in these samples are often underappreciated and could be a common source of variable expression.

One of the fundamental causes of signal heterogeneity is variation in the cellular composition of tissue. As we know, tissue is not a monomorphic entity with -omic expression from a single source. Rather, tissues are made of numerous cells, which have their own expression patterns, and the combination of numerous unique cell types within a tissue is what constitutes the tissue signal. Many studies have established causes of signal variability, and some of the best-characterized causes are disease status, tissue quality, and underlying genetic variability.[5–7]

It is challenging to identify the exact composition of a tissue from its cellular components, and this deconvolution of a tissue signal is rarely performed. However, even among "control" tissues, presumed to be highly similar, there can be substantial heterogeneity. For example, as tissues age, their fibroblast content could increase as fibrosis increases, or they could undergo subclinical atrophy and reduce their epithelial cell component.[8–10] Even the extent of resident chronic inflammatory cells can change, altering an inflammatory signal. Between normal and diseased tissues, particularly malignancies, the overall cellular composition of the tissue can vary even more widely and exert a heavy influence on the interpretation or misinterpretation of the tissue-level gene expression results.[7,11]

Measurements of tissue gene expression are limited by the unknown contribution of cellular composition and other biologic variables to the reported measurements. For example, coexpression of two genes across tissue samples could represent a mechanistic relationship between these two genes, or it could simply mean that these genes are both expressed in the same cell type, and the coexpression is driven by changes in the proportion of that cell type in the tissue samples.

Additionally, substantial spatial heterogeneity can arise from compositional differences throughout a single tissue. A standard RNA-sequencing (RNA-seq) analysis typically uses only a small region of the tissue to obtain RNA, which makes it extremely difficult to ascertain whether the composition of the sample is representative of the entire tissue. Heterogeneity in cellular composition can masquerade as alterations in cellular function (false positives) or obscure true changes in cellular function (false negatives).

Currently, mixture modeling is the primary method of identifying the contribution of individual cell types to the tissue-level gene expression profile. These methods

[1]Department of Biostatistics and Computational Biology, University of Rochester Medical Center, Rochester NY, 14642, USA; [2]Department of Pathology, School of Medicine, Johns Hopkins University, Baltimore MD, 21287, USA
*Correspondence: mccallm@gmail.com (M.N.M.), mhalush1@jhmi.edu (M.K.H.)
http://dx.doi.org/10.1016/j.ajhg.2016.07.007.

model the observed expression of gene i in tissue sample j (Y_{ij}) as a linear combination of cell-type-specific expression (X_{ijk}) of each of the K cell types that make up the tissue:

$$Y_{ij} = \sum_{k=1}^{K} \rho_{jk} X_{ijk}.$$

In this framework, even if we assume that we know the number of mixture components (K), the system of equations is under-determined, meaning that there are fewer equations than parameters. Statistical methods to address this challenge require different types of auxiliary data and/or making various biologically implausible assumptions about the unobserved parameters.[12–16] A different class of mixture-modeling methods require cell-type-specific transcriptomic data.[17] However, data from purified cell types have their own challenges, including residual compositional heterogeneity, the introduction of technical artifacts, or RNA-degradation-related alterations in the transcriptome.[18,19] In the absence of cell expression data, or confidence in ancillary mixture components, a different approach is necessary.

The Genotype-Tissue Expression (GTEx) Project is a resource project designed to develop expression quantitative trait loci on the basis of data from 900 human donors over 53 sampled sites.[20] The GTEx Project obtains samples from rapid harvesting of autopsies of relatively healthy individuals and some materials from "normal" surgical specimens. At an interim point in the project, the GTEx Consortium has made available gene expression data from multiple tissues. This includes lung RNA-seq data from 133 well-characterized individuals. In the sampling process, the GTEx Project also obtains adjacent histologic tissues that can be used for evaluating the presence of histopathology or heterogeneity between samples. These GTEx data are an opportunity to identify new, robust methods of understanding tissue heterogeneity.

The GTEx program has already yielded several interesting findings related to the transcriptome. The GTEx Project compared the transcriptome across tissues and individuals, and a principal-component analysis (PCA) nicely separated the tissues on the basis of characteristic gene signatures.[21] In this study, the contribution of individual phenotypes was considered minor. GTEx data also uncovered that the transcriptome depends on age-related gene expression changes.[22]

We have been interested in understanding cellular expression[11,23,24] and the cellular composition of tissues[25] and determining how these individual components contribute to the complexity of a tissue-level signal.[26–28] Within the GTEx lung RNA-seq dataset, we expected to find variation related to technical factors and hypothesized that we also would be able to identify samples that had more chronic inflammation and/or fibrosis as a feature of inflammatory cell and mesenchymal cell expression. The large and robust nature of the GTEx datasets allowed this attempt to quantify, in specific biological terms, the sources of variation in a tissue transcriptome as a function of cell composition. We set out to use robust statistical methods and a deep understanding of lung tissue composition to find these and other patterns of heterogeneity within the expression data.

Material and Methods

Acquisition of RNA-Seq Datasets and Lung H&E Images

Gene read counts (version 4: GTEx_Analysis_V4_RNA-seq_RNA-SeQCv1.1.8_gene_reads.gct.gz) were downloaded from the GTEx Portal. From the full dataset, we selected the 133 lung tissue samples for further analysis. Sample-level variables (autolysis score, ischemic time, RNA integrity number, RNA-seq date, and RNA extraction date) were obtained from the GTEx Portal (version 4: GTEx_Data_V4_Annotations_SampleAttributesDSlung.txt). Digital lung H&E images, in .svs format, were also obtained from GTEx for 120 samples. Subject-level variables (age, gender, and death classification based on the 4-point Hardy Scale) were obtained from the GTEx Portal (version 4: GTEx_Data_V4_Annotations_SubjectPhenotypes_DD.xlsx).

There were five lung H&E images without corresponding RNA-seq data and 19 RNA-seq samples without corresponding H&E images. There were three RNA-seq samples without a death classification based on the 4-point Hardy Scale.

Ethics Statement

The GTEx Project involves recruitment, institutional-review-board approval, and consent issues for deceased donors and their families. Although the collection of tissues from deceased donors is not legally classified as human subjects research under 45 CFR 46 in the Code of Federal Regulations, written or recorded verbal authorization was obtained from the deceased donor's next of kin. All data included in this manuscript were previously published and made available through the GTEx Portal.

Processing of RNA-Seq Datasets

The gene read counts were processed with the R-Bioconductor software suite (R version 3.2.2; Bioconductor version 3.2). Most subsequent analyses were performed with count data. To identify high-variance genes and perform exploratory data analyses, we normalized counts to account for differences in library size by using size factors estimated via the median-ratio method and subsequently transformed the counts by using a variance-stabilizing transformation based on the dispersion-mean relationship (implemented in DESeq2 version 1.10.1).[29] This yielded normalized and transformed counts that were approximately homoskedastic (constant variance across the range of mean values). From the 55,993 genes represented in the GTEx data, we selected those with a minimum level of average expression (mean of normalized and transformed counts > 5). This threshold was chosen because it appeared to separate background noise from signal (Figure S1). A total of 18,703 genes passed this filter, which seems to be a reasonable estimate of the lung-specific transcriptome. All subsequent analyses, including the high-variance method, DEXUS, SIBER, and PCA, focused on these genes.

PCA

PCA was performed on the normalized, transformed, and filtered counts with the prcomp function from the stats R package. PCA was performed on the centered expression values.

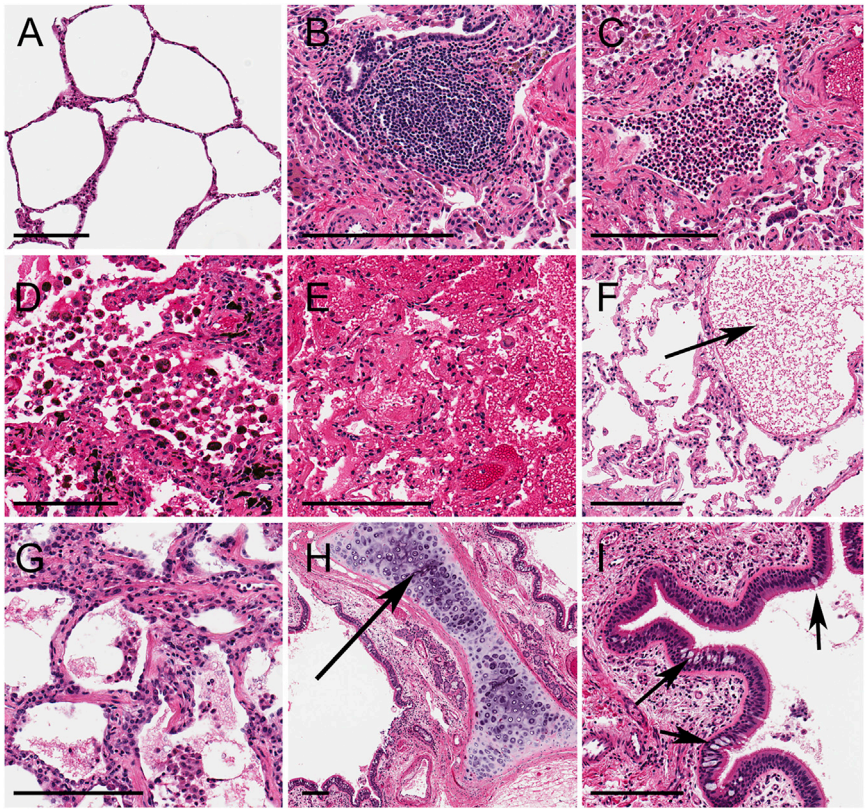


Figure 1. Representative Images for Each Histologic and Histopathologic Entity Seen in the Lung
(A) Normal alveolar lung tissue with thin lung parenchyma lined by type I and type II pneumocytes and clear airspaces.
(B) A lymphocyte cluster.
(C) A collection of neutrophils in a denuded bronchus as a result of pneumonia.
(D) Extensive pulmonary alveolar macrophages with hemosiderin pigment filling the alveolar space.
(E) Extensive hemorrhage into the alveolar airspace.
(F) Fibrin (arrow), associated with edema, collecting in an airspace.
(G) Fibrosis causing thickening of the alveolar parenchymal wall.
(H) Cartilage (arrow), filled with chondrocytes, between two bronchi.
(I) Bronchial epithelium showing both ciliated respiratory epithelium and mucin-secreting goblet cells (arrows).
The stain used is H&E. Scale bars represent 200 μm.

Identification of Highly Variable Genes

We considered three methods to identify genes that varied substantially across the 133 lung tissue samples. The first and most direct method was to identify genes with high variance across samples. We used a threshold of >4 for the variance of normalized and transformed counts. The second method, DEXUS, models RNA-seq counts as a mixture of negative binomial distributions where each mixture component corresponds to an unknown condition. Bimodal genes were those with an informative/non-informative threshold > 0.1, the threshold used in Klambauer et al.[30] The third method, SIBER, also identifies bimodally expressed genes by fitting a mixture model to RNA-seq count data; however, it recommends a mixture of log-normal distributions. Bimodal genes were those with a bimodality index > 1.2, the threshold used in Tong et al.[31]

Identification of Correlated Gene Clusters

We subsequently assessed each group of genes identified by one of the methods above for patterns of between-gene correlation by using the normalized and transformed counts. Specifically, we hierarchically clustered each group by using average linkage and a distance of 1 − Kendall's τ, the Kendall rank-correlation coefficient. We cut the resulting dendrogram at a specific height to yield clusters of genes with an average correlation of at least τ_{crit}, corresponding to a conservative Bonferroni-corrected critical value for statistical significance. These critical values ranged from 0.3 to 0.35 depending on the multiple-testing correction.

Calculation of Average Standardized Gene Expression

To summarize the expression of a cluster of highly correlated genes, we centered and scaled the expression of each gene and then averaged it across genes within the cluster. Specifically, we centered the normalized and transformed gene expression values by subtracting the mean expression and then scaled them by dividing by the median absolute deviation. We then averaged these values across all genes in the cluster to produce sample-specific average expression profiles for a gene cluster.

Analysis of Subject- and Sample-Level Variables

To assess whether variability in gene expression might be due to technical differences between samples or subject-level biological differences, we quantified the pairwise dependence between sample-level variables (e.g., ischemic time), subject-level variables (e.g., gender), and average standardized gene expression by using the normalized and transformed counts.

Analysis of Lung Images for General Histologic and Histopathologic Characteristics

The 120 digitally scanned lung images from the GTEx Consortium were analyzed with the Aperio ImageScope (Leica Biosystems). We (M.K.H.) characterized all 120 images for the extent of fibrosis, the presence of inflammatory cells (neutrophils, pulmonary alveolar macrophages [PAMs], and lymphocytes), hemorrhage (red blood cells within the alveolar spaces), cartilage (seen in bronchi), and excessive fibrin (proteinaceous material) in the airspaces (Figure 1). The extent of fibrosis, the amount of PAMs, the amount of lymphocytes, and the extent of hemorrhage were scored on a 0–3 scale. The presence of neutrophils, fibrin, and cartilage was scored as present or absent.

Analysis of Lung Images for the Percentage of Bronchial Epithelium

A second analysis of the images was to determine the percentage of bronchial epithelium in the slides. Bronchial epithelium is a lung substructure that contains ciliated epithelial cells and mucous-secreting cells of the terminal bronchiole or segmental bronchus.

To determine the percentage of bronchial epithelium, we divided the summed area of all bronchial epithelium (digitally circled) by the total area of lung tissue on the slide (Scanscope, Leica Biosystems).[27]

Analysis of Pneumocyte Ratios

We (P.B.I.) determined type II pneumocyte hyperplasia on all images. Histologically, type I pneumocytes are squamoid in shape and line the alveoli. Type II pneumocytes are slightly larger, cuboidal cells that also line the alveoli (Figure S2).[32] Type II pneumocytes undergo hyperplasia in the setting of ventilator-associated lung injury (VALI) and other injurious conditions to the lung. Lung images were scored as 0 (normal type II pneumocytes), 1 (focal type II pneumocyte hyperplasia), 2 (focally marked or diffusely variable type II pneumocyte hyperplasia), or 3 (diffusely marked type II pneumocyte hyperplasia). All analyses of gene expression in lung tissue were blinded to the pathologists (P.B.I. and M.K.H.).

Online Searchable Databases

To uncover and confirm the meaning of the clustered genes, we used the following tools: the Gene Ontology (GO) Consortium,[33] GeneCards,[34] the Human Protein Atlas,[35] the Molecular Signatures Database,[36] and STRING version 10.0.[37]

Statistical Testing

Association testing between the expression of individual genes and phenotypic or pathologic variables was performed with the DESeq2 R package (version 1.10.1), which takes count data as input. To test for an association between binary variables and gene expression, we performed a Wald test. To test for an association between categorical variables and gene expression, we performed likelihood-ratio tests of change in deviance between a full model including the given variable and a reduced model without it. This latter model could be an intercept-only model or could include other variables that were adjusted for in the full model. This allowed us to compare the goodness of fit of these two models. In other words, it tested whether including the given categorical variable would significantly improve the model fit. Both the Wald and likelihood-ratio tests were implemented in DESeq2.

We used Kendall's τ to assess the correlation between gene expression and subject- and sample-level variables, as well as the percentage of bronchial epithelium or type II pneumocyte hyperplasia. All reported p values were corrected for multiple testing via Holm's method.[38] These are reported in the text in the following format: ($\tau = -0.33$; $p = 1.1 \times 10^{-5}$).

Results

Identification and Clustering of High-Variance Genes

We began our analysis by focusing on extreme variance within the gene expression data and worked backward to establish what technical, biological, phenotypic, and/or sampling-related variables could contribute toward signal heterogeneity. To examine this heterogeneity in the lung, we selected the top 175 high-variance genes (Figure S3 and Table S1) that passed our thresholds for overall expression and residual variance (see Material and Methods). Among these 175 genes, we identified five clear groups of highly correlated genes, subsequently referred to as high-variance clusters A–E, which contained 33, 70, 14, 9, and 9 genes, respectively (Figure S4 and Table S2). An additional 40 genes had high variance but did not robustly cluster into specific sets (Figure S5).

Identification and Clustering of Bimodal Genes

An alternative approach to identifying genes that vary across samples is to select genes with bimodal expression. Two methods of identifying such genes, DEXUS and SIBER, were applied to the GTEx lung data (see Material and Methods).[30,31] DEXUS identified 1,255 bimodal genes, and SIBER identified 1,101 bimodal genes; 494 genes were found in common between the two methods. The majority of high-variance genes were also identified as bimodal, and only 5 of 175 high-variance genes were not identified by either method (Figure S6).

These bimodal methods identified many additional gene clusters. Clustering of the 1,255 genes selected by DEXUS identified 25 clusters of highly correlated genes (Figure S7 and Table S3). Clustering of the 1,101 genes selected by SIBER identified 24 clusters of highly correlated genes (Figure S8 and Table S4). However, many of these gene clusters appeared to be weakly correlated with each other (Figures S7 and S8), suggesting the possibility of a non-biological effect. In the subsequent analyses, the SIBER gene clusters added no information not captured by DEXUS gene clusters; therefore, for the sake of brevity, we will focus on the DEXUS results.

PCA

PCA, a complimentary approach to identifying potential sources of variation, was also applied to these data (see Material and Methods). The top ten principal components accounted for over 50% of the variation in the data. The first principal component (accounting for approximately 15% of the variance) was strongly negatively correlated with high-variance cluster A. The second principal component (accounting for just over 9% of the variance) was strongly correlated with high-variance cluster B. The third principal component (accounting for just under 9% of the variance) was correlated with high-variance cluster E. The ninth principal component (accounting for just over 2% of the variance) was positively correlated with high-variance cluster D and negatively correlated with high-variance cluster C. Other top principal components were not clearly associated with high-variance clusters (Figure S9).

Many of the top principal components were also correlated with DEXUS clusters. The first principal component was strongly negatively correlated with DEXUS clusters H, I, and V and strongly positively correlated with DEXUS clusters C, F, M, and P. Additionally, the first principal component showed moderate or weak correlation with several other DEXUS clusters. Principal components 2–4 were each correlated with one to two DEXUS clusters, whereas other principal components showed generally weaker correlation with DEXUS clusters (Figure S10).

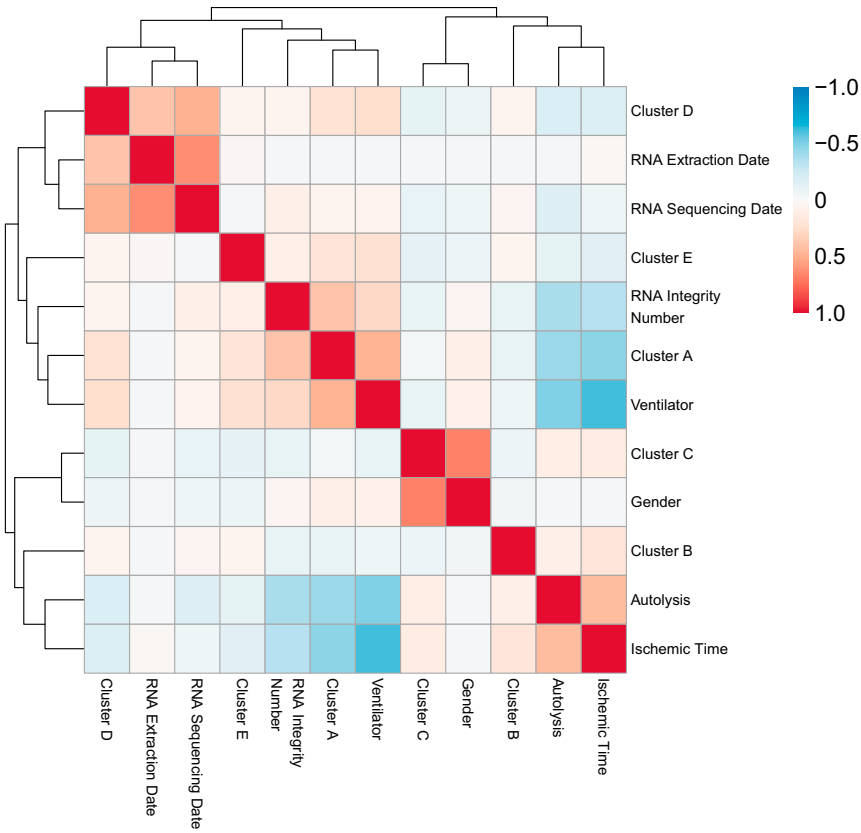


Figure 2. Correlation between Subject- and Sample-Level Variables and Gene Expression Clusters

We used Kendall's τ, the Kendall rank-correlation coefficient, to assess the pairwise dependence between sample- and subject-level variables, as well as the average standardized gene expression across 133 lung tissue samples. Cluster A was strongly positively correlated with ventilation and RNA integrity number and strongly negative correlated with autolysis score and ischemic time. Cluster C was strongly correlated with the gender of the individuals. Cluster D was strongly correlated with RNA-seq date and RNA extraction date. Clusters B and E were not strongly correlated with any other variables.

Analysis of Sample- and Subject-Level Variables

We began by investigating whether the patterns of gene expression covariation in high-variance clusters A–E could be explained by phenotypic differences between individuals (gender and death classification) or technical differences between samples (autolysis score, ischemic time, RNA integrity number, RNA-seq date, and RNA extraction date). We found clear patterns of pairwise dependence among sample-level variables, subject-level variables, and average standardized gene expression within each of the five gene clusters (Figure 2).

Specifically, average gene expression in high-variance cluster A showed significant positive correlation with ventilation prior to death ($\tau = 0.48$; $p = 4.7 \times 10^{-10}$) and RNA integrity number ($\tau = 0.40$; $p = 2.8 \times 10^{-10}$). High-variance cluster A was also significantly negatively correlated with autolysis score ($\tau = -0.41$; $p = 4.8 \times 10^{-8}$) and ischemic time ($\tau = -0.47$; $p = 1.8 \times 10^{-11}$). Not surprisingly, ischemic time was also strongly positively correlated with autolysis score ($\tau = 0.44$; $p = 1.9 \times 10^{-7}$) and negatively correlated with RNA integrity number ($\tau = -0.33$; $p = 1.1 \times 10^{-5}$) and ventilation prior to death ($\tau = -0.61$; $p = 5.5 \times 10^{-13}$). Taken together, these results suggest a possible latent cause of these correlations—individuals on a ventilator prior to death are likely to be in a hospital, resulting in more rapid tissue harvesting post-mortem. Although gene expression in high-variance cluster A could represent a purely technical effect as a result of differences in post-mortem interval, genes within the signature suggest another potential etiology, evaluated below.

High-variance cluster B was not strongly correlated with any of the sample- or subject-level variables examined here ($\tau < 0.25$). High-variance cluster C was strongly correlated with the gender of the subject ($\tau = 0.68$; $p = 6.1 \times 10^{-20}$). High-variance cluster D was strongly correlated with RNA extraction date ($\tau = 0.39$; $p = 4.5 \times 10^{-10}$) and RNA-seq date ($\tau = 0.49$; $p = 1.5 \times 10^{-14}$), which were strongly correlated with each other ($\tau = 0.64$; $p = 2.5 \times 10^{-24}$). This most likely represents a batch effect. High-variance cluster E was not strongly correlated with any of the sample- or subject-level variables examined here ($\tau < 0.25$).

Correlation between Principal Components and Technical Variables

We performed a similar analysis to assess the correlation between the top ten principal components and each of the technical variables (Figure S11). The first principal component was negatively correlated with RNA integrity number ($\tau = -0.40$; $p = 3.8 \times 10^{-10}$) and ventilation prior to death ($\tau = -0.53$; $p = 4.5 \times 10^{-12}$) and positively correlated with autolysis score ($\tau = 0.43$; $p = 1.5 \times 10^{-8}$) and ischemic time ($\tau = 0.53$; $p = 20 \times 10^{-14}$). The ninth principal component was negatively correlated with gender ($\tau = -0.43$; $p = 4.6 \times 10^{-8}$). None of the top ten principal components were correlated with RNA-seq date or RNA extraction date.

Correlation between DEXUS Gene Clusters and Technical Variables

Finally, we examined the relationship between DEXUS gene clusters and technical variables (Figure S12). DEXUS clusters C, D, F, L, M, O, P, and S were generally positively correlated with ischemic time and autolysis score and negatively correlated with RNA integrity number and

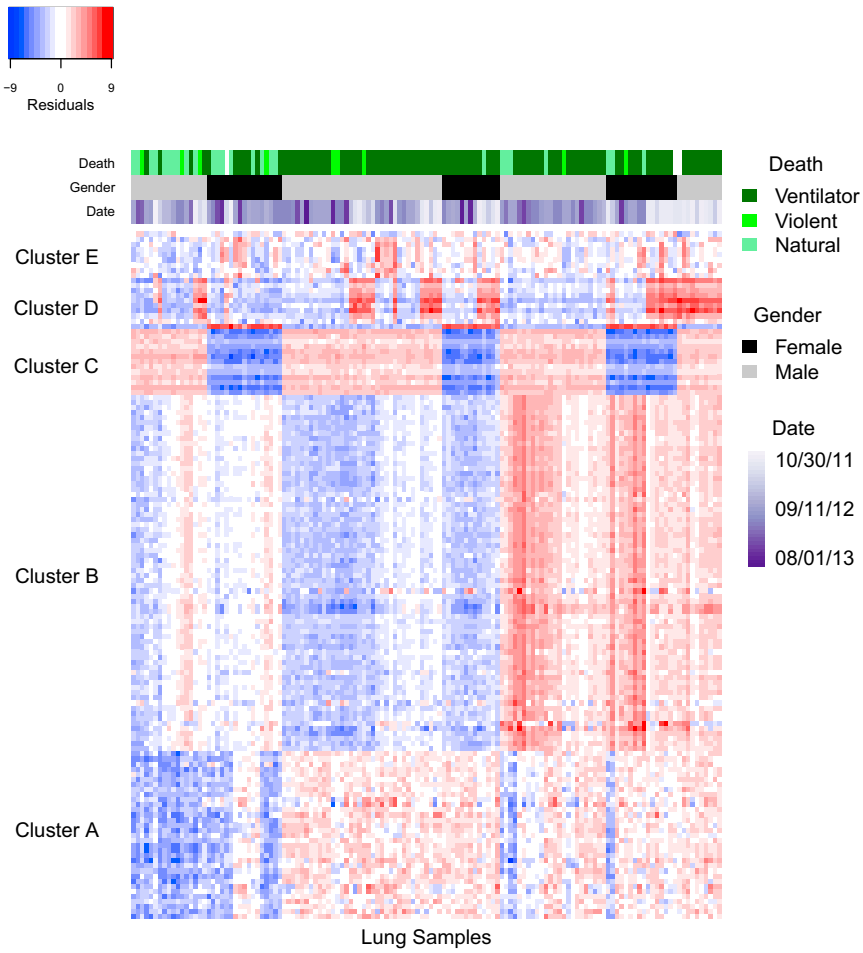


Figure 3. Residual Expression of High-Variance Genes across Lung Tissue Samples

We used the gene-wise residual expression for the 125 genes composing five primary gene clusters across 133 lung tissue samples to cluster samples. These clusters consisted of genes specific to type II pneumocytes (A), the bronchial epithelium (B), the sex of the individual (C), mitochondrial pseudogenes (D), and acute-phase reactants (E). Cluster A was associated with ventilator use. Cluster C was associated with gender. Cluster D was associated with RNA-seq date, indicating a batch effect.

ventilation prior to death; in contrast, DEXUS clusters H and I were generally positively correlated with RNA integrity number and ventilation prior to death and negatively correlated with ischemic time and autolysis score. DEXUS cluster A was correlated with RNA-seq date and RNA extraction date, and DEXUS cluster Y was correlated with gender.

In-Depth Examination of High-Variance Clusters

Gene expression in the five high-variance clusters differed substantially across lung tissue samples (Figure 3). Further examination of these genes and phenotypic data revealed technical, compositional, and biological explanations for the observed patterns of expression. Although high-variance cluster A was already associated with time to harvesting and ventilator-use-related variables, we noted that this group of 33 genes included surfactant genes (*SFTPA1* [MIM: 178630], *SFTPA2* [MIM: 178642], and *SFTPC* [MIM: 178620]) and phospholipase genes (*PLA2G3* [MIM: 611651] and *PLA2G4F*) (Table S2). Surfactant genes are expressed exclusively in type II pneumocytes, which proliferate in VALI.[39,40] Additionally, phospholipases hydrolyze surfactants in acute respiratory-distress syndrome, potentially accounting for their association in this cluster.[41]

Building on the ventilator-associated relationship of this cluster, we found a strong marginal association between elevated levels of five surfactant genes and death on a ventilator (Table 1). Even after adjustment for ischemic time, RNA integrity number, autolysis score, and sequencing date, these surfactant genes remained strongly associated with death on a ventilator (Table 1). Of the 95 subjects who died on a ventilator, 66 were in the lowest ischemic time group (0–3 hr). In contrast, only one subject who did not die on a ventilator was in the lowest ischemic time group. To reduce the confounding between ischemic time and death on a ventilator, we examined the relationship between the surfactant genes and death on a ventilator after removing the group with the lowest ischemic time. Despite a substantial reduction in sample size, all five surfactant genes remained associated with death on a ventilator (Table 1). Finally, we examined the expression of each of these five surfactant genes after stratifying by ventilator and ischemic time (Figure S13). Ventilation was associated with higher median expression for all five surfactant genes in all four ischemic time groups with at least ten samples. Surfactant and phospholipase genes were also among the top drivers of the first principal component. Specifically, *SFTPA1*, *SFTPA2*, and *PLA2G4F* were among the top five genes contributing to principal component 1. Additionally, *SFTPB* (MIM: 178640), *SFTPC*, *SFTPD* (MIM: 178635), *PLA2G1B* (MIM: 172410), and *PLA2G3* (MIM: 611651) were among the top 100 genes contributing to principal component 1.

High-variance cluster B comprised 70 genes, including dynein genes (*DNAH9* [MIM: 603330], *DNAH12* [MIM: 603340], and *DNAH3* [MIM: 603334]) and mucin genes (*MUC5B* [MIM: 600770] and *MUC16* [MIM: 606154]) (Table S2). In lung tissue, these genes are exclusive to the microvilli of respiratory epithelium and goblet cells. A GO analysis of this cluster identified cilium movement and axoneme assembly as enriched biological clusters.

Table 1. Association between SFTP Genes and Death on a Ventilator

Gene	Marginal Effect		Adjusted Effect		Ischemic Time > 4 hr	
	Expression Change (-fold)	Adjusted p Value	Expression Change (-fold)	Adjusted p Value	Expression Change (-fold)	Adjusted p Value
SFTPA1	3.29	1.58×10^{-11}	1.58	1.09×10^{-13}	2.05	6.40×10^{-4}
SFTPA2	3.15	2.89×10^{-13}	1.51	1.43×10^{-11}	1.87	1.92×10^{-3}
SFTPB	2.77	1.35×10^{-12}	1.46	7.74×10^{-11}	1.85	1.53×10^{-3}
SFTPC	2.35	9.02×10^{-7}	1.23	1.79×10^{-4}	1.62	1.14×10^{-2}
SFTPD	3.05	1.36×10^{-13}	1.50	9.41×10^{-12}	1.93	1.20×10^{-3}

Analysis of high-variance cluster B genes in the Human Protein Atlas showed many genes to be distinctly expressed by either respiratory epithelium or goblet cells (Figure S14).[35] This gene signature indicates heterogeneous amounts of bronchus or terminal bronchiole within samples as a result of the sampling location within the lung. Dynein and mucin genes were also among the top drivers of the second principal component. Specifically, *MUC4* (MIM: 158372), *MUC5B*, *MUC16*, *DNAH2* (MIM: 603333), *DNAH3*, *DNAH6* (MIM: 603336), *DNAH7* (MIM: 610061), *DNAH9*, *DNAH10* (MIM: 605884), and *DNAH12* were among the top 150 genes contributing to principal component 2.

High-variance cluster C comprised 13 Y chromosome genes and *XIST* (MIM: 314670), a major effector of the X inactivation process. This cluster perfectly matched each subject's gender (Figure 3 and Table S2). Cluster D comprised primarily mitochondrial pseudogenes and was strongly associated with the date on which the samples were sequenced (Figure 3 and Table S2). Cluster E contained serum amyloid genes (*SAA1* [MIM: 104750], *SAA2* [MIM: 104751], and *SAA2–SAA4*) and other genes that can act as acute phase reactants (Table S2).

In-Depth Examination of DEXUS Clusters

The DEXUS analysis identified all of the high-variance clusters. Additionally, it identified clusters related to fibrosis and fibroblasts (DEXUS clusters D and G), cell division or mitosis (DEXUS cluster J), red blood cells (DEXUS cluster Q), neutrophils (DEXUS cluster X), lymphocytes (DEXUS cluster C and T), inflammatory responses (DEXUS clusters L, O, R, and V), cholesterol biosynthesis (DEXUS cluster H), nucleic acid metabolic processes (DEXUS cluster I), and Jun-Fos signaling (DEXUS cluster K). We were unable to assign clear biologic function or cell-type localization to the other DEXUS clusters (Table S3). Altogether, the inclusion of a greater number of variable genes through the use of DEXUS did increase the number of clusters that could be associated with specific cell types or cell processes and reaffirmed the five high-variance clusters.

Image Analysis for Type II Pneumocyte Ratio

We hypothesized that the gene signature showing high variance in surfactant genes and correlation with clinical ventilator use (high-variance cluster A) would indicate an altered ratio of type II to type I pneumocytes, specifically a proliferation of type II pneumocytes. To address this, we evaluated 114 lung images for the extent of type II pneumocyte hyperplasia on a 0–3 scale (Figure S2) in a blinded fashion and compared these scores with the high-variance cluster A gene expression. We found a significant correlation between type II pneumocyte hyperplasia and the average standardized expression of genes in high-variance cluster A ($\tau = 0.26$; $p = 2.6 \times 10^{-4}$) (Figure 4A). Marginal analysis of the association between individual genes and type II pneumocyte hyperplasia showed a moderate correlation ($\tau = 0.18$–0.32). The presence of type II pneumocyte hyperplasia also correlated with the clinical scenario of ventilator use at the time of death ($\tau = 0.35$; $p = 1.1 \times 10^{-4}$).

Image Analysis for Bronchial Epithelium

The genes in high-variance cluster B implicated bronchial epithelium as a major source of gene expression variability in the lung samples. Therefore, we examined the presence of terminal bronchi in adjacent histologic sections to determine whether these images were able to predict the presence of bronchial epithelium in the adjacent RNA-harvested material. We compared the percentage of bronchial epithelium to the gene expression cluster for the 114 lung samples found in both datasets. The range of the percentage of bronchial epithelium to total lung parenchyma across the 114 images was 0%–6.82% (median 0.66%) (Table S5). The association between the average standardized gene expression in high-variance cluster B and the percentage of bronchial epithelium was minimal and not statistically significantly different from 0 (Figure 4B; $\tau = 0.09$; $p = 0.15$). Marginal analysis of the association between individual genes and the percentage of bronchial epithelium showed a consistently weak relationship ($\tau =$ 0.03–0.14). Thus, we were unable to associate the histologic presence of bronchial epithelium with the gene signature.

Gene Sets for Other Histopathologic Findings: Presence or Absence

After our approach to decipher the expression data by identifying high-variance or bimodal gene expression, we

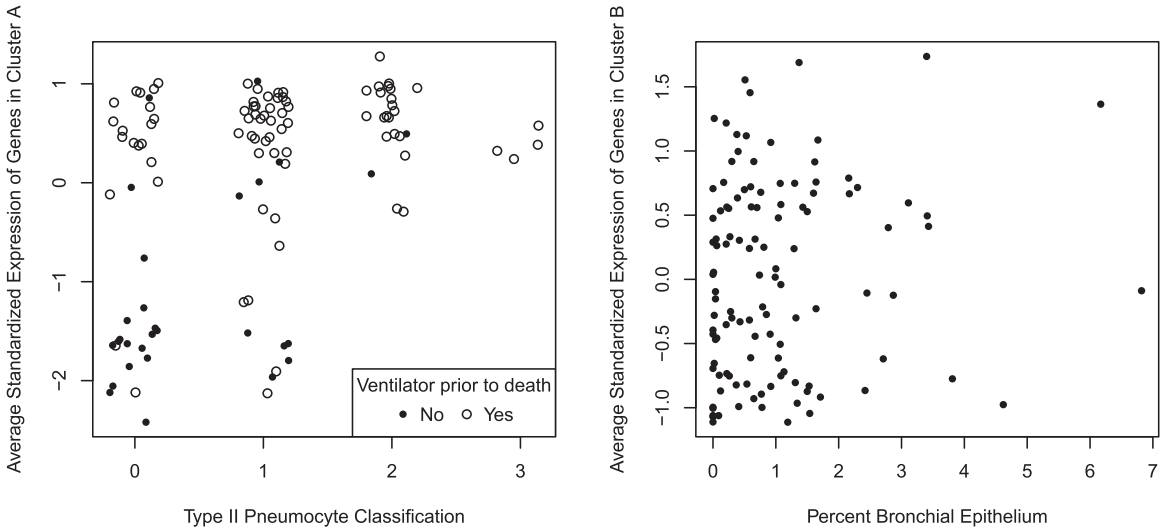


Figure 4. Association between Histology and Gene Expression
Association between type II pneumocyte hyperplasia and the average standardized expression of genes in cluster A and the percentage of bronchial epithelium and average standardized expression of genes in cluster B in 114 matched lung tissue samples. Histologic changes were determined from tissue sections adjacent to the section used for RNA-seq. To compute the average gene expression in both clusters, we first standardized the expression of each gene to a comparable scale and then averaged it across genes. The vast majority of samples with high average expression in cluster A but a type II pneumocyte classification of 0 or 1 were on a ventilator prior to death.

then inverted our investigation to ascertain whether certain histologic or histopathologic findings could be associated with differences in gene expression. To explore potential additional sources of complexity in the tissue signal, we characterized histologic and histopathologic findings in each of the lung samples on the basis of the accompanying digital images. All findings were scored on the basis of presence and severity (Figure 1 and Tables 2, 3, and S5). After adjusting for ischemic time, autolysis score, RNA integrity number, and RNA-seq date, we tested the association between gene expression and each variable to identify genes significantly associated with neutrophils, fibrin, or cartilage (see Material and Methods).

We identified 1,011 genes associated with the presence of neutrophils (adjusted p values < 0.01). These included genes encoding matrix metallopeptidases (*MMP8* [MIM: 120355], *MMP9* [MIM: 120361], *MMP14* [MIM: 600754], and *MMP25* [MIM: 608482]), S100 proteins (*S100A8* [MIM: 123885], *S100A9* [MIM: 123886], *S100A11* [MIM: 603114], and *S100A12* [MIM: 603112]), chemokine ligands (*CCL19* [MIM: 602227] and *CCL22* [MIM: 602957]), and serum amyloid proteins (*SAA1*, *SAA2*, *SAA2–SAA4*) (Table S6). Six of these genes were part of high-variance cluster E, and four others were selected as high variance but did not cluster strongly with other genes. 50 of these genes were selected by DEXUS; however, only one of these was part of DEXUS cluster X (neutrophil genes). We performed a GO analysis on the top 63 genes with $\log_2$ expression changes between 1.0-fold and 2.0-fold.[33] The top biological processes for this gene set were neutrophil chemotaxis, acute-phase response, and neutrophil migration, which was a strong match to the observed histology.

We identified 74 genes that were significantly overexpressed in the presence of increased fibrin (Table S7). Of these genes, four were part of high-variance cluster E. Also, 21 of these genes were selected by DEXUS; however, 14 did not belong to any cluster, and the other seven were spread across four DEXUS clusters with no clear relationship with fibrin. A GO analysis of these genes for biological processes showed enrichment of glucose transmembrane transport. No genes associated with the presence of cartilage were identified. However, the presence of cartilage was associated with higher levels of bronchial epithelium (Wilcoxon rank-sum test p value $= 1.5 \times 10^{-6}$), consistent with a shared localization within the lung.

Gene Sets for Other Histopathologic Findings: Categorical

For the histologic and histopathologic variables that were categorical on the basis of severity (lymphocytes, interstitial fibrosis, pulmonary alveolar macrophages, and hemorrhage), we performed an omnibus test to compare a model including the given variable, as well as ischemic time, autolysis score, RNA integrity number, and RNA-seq date, and a model with only the technical variables (see Material and Methods).

We identified 115 genes that were significantly associated with the extent of lymphocyte infiltration (Table S8). This list included four immunoglobulin genes and 11 cluster of designation (CD) lymphocyte genes. None of these genes were identified as high variance; however, 17 were identified as bimodal by DEXUS. Of these, seven were part of DEXUS cluster T, and five were part of DEXUS cluster C, both previously annotated as lymphocyte clusters. Gene-set analysis identified pathways

Table 2. Presence of Histologic and Histopathologic Findings and Associated Gene Expression

Histologic and Histopathologic Findings	Extent		Associated Genes
	Absent	Present	
Neutrophils	108	12	1,011
Fibrin	109	11	74
Cartilage	107	13	0

related to T cell and B cell activities, consistent with the histology.

We identified 27 genes that were significantly associated with the extent of interstitial fibrosis, and 25 of these showed higher expression in the most severe cases than in the least severe cases (Table S9). Four of these genes were part of high-variance cluster B, 13 were selected by DEXUS, and eight were part of DEXUS cluster B. Both of these gene clusters represent a bronchial epithelium signature. Analysis in MSigDB identified a matrisome signal for the gene set.

The extent of hemorrhage was associated with an elevation of 27 genes (Table S10). Analysis of the genes identified no pathways that associated with red blood cell activities, and none of these genes were present in DEXUS cluster Q (red blood cell associated).

Only seven genes were associated with an increase in pulmonary alveolar macrophages (Table S11). None of these genes were identified as high variance, and only one of these was identified by DEXUS. MSigDB matched these seven to sets of genes modulated in monocytes.

In summary, to compare estimates of cellular composition from the lung images with gene expression from RNA-seq, we scored seven histologic and histopathologic aspects of 114 lung tissue samples and linked these with changes in gene expression (Tables 2 and 3). The presence of neutrophils and the extent of lymphocytes, pulmonary alveolar macrophages, and fibrosis correlated well with the observed gene expression alterations, but no clear gene expression alterations were associated with the presence or severity of hemorrhage, fibrin, or cartilage.

Discussion

We performed an extensive analysis of 133 lung samples from the GTEx Project as a means of understanding the roles of tissue complexity and technical variation in gene expression heterogeneity. Surprisingly, we discovered that two of the strongest signals of expression variance in the samples were the result of sampling location (bronchial epithelium) and an end-of-life treatment modality (ventilator usage). Bronchial epithelium is generally increased in the medial lung, indicating variable sampling localization or random chance in capturing bronchial epithelium within the small sample used for RNA generation. The surfactant signal associated with high-variance cluster A and ventilator usage appeared to be the result of VALI and type II pneumocyte hyperplasia because it was strongly correlated with both. High-variance cluster E implicated acute phase reactants, also implicating non-steady-state (homeostatic) expression alteration. Gender (high-variance cluster C) and sequencing date (high-variance cluster D) affected only a few high-variance and bimodal genes.

Additionally, a large source of technical variation, associated with prolonged ischemic time and an RNA degradation artifact, appears to represent the effect of the post-mortem interval at lung harvesting.[42] This effect was partially captured by the first principal component, resulted in diffuse moderate correlation between roughly half of the DEXUS clusters, and most likely explains the variation of some genes in high-variance cluster A. However, the more rapid harvesting of tissue from subjects on ventilators, due to hospitalization at the time of death, has led to confounding between ventilation- and post-mortem-interval-associated technical variables. When technical and biological variables are confounded, principal components often represent a complex combination of technical and biological effects. In this case, principal component 1 appears to represent a combination of the effects of type-II-pneumocyte-specific expression, ischemic time, autolysis, and RNA integrity.

We originally hypothesized that we would observe sample heterogeneity related to fibrosis, chronic inflammation, RNA-seq date, and RNA extraction date. With our high-variance threshold (>4), we were unable to uncover fibrosis-related gene signatures in the sample. However, with the broader DEXUS approach, we were able to find two small fibrosis-related gene clusters. Gene expression alterations due to inflammation, both acute (neutrophils) and chronic (lymphocytes), were detected but showed less variability across the samples than did the bronchial epithelium and type II pneumocyte signals. The presence of neutrophils is consistent with end-of-life pneumonia and, like the VALI-related data, is a gene expression alteration that does not reflect the tissue's steady state. The lymphocyte and neutrophil data also represent a limitation of our high-variance approach, necessitating the use of complimentary methods to capture these signals. Lastly, to our surprise, we did not find a large technical effect associated with RNA-seq date or RNA extraction date. Sequencing date is often used as a surrogate for numerous batch effects, but in these data it appears to have had a substantial effect on only a small number of genes.

Our findings establish that among the 175 most variable genes, the cellular composition of the tissues appears to be a major source of the observed variation. This is a significant departure from the general interpretation of tissue expression variation, which is that most variation is reported as the up- or downregulation of genes. It is worth noting that cellular composition, as a source of tissue heterogeneity, was not uncovered or reported in the first large, standard analysis of the GTEx dataset, which included

Table 3. Extent of Histologic and Histopathologic Findings and Associated Gene Expression

Histologic and Histopathologic Findings	Extent				Associated Genes
	Absent or Minimal	Mild	Moderate	Severe	
Lymphocytes	85	20	14	1	115
Interstitial fibrosis	83	14	15	8	27
Pulmonary alveolar macrophages	95	5	14	6	7
Hemorrhage	101	6	6	7	27

these lung samples. That analysis focused specifically on disease variables (diabetes, hypertension, etc.) and tissue quality (autolysis score, RNA integrity number, etc.) and did not take into account cellular heterogeneity of the tissue.[21] It is clear that more attention should be paid to tissue heterogeneity, even among "normal" and "control" tissues found in the GTEx Project.

Our discoveries concerning the complexity of sample heterogeneity have led us to generate some broad conclusions about evaluating -omic signals from human tissues. The first is that not all sources of gene expression variability can be detected from adjacent histology. For organs that vary in composition in a regional way, the expression signal is strongly affected by subsampling of tissue used for RNA experiments, which might not randomly capture a meaningful amount of any given interposed organ substructure (i.e., bronchial epithelium, larger blood vessel, nerve, adipose cluster, or lymphoid aggregate) even if that feature is in the general area according to the adjacent histology. That was particularly true for the lack of correlation between bronchial epithelium in the histology section and the matched gene signature. As seen in Figure S15, this is a general feature of spatial and sampling heterogeneity when a given feature is not homogeneously present. This will greatly affect mixture-modeling methods that rely on adjacent histology to estimate the mixture proportions. Where the adjacent histology appears useful is in the identification of a more homogeneous infiltrate of non-intrinsic cells. We were able to identify lymphocyte and neutrophil signals after easily characterizing their presence in lung tissue. However, the identification of type II pneumocyte hyperplasia is subtle and required a specialized pulmonary pathologist (P.B.I.) because this signal was poorly interpreted by a non-pulmonary pathologist (M.K.H.).

We were able to detect expression signatures for fibrosis but only through the more permissive DEXUS method. We were initially surprised that there was no chondrocyte (cartilage cell) signal similar to a bronchial epithelial signal. To best explain this finding, we hypothesize that the standard RNA isolation is unlikely to break down the surrounding cartilage to liberate chondrocyte RNA.

To uncover the sample heterogeneity, our primary approach was to utilize the highly expressed genes that showed substantial variability across samples. We then compared these genes to all available phenotypic and technical data. After this routine analysis failed to fully explain the observed variability, we added DEXUS and SIBER and then focused on highly correlated gene clusters and sought to identify the biological variable or cell type they represented. This process is not exhaustive—there will be undetected sources of variation as described above. However, by coupling data-driven discovery of residual variation with knowledge-based examination of the data, we were able to uncover both anticipated and unanticipated sources of variation in these data. Only by understanding the underlying causes of variation and incorporating them into subsequent analyses can one assess biologically relevant gene expression.

We suggest that an approach similar to that reported herein become standard for all tissue-level studies of sampled organ tissues to elucidate the full complexity of tissue composition. Specifically, one should carefully examine shared patterns of residual variation to identify groups of genes that differ consistently between samples. Often, expert knowledge is sufficient for identifying a probable cause of the observed variability, but otherwise one can rely on a number of bioinformatics tools to identify associated biologic pathways of the gene clusters.[33–37] Finally, changes in cellular composition can occur at both the sample level (as a result of the specific region within the tissue from which the RNA was obtained) and the tissue level (as a result of changes in the cellular composition of the tissue itself). Understanding which is relevant in a given situation is crucial to determining how one should approach the subsequent analysis of such data. For example, disease-related changes in cellular composition might not represent a biologically meaningful finding.

The statistical approach described here is related to work seeking to identify groups of genes affected by technical sources of variation, commonly referred to as batch effects.[43] Methods such as surrogate variable analysis[44,45] could be used to identify groups of genes associated with certain changes in the cellular composition of tissue samples. However, unlike batch effects, many of the sources of variation here are fundamentally biological in nature and often do not present as discrete batches of samples.

If cellular composition of a tissue is among the largest drivers of sample variability, much of our tissue-level expression studies must be reconsidered.[26–28] The largest changes in signals between diseased or cancer tissues

and normal tissue are most likely due to the presence of cell types that differ between samples and can mask subtler yet more biologically important cellular expression alterations.

There are certain meaningful limitations to this study. It is unknown how generalizable the findings that we propose in the lung are to other tissues. Some organs, such as the kidney, are much more complex with many more cell types and important substructures (e.g., glomeruli). Other organs, such as the liver, are less complex and generally more homogeneous. Thus, it will take additional studies in these tissues to make more global pronouncements regarding the role of sampling in tissue heterogeneity. We were fortunate to have a pulmonary pathologist evaluate the lung slides for the presence of type II pneumocytes. It would be challenging to evaluate this subtle histologic picture in most laboratory settings; however, using immunohistochemistry or immunofluorescence markers to differentiate type I and type II pneumocytes to evaluate hyperplasia would be possible. For example, TTF1, the protein product of *NKX2-1* (MIM: 600635), is found exclusively in type II pneumocytes.[46]

In conclusion, we have conducted a study to capture the complex and diverse sources of tissue heterogeneity, specifically histologic and histopathologic characterization in a "normal" lung tissue dataset. In this analysis, we uncovered unexpected extensive signal heterogeneity that was the result of sampling-related or treatment-related changes in the cellular composition of tissue. This cellular composition was the most important cause of unexpected tissue-level gene expression variation. We must all be cognizant of this underlying variability as a potential cause of differences in our tissue studies across all -omic platforms.

Supplemental Data

Supplemental Data include 15 figures and 11 tables and can be found with this article online at http://dx.doi.org/10.1016/j.ajhg.2016.07.007.

Acknowledgments

The authors are grateful to David Tabor for his assistance in obtaining the lung image and many helpful conversations. The authors are grateful to Jeff Struewing for his assistance and encouragement in this project. The authors thank Dan Arking, Alexander Baras, Dongwon Lee, and Aravinda Chakravarti for their helpful critiques of the manuscript. M.N.M. is supported by the NIH (HG006853), and M.K.H. is supported by the American Heart Association (13GRNT16420015).

Received: January 28, 2016
Accepted: July 8, 2016
Published: September 1, 2016

Web Resources

GeneCards, http://www.genecards.org/
Gene Ontology Consortium, http://geneontology.org/
GTEx Portal, http://www.gtexportal.org/
Molecular Signatures Database (MSigDB), http://software.broadinstitute.org/gsea/msigdb/index.jsp
OMIM, http://www.omim.org/
STRING 10.0, http://string-db.org/
The Human Protein Atlas: http://www.proteinatlas.org/

References

1. Heller, R.A., Schena, M., Chai, A., Shalon, D., Bedilion, T., Gilmore, J., Woolley, D.E., and Davis, R.W. (1997). Discovery and analysis of inflammatory disease-related genes using cDNA microarrays. Proc. Natl. Acad. Sci. USA *94*, 2150–2155.
2. Volinia, S., Calin, G.A., Liu, C.G., Ambs, S., Cimmino, A., Petrocca, F., Visone, R., Iorio, M., Roldo, C., Ferracin, M., et al. (2006). A microRNA expression signature of human solid tumors defines cancer gene targets. Proc. Natl. Acad. Sci. USA *103*, 2257–2261.
3. Kokai, Y., Cohen, J.A., Drebin, J.A., and Greene, M.I. (1987). Stage- and tissue-specific expression of the neu oncogene in rat development. Proc. Natl. Acad. Sci. USA *84*, 8498–8501.
4. Alon, U., Barkai, N., Notterman, D.A., Gish, K., Ybarra, S., Mack, D., and Levine, A.J. (1999). Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. Proc. Natl. Acad. Sci. USA *96*, 6745–6750.
5. Pritchard, C.C., Hsu, L., Delrow, J., and Nelson, P.S. (2001). Project normal: defining normal variance in mouse gene expression. Proc. Natl. Acad. Sci. USA *98*, 13266–13271.
6. Lipska, B.K., Deep-Soboslay, A., Weickert, C.S., Hyde, T.M., Martin, C.E., Herman, M.M., and Kleinman, J.E. (2006). Critical factors in gene expression in postmortem human brain: Focus on studies in schizophrenia. Biol. Psychiatry *60*, 650–658.
7. Bakay, M., Chen, Y.W., Borup, R., Zhao, P., Nagaraju, K., and Hoffman, E.P. (2002). Sources of variability and effect of experimental approach on expression profiling data interpretation. BMC Bioinformatics *3*, 4.
8. Brack, A.S., Conboy, M.J., Roy, S., Lee, M., Kuo, C.J., Keller, C., and Rando, T.A. (2007). Increased Wnt signaling during aging alters muscle stem cell fate and increases fibrosis. Science *317*, 807–810.
9. Majumdar, A.P., Jaszewski, R., and Dubick, M.A. (1997). Effect of aging on the gastrointestinal tract and the pancreas. Proc. Soc. Exp. Biol. Med. *215*, 134–144.
10. Damato, B.E., Allan, D., Murray, S.B., and Lee, W.R. (1984). Senile atrophy of the human lacrimal gland: the contribution of chronic inflammatory disease. Br. J. Ophthalmol. *68*, 674–680.
11. Kent, O.A., McCall, M.N., Cornish, T.C., and Halushka, M.K. (2014). Lessons from miR-143/145: the importance of cell-type localization of miRNAs. Nucleic Acids Res. *42*, 7528–7538.
12. Venet, D., Pecasse, F., Maenhaut, C., and Bersini, H. (2001). Separation of samples into their constituents using gene expression data. Bioinformatics *17* (*Suppl 1*), S279–S287.
13. Quon, G., and Morris, Q. (2009). ISOLATE: a computational strategy for identifying the primary origin of cancers using high-throughput sequencing. Bioinformatics *25*, 2882–2889.
14. Stuart, R.O., Wachsman, W., Berry, C.C., Wang-Rodriguez, J., Wasserman, L., Klacansky, I., Masys, D., Arden, K., Goodison, S., McClelland, M., et al. (2004). In silico dissection of

cell-type-associated patterns of gene expression in prostate cancer. Proc. Natl. Acad. Sci. USA *101*, 615–620.
15. Lähdesmäki, H., Shmulevich, L., Dunmire, V., Yli-Harja, O., and Zhang, W. (2005). In silico microdissection of microarray data from heterogeneous cell populations. BMC Bioinformatics *6*, 54.
16. Shen-Orr, S.S., Tibshirani, R., Khatri, P., Bodian, D.L., Staedtler, F., Perry, N.M., Hastie, T., Sarwal, M.M., Davis, M.M., and Butte, A.J. (2010). Cell type-specific gene expression differences in complex tissues. Nat. Methods *7*, 287–289.
17. Newman, A.M., Liu, C.L., Green, M.R., Gentles, A.J., Feng, W., Xu, Y., Hoang, C.D., Diehn, M., and Alizadeh, A.A. (2015). Robust enumeration of cell subsets from tissue expression profiles. Nat. Methods *12*, 453–457.
18. Okaty, B.W., Sugino, K., and Nelson, S.B. (2011). A quantitative comparison of cell-type-specific microarray gene expression profiling methods in the mouse brain. PLoS ONE *6*, e16493.
19. Debey, S., Schoenbeck, U., Hellmich, M., Gathof, B.S., Pillai, R., Zander, T., and Schultze, J.L. (2004). Comparison of different isolation techniques prior gene expression profiling of blood derived cells: impact on physiological responses, on overall expression and the role of different cell types. Pharmacogenomics J. *4*, 193–207.
20. Consortium, G.T.; GTEx Consortium (2013). The Genotype-Tissue Expression (GTEx) project. Nat. Genet. *45*, 580–585.
21. Melé, M., Ferreira, P.G., Reverter, F., DeLuca, D.S., Monlong, J., Sammeth, M., Young, T.R., Goldmann, J.M., Pervouchine, D.D., Sullivan, T.J., et al.; GTEx Consortium (2015). Human genomics. The human transcriptome across tissues and individuals. Science *348*, 660–665.
22. Yang, J., Huang, T., Petralia, F., Long, Q., Zhang, B., Argmann, C., Zhao, Y., Mobbs, C.V., Schadt, E.E., Zhu, J., and Tu, Z.; GTEx Consortium (2015). Synchronized age-related gene expression changes across multiple tissues in human and the link to complex diseases. Sci. Rep. *5*, 15145.
23. McCall, M.N., Kent, O.A., Yu, J., Fox-Talbot, K., Zaiman, A.L., and Halushka, M.K. (2011). MicroRNA profiling of diverse endothelial cell types. BMC Med. Genomics *4*, 78.
24. McCall, M.N., Jaffee, H.A., Zelisko, S.J., Sinha, N., Hooiveld, G., Irizarry, R.A., and Zilliox, M.J. (2014). The Gene Expression Barcode 3.0: improved data processing and mining tools. Nucleic Acids Res. *42*, D938–D943.
25. Cornish, T.C., Chakravarti, A., Kapoor, A., and Halushka, M.K. (2015). HPASubC: A suite of tools for user subclassification of human protein atlas tissue images. J. Pathol. Inform. *6*, 36.
26. Kim, M.S., Pinto, S.M., Getnet, D., Nirujogi, R.S., Manda, S.S., Chaerkady, R., Madugundu, A.K., Kelkar, D.S., Isserlin, R., Jain, S., et al. (2014). A draft map of the human proteome. Nature *509*, 575–581.
27. Halushka, M.K., Cornish, T.C., Lu, J., Selvin, S., and Selvin, E. (2010). Creation, validation, and quantitative analysis of protein expression in vascular tissue microarrays. Cardiovasc. Pathol. *19*, 136–146.
28. Macgregor, A.M., Eberhart, C.G., Fraig, M., Lu, J., and Halushka, M.K. (2009). Tissue inhibitor of matrix metalloproteinase-3 levels in the extracellular matrix of lung, kidney, and eye increase with age. J. Histochem. Cytochem. *57*, 207–213.
29. Anders, S., and Huber, W. (2010). Differential expression analysis for sequence count data. Genome Biol. *11*, R106.
30. Klambauer, G., Unterthiner, T., and Hochreiter, S. (2013). DEXUS: identifying differential expression in RNA-Seq studies with unknown conditions. Nucleic Acids Res. *41*, e198.
31. Tong, P., Chen, Y., Su, X., and Coombes, K.R. (2013). SIBER: systematic identification of bimodally expressed genes using RNAseq data. Bioinformatics *29*, 605–613.
32. Young, B., Lowe, J.S., Stevens, A., and Heath, J.W. (2006). Wheater's Functional Histology (China: Elsevier).
33. Gene Ontology, C.; Gene Ontology Consortium (2015). Gene Ontology Consortium: going forward. Nucleic Acids Res. *43*, D1049–D1056.
34. Rebhan, M., Chalifa-Caspi, V., Prilusky, J., and Lancet, D. (1997). GeneCards: integrating information about genes, proteins and diseases. Trends Genet. *13*, 163.
35. Uhlen, M., Oksvold, P., Fagerberg, L., Lundberg, E., Jonasson, K., Forsberg, M., Zwahlen, M., Kampf, C., Wester, K., Hober, S., et al. (2010). Towards a knowledge-based Human Protein Atlas. Nat. Biotechnol. *28*, 1248–1250.
36. Subramanian, A., Tamayo, P., Mootha, V.K., Mukherjee, S., Ebert, B.L., Gillette, M.A., Paulovich, A., Pomeroy, S.L., Golub, T.R., Lander, E.S., and Mesirov, J.P. (2005). Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc. Natl. Acad. Sci. USA *102*, 15545–15550.
37. Szklarczyk, D., Franceschini, A., Wyder, S., Forslund, K., Heller, D., Huerta-Cepas, J., Simonovic, M., Roth, A., Santos, A., Tsafou, K.P., et al. (2015). STRING v10: protein-protein interaction networks, integrated over the tree of life. Nucleic Acids Res. *43*, D447–D452.
38. Holm, S. (1979). A Simple Sequentially Rejective Multiple Test Procedure. Scand. J. Stat. *6*, 65–70.
39. Rooney, S.A., Young, S.L., and Mendelson, C.R. (1994). Molecular and cellular processing of lung surfactant. FASEB J. *8*, 957–967.
40. Tsuno, K., Miura, K., Takeya, M., Kolobow, T., and Morioka, T. (1991). Histopathologic pulmonary changes from mechanical ventilation at high peak airway pressures. Am. Rev. Respir. Dis. *143*, 1115–1120.
41. Kitsiouli, E., Nakos, G., and Lekka, M.E. (2009). Phospholipase A2 subclasses in acute respiratory distress syndrome. Biochim. Biophys. Acta *1792*, 941–953.
42. Gupta, S., Halushka, M.K., Hilton, G.M., and Arking, D.E. (2012). Postmortem cardiac tissue maintains gene expression profile even after late harvesting. BMC Genomics *13*, 26.
43. Leek, J.T., Johnson, W.E., Parker, H.S., Jaffe, A.E., and Storey, J.D. (2012). The sva package for removing batch effects and other unwanted variation in high-throughput experiments. Bioinformatics *28*, 882–883.
44. Leek, J.T., and Storey, J.D. (2007). Capturing heterogeneity in gene expression studies by surrogate variable analysis. PLoS Genet. *3*, 1724–1735.
45. Leek, J.T. (2014). svaseq: removing batch effects and other unwanted noise from sequencing data. Nucleic Acids Res. *42*, e161.
46. Stahlman, M.T., Gray, M.E., and Whitsett, J.A. (1996). Expression of thyroid transcription factor-1(TTF-1) in fetal and neonatal human lung. J. Histochem. Cytochem. *44*, 673–678.

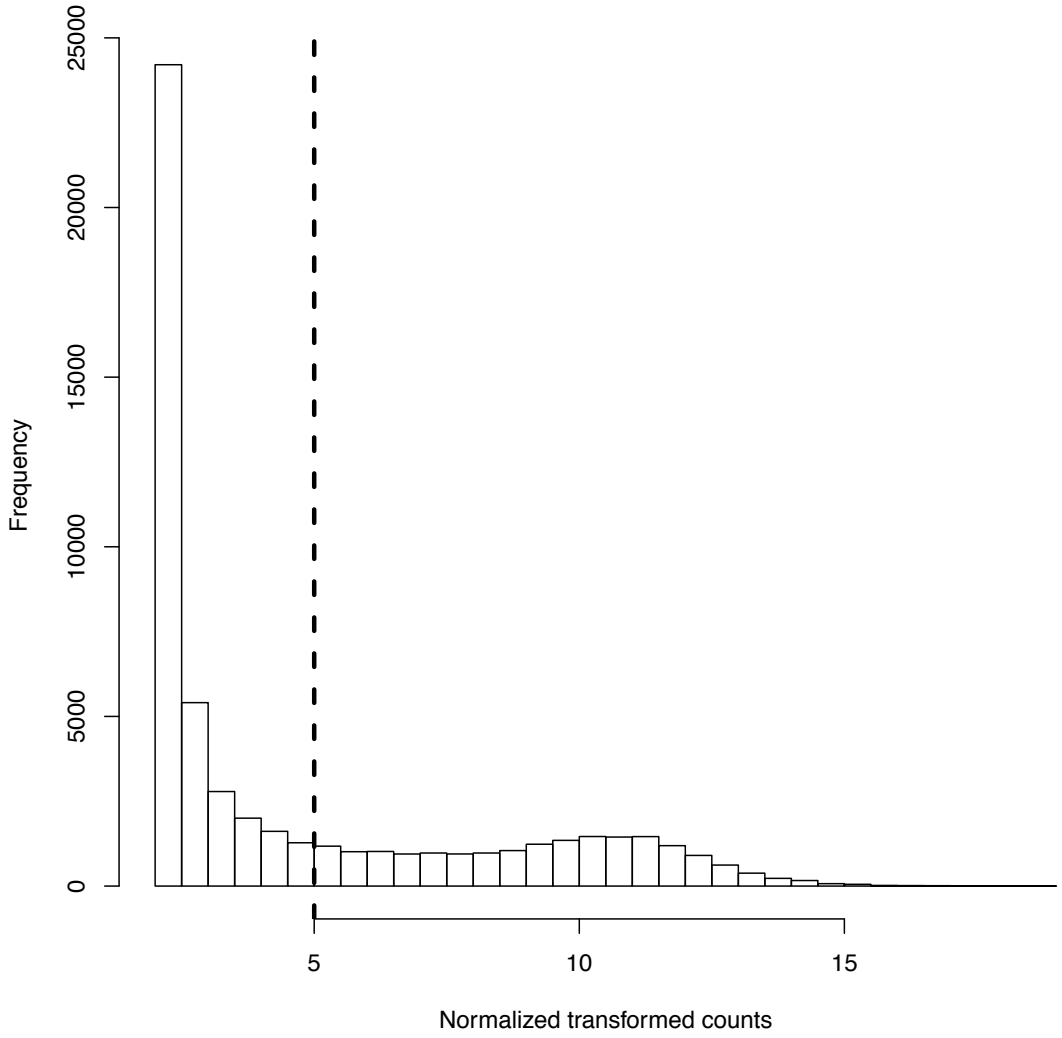
The American Journal of Human Genetics, Volume *99*

Supplemental Data

Complex Sources of Variation in Tissue Expression Data: Analysis of the GTEx Lung Transcriptome

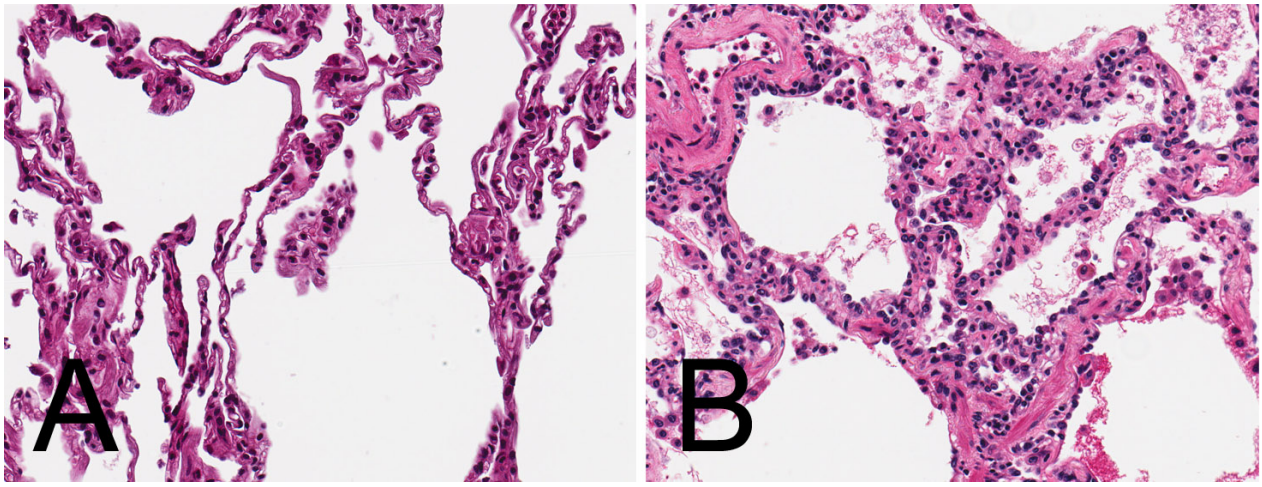
Matthew N. McCall, Peter B. Illei, and Marc K. Halushka

Figure S1: Histogram of average normalized and transformed counts



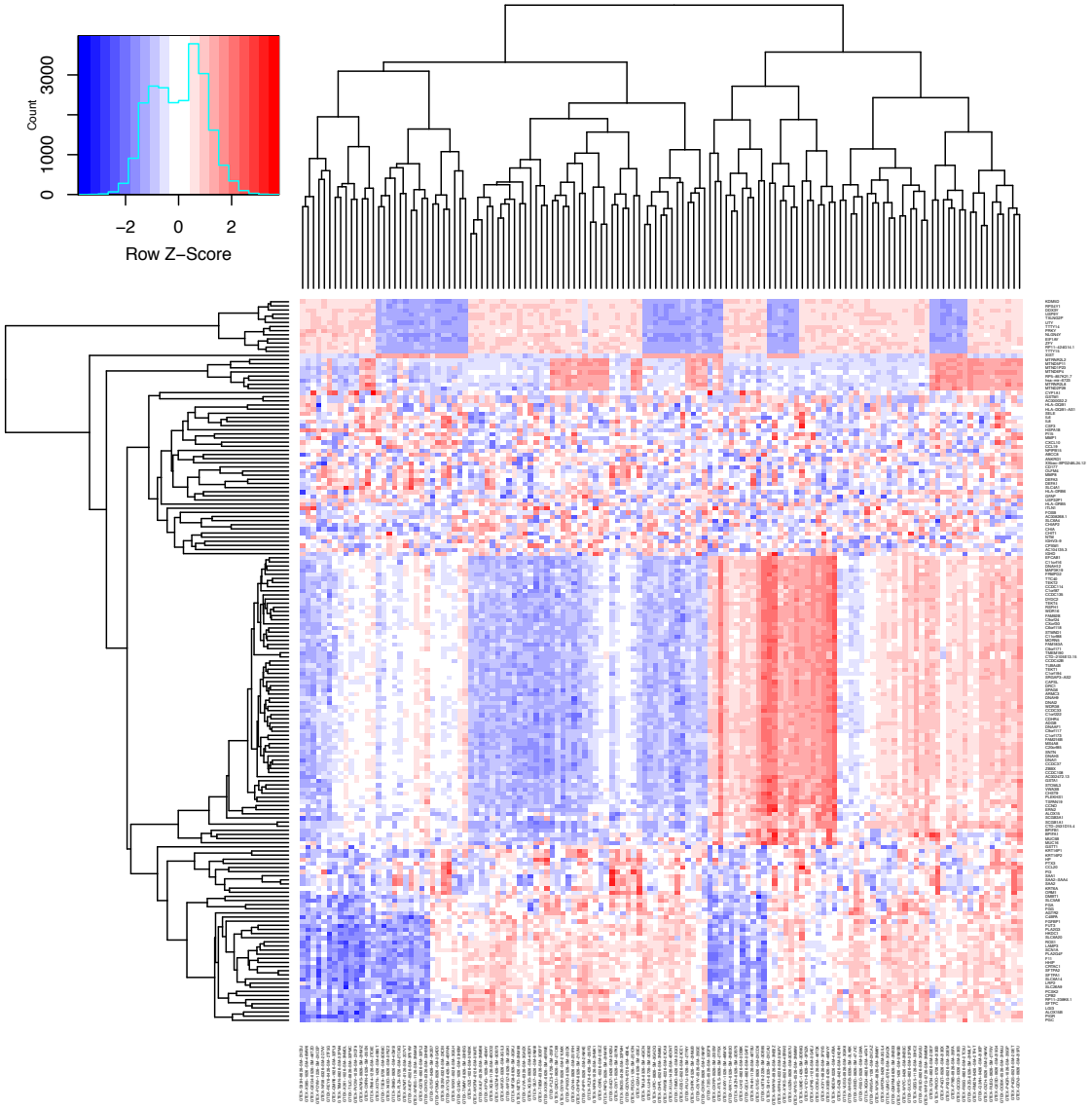
Gene counts were normalized to account for differences in library size using size factors estimated via the median ratio method and subsequently transformed using a variance stabilizing transformation based on the dispersion-mean relationship. The average normalized transformed counts were computed for each of the 55,993 genes represented in the GTEx data across the 133 lung samples. The histogram of these values shows the typical mixture of noise and signal. To filter genes that appear to be primarily noise, we removed those with average normalized transformed counts < 5 (dashed line). 18,703 genes passed this filter.

Figure S2: Type II pneumocyte hyperplasia



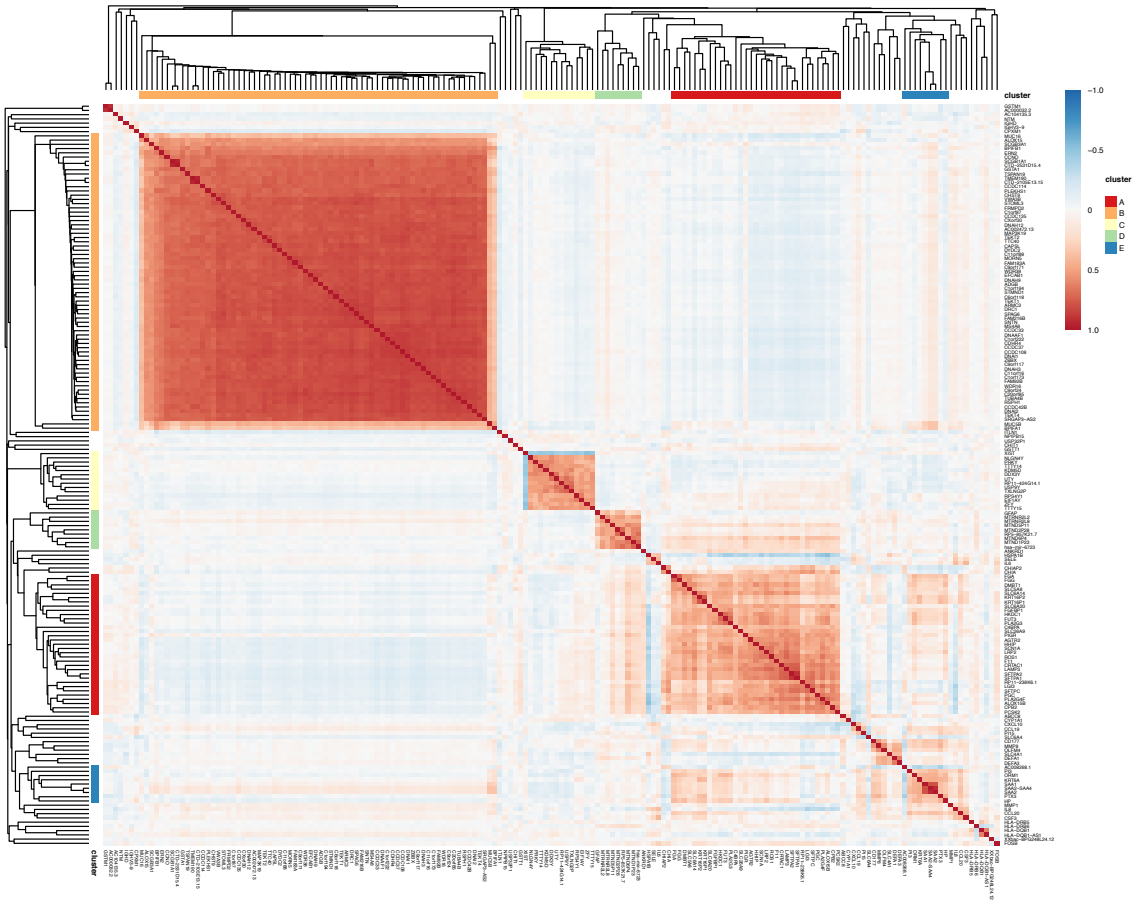
(A) A normal ratio of type I to type II pneumocytes. (B) Type II pneumocyte hyperplasia as noted by more cuboidal rather than squamoid epithelium lining the alveolar walls.

Figure S3: Residual Expression of High Variance Genes across Lung Tissue Samples



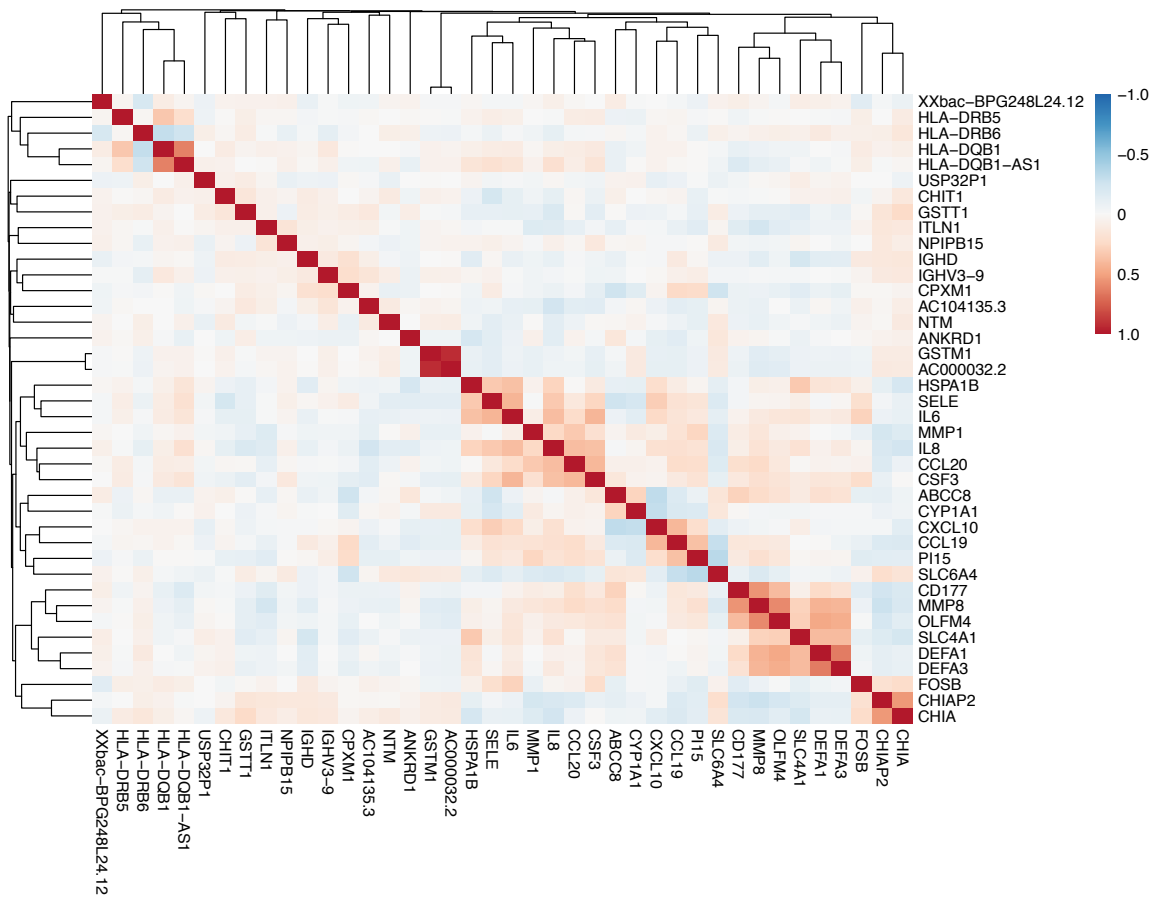
The gene-wise residual expression for the 175 genes with the highest variance (>4) across 133 lung tissue samples was used to cluster both genes and samples.

Figure S4: Between gene correlation of 175 high variance genes in lung tissue samples



Kendall rank correlation coefficient between each pair of the 175 genes shown in Figure S3 was used to cluster genes. White corresponds to correlation of zero; darker shades of red represent increasing positive correlation; darker shades of blue represent increasing negative correlation. One can observe five clear groups of highly correlated genes labeled A-E and shown with colored bars on both axes. Each of these gene sets was examined further to identify the commonality between genes in each of them.

Figure S5: Between gene correlation of the 40 high variance genes that showed poor overall clustering



Kendall rank correlation coefficient between each pair of the genes was used to cluster genes. White corresponds to correlation of zero; darker shades of red represent increasing positive correlation; darker shades of blue represent increasing negative correlation. Using a suite of functional analysis tools, these 40 genes were found to be generally immune response related. Among the 40 genes, there are rare, tight clusters of 2-3 genes.

Figure S6: Overlap between high variance genes and bimodal genes selected by SIBER or DEXUS

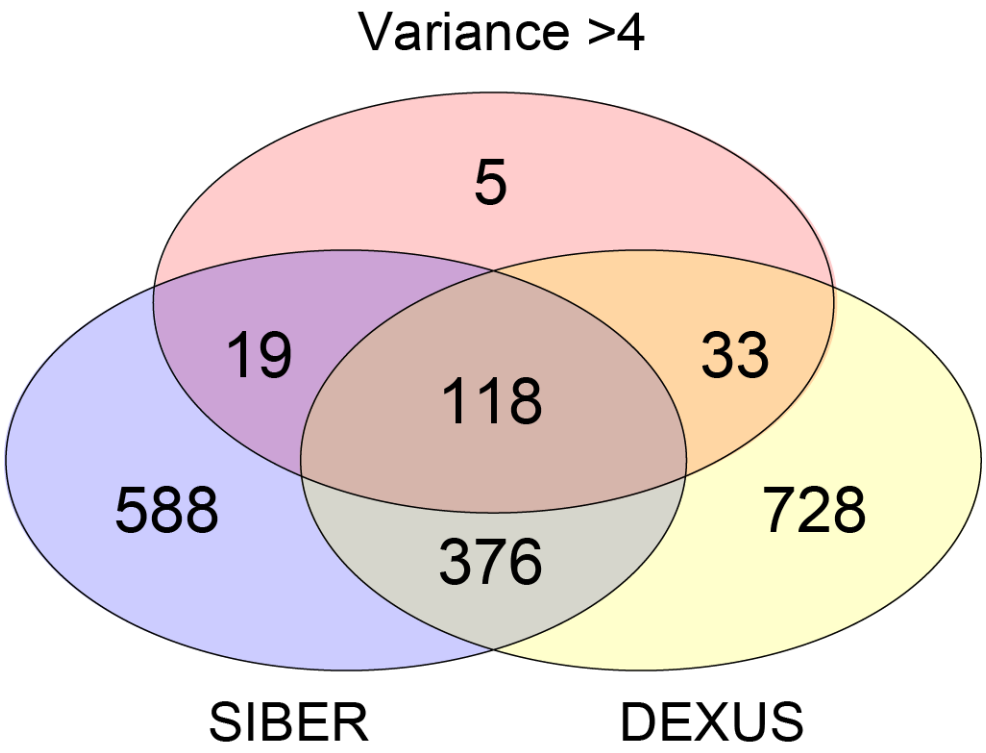
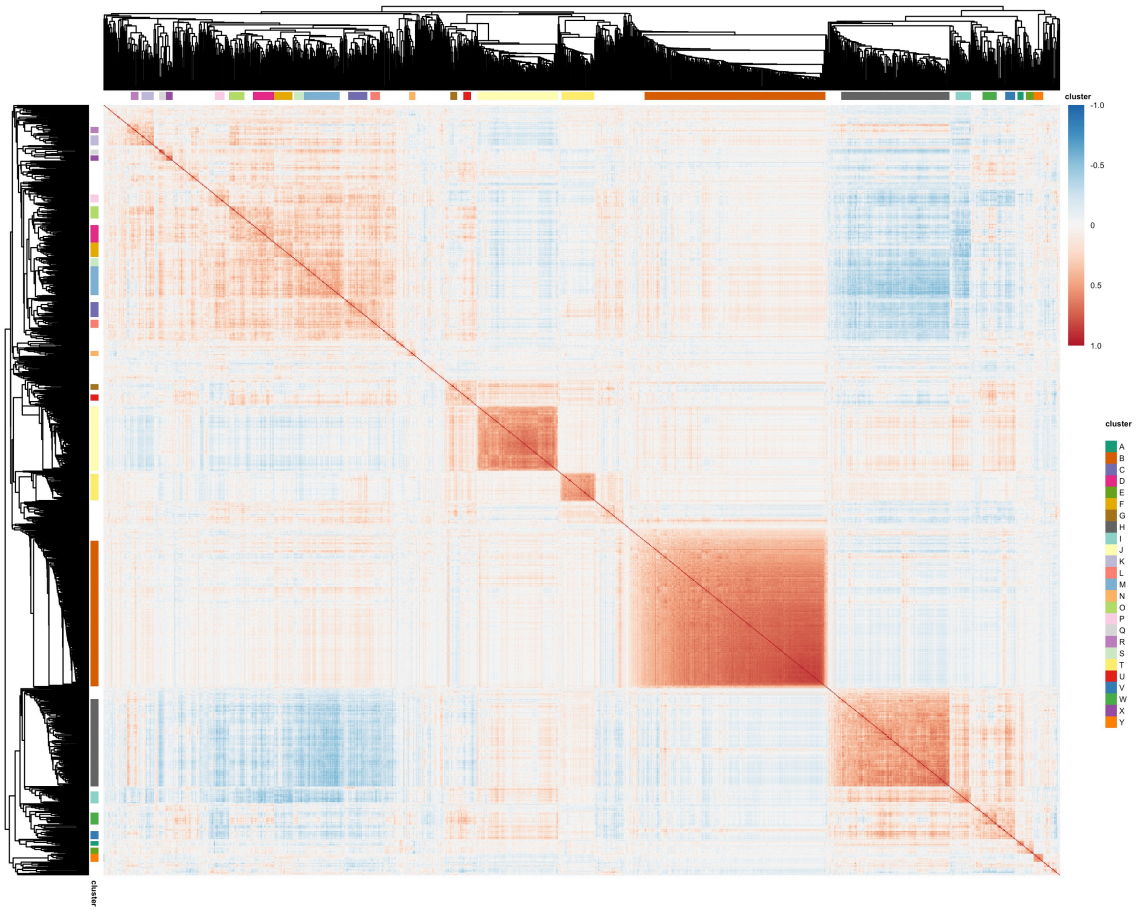
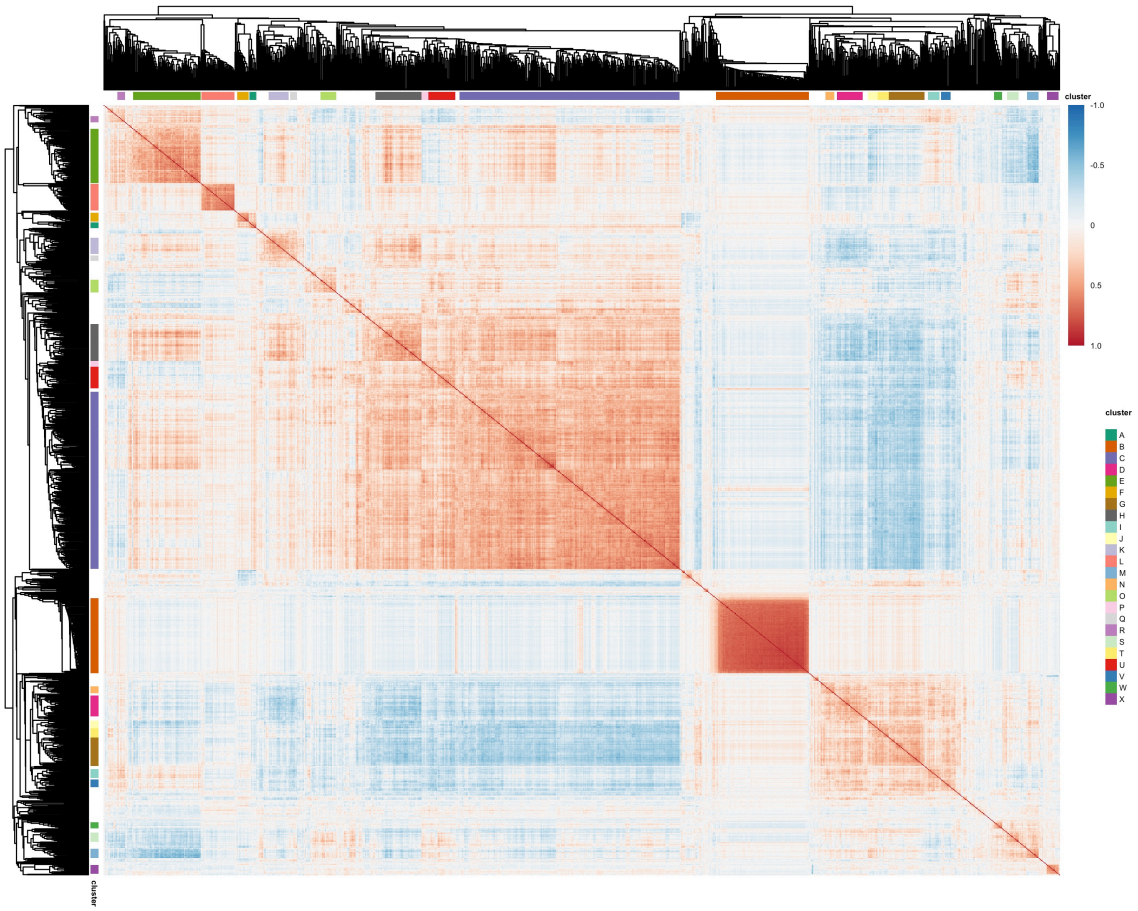


Figure S7: Between gene correlation of DEXUS bimodal genes in lung tissue samples



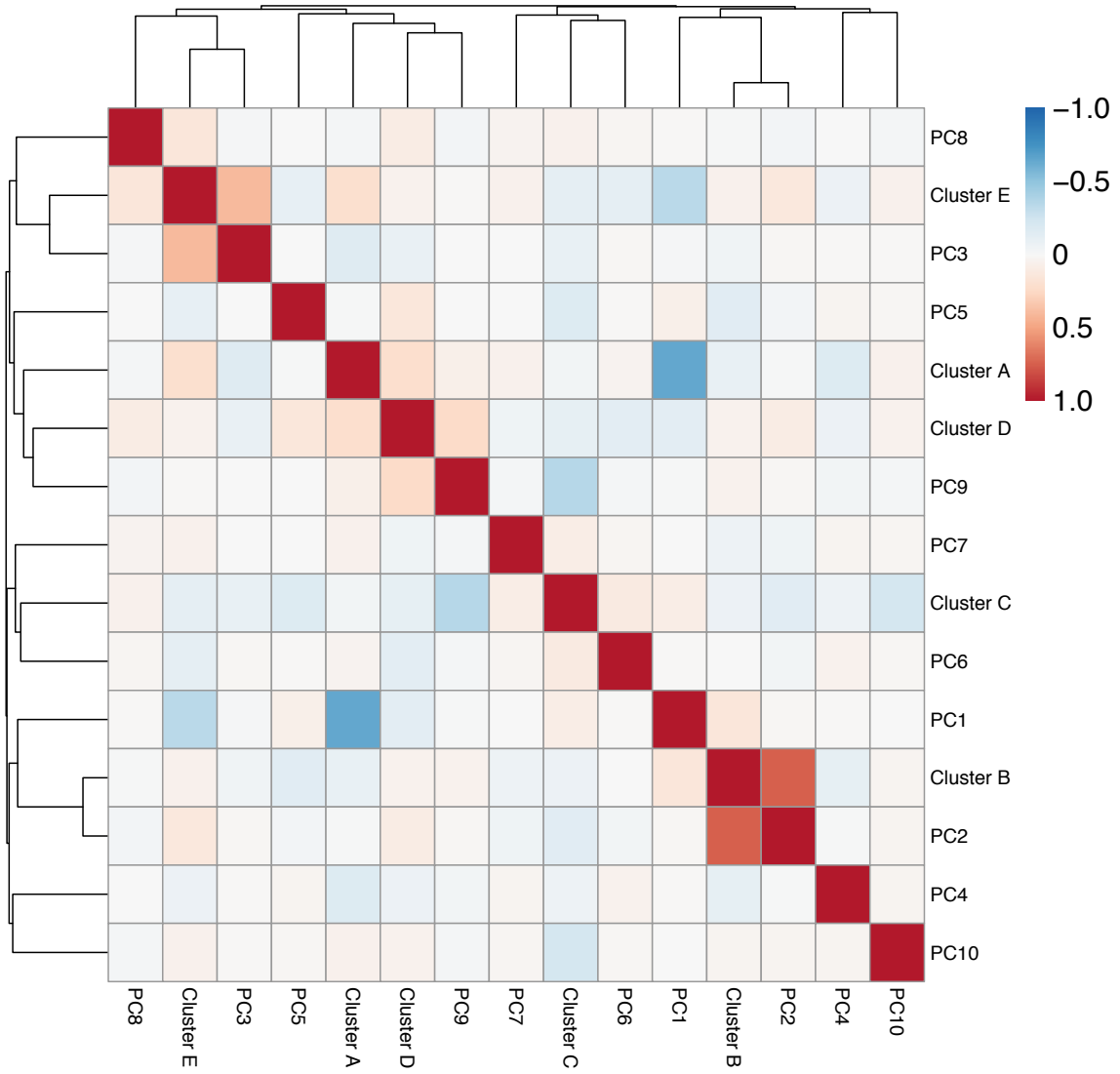
Kendall rank correlation coefficient between each pair of the 1,255 genes identified as bimodal by DEXUS was used to cluster genes. White corresponds to correlation of zero; darker shades of red represent increasing positive correlation; darker shades of blue represent increasing negative correlation. Groups of highly correlated genes are labeled A-Y and shown with colored bars on both axes. Each of these gene sets was examined further to identify the commonality between genes in each of them.

Figure S8: Between gene correlation of SIBER bimodal genes in lung tissue samples



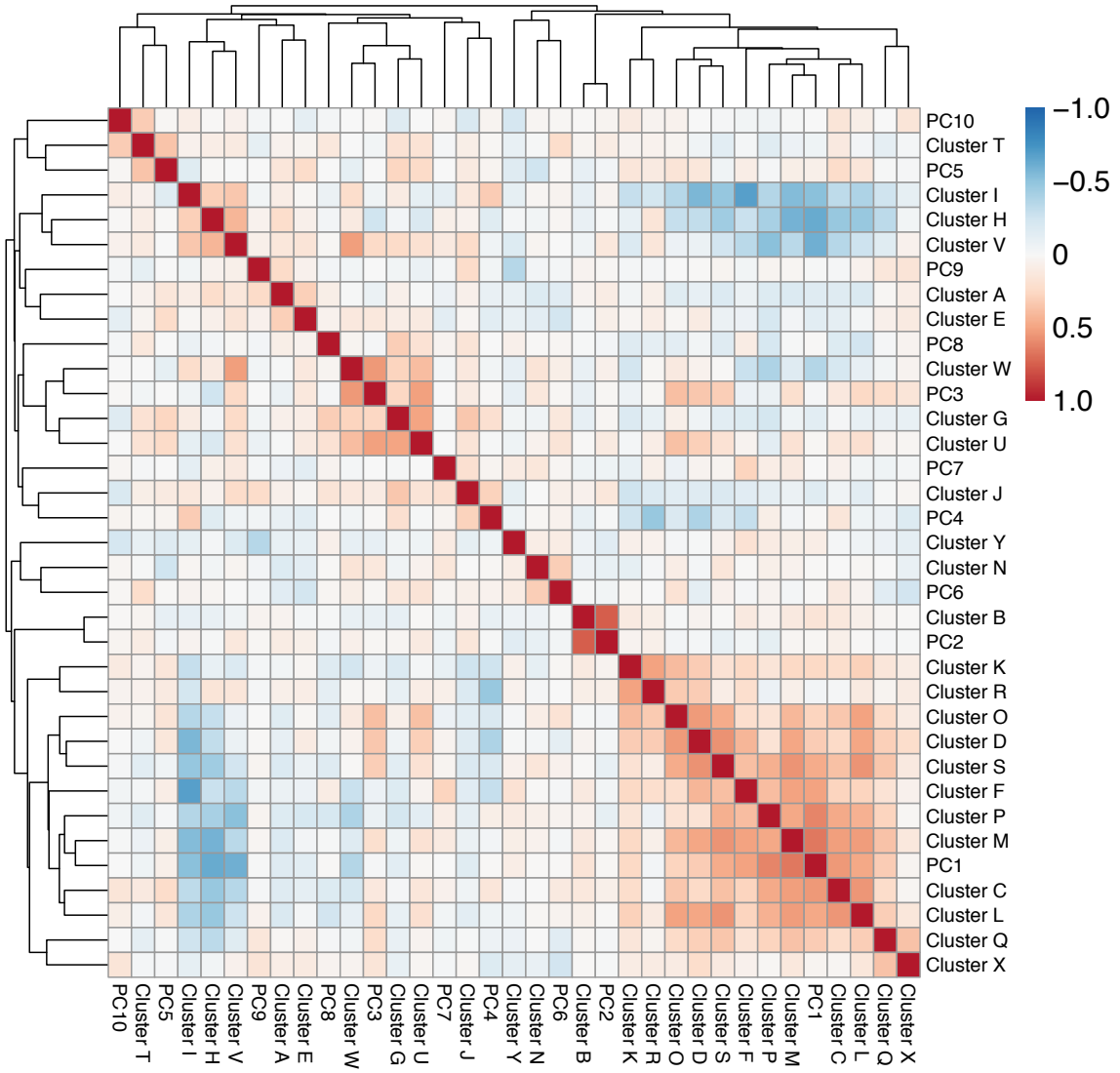
Kendall rank correlation coefficient between each pair of the 1,101 genes identified as bimodal by SIBER was used to cluster genes. White corresponds to correlation of zero; darker shades of red represent increasing positive correlation; darker shades of blue represent increasing negative correlation. Groups of highly correlated genes are labeled A-X and shown with colored bars on both axes.

Figure S9: Correlation between high variance clusters and principal components



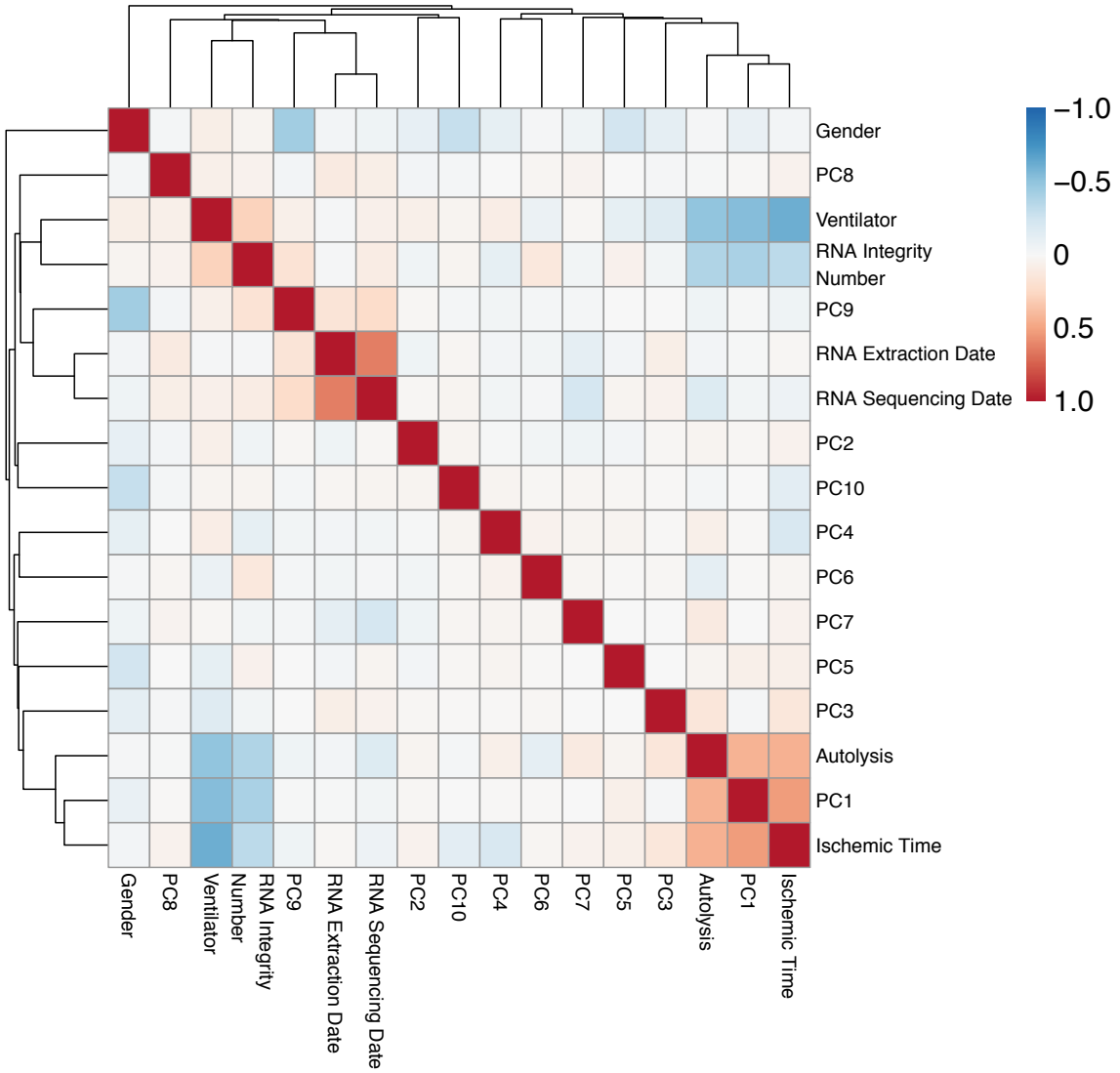
The Kendall rank correlation coefficient, Kendall's τ, was used to assess the pair-wise dependence between average standardized gene expression for clusters A-E from the high variance gene sets and the top 10 principal components across 133 lung tissue samples. PC1 was strongly negatively correlated with Cluster A ($\tau = -0.64$; p-value < 1.4×10^{-26}). PC2 was strongly correlated with Cluster B ($\tau = 0.74$; p-value < 1.5×10^{-35}). PC3 was correlated with Cluster E ($\tau = 0.40$; p-value < 1.1×10^{-10}). PC9 was positively correlated with Cluster D ($\tau = 0.24$; p-value = 9.9×10^{-4}) and negatively correlated with Cluster C ($\tau = -0.35$; p-value = 2.8×10^{-8}).

Figure S10: Correlation between principal components and DEXUS clusters



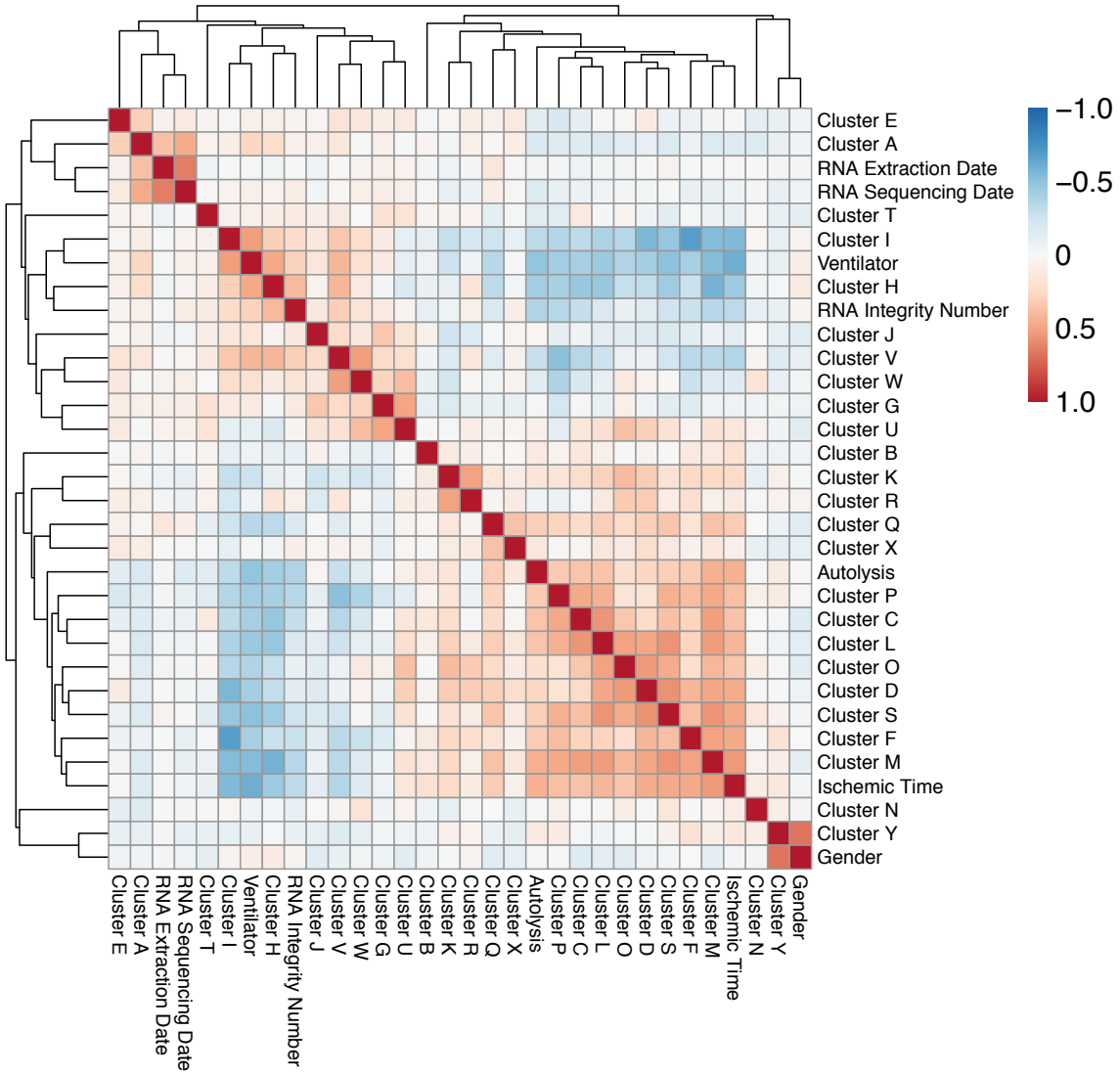
The Kendall rank correlation coefficient, Kendall's τ, was used to assess the pair-wise dependence between the top 10 principal components and 25 DEXUS gene clusters across 133 lung tissue samples. PC1 was strongly negatively correlated with Cluster H ($\tau = -0.63$; p-value < 4.5×10^{-25}), Cluster I ($\tau = -0.52$; p-value < 2.2×10^{-17}), and Cluster V ($\tau = -0.60$; p-value < 2.3×10^{-23}), and strongly positively correlated with Cluster C ($\tau = 0.54$; p-value < 8.2×10^{-19}), Cluster F (τ=0.50; p-value < 5.4×10^{-16}), Cluster M ($\tau = 0.66$; p-value < 6.6×10^{-28}), and Cluster P ($\tau = 0.62$; p-value < 3.2×10^{-24}). PC2 was strongly positively correlated with Cluster B ($\tau = 0.75$; p-value < 7.2×10^{-36}). PC3 was strongly positively correlated with Cluster U ($\tau = 0.51$; p-value < 1.4×10^{-16}) and Cluster W ($\tau = 0.54$; p-value < 1.5×10^{-18}). The fourth principal component was negatively correlated with Cluster R ($\tau = -0.47$; p-value < 2.7×10^{-14}).

Figure S11: Correlation between principal components and technical variables



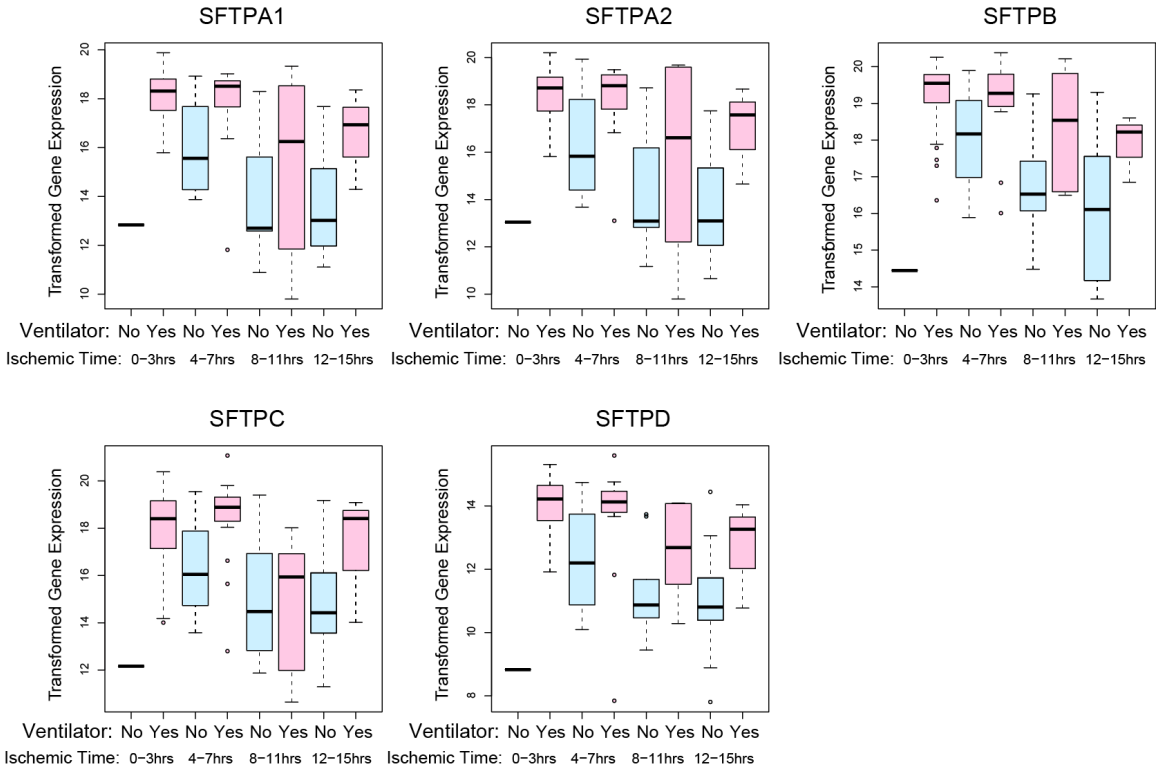
The Kendall rank correlation coefficient, Kendall's τ, was used to assess the pair-wise dependence between the top 10 principal components and 7 technical variables across 133 lung tissue samples.

Figure S12: Correlation between DEXUS clusters and technical variables



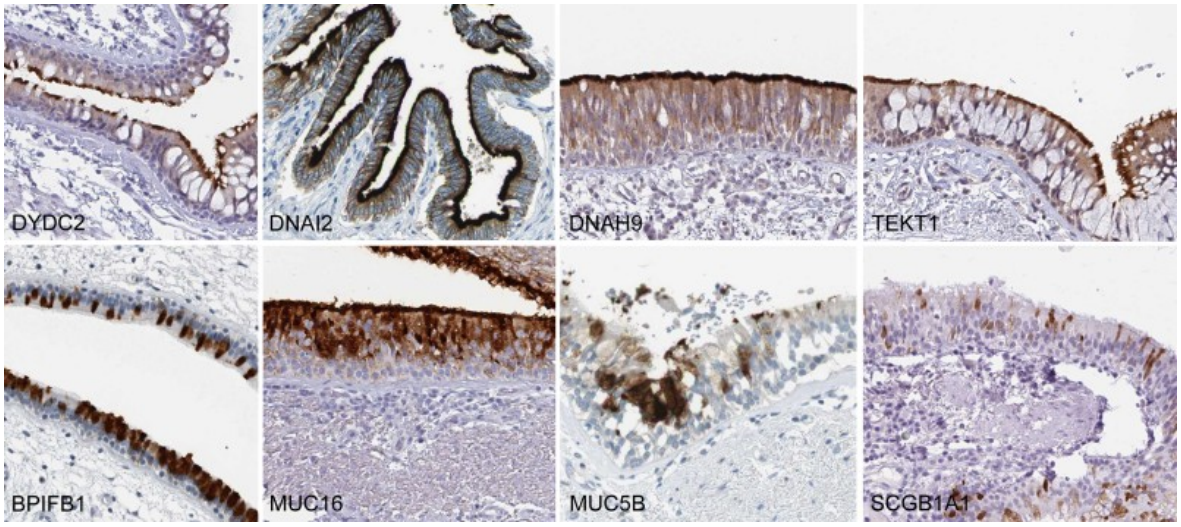
The Kendall rank correlation coefficient, Kendall's τ, was used to assess the pair-wise dependence between the 25 DEXUS gene clusters and 7 technical variables across 133 lung tissue samples. Clusters C, D, F, L, M, O, P, and S were generally positively correlated with ischemic time and autolysis score and negatively correlated with RNA integrity number and ventilation prior to death; in contrast, Clusters H & I were generally positively correlated with RNA integrity number and ventilation prior to death and negatively correlated with ischemic time and autolysis score. Gender was positively correlated with Cluster Y ($\tau = 0.68$; p-value = 1.2×10^{-19}). RNA extraction date was positively correlated with Cluster A ($\tau = 0.37$; p-value = 1.6×10^{-8}), and RNA sequencing date was also positively correlated with Cluster A ($\tau = 0.46$; p-value = 1.4×10^{-12}).

Figure S13: Surfactant gene expression stratified by ventilation and ischemic time



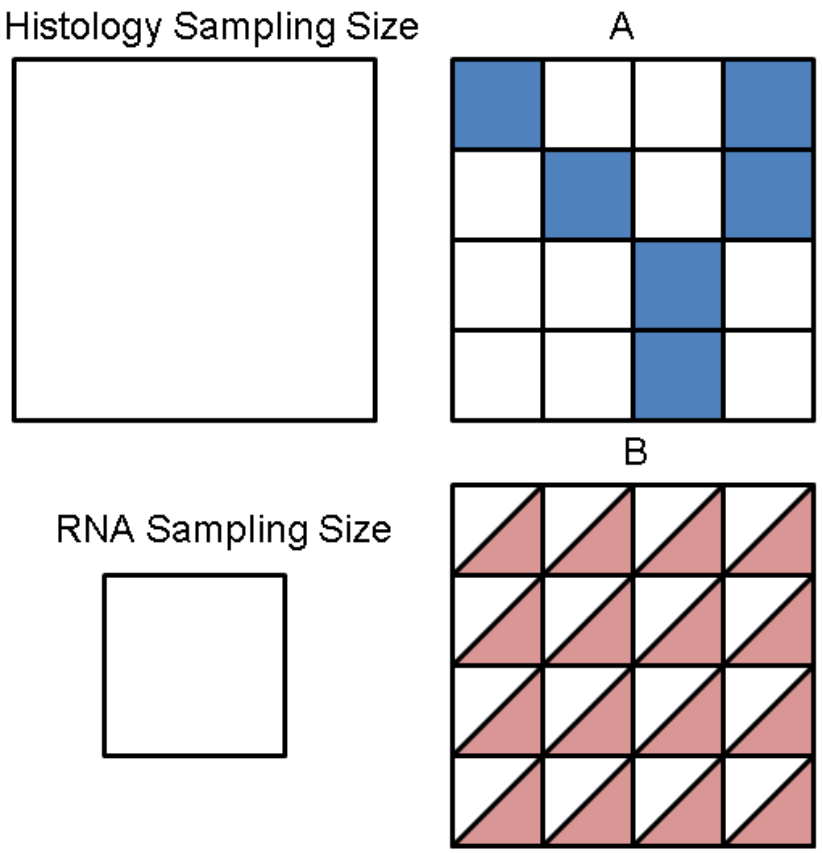
Transformed gene expression of five surfactant genes (SFTPA1, SFTPA2, SFTPB, SFTPC, and SFTPD) was stratified based on ventilation prior to death and ischemic time. The three highest ischemic time groups, which had fewer than 10 samples, were omitted from these figures. Ventilation was associated with higher median expression for all five genes in all four ischemic time groups.

Figure S14: Bronchial Epithelial Staining Patterns of Cluster B proteins



Eight Cluster B proteins, localized by immunohistochemistry are demonstrated as representative of the cluster. DYDC2, DNAI2, DNAH9 and TEKT1 are present on the cilia of respiratory epithelial cells (dark brown staining along the apical edge of the cells). Goblet cells appear as clear cells between these epithelial cells. BP1FB1, MUC16, MUC5B and SCGB1A1 all stain intervening goblet cells (scattered brown staining). MUC16 also staining a mucoid layer above the cells. Respiratory epithelial cells and goblet cells constitute the bronchial epithelium. Images from Human Protein Atlas.[35] Original magnification 200x.

Figure S15: The effect of subsampling on signal spatial and sampling heterogeneity



The heterogeneity or homogeneity of cellular differences and sampling sizes in tissue affect correlations between histologic and gene discoveries in tissues. A) A heterogeneous element (blue squares, ex. bronchial epithelium) would represent 37.5% of the feature in a histology sample of 4x4 units. However, by subsampling for RNA as a 2x2 unit size, one has a 66% chance of getting 50% element signal, a 22% chance of 25% element signal and an 11% chance of getting 0% element. B) For a homogeneous element (pink triangles, ex. neutrophil infiltrate) both sampling methods would capture 50% signals of the element.